



UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG
FACULTY OF COMMERCE, LAW AND MANAGEMENT

**A REGULATORY CAPTURE EXPLANATION OF
SOUTH AFRICA'S PRIVATE HEALTH INSURANCE
LEGISLATION**

Name of Applicant: Hugh-David Hutcheson

Student Number: 9403189P

Degree: Doctor of Philosophy (PhD)

School: School of Economic and Business Sciences

Supervisor: Professor Robert W Vivian

A thesis submitted to the Faculty of Commerce, Law and Management, University of the Witwatersrand, in fulfilment of the requirements for the degree of Doctor of Philosophy (PhD).

Johannesburg, South Africa

September 2011

A Regulatory Capture Explanation of South Africa's Private Health Insurance Legislation

ABSTRACT

Private healthcare financing in South Africa has undergone several regulatory reforms, the most recent of which saw the enactment of the Medical Schemes Act No. 131 of 1998. The stated reforms, most especially open enrolment and community rating, were touted by the government as necessary to address the undesirable effects of adverse selection. However, it was never questioned whether in fact adverse selection is a feature of the South African medical schemes landscape. Adverse selection is found to be absent. Thus, government's supposition that adverse selection, as a consequence of the deregulation that took place during the late 1980s and early 1990s, is responsible for the deterioration in medical scheme coverage for the elderly, unhealthy or poor is fallacious. Since the ostensible reason for the current legislation does not stand up to scrutiny, regulatory capture is offered as the plausible alternative explanation for the promulgation of the current legislation governing medical schemes business.

Keywords: Insurance, adverse selection, health insurance, information asymmetry, regulation, regulatory capture.

JEL Classification: G22, D72, D82, I11, I18.

DECLARATION

I, Hugh-David Hutcheson, declare that this thesis is my own, unaided work. It is submitted in fulfilment of the requirements for the degree of Doctor of Philosophy (PhD) at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination in this or any other university.

Hugh-David Hutcheson

20 September 2011

ACKNOWLEDGEMENTS

I wish to express my sincere thanks to following people:

1. My parents, John and Cynthia Hutcheson, who have been a significant source of support and encouragement throughout my life. I must thank them for instilling within me the principles and values that dictate every choice, and decision that I have made and will continue to make into the future. In particular, the conscious decision to always strive for excellence and aim to reach one's potential has indeed guided me to value most highly the pursuit of knowledge through education. Without my parents, this thesis and indeed all my prior qualifications would not have been possible. To my siblings; Yvonne Hutcheson, Nola Fischer, Dale Hutcheson and Sheryl Walters, I extend my sincere gratitude for always making life such a wonderful experience for me and for always being there for me.

2. My supervisor, Professor Robert Vivian, who has been an inspiration to me, not only in completing my doctorate, but also for encouraging me to pursue the most rewarding career that of becoming an academic, lecturer and researcher. Thank you for all your time, effort, guidance and patience.

3. My fellow colleagues at the University of the Witwatersrand: thank you so much for your support and guidance. The discussions that I have had with you and the feedback that you have provided have certainly assisted me immensely in guiding my research.

TABLE OF CONTENTS

1	INTRODUCTION	1
1.1.	DEMAND FOR HEALTH AND HEALTH CAPITAL	7
1.2.	ROLE OF HEALTHCARE AND MEDICINE	9
1.3.	CHAPTER OUTLINE	10
1.4.	CONTRIBUTION OF THE THESIS	13
1.4.1	<i>Objective of the Thesis</i>	14
2	SOUTH AFRICA'S PRIVATE HEALTH INSURANCE SECTOR	16
2.1	BACKGROUND	16
2.1.1	<i>Medical Schemes Prior to the Deregulation</i>	19
2.1.2	<i>Deregulation during the 1990s</i>	22
2.1.3	<i>Rationale behind the enactment of the Medical Schemes Act No. 131 of 1998</i>	28
2.2	THE MEDICAL SCHEMES ACT NO. 131 OF 1998	38
2.2.1	<i>Three Core Pillars of the Medical Schemes Act No. 131 of 1998</i>	41
2.2.2	<i>Community Rating and the Adverse Selection Death Spiral</i>	46
2.2.3	<i>Risk-Related Cross-Subsidisation and the Risk Equalisation Fund</i>	48
2.2.4	<i>Government's Initiatives to Extend Medical Scheme Cover to Low-Income Earners and the Proposed National Health Insurance System</i>	57
2.2.5	<i>Brief Assessment of the South African Private Health Insurance Industry</i>	68
2.3	REGULATION RELATING TO HEALTHCARE PROVISION	81
2.4	COST CONTAINMENT	105
2.4.1	<i>Is Medical Scheme Coverage becoming more affordable?</i>	107
2.4.2	<i>Medical Scheme Cost Drivers</i>	116
2.4.3	<i>Strategies to Contain Healthcare Costs</i>	122
2.4.4	<i>Principle of Managed Care</i>	126
2.5	CONCLUSION	144
3	THEORY OF ADVERSE SELECTION.....	147
3.1.	IMPERFECT INFORMATION IN COMPETITIVE INSURANCE MARKETS	156
3.1.1.	<i>Lemons Principle</i>	157
3.1.2.	<i>Gresham's Law</i>	170
3.1.3.	<i>Rothschild and Stiglitz (1976)'s Model</i>	173
3.2.	TESTING FOR THE COVERAGE-RISK CORRELATION	219
3.2.1.	<i>Formulation of the Coverage-Risk Specification</i>	222
3.2.2.	<i>Contradictory Evidence</i>	226
3.3.	CONCLUSION	233
4	EMPIRICAL RESEARCH METHODOLOGY AND EMPIRICAL RESULTS.....	235
4.1.	DATA AND PANEL STRUCTURAL DESIGN	237
4.2.	EMPIRICAL FINDINGS	250
4.3.	REASONS FOR THE ABSENCE OF THE POSITIVE COVERAGE-RISK CORRELATION	260
4.3.1.	<i>Absence of Private Information</i>	262
4.3.2.	<i>Private Information Distributed Unevenly Amongst Policyholders</i>	265
4.3.3.	<i>Failure to Utilise Private Information</i>	265
4.3.4.	<i>Insurers possess Superior Information</i>	266
4.3.5.	<i>Interaction between Risk and Risk Aversion (Propitious Selection)</i>	268
4.3.6.	<i>Institutional and Regulatory Factors</i>	269
4.4.	CONCLUSION	270
5	REGULATORY CAPTURE	272
5.1	THEORIES OF ECONOMIC REGULATION.....	274
5.1.1	<i>Public Interest Theory</i>	275
5.1.2	<i>Capture Theory</i>	289

5.1.3	<i>Economic Theory of Regulation</i>	292
5.2	CHICAGO SCHOOL OF REGULATORY CAPTURE.....	298
5.3	VIRGINIA SCHOOL OF REGULATORY CAPTURE	305
5.4	TOLLBOOTH THEORIES	309
5.5	TOULOUSE SCHOOL OF REGULATORY CAPTURE	311
5.6	CONCLUSION	318
6	DISCUSSION AND RECOMMENDATIONS	319
6.1.	CONTRIBUTIONS AND THEORETICAL IMPLICATIONS	319
6.1.1	<i>Regulatory Capture as Alternative Explanation</i>	323
6.2.	LIMITATIONS OF THE RESEARCH.....	327
6.3.	RECOMMENDATIONS FOR FUTURE RESEARCH.....	328
6.4.	CONCLUDING REMARKS	332
7	REFERENCES.....	335

LIST OF TABLES

TABLE 2-1: SUGGESTED LIST OF HEALTHCARE SERVICES COVERED UNDER THE PROPOSED STANDARD NHI BENEFIT PACKAGE	64
TABLE 2-2: PRESCRIBED SOLVENCY LEVELS AND THE NUMBER OF MEDICAL SCHEME BENEFICIARIES.....	74
TABLE 2-3: DEGREE OF COMPETITION WITHIN SOUTH AFRICA'S PRIVATE HEALTH INSURANCE SECTOR (1993 - 2009) ..	80
TABLE 2-4: SUMMARY OF PUBLIC VERSUS PRIVATE SECTOR MIX - HEALTHCARE PROFESSIONALS (WADEE & KAHN, 2007)	93
TABLE 2-5: SUMMARY OF PUBLIC VERSUS PRIVATE SECTOR MIX - HEALTHCARE PROFESSIONALS (2010).....	93
TABLE 2-6: UNITED STATES EMPLOYER-SPONSORED HEALTH INSURANCE - AVERAGE EMPLOYEE CONTRIBUTION TOWARDS THE COST OF COVERAGE	136
TABLE 3-1: POSSIBLE OUTCOMES WHEN PURCHASING A MOTOR VEHICLE	158
TABLE 3-2: WEALTH-COMBINATIONS WITH AND WITHOUT INSURANCE	177
TABLE 4-1: FISHER-TYPE UNIT-ROOT TEST RESULTS (BASED UPON AUGMENTED DICKEY-FULLER TESTS).....	251
TABLE 4-2: SUMMARY OF THE DESCRIPTIVE STATISTICS FOR PANEL DATA VARIABLES.....	252
TABLE 4-3: SUMMARY OF <i>WITHIN</i> AND <i>BETWEEN</i> VARIATION FOR A SELECTION OF THE PANEL DATA VARIABLES UTILISED IN BOTH THE FE AND RE ESTIMATION MODELS	253
TABLE 4-4: REGRESSION OUTPUT - FIXED-EFFECTS MODEL.....	254
TABLE 4-5: REGRESSION OUTPUT - RANDOM-EFFECTS MODEL.....	257
TABLE 4-6: SUMMARY OF FE AND RE ESTIMATION RESULTS	259
TABLE 4-7: HAUSMAN TEST - FIXED-EFFECTS VERSUS RANDOM-EFFECTS	260

LIST OF FIGURES

FIGURE 2-1: PROPOSED FUNDING ARRANGEMENT FOR THE SOUTH AFRICAN NHI SYSTEM	63
FIGURE 2-2: TRENDS IN THE TOTAL NUMBER OF REGISTERED MEDICAL SCHEMES, NUMBER OF OPEN AND RESTRICTED MEDICAL SCHEMES (1993 – 2009).....	70
FIGURE 2-3: TOTAL NUMBER OF MEDICAL SCHEME BENEFICIARIES – MEMBERS AND DEPENDENTS (1993 – 2009).....	73
FIGURE 2-4: INDUSTRY SOLVENCY TRENDS FOR ALL REGISTERED MEDICAL SCHEMES (2000 - 2009)	76
FIGURE 2-5: INDUSTRY SOLVENCY TRENDS FOR OPEN MEDICAL SCHEMES (2000 - 2009)	76
FIGURE 2-6: INDUSTRY SOLVENCY TRENDS FOR RESTRICTED MEDICAL SCHEMES (2000 - 2009)	77
FIGURE 2-7: HERFINDAHL-HIRSCHMAN INDEX (HHI) FOR SOUTH AFRICA'S MEDICAL SCHEME INDUSTRY (1993-2009)	81
FIGURE 2-8: GROWTH IN MEDICAL PRACTITIONERS (GPs AND SPECIALISTS), 2002-2010	90
FIGURE 2-9: TOTAL NET CONTRIBUTION INCOME VERSUS TOTAL NET CLAIMS INCURRED (1993 – 2009) (IN 2009 PRICES)	111
FIGURE 2-10: NET CONTRIBUTION INCOME VERSUS NET CLAIMS INCURRED (1993 – 2009) (PABPM) (IN 2009 PRICES)	112
FIGURE 2-11: ANNUAL REAL GROWTH RATE IN NET CONTRIBUTION INCOME AND NET CLAIMS INCURRED (PABPM) (1993 - 2009)	113
FIGURE 2-12: DIFFERENCE BETWEEN THE AVERAGE REAL MONTHLY NET CONTRIBUTIONS AND NET CLAIMS INCURRED (PABPM) (1993 - 2009) (IN 2009 PRICES)	114
FIGURE 2-13: CLAIMS RATIO OF SOUTH AFRICA'S PRIVATE HEALTH INSURANCE SECTOR (1993 - 2009).....	115
FIGURE 2-14: ANNUAL REAL GROWTH RATE IN NET CONTRIBUTION INCOME AND NET CLAIMS INCURRED (1993 - 2009)	116
FIGURE 2-15: GROSS ADMINISTRATION EXPENDITURE (PABPM) (1993-2009) (IN 2009 PRICES).....	118
FIGURE 2-16: DISTRIBUTION OF HEALTHCARE RISK BENEFITS PAID TO HEALTHCARE PROVIDERS (1993-2009) (PABPM) (IN 2009 PRICES).....	119
FIGURE 2-17: UNITED STATES EMPLOYER-SPONSORED HEALTH INSURANCE - AVERAGE ANNUAL HEALTH INSURANCE PREMIUM - SINGLE COVERAGE (2000-2010)	133
FIGURE 2-18: UNITED STATES EMPLOYER-SPONSORED HEALTH INSURANCE - AVERAGE ANNUAL HEALTH INSURANCE PREMIUM - FAMILY COVERAGE (2000-2010)	133
FIGURE 2-19: UNITED STATES EMPLOYER-SPONSORED HEALTH INSURANCE - AVERAGE ANNUAL HEALTH INSURANCE PREMIUM - EMPLOYER-EMPLOYEE CONTRIBUTIONS FOR SINGLE COVERAGE, 2000 VERSUS 2010	134
FIGURE 2-20: UNITED STATES EMPLOYER-SPONSORED HEALTH INSURANCE - AVERAGE ANNUAL HEALTH INSURANCE PREMIUM - EMPLOYER-EMPLOYEE CONTRIBUTIONS FOR FAMILY COVERAGE, 2000 VERSUS 2010.....	135
FIGURE 2-21: DISTRIBUTION OF HEALTH CARE PLAN ENROLMENT FOR COVERED EMPLOYEES IN THE UNITED STATES, 1988-2010	138
FIGURE 2-22: AVERAGE ANNUAL CONTRIBUTION (PREMIUM) FOR COVERED EMPLOYEES BY MANAGED HEALTH CARE PLAN TYPE IN THE UNITED STATES FOR SINGLE AND FAMILY COVERAGE, 2010.	140
FIGURE 3-1: PROBABILITY DENSITY FUNCTION FOR THE AVERAGE QUALITY M	163
FIGURE 3-2: EQUILIBRIUM WITH PERFECT INFORMATION (ROTHSCHILD & STIGLITZ, 1976)	189
FIGURE 3-3: POOLING EQUILIBRIUM WITH IMPERFECT INFORMATION (ROTHSCHILD & STIGLITZ, 1976)	202
FIGURE 3-4: TRADE-OFF BETWEEN WEALTH IN THE NO ACCIDENT STATE AND WEALTH IN THE ACCIDENT STATE FOR A LOW-RISK INDIVIDUAL.....	203
FIGURE 3-5: TRADE-OFF BETWEEN WEALTH IN THE NO ACCIDENT STATE AND WEALTH IN THE ACCIDENT STATE FOR A HIGH-RISK INDIVIDUAL	204
FIGURE 3-6: SEPARATING EQUILIBRIUM WITH IMPERFECT INFORMATION (ROTHSCHILD & STIGLITZ, 1976).....	212
FIGURE 5-1: ADAMS AND TOWER (1994: 165)'S PUBLIC INTEREST THEORY – CONCEPTUAL MODEL.....	276
FIGURE 5-2: ADAMS AND TOWER (1994: 168)'S POLITICAL SCIENTIST VIEW OF CAPTURE THEORY – CONCEPTUAL MODEL	291
FIGURE 5-3: ADAMS AND TOWER (1994: 171)'S ECONOMIC THEORY OF REGULATION – CONCEPTUAL MODEL	293

LIST OF EQUATIONS

EQUATION 2-1: GENERAL FORMULA FOR THE DETERMINATION OF THE SUBSIDY PER MEDICAL SCHEME BENEFICIARY UNDER THE REF FRAMEWORK (GROBLER, THERON & COOPER, 2003)	54
EQUATION 2-2: GENERAL FORMULA FOR THE HERFINDAHL-HIRSCHMAN INDEX (HHI) (HIRSCHMAN, 1964; BROWN & WARREN-BOULTON, 1988)	79
EQUATION 3-1: BUYERS' WEIGHTED-AVERAGE VALUATION OF USED CARS ONCE A SINGLE MARKET EMERGES	160
EQUATION 3-2: UTILITY FUNCTION OF GROUP ONE (AKERLOF, 1970)	161
EQUATION 3-3: UTILITY FUNCTION OF GROUP TWO (AKERLOF, 1970)	162
EQUATION 3-4: DEMAND FOR USED CARS BY GROUP ONE (AKERLOF, 1970)	163
EQUATION 3-5: DEMAND FOR USED CARS BY GROUP TWO (AKERLOF, 1970)	164
EQUATION 3-6: TOTAL DEMAND FOR USED CARS BY BOTH GROUPS (AKERLOF, 1970)	166
EQUATION 3-7: TOTAL DEMAND FOR USED CARS BY BOTH GROUPS SUBSTITUTING FOR THE $E\mu$	167
EQUATION 3-8: INDIVIDUAL'S EXPECTED UTILITY FUNCTION (ROTHSCHILD & STIGLITZ, 1976)	179
EQUATION 3-9: INDIVIDUAL'S EXPECTED UTILITY FUNCTION UNDER INSURANCE (ROTHSCHILD & STIGLITZ, 1976)	179
EQUATION 3-10: INDIVIDUAL'S EXPECTED UTILITY FUNCTION UNDER NO INSURANCE (ROTHSCHILD & STIGLITZ, 1976)	180
EQUATION 3-11: INDIVIDUAL'S EXPECTED UTILITY FUNCTION UNDER INSURANCE (WITH TWO RISK CLASSES)	181
EQUATION 3-12: INDIVIDUAL'S EXPECTED UTILITY FUNCTION UNDER NO INSURANCE (WITH TWO RISK CLASSES)	182
EQUATION 3-13: INSURER'S EXPECTED PROFIT FUNCTION FOR EACH α SOLD (ROTHSCHILD & STIGLITZ, 1976)	182
EQUATION 3-14: INSURER'S EXPECTED PROFIT FUNCTION UNDER COMPETITIVE CONDITIONS – EXPECTED PROFIT CONVERGES TOWARDS ZERO (ROTHSCHILD & STIGLITZ, 1976)	190
EQUATION 3-15: REGRESSION SPECIFICATION FOR TESTING THE RELATIONSHIP BETWEEN RISK AND COVERAGE (COHEN & SIEGELMAN, 2010)	223
EQUATION 3-16: COVERAGE SPECIFICATION UNDER CHIAPPORI AND SALANIÉ (1997, 2000)'S BIVARIATE REGRESSION MODEL (COHEN & SIEGELMAN, 2010)	224
EQUATION 3-17: RISK SPECIFICATION UNDER CHIAPPORI AND SALANIÉ (1997, 2000)'S BIVARIATE REGRESSION MODEL (COHEN & SIEGELMAN, 2010)	224
EQUATION 4-1: BIAS ADJUSTMENT TO INFLATION-ADJUSTED NET ANNUAL CLAIMS INCURRED	241
EQUATION 4-2: BIAS ADJUSTMENT TO INFLATION-ADJUSTED NET ANNUAL CONTRIBUTION INCOME	241
EQUATION 4-3: BIAS ADJUSTMENT TO INFLATION-ADJUSTED GROSS ANNUAL ADMINISTRATION EXPENSES	241
EQUATION 4-4: STANDARD SPECIFICATION UNDER A PANEL DATA ESTIMATION FRAMEWORK (BALTAGI, 2008)	243
EQUATION 4-5: ONE-WAY ERROR COMPONENT MODEL FOR THE DISTURBANCE TERM (BALTAGI, 2008)	243
EQUATION 4-6: LEVIN ET AL. (2002)'S INITIAL SPECIFICATION TO TEST FOR UNIT ROOTS IN PANEL DATA STUDIES (BALTAGI, 2008)	245
EQUATION 4-7: FISHER'S INVERSE CHI-SQUARED TEST STATISTIC TO TEST FOR UNIT ROOTS IN PANEL DATA (MADDALA & WU, 1999; CHOI, 2001; BALTAGI, 2008)	246
EQUATION 4-8: FIXED-EFFECTS PANEL DATA SPECIFICATION TO TEST FOR A POSITIVE COVERAGE-RISK CORRELATION ..	249
EQUATION 4-9: RANDOM-EFFECTS PANEL DATA SPECIFICATION TO TEST FOR A POSITIVE COVERAGE-RISK CORRELATION	249
EQUATION 4-10: FIXED-EFFECTS ESTIMATION RESULTS	256
EQUATION 4-11: RANDOM-EFFECTS ESTIMATION RESULTS	258
EQUATION 4-12: EXPECTED LOSS COST FOR INDIVIDUAL i UNDER INSURANCE (COHEN & SIEGELMAN, 2010)	262
EQUATION 5-1: EXPECTED NUMBER OF VOTES FOR THE LEGISLATOR (PELTZMAN, 1976)	300
EQUATION 5-2: PROBABILITY OF SUPPORT (PELTZMAN, 1976)	301
EQUATION 5-3: TOTAL AMOUNT TRANSFERRED TO THE BENEFICIARY GROUP (PELTZMAN, 1976)	301
EQUATION 5-4: PROBABILITY OF OPPOSITION (PELTZMAN, 1976)	302

GLOSSARY OF TERMS

Derived from Folland, Goodman and Stano (2010: 568-574) unless otherwise stated.

Actuarially Fair: Insurance under which the expected pay-outs (claims) are equal to the premiums paid by policyholders.

Adverse Selection: refers to the situation whereby high-risk individuals demonstrate a greater propensity to purchase insurance than low-risk individuals (Rejda, 2005). Adverse selection arises in insurance markets when the purchaser of insurance has more information about either the probability of a loss or the distribution of the magnitude of the loss should the loss event occur, than the insurance company.

Capitation: under managed care health plans, a fixed amount is paid to a healthcare provider per enrollee over a given time period regardless of the amount of healthcare services rendered.

Chronic Disease List (CDL): a list of chronic diseases, outlined in the Regulations to the Medical Schemes Act No. 131 of 1998, that medical schemes are required to cover in full; including the diagnosis, treatment and medication prescribed with these chronic diseases (McLeod, Matisonn, Fourie, Grobler, Mynhardt, & Marx, 2004).

Claims Review: the procedure by which an enrollee's healthcare claims are reviewed prior to reimbursement. The objective is to establish the medical necessity for the healthcare services to be provided to the enrollee and to ensure that the cost of any healthcare services provided to the enrollee is not excessive.

Community Rating: requires that medical schemes offer cover for healthcare expenditure at effectively the same price to a heterogeneous group of members irrespective of their risk profile (health status) (Neuhaus, 1995). In other words, medical underwriting is abandoned in favour of the application of pre-defined criteria, which are often not directly related to risk.

Co-payment: the portion of a claim in relation to healthcare expenditure that a medical scheme member is required to pay out of his or her own pocket.

Council for Medical Schemes: the statutory body responsible for medical schemes supervision in South Africa. Section 7 of the Medical Schemes Act No. 131 of 1998 outlines the statutory functions and duties of the Council for Medical Schemes.

Gatekeeper: a primary healthcare provider, usually a general practitioner, who is responsible for overseeing and coordinating all aspects of a patient's healthcare.

Fee-for-Service (FFS): A method of payment under which the healthcare provider is paid for each medical procedure that is provided to the medical scheme beneficiary.

Health Maintenance Organisation (HMO): a form of managed care organisation that provides enrolees with relatively comprehensive healthcare services, rather than reimbursements for healthcare expenditure. Therefore, for a fixed (or capitation) fee, enrolees face few out-of-pocket expenses and are covered for most inpatient and outpatient treatment, which includes medical consultation (diagnosis), surgery and prescribed medication.

Herfindahl-Hirschman Index (HHI): A measure of market concentration that incorporates the market share of the largest firms within an industry or sector. The HHI is defined as the sum of the squares of the market shares of the fifty largest firms within an industry, where the market share is expressed as a proportion of the total market share.

Independent Practice Association (IPA): Healthcare providers who operate within independent, sole or small group practices that contract out their healthcare services to a health maintenance organisation (HMO).

Managed Care: Managed care refers to an organised delivery system whereby there is a network of organisations (hospitals, physicians, clinics and other healthcare providers such as hospices) that provides or arranges to provide a coordinated continuum of services (from primary healthcare to emergency surgery) to a defined group of individuals. The managed care system is held clinically (treatment) and fiscally (financially) accountable for the outcomes and/or health status of its members (Reekie, 1999).

Medical Aid Schemes: are akin to what nowadays one would refer to simply as medical schemes. In return for a regular contribution fee, medical aid schemes pay healthcare providers on a fee-for-service basis based upon a pre-determined or pre-negotiated tariff schedule. Historically, members of medical aid schemes were permitted to select their own healthcare providers.

Medical Benefit Schemes: Members of medical benefit schemes are restricted to a panel of healthcare providers. In return for a regular contribution fee, medical benefit schemes pay healthcare providers either on a salaried basis or on a capitation basis (a pre-determined fixed monthly or annual fee based upon the number of scheme members and their dependents receiving treatment from the relevant healthcare provider). Medical benefit schemes are thus more akin to managed care type organisations.

Medical Savings Account (MSA): A member of a healthcare plan with a Medical Savings Account (MSA) pays a monthly contribution fee to the health insurer that consists of both a traditional health insurance (risk) premium and a fixed amount that is set aside ‘within’ the person’s medical savings account. Funds that accumulate within the MSA are designed to cover day-to-day healthcare expenses.

Medical Scheme Business: The Medical Schemes Act No. 131 of 1998 defines the ‘business of a medical scheme’ as the business of undertaking an obligation to reimburse healthcare providers for services received by a member in return for a contribution fee.

Moral Hazard: refers to the tendency of insurance coverage to alter the behaviour of the insured in such a way that it may change the probability of loss, upon which the insurance company has relied in the determination of the appropriate premium and extent of coverage to offer to the insured (Shavell, 1979).

National Health Insurance (NHI): is the proposed South African government-sponsored national health insurance programme whereby a system of compulsory contributions (funded via multiple sources including additional taxation) is pooled into a designated fund, from which reimbursements are made to both public and private healthcare providers such that the population-at-large receives prescribed healthcare benefits free at the point of service.

Open Enrolment: Under the principle of open enrolment, open medical schemes have to accept any individual who wishes to join their medical scheme and are required to charge them according to the standard schedule of contribution rates offered to current members.

Open Medical Schemes: are medical schemes that permit new members from the public to join the scheme, subject to a mandatory waiting period and possible penalties.

Preferred Provider Organisation (PPO): An arrangement under which an enrolee is given appropriate financial incentives, such as no co-payments, to seek healthcare from selected healthcare providers with which the payer has contracted.

Prescribed Minimum Benefits (PMBs): The Prescribed Minimum Benefits (PMBs) set a minimum package of healthcare benefits that must be offered by all medical schemes. Annexure A to the Regulations of the Medical Schemes Act No. 131 of 1998 defines the PMBs in terms of some 270 diagnosis-treatment pairs.

Restricted Membership Schemes: are medical schemes that only permit individuals that meet the specified criteria of the scheme to become members. The basis of the restriction can only be employment or former employment in a profession, trade, industry or calling, or by a particular employer or class of employer.

Risk Equalisation Fund (REF): receives funds from medical schemes with lower risk profiles and pays out funds to medical schemes with a higher risk profile (McLeod et al., 2004). The REF serves as a mechanism that equalises contribution schedules (or tables) across medical schemes.

Utilisation Review: the procedure by which an enrolee's utilisation of healthcare services as well as their appropriateness, and quality of care are reviewed and controlled.

1 INTRODUCTION

Regulation is widely considered to be a necessary intervention by the government to address market failures and enhance social welfare. According to Pigou (1932), the intervention ought to be based upon the identification of market failures requiring the government to intervene through a policy aligned to correct these failures. Regulation may also be required to protect property rights. Thus, it is not a response to correct some form of market failure or imperfection, but rather a mechanism to entrench certain *natural rights*. Another important reason for regulation to be enacted is as a framework of codification. It codifies the rules and regulations that individuals and firms have to operate within, defining appropriate (prudent) behaviour and activities that participants may or may not undertake. Regulatory capture occurs when an interest group within a market uses its influence or resources to acquire a favourable regulatory decision, or for that matter regulatory indecision. It may be possible for interest groups to acquire a series of regulations or an entire regulatory framework. This is undertaken so that the group can extract a benefit for themselves rather than benefit society as a whole.

The twenty-eighth President of the United States, Woodrow Wilson, described the regulatory capture process coarsely as:

"If the government is to tell big businessmen how to run their business, then don't you see that big businessmen have to get closer to the government even than they are now? Don't you see that they must capture the government, in order not to be restrained too much by it? Must capture the government? They have already captured it" (Wilson, 1913: 201-2).

Private financing of healthcare expenditure in South Africa has traditionally been provided through medical schemes. These institutions have effectively provided insurance cover for healthcare expenditures to a select minority of South Africans that could afford cover in the absence of a coordinated public sector health insurance

system (Reekie, 1999). The primary function of a medical scheme is to provide members with protection against the healthcare expenditure that they (or indeed their dependents) may incur due to the diagnosis, treatment or long-term care of a medical condition (incident) that can include illness, bodily injury, congenital disease or hereditary condition from which the member suffers a deterioration in their health or general well-being. The member is required to compensate the medical scheme through a monthly contribution paid over to the scheme timeously and in return the member can expect the medical scheme to reimburse healthcare providers for the healthcare expenditure that the member or their dependants incur according to a pre-defined schedule of benefits.

Medical schemes in South Africa are currently regulated by the Medical Schemes Act No. 131 of 1998,¹ and they operate on a not-for-profit basis. Several entities that interact with medical scheme, healthcare providers and members do operate under the profit motive. These entities provide administration, marketing, managed care; consulting and advisory services to medical schemes (McLeod & Ramjee, 2007). It is precisely the high-profile medical scheme administrators that tend to attract much of the attention concerning the alleged exploitation of medical scheme members through higher contribution rates and lower benefit coverage levels. While medical scheme administrators do not directly charge members for health insurance cover, administrators indirectly extract a fee from members through the fees that they charge medical schemes for their services.

According to Pearmain (2000), private financing of healthcare expenditure in South Africa is not a new phenomenon. De Beers Consolidated Mines established a medical scheme for their employees, in as early as 1889 (Pearmain, 2000). After World War II, medical aid schemes proliferated and from 1956, medical aid schemes (including medical benefit schemes) had to register as *friendly societies* in terms of the Friendly Societies Act No. 25 of 1956. By 1960, there were 169 schemes covering approximately 1.5 million individuals. One might anticipate that the number of

¹ The Medical Schemes Act No. 131 of 1998 refers to the original Act, subsequent amendments and regulations.

medical schemes and their members (and beneficiaries) would have risen sharply over the past five decades. However, surprisingly quite the opposite has taken place.

At the time when Reekie (1999) concluded his examination of the medical scheme industry in South Africa, the number of beneficiaries (principal members and dependents) covered by medical schemes had increased to just under 7 million individuals whilst the number of medical schemes (open and restricted schemes) had fallen to 160. By the end of 2009, the total number of beneficiaries covered by medical schemes stood at 8,068,505 individuals. As a proportion of the total South African population it represents approximately 16% of all individuals in the country. The number of registered medical schemes had declined even further to 110 as at 31 December 2009. One of the key objectives of the Medical Schemes Act No. 131 of 1998 was to increase access to affordable private healthcare through the principles of *community rating* and *open enrolment*.² However, the number of individuals covered by medical schemes has only risen by around 18.15 percent over the entire ten-year period ending 2009. This translates into a compound per annum growth rate of 1.4 percent.

Reekie (1999) describes the Medical Schemes Act No. 131 of 1998 as an indictment by the South African government of the manner in which the local medical schemes industry had responded to the legislative framework prior to the implementation of the Medical Schemes Act No. 131 of 1998. More specifically, how the medical scheme industry had responded to the Medical Schemes Amendment Act No. 19 of 1989 and the Medical Schemes Amendment Act No. 23 of 1993. Before the 1990s, medical schemes reported frequent underwriting losses, which meant that the industry was unable to maintain prudential solvency ratios. The author argues that a financial turnaround was only possible due to the structural changes that followed the 1989 and

² In terms of the Medical Schemes Act No. 131 of 1998, *community rating* requires medical schemes to charge contributions that are differentiated only on the basis of income and number of dependants, and not on the age or the health status of the member. *Open enrolment* makes it compulsory for every open scheme to accept all eligible applicants.

1993 Medical Schemes Amendment Acts.³ Doherty and McLeod (2003) claim that these amendments were in response to calls from the medical scheme industry to free it from restrictions that the industry claimed limited its ability to deal with cost escalation. McIntyre, Bloom, Doherty and Brijlal (1995) point out that between 1982 and 1992, medical scheme contributions as a proportion of the average monthly salary, had risen steadily from 7.1 percent to 15.2 percent.⁴

In South Africa, the process of deregulation that followed the 1989 and 1993 Medical Schemes Amendment Acts, permitted medical schemes to engage in opportunistic behaviour, through contribution loadings and the restructuring of benefit plans, to attract the young and healthy. By redesigning the menu of medical cover options it effectively reduced benefits for the elderly and unhealthy segment of the population. Senior citizens found it practically impossible and unaffordable to maintain the level of cover that they had previously enjoyed. In addition, policies were designed to actively encourage a healthy lifestyle by offering incentives to members to exercise regularly and refrain from unhealthy activities. Thus, government correctly identified that senior citizens and the unhealthy had seen their benefits greatly reduced as a direct result of the deregulation that took place during the 1990s. However, the deterioration in coverage and to all intents and purposes higher contribution rates for the elderly and unhealthy was misinterpreted by the South African government as a manifestation of adverse selection.

Under the prior regulatory environment, it was argued that low-risk individuals (the young and healthy) were essentially driving out the high-risk individuals (the elderly and unhealthy) from the market. Through the mechanism of risk rating, medical schemes would offer lower premiums to low-risk individuals, while requiring those who are high risk, such as the elderly or unhealthy, to pay significantly higher premiums. The result is a fragmentation of members into ‘homogenous’ risk pools and the menu of policies (options) would be differentiated across the various risk

³ According to Reekie (1999), the application of sound insurance and risk management principles only became a reality once *managed care* became a legal possibility.

⁴ This occurred while salaries in real terms remained fairly constant.

classes. Consequently, health insurance cover became extremely expensive for those who were expected to need it the most. From the South African government's perspective, the prior regulatory framework translated into an increasing number of the elderly and unhealthy requiring healthcare services from the public healthcare system.

Government's argument that adverse selection was at the heart of the failure of the prior regulatory framework to provide access to affordable private healthcare to those who required it most is fallacious (Pearmain, 2000; Reekie, 1999). The observation that risk rating induces adverse selection is erroneous. Risk rating is a mechanism that reduces information asymmetry and therefore mitigates the negative consequences of adverse selection. Akerlof (1970)'s *Lemons Principle* describes the situation whereby adverse selection occurs within the automobile market. In the presence of quality uncertainty, 'bad' cars (or lemons) drive out the good cars (Akerlof, 1970). In the insurance context, this would translate into the fact that in the presence of information asymmetry, high risks drive out low risks. Rothschild and Stiglitz (1976) demonstrate that in the presence of imperfect information, equilibrium may not exist within a competitive insurance market or at least equilibrium may have strange properties. The theory suggests that in competitive insurance markets adverse selection results in risk pools forming and if these are unable to form, the pervasive presence of adverse selection may result in the market (or parts of it) not forming.⁵ From Rothschild and Stiglitz (1976)'s analysis it is possible to infer that when medical schemes are prohibited from using risk-related characteristics to assess the risk profile of an individual and thereby charge differentiated risk-based contribution rates, an enforced asymmetry of information ensues which prevents risk pools from forming. It is under these circumstances that the theory would predict adverse selection to be pervasive and lead to market failure.

⁵ This was a feature in the United States with so-called Redlined areas. Redlining is the practice of denying, or increasing the cost of insurance coverage to residents in certain predetermined areas. It describes the practice of delineating geographical areas where insurers would often deny coverage to individuals living within those areas. Redlining was often seen as a mechanism to racially profile segments of the population and subsequently the term has been applied to discrimination against any particular demographic profile irrespective of geographic location.

Therefore, government's rationale behind the repeal of the Medical Schemes Act No. 72 of 1967, the Medical Schemes Amendment Act No. 19 of 1989 and the Medical Schemes Amendment Act No. 23 of 1993, and subsequent promulgation of the Medical Schemes Act No. 131 of 1998 was to overcome adverse selection within the private health insurance sector is questionable. According to the government, the new legislation was sought to address the "destructive" practices of risk rating and reinstate the statutory minimum benefits (Pearmain, 2000; Reekie, 1999).⁶ Hence, the Medical Schemes Act No. 131 of 1998 prohibits risk-rating on the basis of age, prior or current health status. In addition, the legislation requires medical schemes to offer a minimum benefits package.⁷ In terms of the Medical Schemes Act No. 131 of 1998, the Prescribed Minimum Benefits (PMB) cover the healthcare costs related to the diagnosis, treatment and care of any emergency medical condition⁸, a limited set of 270 medical conditions and 25 chronic conditions.⁹

⁶ In addition, Section 68(1) of the Medical Schemes Act No. 131 of 1998 repealed the Medical Schemes Amendment Act No. 95 of 1969, Medical Schemes Amendment Act No. 49 of 1972, Medical Schemes Amendment Act No. 43 of 1975, Medical Schemes Amendment Act No. 51 of 1978, Medical Schemes Amendment Act No. 42 of 1980, Medical Schemes Amendment Act No. 72 of 1981 and Medical Schemes Amendment Act No. 59 of 1984.

⁷ According to the Council for Medical Schemes (2010) the Prescribed Minimum Benefits (PMB) is a set of defined benefits that all medical schemes have to offer to their members. It ensures that all medical scheme members and their dependants have access to certain minimum healthcare services with respect to a set of pre-defined conditions. Access to the Prescribed Minimum Benefits is regardless of the benefit option that members have selected. Beneficiaries must be covered in full for these conditions with no limits or co-payments. Medical schemes may insist on the use of a contracted network of providers and formularies (medication) in the treatment of the set of conditions defined under the Prescribed Minimum Benefits.

⁸ An emergency medical condition refers to the sudden and unexpected onset of a medical condition that requires immediate medical treatment and/or an operation. If the treatment is not administered timeously, the emergency could result in weakened bodily functions, serious and permanent damage to organs, limbs or other body parts, or ultimately death (Council for Medical Schemes, 2008).

⁹ In Annexure A of the Regulations to the Medical Schemes Act No. 131 of 1998 a list of conditions identified as Prescribed Minimum Benefits are presented in the form of Diagnosis and Treatment Pairs (DTPs). The Chronic Disease List (CDL) specifies the 25 chronic diseases. Medical schemes are required to cover the healthcare costs associated with both the medication and treatment of these chronic diseases (Council for Medical Schemes, 2008).

1.1. Demand for Health and Health Capital

It is important to acknowledge that individuals do not demand healthcare for the same reasons that they would demand other goods or services, such as food or public transport. Arrow (1963: 948) describes the demand for medical services as being “not steady in origin”, but rather “irregular and unpredictable”. Generally, individuals seek healthcare only as a response to the deterioration in their health (or health capital). It is rather health *per se* that is sought by the individual. Grossman (1972), Fuchs (1982) and Becker (1993, 1995 and 2006) have all used the theory of human capital to explain the demand for health and healthcare. Individuals are seen as rational economic agents that invest in themselves through education, skill enhancement and health to achieve greater returns from employment. These returns translate into the capacity to obtain goods and services that they need and desire. But further it allows individuals to access leisure time and once again invest in their health capital.

Grossman (1972) explains that individuals use health capital to produce ‘healthy days’, which in turn can be allocated to either work or leisure. Work enables individuals to generate income and this permits the purchase of goods and services, but also allows for further health investment. The greater allocation of resources (both time and funding) towards health investment augments health capital and this generates a greater number of ‘healthy days’. This self-enforcing cycle describes the demand for healthcare services. To increase the stock of health capital allows one to produce more time (‘healthy days’) that are available for work and leisure pursuits.

He summarises the important elements of health demand that differ from the traditional approach to demand:

1. Individuals seek health rather than healthcare *per se*. The demand for healthcare is a derived demand for an input to produce health.

2. Individuals do not simply transact within the market to purchase health, but also actively produce health. This is achieved through the purchase of healthcare inputs and allocating time to health-enhancing efforts.
3. Health can be seen as a capital good as it exists for more than one period. Although, it does depreciate over time, it does not do so instantaneously.
4. Health can be treated as both a consumption good and investment good. As a consumption good, health is sought since one derives utility from feeling healthy. In relation to an investment good, health is desired as it increases the number of 'healthy days' produced, thereby allowing one to work more and earn a greater income.

Health capital depreciates over time and although one is able to increase one's stock of health capital through health investment, eventually the rate at which health capital depreciates exceeds the rate at which one is able to augment health capital through health investment. The theory does not preclude the possibility that the rate at which a person's health stock depreciates can vary over time. For example, it is possible that during periods of great stress, a person's depreciation rate increases and then falls once the stressful situation has dissipated. Nevertheless, Grossman (1972)'s model suggests that as one ages, the depreciation rate has a greater propensity to increase, which causes the optimal level of health capital to fall. Higher depreciation rates increase the cost of holding health capital. Therefore, over time individuals have no choice but to allow their health capital to deteriorate to a minimum level, H_{min} , where an individual is no longer able to produce 'healthy days', thereby they are unable to sustain life and their life terminates (Grossman, 1972).

The phenomenon can be described as follows:

“Gross investment's life cycle profile would not, in general, simply mirror that of health capital...This follows because a rise in the rate of depreciation not only reduces the amount of health capital demanded by consumers but also reduces the amount of capital supplied to them by a given amount of gross investment” (Grossman, 1972: 238).

Extensions to Grossman (1972)'s model imply that people are more likely to increase their investment in health as they age, which confirms our general observations that the elderly demand more healthcare services than young individuals.

1.2. Role of Healthcare and Medicine

The private financing of healthcare is a more contemporary phenomenon when compared to other traditional lines of insurance such as marine or fire insurance. In South Africa, it was only after the Second World War that medical schemes became commonplace. This is partly due to the fact that the healthcare services that society has become accustomed to, including medical interventions such as effective medicines and modern surgical treatment, have only been readily available in the last 60 years or so. Global access to what could be referred to as 'effective' healthcare is in fact a recent occurrence. By implication, it suggests that the improvement in the health status¹⁰ of the general population would only be observable from 1950 onwards. What is more surprising is that research, such as that undertaken by McKeown, Brown and Record (1972) and McKeown (1976) suggest that historical downward trends in mortality rates in England and Wales cannot be explained by practitioner-provided healthcare.

The role of healthcare in society can be viewed as a production question. The production function describes the relationship between inputs and outputs. When examining the health production function, it is necessary to investigate the relationship between healthcare inputs and health. A primary concern for public policymakers and medical historians alike is what contribution has healthcare, and thus medical interventions, made to the overall health status of the population. To

¹⁰ Measuring health status is not as straightforward as one might believe. Various measures have been suggested, including the number of healthy days experienced by the population per capita or the reciprocal of a reverse indicator such as mortality or morbidity rates.

assess the impact upon health status, it is common to focus upon population growth to assess the success that medical interventions have had in lowering mortality. Population increases are the result of three factors; increased birth rates, reduced mortality or increased net-migration. In the case of England and Wales, the sudden rise in the population after 1750 can largely be attributed to a reduction in mortality. McKeown, Brown and Record (1972) and McKeown (1976) examine the historical trends in mortality and investigate specifically which diseases contribute to the patterns observed in mortality rates. Interestingly, airborne infectious diseases account for the largest single portion of mortality reduction. In particular, the decline in mortality rates related to the following diseases tuberculosis, bronchitis, pneumonia and influenza, can be viewed as the primary contributors to the overall fall in mortality (McKeown, 1976).

McKeown (1976) point out that many would believe that medical interventions are the main reasons for the observed decline in mortality rates with respect to these diseases; however, effective medical interventions were not available until much later in the period examined, well after the most significant decline in mortality had already been observed. Rather factors such as improved nutrition, public health measures and improvements in the environment in which people lived and worked, appear to be the principle reasons for the marked decline in mortality since 1750.

1.3. Chapter Outline

To test whether adverse selection is indeed a feature of the South African private health insurance sector and to refute government's claim that the current legislation governing medical scheme business was a necessary intervention to address the destructive consequences of adverse selection, the following chapters will be presented in the thesis.

Chapter Two provides a comprehensive review of the South African private health insurance sector. This includes the background into how medical schemes came about and particular, the evolution of the regulatory environment. Pearmain (2000) asserts that the deregulation that occurred during the 1990s in South Africa's private health insurance sector, allowed medical aid schemes to actively discourage membership by high-risk individuals. At the same time, government undertook an ambitious plan to provide free primary health care, at the point-of-service, to those who did not have access to private healthcare.¹¹ As a result there was an increased demand for public sector healthcare services by those that had fallen outside of the private-sector medical scheme net. With the promulgation of the Medical Schemes Act No. 131 of 1998, some may argue that the government acted in the public interest to arrest the benefit deterioration observed for the elderly and poorly. However, a more plausible explanation is that the South African government hastily sought to relinquish its social responsibility to provide healthcare to those citizens that were unable to obtain medical cover within the private sector (Reekie, 1999).

The Medical Schemes Act No. 131 of 1998 saw the introduction of the concepts of *community rating* and *open enrolment*. These concepts were part of an attempt by the South African government to undertake an ambitious social health engineering exercise, whereby a national health system is established through private sector medical schemes (Pearmain, 2000).¹² Discrimination on the basis of age, gender, medical history or current health status was prohibited under the new regulatory framework. Moreover, contribution rates can only be determined on the basis of income level and/or number of dependents. Access to medical schemes is ensured through the principle of open enrolment, whereby medical schemes may not exclude any applicants or their dependents for membership except under certain prescribed conditions. The Medical Schemes Act No. 131 of 1998 does provide for some protection against *anti-selection* by members through the imposition of waiting

¹¹ Access to free primary health care is offered at public hospitals and clinics.

¹² A national health system was not an outcome of the new regulatory framework, which suggests that industry interest groups were successful in protecting their interests. The failure to establish a national health system has mobilised politically-charged organisations, such as the Congress of South African Trade Unions (COSATU), to start advocating in 2009, the establishment of a National Health Insurance (NHI) system.

periods and late-joiner penalties under predefined circumstances. The chapter also includes an assessment of the medical schemes industry and specifically, considers the competitiveness of the market by examining the degree of concentration over the period 1993 to 2009.

Chapter Three examines the seminal contributions of Akerlof (1970), Spence (1973, 1974) and Rothschild and Stiglitz (1976) in relation to theory of adverse selection. In particular, the thesis will discuss how competitive insurance markets respond in the presence of imperfect information. This requires a detailed discussion of both Akerlof (1970)'s *Lemons Principle* and Rothschild and Stiglitz (1976)'s well-renowned model that explains how both insurers and insureds behave in the presence of information asymmetry. The chapter also includes a discussion of the primary indicator often utilised to empirically detect the presence of adverse selection, which involves detecting a positive coverage-risk correlation. A review of the empirical literature testing the theoretical predictions of adverse selection are presented, which highlight the contradictory evidence relating to the presence of adverse selection within selected insurance lines and markets.

Chapter Four outlines the empirical methodology to test for the presence of adverse selection. Testing for a positive coverage-risk relationship within South Africa's private health insurance sector would be the ideal avenue to detect the presence of adverse selection, as a positive coverage-risk correlation is a necessary condition for adverse selection to be present. However, due to data limitations, specifically the aggregated data constraint, an innovative empirical procedure is devised based upon Akerlof (1970)'s *Lemon Principle* and Rothschild and Stiglitz (1976)'s analysis to directly test for the presence of adverse selection in South Africa's medical scheme sector. The empirical results are then presented and discussed. In addition, within this chapter, several reasons why the absence of a positive coverage-risk correlation may be observed in insurance markets are examined. This is necessary to highlight the important aspects that may explain the absence of adverse selection within South Africa's private health insurance sector.

Chapter Five outlines the theories of economic regulation that explain the observed behaviour of government intervention. This is done by reviewing the theoretical literature in this area including Stigler (1971), Posner (1974), Priest (1993), Adams and Tower (1994), Laffont (1994) and Shleifer (2005). The four distinct schools of thought regarding regulatory capture will be discussed extensively.

Chapter Six provides a discussion into the findings of the thesis including an argument that provides an alternative explanation for the current legislation governing medical scheme business – namely, regulatory capture. The limitations of the research and recommendations for future research are also discussed in the concluding chapter.

1.4. Contribution of the Thesis

The South African government considered that the Medical Schemes Act No. 72 of 1967 and in particular, the subsequent amendments, inconsistent with the objective to provide universal access to affordable healthcare for all South Africans. Hence, a comprehensive reformulation of the legislation governing private health insurance sector was called for by the mid-1990s. Although the Medical Schemes Act No. 72 of 1967 had survived a number of amendments, it was finally repealed in 2000 with the enactment of the Medical Schemes Act No. 131 of 1998. The stated reforms, most especially *open enrolment* and *community rating*, were touted by the government as necessary to address the undesirable effects of adverse selection. It is precisely the contribution of this thesis in that it offers an alternative explanation as to why government reformulated the regulatory environment governing medical scheme business – namely, regulatory capture.

1.4.1 Objective of the Thesis

The thesis highlights the fact that the government had completely misinterpreted the concept of adverse selection, but importantly, the question was never raised whether or not in fact adverse selection is a feature of the South African private health insurance market. A significant body of empirical literature indicates that adverse selection may not be present in all insurance lines or markets.¹³ The key test utilised to detect adverse selection is to establish the presence of a positive coverage-risk relationship. The evidence has been mixed with a coverage-risk correlation present within some insurance markets but absent in others (Löfgren, Persson & Weibull, 2002). It is necessary to answer the question whether adverse selection is indeed a feature of the South African private health insurance market. This is achieved by empirically testing for the presence of adverse selection utilising a panel dataset of all registered medical schemes in South Africa over the period 1993 to 2009. Notwithstanding, the new regulatory framework, the South African medical scheme industry remains sufficiently competitive. Therefore, it is possible to utilise an innovative empirical procedure based upon Akerlof (1970)'s *Lemon Principle* and Rothschild and Stiglitz (1976)'s analysis to directly test for the presence of adverse selection in South Africa's medical scheme sector. If adverse selection is found to be absent then government's supposition that adverse selection (as a consequence of the deregulation that took place during the late 1980s and early 1990s) is responsible for the deterioration in medical scheme coverage for the elderly, unhealthy or poor would prove to be fallacious.

Once it is established that the ostensible reason for the current legislation does not stand up to scrutiny, the primary objective of this thesis is to offer a plausible alternative explanation for the reasoning behind the promulgation of the current legislation. The stated objective of the South African government in the provision of private healthcare is for certain members (young, healthy or wealthy) to subsidise others (elderly, unhealthy or poor). Cross-subsidisation within the private health

¹³ Refer to Chiappori and Salanié (2000), Cardon and Hendel (2001), Finkelstein and Poterba (2004), Cohen (2005), Saito (2006), and Fang, Keane and Silverman (2008).

insurance market as envisaged by the government is simply not amenable to the functioning of a competitive insurance market (Reekie, 1999). Therefore, the key question is whether this objective is inconsistent with the concept of in the public interest. A detailed account of the theories of economic regulation is presented to provide the framework for the exposition of an alternative explanation behind the implementation of the current legislation. It is the view of this thesis that the principal argument of Stigler (1971) is correct and therefore, regulatory capture is the more plausible explanation behind the implementation of the Medical Schemes Act No. 131 of 1998 and accompanying regulations.

2 SOUTH AFRICA'S PRIVATE HEALTH INSURANCE SECTOR

In this chapter a comprehensive review of the South African private health insurance sector is presented. An overview of the regulatory evolution governing medical scheme business is discussed. The Medical Schemes Act No. 131 of 1998 represented the South African government's attempt to reformulate the legislation so as to address what it identified as the deterioration in private health insurance coverage for the elderly, unhealthy and poor.

Also broadly discussed in this chapter is an assessment of the performance of the medical scheme market including an analysis of the competitiveness of the sector. Explanations are sought in relation to the higher healthcare costs observed since the implementation of the Medical Schemes Act No. 131 of 1998 in 2000.

2.1 Background

In South Africa, medical schemes are the principal vehicle for providing insurance cover against healthcare expenditure incurred within the private health sector (McLeod & Ramjee, 2007). A medical scheme provides members with protection against the healthcare expenditure that they (or their dependents) may incur due to the diagnosis, treatment or long-term care of a medical condition that can include illness, bodily injury, congenital or hereditary disease from which the member suffers a deterioration in their health or general well-being. The member is required to compensate the medical scheme by means of monthly contribution paid over to the scheme timeously and in return the member can expect the scheme to reimburse healthcare providers for the healthcare expenditure that the member or their dependents incur according to a predefined schedule of benefits. Medical schemes operate on a not-for-profit basis and are regulated in South Africa by the Medical

Schemes Act No. 131 of 1998. A medical scheme acts essentially as a mutual society that is managed by a board of trustees (McLeod & Ramjee, 2007).

The two broad types of medical schemes that operate within South Africa are open and restricted schemes. Open medical schemes permit new individuals from the general public to become members, subject to a mandatory waiting period and in certain cases late joiner penalties. Restricted membership schemes only permit individuals that meet the specified criteria of the scheme to become members. The basis of the restriction can only be employment or former employment in a profession, trade, industry or calling, or by a particular employer or class of employer.

The Council for Medical Schemes is the statutory body responsible for medical schemes supervision in South Africa. Section 7 of the Medical Schemes Act No. 131 of 1998 outlines the functions of the Council for Medical Schemes. Pearmain (2000: 186) points out that these functions are “...far more purposeful and consumer-oriented” than the Council’s functions under the prior legislation. Pearmain (2000: 186) describes that under the Medical Schemes Act No. 131 of 1998, the Council for Medical Schemes is required to:

“Protect the interests of [medical scheme] members at all times;

Control and co-ordinate the functioning of medical schemes in a manner that is complementary with the national health policy;

Make recommendations to the Minister [of Health] on criteria for the measurement of quality and outcomes of the relevant health services provided for by medical schemes, and such other services as the Council may from time to time determine;

Investigate complaints and settle disputes in relation to the affairs of medical schemes as provided for in the [Medical Schemes] Act [No. 131 of 1998];

Collect and disseminate information about private health care;

Make rules, not inconsistent with the provisions of the [Medical Schemes] Act [No. 131 of 1998], for the purpose of the performance of its functions and the exercise of the Council's powers;

Advise the Minister [of Health] on any matter concerning medical schemes; and

Perform any other functions conferred on the Council by the Minister [of Health] or by the [Medical Schemes] Act [No. 131 of 1998]”.

The Council for Medical Schemes is required to measure the delivery, quality and outcomes of healthcare services that medical scheme members receive from private healthcare providers (Pearmain, 2000). The Council for Medical Schemes has the authority to collect and disseminate information concerning the operation of both open and restricted medical schemes in the country. In addition, it is able to assess the quality of private healthcare provision to medical scheme members.

Whilst short and long-term insurers are permitted to sell health insurance policies, the demarcation between medical scheme business and health insurance sold by short and long-term insurers was clearly redefined within the Long-Term Insurance Act No. 52 of 1998, Short-Term Insurance Act No. 53 of 1998 and Medical Schemes Act No. 131 of 1998. Health insurance policies are not allowed to indemnify policyholders against actual healthcare expenditure, but must rather offer a defined sum assured in advance of any healthcare provision. Further, health insurance policies may not directly reimburse healthcare providers. Even though there is this distinction between medical scheme business and health insurance in law, for the most part medical scheme business is nevertheless by its nature also health insurance. After all medical schemes are indemnifying members for uncertain future healthcare expenditure in return for a contribution fee (premium). Unlike medical scheme policies, health insurance policies sold by short and long-term insurers in South Africa are supervised by the Financial Services Board (FSB).

The legislative framework that has governed medical scheme business in South Africa can be segmented into three distinct regimes; namely, the pre-deregulation, deregulation and current regulatory periods.

2.1.1 Medical Schemes Prior to the Deregulation

De Beers Consolidated Mines first introduced the forerunner of today's medical schemes in 1889 for its employees (Pearmain, 2000). After the Second World War, medical aid schemes proliferated and from 1956, medical aid schemes (including medical benefit schemes) had to register as *friendly societies* in terms of the Friendly Societies Act No. 25 of 1956. Medical aid schemes were akin to what nowadays one would refer to simply as medical schemes. In return for a regular contribution fee, medical aid schemes paid healthcare providers on a fee-for-service basis based upon a pre-determined or pre-negotiated tariff schedule. Members of medical aid schemes were allowed to choose their own healthcare providers. Medical benefit schemes did not operate on a fee-for-service basis, but rather members were restricted to a panel of healthcare providers that were paid by the medical benefit scheme either on a salaried basis or on a capitation basis.¹⁴ Medical benefit schemes were thus more akin to managed care type organisations. Reekie (1999) argues that medical schemes provided insurance cover for healthcare expenditures to a select minority of South Africans that could afford cover in the absence of a coordinated public sector health insurance system.

The first key piece of legislation that was introduced to regulate and coordinate the functioning of the medical aid schemes and medical benefit schemes was the Medical Schemes Act No. 72 of 1967. By 1967, there were 256 medical schemes operating in South Africa, covering over 1.87 million individuals. Between 1967 and 1975, medical schemes were concurrently regulated by the Ministries of Finance and Health

¹⁴ Under the capitation basis, healthcare providers received a pre-determined fixed monthly or annual fee based upon the number of benefit scheme members and their dependants to whom they provided treatment.

in terms of the Friendly Societies Act No. 25 of 1956 and the Medical Schemes Act No. 72 of 1967 respectively. However, through the enactment of the Medical Schemes Amendment Act No. 43 of 1975, medical schemes became only subject to the regulations provided within the Medical Schemes Act No. 72 of 1967 (as amended). The Medical Schemes Act No. 72 of 1967 survived a number of amendments, the most important of which included the Medical Schemes Amendment Act No. 19 of 1989 and the Medical Schemes Amendment Act No. 23 of 1993. These amendments signalled a significant shift in direction by the government at the time, whereby the medical scheme business was deregulated.

Prior to the deregulation that took place during the late 1980s and early 1990s, medical schemes operated largely on a pay-as-you-go basis. There was little or no provision made for the funding of future events beyond the financial accounting year (Reekie, 1999). Reekie (1999) indicates that prior to 1993, contribution fees were usually determined on a historic claims experience basis. Thus, medical schemes failed to account for future liabilities that arose from certain medical conditions over short-term horizons, or even to consider the fact that the elderly once retired would expect to pay lower contribution fees whilst submitting more frequent and/or larger claims to the medical scheme.

Reekie (1999) points out that even though contribution revenue had risen between the years of 1982 and 1989, it was insufficient to meet the higher costs experienced by medical schemes. The result was annual underwriting losses that occurred on no fewer than eight occasions during the nine year period between 1982 and 1990. Hence, as he explains these underwriting losses were of great concern, especially since the cost increases could not be explained simply by the growth in the number of dependants or by the depreciation of the local currency as measured by general price inflation. It was suggested by Reekie (1999) that the problem lay with the regulatory environment that medical schemes operated within at the time whereby incentives were misplaced. Since healthcare providers received a statutory guaranteed fee-for-service from medical schemes and medical schemes were legally obligated to provide

full cover (first rand cover) to members, the regulatory framework diminished the incentives to contain costs (Reekie, 1999). As the author articulated the regulatory framework governing the private financing of healthcare expenditure encouraged the production and consumption of healthcare, rather than assisting the industry to economise upon scarce resources.

During the 1980s, the largest single item of expenditure for medical schemes was pharmaceuticals, which constituted approximately 30% of all reimbursements (Reekie, 1999). Almost all medicines dispensed to the private sector passed through a limited number of wholesalers to retail pharmacies. The author notes that price competition was rare as pharmacies were prohibited from advertising price differentials. In addition, the Pharmacy Act No.53 of 1976 stipulated that only pharmacists could own shares in businesses that operated as retail pharmacists. Hence, the expansion of retail pharmacies was severely restricted, since pharmacists could only gain access to additional capital by approaching other pharmacists. Hence, according to Reekie (1999: 3), medical schemes and effectively their members had to:

“...pay for a retail structure which offered the worst of all possible worlds, namely; a multiplicity of high-cost, small-scale retailers which, as a whole, operated as a government-protected monopoly”.

Despite the relatively high margins applied by pharmacists to medicines, it did not translate into significant profits for the retail pharmacies. The reason being that the turnover generated by retail pharmacies was simply too low due to the excessive number of pharmacies present within the market (Reekie, 1999). He suggests the retail pharmaceutical market's incentive to innovate was all but eliminated by the regulatory framework. Reekie (1999) points out that the system prevented the adoption of innovative retailing technologies, which would translate into large-scale retail pharmacy chain stores that would offer a more cost-effective service. This would include more affordable medication to medical scheme members as monopoly mark-ups would dissipate. The author believed that it was inevitable that monopolistic mark-ups in the pharmaceutical retail sector, reinforced by regulatory protection of distribution channels, would give rise to new distribution mechanisms that would

escape the regulatory net. At the beginning of the 1990s, medical practitioners took advantage of the exorbitant mark-ups in the pharmaceutical retail sector and began dispensing medication directly to patients.¹⁵

Before the reforms of the 1990s, medical schemes had to fend off a direct threat to their very existence – the emergence of new health insurance policies offered by insurance companies. Large-scale insurers recognised the restricted environment that medical schemes operated within and therefore identified an opportunity to offer a set of cost-effective health insurance policies that would not be subject to the legally-constrained framework governing medical schemes (Reekie, 1999). The dire situation that faced medical schemes towards the end of the 1980s prompted the government to reevaluate its regulatory framework with a view to radically reform the legislation to address the incentive mechanisms that had led to the observed cost escalation.

2.1.2 Deregulation during the 1990s

The Medical Schemes Amendment Act No. 19 of 1989 and the Medical Schemes Amendment Act No. 23 of 1993 was a direct response to calls by the medical schemes industry to free it from the controls that it claimed hampered its ability to address cost escalation (Doherty & McLeod, 2003). The Pharmacy Amendment Act No. 88 of 1997 also assisted in addressing the cost escalation observed with respect to medicines. Medicines had been a significant cost constituent in terms of the reimbursements made by medical schemes on behalf of their members. From Reekie (1999)'s perspective, the underwriting losses experienced during the 1980s were untenable and it was acknowledged by the government that at the time a drastic reformulation of the regulatory framework governing medical scheme business was required.

¹⁵ At the beginning of the 1980s almost 100% of private sector sales of medication passed through retail pharmacies, but by 1993 this had diminished to 41.3% in terms of value (Reekie, 1999).

According to Reekie (1999) the financial turnaround experienced by the medical schemes industry after 1993 can be attributed to the introduction of the principle of managed care. Under the Medical Schemes Amendment Act No. 19 of 1989, medical schemes were now able to apply the concept of risk-rating in their computation of contribution rates and benefit coverage options. This represented a key reform whereby medical schemes could align the expected claims experience of members (benefits that they would receive under the medical scheme policy) to the level of their contributions. As a consequence, medical schemes were permitted to charge higher contribution rates to those individuals who represented a higher risk to the scheme. In effect, it would eliminate the cross-subsidisation that had existed between the young and old, and between the healthy and unhealthy.

The Medical Schemes Amendment Act No. 23 of 1993 removed the statutory minimum benefits and guaranteed payment of claims. Thus, medical schemes were allowed to ration coverage or even exclude coverage completely for certain procedures. Furthermore, the Medical Schemes Amendment Act No. 23 of 1993 permitted medical schemes to provide healthcare services directly to their members through a wide spectrum of options. Reekie (1999) suggests that the fee-for-service based upon a member's free choice of healthcare provider remained an option at the one extreme, alternatively medical schemes could specify an approved list of healthcare providers (Preferred Provider Organisations – PPOs), enter into partnerships with healthcare providers (Health Maintenance Organisations – HMOs), or simply employ healthcare providers by establishing their own healthcare clinics, hospitals and pharmacies. The Medical Schemes Amendment Act No. 23 of 1993 was based upon revised legislation and practices in the United States (Reekie, 1999). He suggested that the competitive nature of health insurance providers in the United States generally ensured an efficient system of healthcare provision, with providers minimising costs to maximise their residual surplus.

The Pharmacy Amendment Act No. 88 of 1997 repealed the ban on the corporate ownership of pharmacies thereby opening up access to capital for the retail pharmaceutical industry that would permit the creation of larger retail outlets. This in turn would translate into lower medicine prices at the point-of-service. Reekie (1999) also highlights an important income redistributive mechanism that takes place between the public sector (government) and the private sector (medical schemes) with respect to the purchasing of pharmaceuticals. In South Africa, the public sector purchases pharmaceuticals through competitive tender, whilst the private sector's purchases of pharmaceuticals are initiated by individual prescribing healthcare practitioners. With approximately 8,068,505 individuals serviced by the private sector through medical scheme membership, at least four out of every five South Africans remain reliant upon the public sector for receipt of their prescribed medicines. The author points out that the pharmaceutical manufacturers sell roughly two-thirds of all medicines to the state and one-third to the private sector. In monetary terms, this proportion is reversed.

Pharmaceutical manufacturers simply apply price discrimination principles in response to varying price elasticities of demand. Individuals serviced by the private sector are generally in full employment and are less price sensitive than poorer individuals serviced by the state (Reekie, 1999). In South Africa, pharmaceutical manufacturers have merely responded to the observable price sensitivities by pricing at a lower level in the more price sensitive sector (Reekie, 1999). This is referred to as third-degree price discrimination or market segmentation. Through pursuing their own self-interest and attempting to maximise expected profits, pharmaceutical companies achieve a redistribution of income across the public and private sectors.

For price discrimination to be effective it requires that firms must be able to identify market segments according to their relative price elasticities and further be able to enforce such price discrimination (Samuelson & Marks, 2006). Samuelson and Marks (2006) consider the airline industry, where airlines routinely segment their product

offerings into various travel classes. Airlines charge customers with relatively inelastic demand (first and business-class passengers) significantly higher ticket prices, whilst offering discount ticket prices to customers with relatively elastic demand (economy-class passengers). Although product differentiation is clearly applicable, it does not adequately translate into the price differential observed. To discourage business travellers from purchasing discounted economy-class tickets, airlines often impose advance ticketing requirements or minimum stay requirements - conditions that would prove difficult for the average business traveller to adhere to (Samuelson & Marks, 2006).

Summary of Reforms and Risk Rating

A key objective of the reforms was to permit medical schemes to provide healthcare services themselves with no restrictions as to the required minimum benefit level that could be offered to members (Reekie, 1999). In addition, first rand cover was no longer a legal requirement (Reekie 1999). According to Pearmain (2000), the reforms had a direct impact upon medical scheme operations by:

- Abolishing compulsory direct payment to healthcare providers;
- Permitting medical schemes, at their discretion, to vary benefit levels and options;
- Removing the statutory status of the scale of benefits;
- Introducing risk-rating in the computation of contribution rates and coverage levels; and
- Letting medical schemes operate pharmacies, hospitals and other healthcare facilities.

The reforms and in particular the Medical Schemes Amendment Act No. 23 of 1993, effectively removed unnecessary regulation from the medical scheme industry and allowed a more competitive environment to flourish (Reekie, 1999). Pearmain (2000: 185) identifies that the deregulation:

“...allowed detailed and individual specific risk rating and many variations in both the level and structuring of benefits”.

Pearmain (2000) was less enthusiastic by the deregulation that took place during the 1990s as she viewed the new reforms as destructive and harmful to those who needed access to affordable private healthcare the most; namely, the elderly and those individuals in poor health. According to the author, the Medical Schemes Amendment Act No. 23 of 1993 Act had adverse effects for the elderly and ailing whereby the benefits offered to them by medical schemes diminished significantly.

Pearmain (2000) provides a useful explanation of risk rating in the context of the South African health insurance sector. According to the author, risk rating principles dictate that medical schemes would analyse the historical claims experience of relatively homogenous risk groups, and in addition, schemes would examine the medical history and current health status of members. New applicants to an open medical scheme would be assessed and classified according to the relevant risk group that best-described their risk characteristics. Based upon the risk classification procedure the appropriate contribution rate can be determined whereby the:

“standard or average contribution levied by [the medical scheme] is loaded...in proportion to the increased risk to the scheme represented by the group or individual applying for membership” (Pearmain, 2000: 185).

Moreover, she highlights that risk rating can be undertaken on a continual basis instead of a single point in time – when the member is admitted into the medical scheme. An example by Pearmain (2000) illustrates; if an open scheme admits a group of employees from a single employer and the risk classification procedure is based upon the risk profile of the group as a whole, then it is possible for the medical scheme to re-examine the group’s claims experience dynamically to determine if the “...health profile of the group has improved or deteriorated [over] time” (Pearmain, 2000: 185). If the health (risk profile) of the group deteriorates then the medical scheme can conceivably adjust contributions upwards to reflect the increase in risk (Pearmain, 2000). Thus, Pearmain (2000: 185) emphasises that it is possible where risk rating is applied within an open scheme:

“...different employer groupings might end up paying very different levels of contributions for the same level of benefits depending [upon] their risk profiles”.

Medical schemes responded to the deregulation by structuring their benefits in such a way, as to attract young and healthy individuals, and also in some cases actively discouraging membership by high-risk individuals through benefit restructuring and contribution loadings on the basis of risk profile (Pearmain, 2000). Consequently, Pearmain (2000) argues that this form of risk rating would result in an increase in the demand on the public healthcare service by the elderly and those individuals who were in poor health since they could no longer access affordable private healthcare cover. She contends that risk rating diminishes or even eliminates the operation of cross-subsidization and the consequence of this is that low-risk individuals benefit from lower premiums, whilst premiums for high-risk individuals (elderly and unhealthy) rise in line with their corresponding risk status. The author reiterates the concern that under these circumstances health insurance cover becomes unaffordable for those who are precisely in need of such cover. However commendable Pearmain (2000)'s disparaging comments concerning risk rating may be in a social context, it is easy to overlook an important objective of risk rating; namely, to align the contributions charged to members and the benefits received by members.

Pearmain (2000) does not adequately consider an important feature of insurance markets, namely; adverse selection, when discussing the application of risk rating. The principle of adverse selection is inextricably linked to the application of risk rating within the health insurance market. Key to the discussion of adverse selection is the notion that the premiums charged to an individual (or homogenous risk group that they belong to) should reflect their risk status – and thereby, the benefits that members are expected to receive under a medical scheme policy. If this is not the case, individuals who end up subsidising other individuals that have a higher propensity to require healthcare services than themselves, will be more likely to opt out of insurance – the old adage where the “lemons” drive out the good (Akerlof, 1970).

2.1.3 Rationale behind the enactment of the Medical Schemes Act No. 131 of 1998

A serious consequence of the *apartheid* government's reign of oppression was a legacy of inequality. If one considers, measures of income inequality, such as the Gini coefficient,¹⁶ it is apparent that there is a significant disparity between the wealthy and poor in South Africa. By 1993, South Africa's Gini coefficient was 0.5933 (World Bank, 2009).¹⁷ Most developed European countries, Australia, Japan and Canada all have Gini coefficients ranging between 0.25 and 0.36 (Firebaugh, 2003; World Bank, 2009). South Africa's real Gross Domestic Product (GDP) per capita¹⁸ had risen only marginally from \$2,984 in 1960 to \$3,217 by 1993 (United Nations Development Programme, 1996). If one considers more developed countries, South Africa's real GDP per capita was considerably lower - Australia (\$18,530), United Kingdom (\$17,230), United States (\$24,680) and the Republic of Ireland (\$15,120) (United Nations Development Programme, 1996). In 1993, South Africa also compared unfavourably against other emerging economies; such as Brazil (\$5,500), Poland (\$4,702), Russian Federation (\$4,760) and Botswana (\$5,220) (United Nations Development Programme, 1996).

The patterns of economic inequality in South Africa were also mirrored in relation to the patterns observed with respect to healthcare inequality (McIntyre, Thiede, Nkosi, Mutyambizi, Castilo-Riquelme, Gilson, Erasmus & Goudge, 2007). Life expectancy at birth in South Africa continued to lag behind most developed economies during the

¹⁶ The Gini coefficient, as formulated by Corrado Gini and published in 1912, is a measure of the inequality of a distribution (Gini, 1912, 1937). The Gini coefficient can be calculated by considering a two-dimensional graphical representation of the Lorenz curve, which plots the proportion of the total income of a country's population (*y-axis*) that is cumulatively earned by the bottom *x%* of the country's population (Firebaugh, 2003). A 45-degree line denotes perfect equality, and thus the coefficient can be determined by considering the ratio of the area between the 45-degree line and the Lorenz curve. The Gini coefficient can take on a value between 0 and 1 (in some cases, the coefficient is multiplied by 100, to scale the coefficient to lie between 0 and 100). A low Gini coefficient indicates a more equitable income distribution, with a value of 0 corresponding to perfect equality.

¹⁷ Interestingly, the disparity between the wealthy and poor in South Africa, has deteriorated further since the first democratic elections held in 1994. By 2009, South Africa's Gini coefficient stood at 0.679 (EuropeAid Development and Cooperation, 2010).

¹⁸ Real GDP per capita is converted into US dollars on the basis of purchasing power parity (PPP) relative to the domestic currency.

late 1980s and early 1990s.¹⁹ Life expectancy at birth increased only modestly from 61 years in 1987 to 63.2 years in 1993 (United Nations Development Programme; 1990, 1996).²⁰ Gilson and McIntyre (2001) observe that the ratio of the infant mortality rate between the poorest and richest quintiles was as high as 2.9 during the early 1990s.

Bradshaw and Steyn (2001) go further to suggest that the relationship between health and socio-economic status is so strong in South Africa. For example, so-called lifestyle diseases, such as cardiovascular disease, which traditionally affect mainly higher-income individuals, have impinged upon all socio-economic groupings in the country (Bradshaw & Steyn, 2001). Moreover, the authors indicate that poorer communities were inadequately serviced with respect to preventative care and the treatment of both lifestyle diseases and chronic illnesses. By 1994, hospital-based curative care rather than preventative care was particularly favoured within the public healthcare sector as reflected in a high hospital bed to population ratio concentrated in urban areas and at higher levels of care (McIntyre et al., 2007). McIntyre et al. (2007) suggest that district hospital capacity was inadequate and primary healthcare had been systematically neglected throughout the 1980s and early 1990s.

Health policy under the *apartheid* government created a healthcare system that ensured that the minority white population benefitted from a higher-quality of care, through primarily the private healthcare sector, but also through a carefully constructed system of public healthcare provision that favoured the white population. Until the late 1980s, access to public healthcare was delineated along racial lines, thereby creating effectively a two-tiered public healthcare sector. Preventative and

¹⁹ In 1993, the life expectancy at birth in the following developed and emerging economies were - Australia (77.8 years), United Kingdom (76.3 years), United States (76.1 years), the Republic of Ireland (75.4 years), Brazil (66.5 years), Poland (71.1 years), Russian Federation (67.4 years) and Botswana (65.2 years) (United Nations Development Programme, 1996).

²⁰ By 2010, South Africa's life expectancy at birth had deteriorated significantly to only 52 years (United Nations Development Programme, 2010). The rapid deterioration in life expectancy has been attributed to the HIV/AIDS epidemic and the related issue of a significant growth in multi-drug and extreme-drug resistant tuberculosis. According to Dorrington, Bradshaw, Johnson and Daniel (2006), approximately 5.4 million individuals in South Africa were HIV positive. This constituted more than 11 percent of the total South African population.

curative primary healthcare services were provided in separate healthcare facilities (hospitals and clinics) and administered by different healthcare authorities for the white and non-white population groupings (McIntyre & Gilson, 2002). McIntyre and Gilson (2002) note that the considerable inequalities established through the *apartheid* legislation were further exacerbated by a funding bias towards public healthcare facilities that were located in white-dominated geographical areas. Thus, the majority non-white population had to rely upon a deliberately under-resourced and effectively second-tier public healthcare sector that delivered a lower quality of care (McIntyre et al., 2007).

The provision of private healthcare has an extensive history in South Africa. Healthcare professionals have always been permitted to operate within a private practice setting. Naylor (1988) points out that by the early 1980s, approximately 40 percent of all general practitioners operated within a single or group private practice. Rispel and Behr (1992) indicate that by 1990, 62 percent of all general practitioners and 66 percent of medical specialists were in private practice. Private hospitals were initially limited to non-profit mission healthcare facilities that served rural communities and industry-aligned onsite hospitals, such as those healthcare facilities located at large mines (McIntyre et al., 2007). As the government sought to contract out long-term tuberculosis and psychiatric care in the first half of the 19th century there was an initial spurt in the growth of for-profit private hospitals (McIntyre et al., 2007). But it is only since the early 1980s that for-profit private hospitals have become a more prominent feature of the healthcare provision landscape. McIntyre et al. (1995) report that private hospital beds nearly doubled between 1988 and 1993. The rapid increase in private hospital beds exacerbated the exodus of medical specialists into private practice. The authors explain that due to the increased availability of private hospital beds, medical specialists would be encouraged to move into private practice since typically they operate their consulting practices within private hospitals and subsequently admit their patients to these hospitals.

The voluntary financing of private healthcare through medical schemes served as the complementary device that augmented demand for private healthcare provision in South Africa. Between the Second World War and the early 1990s, the number of medical schemes grew significantly, which raised concerns about the sustainability of the private health financing mechanism (McIntyre et al., 2007). From the mid-1980s until the early 1990s, they note that medical scheme expenditure per beneficiary rose sharply. The rise in expenditure was attributed to increases in expenditure on medication and private hospital care. According to McIntyre et al. (2007: 15), "... [b]y the late 1980s, increases in medical scheme expenditure, and in contribution rates, were of the order of 30 [percent] per annum in nominal terms".

Medical schemes' fee-for-service reimbursement arrangements with private hospitals and healthcare professionals encouraged utilisation as earnings were directly linked to the volume generated (McIntyre et al., 2007). The authors point out one clear-cut example is the purchase of expensive advanced diagnostic equipment, such as magnetic resonance imaging (MRI) scanners.²¹ Owners of private healthcare facilities that purchased advanced diagnostic equipment would actively encourage healthcare practitioners to utilise the equipment so as to generate greater revenue (McIntyre et al., 2007). Since many private healthcare professionals benefit from ownership stakes in private healthcare practices or similar profit-incentive programmes, private healthcare professionals had every incentive to increase utilisation.²²

McIntyre (1997) argue that the underlying incentive structure was a significant contributor for the observable trends in higher levels of hospitalisation and the increased utilisation of advanced diagnostic equipment. Unlike healthcare expenditure in the public sector, healthcare expenditure in the private sector continued to increase

²¹ An MRI machine utilises the property of nuclear magnetic resonance to generate images of the nuclei of atoms within the human body. This is achieved by utilising a powerful magnetic field, which triggers the nuclei of atoms within the body to produce a rotating magnetic field, which is detectable by the MRI scanner.

²² Van den Heever (2007) discusses other examples of financially beneficial relationships that can exist between the private healthcare institutions and their contracted healthcare practitioners. For example, rent-free or subsidised consulting rooms. In particular, medical specialists would be offered rent-free or subsidised consulting rooms if they achieved certain performance targets.

unabated during the 1980s and early 1990s - far exceeding consumer inflation (McIntyre et al., 1995). Until the promulgation of the Medical Schemes Amendment Act No. 19 of 1989 and the Medical Schemes Amendment Act No. 23 of 1993, medical schemes could only differentiate contributions on the basis of the principal member's income and their number of dependents. Thus, excessive private healthcare cost escalation (exacerbated by higher levels of private healthcare utilisation) coupled with the inability of medical schemes to align the expected claims experience of members to their contributions caused medical schemes to suffer successive underwriting losses (Reekie, 1999). Medical schemes did respond by increasing contributions, but as these contributions were community-based rather than risk-based, it did not address the underlying causes of the financial distress experienced by many medical schemes.²³ Therefore, out of necessity, medical schemes lobbied successfully for the deregulation of the medical scheme industry.

For Reekie (1999) the deregulation that took place during the late 1980s and early 1990s was a significant move in the right direction to create a regulatory framework that:

“...effectively removed a host of unnecessary regulations...and introduced a much more competitive environment than before” (Reekie, 1999: 4).

Reekie (1999) suggests that the reforms resulted in:

“...improved solvency and financial stability and increased coverage across the whole spectrum of society, the [healthcare] cost explosion had been reigned in” (Reekie, 1999: 4).

Reekie (1999) believed that the Medical Schemes Amendment Act No. 19 of 1989 and the Medical Schemes Amendment Act No. 23 of 1993 created a regulatory environment that promoted competition, and removed the obstacles that hindered the application of insurance and risk management principles within the private healthcare financing sector. The regulatory changes permitted medical schemes to utilise risk-rating to determine contributions, which according to van den Heever (1997) were

²³ Contributions during the 1980s and early 1990s increased on average at a rate in excess of 25 percent per annum in real terms (McIntyre et al., 1995).

along the lines of private health insurance in the United States. Minimum benefits were also abolished with the Medical Schemes Amendment Act No. 23 of 1993 and medical schemes were permitted to ration coverage.

The deregulation that took place during the late 1980s and early 1990s did not address the inequalities and disparities in healthcare provision between the public and private healthcare sectors. Most importantly, the first democratically-elected government inherited a healthcare system that was deeply fragmented and reflected broader socio-economic inequalities (McIntyre et al., 2007). According to McIntyre et al. (2007), the major challenge facing the South African health system was to address the inefficient and inequitable distribution of resources between the public and private healthcare sectors. The authors emphasise the *apartheid* government's policy of privatisation and deregulation as key determinants in exacerbating a public-private healthcare sector mix that did not reflect the new government's objective of creating a high-quality healthcare system that was accessible to all South Africans (Department of Health, 2004). The government sought to improve the health status of individuals by focussing upon preventative care, the promotion of healthier lifestyles and providing an equitable healthcare delivery system (McIntyre et al., 2007).

Despite Reekie (1999: 4)'s assurances that "...the [healthcare] cost explosion had been reigned in" once the principle of managed care had become a legal possibility with the deregulation of the late 1980s and early 1990s; Doherty and McLeod (2003) observed that the healthcare cost escalation lingered well into the first term of office of the *post-apartheid* government. Reekie (1999) argued that the regulatory reforms of the late 1980s and early 1990s would facilitate greater access to medical scheme coverage for all South Africans. However, the deregulation that occurred over this period did not translate into greater medical scheme coverage; in fact medical scheme membership remained relatively stagnant until the end of the 1990s. Doherty and McLeod (2003) argued that medical scheme practices, such as the application of the concept of risk-rating, had a negative impact upon medical scheme coverage. According to the authors, high risk individuals, such as the elderly and/or unhealthy,

had their monthly contributions loaded as they represented a greater risk to the medical scheme due to their higher expected claims experience. Medical schemes also imposed life-long exclusions for pre-existing conditions and in certain cases, the elderly and/or unhealthy were simply excluded from coverage altogether.

Another practice that intensified after the deregulation of the late 1980s and early 1990s “...was the ‘dumping’ of private patients on the public healthcare sector once their (now more limited) benefits had been exceeded” (Doherty & McLeod, 2003: 42). Rationing coverage and excluding coverage altogether were, according to Doherty and McLeod (2003), simply a means of cost avoidance as opposed to improvements in efficiency as implied by Reekie (1999). The result was that these measures merely reduced medical scheme coverage for those private patients who required healthcare services the most. As a result, the public healthcare sector was left to bear the additional burden of private patients who had exhausted their medical scheme benefits (Doherty & McLeod, 2003).

McIntyre et al. (2007) maintain that the application of risk-rating undermined the principle of cross-subsidisation between high-risk and low-risk individuals, which according to the authors was a central tenet behind the initial design of medical schemes in South Africa. When De Beers Consolidated Mines was set up as the forerunner of today’s medical schemes back in 1889, the underlying reason was to fund employee healthcare in the face of what was often risky mining activities. Cross-subsidisation was a consequence of the funding arrangement under the scheme rather than a deliberate design feature. Although, cross-subsidisation can serve a beneficial function within insurance markets, it may have unintended adverse consequences. If one considers Dahlby (1983)’s analysis into the prohibition of statistical discrimination with respect to gender in the Canadian automobile insurance market at the end of 1970s, the outcome of regulatory-imposed cross-subsidisation is proved to be less than favourable. Dahlby (1983) illustrates that once insurers are no longer permitted to discriminate against males by charging them higher premiums than females *ceteris paribus*, society is worse off since the discount received by males with

respect to their premiums for collision insurance is smaller than the increase in premiums experienced by females.²⁴ Thus, cross-subsidisation leads to the situation whereby society pays more for the identical level of collision insurance coverage. Moreover, Dahlby (1983) demonstrates that the proportion of females purchasing collision insurance within the different classes would decline once statistical discrimination is prohibited. Lower-risk females are likely to opt out of collision insurance as they are not prepared to subsidise the male population that has an overall higher-associated risk. Not surprisingly, she finds that the proportion of males purchasing collision insurance rises as their premiums are being subsidised by the females. However, the author illustrates that the overall demand for collision insurance is expected to fall.

In South Africa, the desire to ensure that private health insurance remains affordable for particularly the elderly and unhealthy does have moral traction. Observing that certain high-risk individuals including senior citizens were unable to obtain affordable medical scheme cover, the South African government inferred that a form of adverse selection was taking place whereby medical schemes were ‘*cherry-picking*’ or self-selecting low-risk individuals (Reekie, 1999). Despite the validity of this observation that the elderly and unhealthy found it progressively more difficult to maintain their levels of medical scheme coverage once the deregulation took place with the promulgation of the Medical Schemes Amendment Act No. 19 of 1989 and the Medical Schemes Amendment Act No. 23 of 1993, it is unfortunate that the principle of adverse selection is ascribed to this phenomenon. Rejda (2005) describes the principle of adverse selection as the observation that high-risk individuals are more likely to demand insurance than low-risk individuals. Adverse selection arises in insurance markets when the purchaser of insurance has more information about either the probability of a loss or the distribution of the magnitude of the loss should the loss event occur, than the insurance company. For the party that possesses superior information, it is advantageous to withhold such information from the other party.

²⁴ Premiums within the different classes (single versus married and age category) for males and females converge to a single level *ceteris paribus*, as insurers cannot charge differential premiums to reflect the different risk associated with gender characteristics (Dahlby, 1983). Without a significant loss in generality, it is assumed that the number of males and females purchasing collision insurance following the prohibition of statistical discrimination.

Under health insurance, each individual would select amongst a set of contracts based upon their expected probability of utilising healthcare services in the future. High-risk individuals (elderly and/or unhealthy) opt for more generous healthcare plans as opposed to low-risk individuals that anticipate requiring limited utilisation of healthcare services.

Government viewed the risk selection or '*cherry-picking*' practices by medical schemes as a manifestation of adverse selection (Reekie, 1999). However, adverse selection would in fact advocate the opposite. Where insurers find it difficult to distinguish between high and low-risk individuals due to information asymmetry or through regulation that forces insurers to exclude certain risk factors in the determination of premiums, low-risk individuals are likely to exit the private health insurance market as they are unwillingly to subsidise the high-risk individuals. At the same time, high-risk individuals would be induced to continue with medical scheme coverage and in some cases even increase coverage levels (move to more generous medical scheme plans) as they are effectively purchasing coverage at a discount. Whilst it is acknowledged that high-risk individuals were either reducing existing medical scheme benefits, opting out of medical scheme coverage or simply being excluded from coverage altogether, this is contrary to what the theory of adverse selection would predict (Akerlof, 1970). Therefore, government's proposition that the exodus of high-risk individuals from the medical scheme market can somehow be ascribed to adverse selection is a complete misinterpretation of the concept (Reekie, 1999).

A key concern that emerged after the promulgation of the Medical Schemes Amendment Act No. 19 of 1989 and the Medical Schemes Amendment Act No. 23 of 1993 was that medical scheme coverage became increasingly unaffordable for those who required healthcare services the most. This in turn, lead to what Doherty and McLeod (2003: 42) referred to as the "dumping" of private patients on the public healthcare sector, which further entrenched the disparities between the public and private healthcare sectors. Throughout the 1990s, real per capita spending in the

public healthcare sector remained relatively constant, whilst there was a significant rise in real per capita spending on medical scheme members (McIntyre et al., 2007). Even by 2005, the authors point out that the disparities between the public and private healthcare sectors has yet to be resolved. On average, spending per medical scheme member per annum was almost R9500 in 2005; whereas for those who rely solely upon the public healthcare sector for both hospital and primary care, spending was less than R1300 per person per annum (McIntyre et al., 2007).²⁵ The disparities also extend to the availability of healthcare professionals between each sector. For example, in 2005, the ratio of medical specialists to the number of individuals serviced by the private sector was one in less than 500 versus one in almost 11000 in the public sector (McIntyre et al., 2007). Despite the apparent advantageous position that individuals serviced by the private sector find themselves in versus their public sector counterparts, Doherty and McLeod (2003) refer to the *World Health Report* published by the World Health Organisation in 2000, where they identify the poor value for money that South Africa's private healthcare sector offers to consumers. Thus, the World Health Organisation (2000) promotes a greater role for the South African government in stewarding the private provision and financing of healthcare to achieve social objectives, which includes the affordability of medical scheme contributions and equitable outcomes with respect to healthcare delivery (Doherty & McLeod, 2003).

McIntyre et al. (2007) argue that the quintessential challenge for the South African healthcare system was to develop a mechanism to promote cross-subsidisation between the healthy, younger (and wealthier) population and the unhealthy, elderly (and poor) population. Thus, they advocate that it was necessary to reformulate the healthcare regulatory environment and in particular the legislation governing medical scheme business, to create a legislative framework that would function as a redistribution device to address the inequitable distribution of resources (and quality of care delivered) between the public and private healthcare sectors.

²⁵ The spending per capita in the private healthcare sector may exaggerate the disparity, since it based upon spending per medical scheme member rather than per medical scheme beneficiary.

2.2 The Medical Schemes Act No. 131 of 1998

McIntyre et al. (2007) argues that medical schemes started out as having community-rated contributions and initially, the private financing of healthcare expenditure essentially adhered to the principle of *solidarity*. However, through the enactment of the Medical Schemes Amendment Act No. 19 of 1989 and the Medical Schemes Amendment Act No. 23 of 1993, the government at the time embarked upon a process that significantly deregulated the medical scheme industry. As discussed previously, the deregulation permitted detailed and individual risk rating, and provided for variations in both the level and structure of benefits. Hence, as Pearmain (2000) asserts the average contribution levied by a particular scheme for a representative risk pool is loaded or increased in proportion to the increased risk to the scheme represented by the group or individual applying for membership. According to the author, risk rating severely hinders the operation of cross-subsidisation within a medical scheme. Nevertheless, this represents a narrow view of the nature of cross-subsidisation and its application within insurance markets. Whilst it is true that risk classification does indeed permit insurers to offer lower premiums to individuals that represent a lower risk to the medical scheme, it does not preclude medical schemes from engaging in cross-subsidisation should they choose to do so. Medical schemes can still elect to charge low-risk individuals more for coverage than their risk status would dictate. The success of such a strategy would depend upon a number of factors, not least of all the principle of adverse selection - a distinct possibility can occur whereby the low-risk members would exit the health insurance market as they are unwillingly to effectively subsidise the high-risk members (Akerlof, 1970). This is indeed a strong argument in favour of risk-rating. Yet, other factors such as risk aversion and the degree to which people are willing to sacrifice health insurance cover may curtail the adverse selection effect.²⁶

Pearmain (2000)'s view that cross-subsidisation and risk classification are mutually exclusive mechanisms is mistaken. Eggleston (2000) puts forward the idea of cross-subsidisation across time (or *dynamic cross-subsidisation*) and he argues that it is

²⁶ Refer to a more detailed discussion about adverse selection in Chapter 3 and Chapter 4.

indeed prudent for health insurers to incorporate cross-subsidisation across time. The author examines how the design of a health insurance financing system can have a significant impact with respect to the equity and efficiency to which healthcare is delivered to the public at large. Medical scheme cover, as in any other line of insurance, does create a propensity for moral hazard to occur whereby fully-covered members “tend to over-utilise services that appear to be “free” or are heavily-subsidised” (Eggleston, 2000: 174). Hence, medical schemes argue that it is necessary to introduce demand-side cost sharing mechanisms such as medical savings accounts, deductibles and co-insurance to discourage over-utilisation, and thus reduce the problem of moral hazard. Whilst, it is precisely the elderly that increase their utilisation of healthcare and thus demand greater healthcare cover, medical schemes have failed historically to utilise risk financing across time. Young and healthy members’ individual contributions should be seen as contributions to finance their future demand for healthcare. Therefore, medical schemes can inflate the contributions of low-risk members, such as the young and healthy, with the view that a portion of these contributions are allocated towards funding the low-risk’s future healthcare expenditure.

Pearmain (2000: 185) goes further to criticize the utilisation of risk rating:

“It facilitates offers of attractively low premiums to persons who are good risks, [for example] the young and healthy, while requiring those who are old or sick to pay greatly increased premiums which they often cannot afford”.

Despite the fact that it is precisely the objective of risk rating to segment any insurance market into ‘homogenous’ risk pools with the members of a pool having essentially the same risk profile²⁷, this is in line with the definition of equity that leads

²⁷ Therefore, members of a pool that represent a higher risk to the medical scheme can expect to pay higher contributions than the contributions charged to members from a pool with a lower risk.

to efficient outcomes (Eggleston, 2000).²⁸ Eggleston (2000) asserts that risk classification (and as a consequence risk rating) can improve the trade-off between inefficient utilisation and inequitable coverage. In fact, risk classification and risk rating is a response by medical schemes to overcome the problem of adverse selection (Reekie, 1999).

Hence, after the first-democratic elections in 1994, the South African government began redrafting the legislation governing medical scheme business, with the view to bring it in line with the government's objective of providing universal access to affordable private healthcare for all South Africans. The Medical Schemes Act No. 131 of 1998 represented a return to the principle of solidarity and according to Pearmain (2000: 185):

“...the Medical Schemes Act No. 131 of 1998 and its regulations is substantially different, both in terms of focus and government policy, from the preceding law”.

According to McIntyre et al. (2007) the primary focus of the new legislation was to ensure that medical schemes remain financially sound and to ensure that their benefit offerings continue to be sustainable. In addition, the Act promotes the effective governance of medical schemes, through a Board of Trustees (McIntyre et al., 2007). Trustees of a medical scheme are required to have the requisite skills to manage the scheme and to ensure that the interests of the medical scheme members and their beneficiaries are safeguarded. The Medical Schemes Act No. 131 of 1998 established an Office of the Registrar of Medical Schemes, which serves at the executive body of the Council for Medical Schemes (McIntyre et al., 2007). Under the prior legislative framework, the Registrar of Medical Schemes had only a small staff complement and the Office served merely as a deputy directorate under the Department of Health.

²⁸ As discussed previously, this definition of equity proposes that an individual's medical scheme contribution should be aligned with the benefits that the member receives. A person who is high-risk has a greater likelihood of claiming under a health insurance policy (claiming against the medical scheme for healthcare expenditure incurred) and the expected value of any claim is also likely to be larger in comparison to a person with a lower risk profile. Therefore, a high-risk individual is likely to receive greater and more frequent claim reimbursements from the medical scheme for healthcare expenditure incurred than an individual with a lower risk profile – the expected benefits are greater and thus, so should the high-risk person's contributions be greater.

Under the new legislation, the Council for Medical Schemes is a statutory body that has the power to approve, suspend or revoke the registration of medical schemes and medical scheme administrators (McIntyre et al., 2007). As stated by McIntyre et al. (2007: 45):

“...the creation of a strong regulatory authority, in the form of the Council for Medical Schemes, has been critical to the successful implementation of these regulations. Unfortunately, to date, no comprehensive assessment of the effectiveness of these regulations has been undertaken”.

McLeod and Ramjee (2007) point out that under the Act; the Council for Medical Schemes is tasked with protecting the interests of members of medical schemes and their dependents rather than those of the medical scheme industry. The Council for Medical Schemes is funded by a levy on medical scheme members. Besides regulating medical schemes, the Council for Medical Schemes is also responsible for accrediting medical scheme intermediaries, administrators and managed care organisations that provide services to medical schemes (McLeod & Ramjee, 2007).

2.2.1 Three Core Pillars of the Medical Schemes Act No. 131 of 1998

The Medical Schemes Act No. 131 of 1998 and accompanying regulations that were published in 1999 came into effect after January 2000. Three key issues were identified by McLeod and Ramjee (2007) as being at the core of the new legislation and they can be summarised as follows:

- *Open enrolment:* Medical schemes are required to accept all eligible applicants who wish to enrol as a member of the scheme. The Act also requires medical schemes to allow members who retire from employment to continue with membership of the medical scheme.
- *Community-rating:* Contributions charged to members may only be differentiated according to income and the number of dependents. Thus, community-rating effectively replaced risk rating as medical schemes could no longer discriminate

against members on the basis of age, current health status and prior health status (Doherty & McLeod, 2003).

- *Prescribed Minimum Benefits (PMBs)*: The Prescribed Minimum Benefits (PMBs) set a minimum package of healthcare benefits that must be offered by all medical schemes. Annexure A to the Regulations of the Medical Schemes Act No. 131 of 1998 defines the PMBs in terms of some 270 diagnosis-treatment pairs. McLeod, Matisonn, Fourie, Grobler, Mynhardt and Marx (2004) discuss the subsequent Regulations published in November 2002, which provided much-needed clarification with respect to PMB requirements, and further extended the PMB package to include emergency procedures. The authors point out that in the case of PMBs, medical schemes may utilise managed care principles to ration care. These include pre-authorisation, the stipulation of a network of healthcare providers and medicine formularies that members and their dependents would be required to use in relation to the diagnosis-treatment pairs covered under the PMBs. Members and their dependents must be covered in full with respect to a list the diagnosis-treatment pairs and emergency procedures with no limits or co-payments permitted (McLeod & Ramjee, 2007).²⁹ From 1 January 2004, the PMBs were further extended to include a list of 25 chronic disease conditions (Chronic Disease List – CDL).³⁰ The Regulations pertaining to the Chronic Disease List (CDL) require that medical schemes cover, in full, the diagnosis, treatment and medication prescribed with these chronic diseases (McLeod et al., 2004).

General Waiting Periods

Although, the Act permits individuals to switch medical schemes at any time due to for example a change in employment, the medical scheme is permitted to impose a waiting period of up to three months for new members and their dependents (McLeod

²⁹ Co-payments may be levied if a member voluntarily elects to use a healthcare provider outside of medical scheme's list of designated healthcare providers (McLeod et al., 2004).

³⁰ The CDL has subsequently been extended to include 27 chronic disease conditions.

& Ramjee, 2007).³¹ Hence, the medical scheme will not reimburse healthcare providers for services rendered to the new member or their dependents, during the waiting exclusion period. Within the Act, a provision permits the medical scheme to waive the waiting period if the new member pays the medical scheme a fee as determined by the medical scheme. In addition, new members and their dependents may not be penalised with a waiting period stipulation if and only if:

1. At the date of the application to the new medical scheme, the applicant had previously belonged to a medical scheme for a continuous period of two years or more; and
2. The application for membership to the new medical scheme must take place within a period of three months since the applicant left their previous medical scheme.

Pre-Existing Illness and Medical Conditions

Under the current legislation, in the case of a pre-existing illness or medical condition, a medical scheme may impose a waiting period of up to 12 months on new applicants and their dependents. The waiting period is only applicable if at the date of application, the applicant or any of their dependents had a specific medical condition, and if the diagnosis or treatment for the aforesaid condition was sought or received during the past year (Pearmain, 2000; Doherty & McLeod, 2003; McLeod & Ramjee, 2007). However, the Act does not permit medical schemes to enforce any such waiting period if the pre-existing medical condition relates to a diagnosis-treatment pair covered under the PMB package.

Penalties for "late joiners"

To provide some protection against so-called *anti-selection*, whereby individuals delay membership of a medical scheme until they are elderly or unhealthy, the Act permits medical schemes to impose penalties upon persons who decide to join a

³¹ Note if a child is born to a member of a medical scheme during their period of membership, cover will automatically extend to the child.

medical scheme, as a member or a dependent of a member, much later in life (Pearmain, 2000; McLeod & Ramjee, 2007). The starting age for penalties is 35 years old. The penalty is based on the number of years after age 30 an applicant was without medical scheme cover. If left unchecked, *anti-selection* can impose a significant cost upon the medical scheme, since at outset; the newly-joined member (or dependent) is likely to make more frequent and extensive claims against the scheme without having made a meaningful contribution to the funding of the scheme. In other words, the *late-joiner* penalties are necessary to discourage opportunistic behaviour on the part of individuals seeking to delay their medical scheme contributions to a time when they are more likely to receive benefits under the medical scheme policy. As Eggleston (2000) argues, ideally people should start contributing towards their medical schemes whilst they are young and healthy, and medical schemes should then set aside the contributions to pay for future claims.

Medical Savings Accounts (MSAs)

As discussed previously, medical scheme cover, as any in other insurance policy, may diminish the incentives for the member to prevent loss (in this case incur healthcare expenditure) and therefore medical scheme membership may alter the behaviour of the member or their dependents in such a way that encourages over-utilisation of healthcare services (Shavell, 1979). In other words, medical scheme membership creates a propensity for moral hazard to occur whereby fully-covered members “tend to over-utilise services that appear to be “free” or are heavily-subsidised” (Eggleston, 2000: 174). Since the full cost of healthcare is not paid for by the member at the point-of-service, there is little incentive to curtail utilisation. Hence, medical schemes argue that it is necessary to introduce demand-side cost sharing mechanisms such as medical savings accounts, deductibles and co-insurance to discourage over-utilisation, and thus reduce the problem of moral hazard.

Medical Savings Accounts (MSAs) became a feature of the medical scheme landscape in South Africa during the 1990s, but only with the enactment of the Medical Schemes Act No. 131 of 1998 were they formally defined as a financing mechanism

for healthcare expenditure.³² In South Africa, a member of a medical scheme plan with a MSA pays a monthly contribution to the medical scheme that consists of both a traditional health insurance (risk) premium and a fixed amount that is set aside 'within' the person's medical savings account. Funds that accumulate within the MSA are designed to cover day-to-day healthcare expenses, which includes general practitioner consultations, acute treatment and out-patient hospital care. Treatment related to a chronic medical condition or in-hospital care would be covered by the 'insurance' portion of the member's contribution. Therefore, MSAs are simply viewed as a mechanism whereby medical scheme members deposit funds into their medical savings account, which is held by the medical scheme and used by members to pay for healthcare expenditure without risk pooling. Once the funds in a MSA are exhausted, the member becomes responsible for the day-to-day healthcare expenses. If at the end of the year, the member's medical savings account is in a credit balance, the funds are carried over into the subsequent year to cover future healthcare expenditure not covered by their 'risk' portion of their medical scheme policy.

According to McLeod and Ramjee (2007), prior to 2006 many benefit options allowed members to choose their level of medical savings account contributions (subject to the regulatory maximum) or annual routine benefit. This effectively allowed members to tailor their benefits according to their health needs and to pay differential contribution rates (McLeod & Ramjee, 2007). This was identified as a means to effectively risk-rate members and consequently this practice was disallowed. Members who change medical schemes can transfer the funds currently in their MSA to the new medical scheme. If a member terminates their medical scheme membership and does not subsequently join another scheme, the accumulated funds in their MSA are payable to the member, subject to paying the appropriate income tax to the South African Revenue Service.

³² MSAs were permitted under the Medical Schemes Act No. 131 of 1998 subject to limitations with respect to the value of a member's annual contribution. According to the Regulations attaching to the Medical Schemes Act No. 131 of 1998, the amount allowed to accumulate in a member's medical savings account has been capped at 25 percent of his or her total annual contribution. an individual's annual contributions to a MSA may not exceed 25 percent of the total annual contribution paid to the medical scheme.

2.2.2 Community Rating and the Adverse Selection Death Spiral

Observable characteristics such as age, gender and health status are statistically significant determinants of the risk that an individual is likely to require healthcare within a given period of time. Thus, at the outset it is easy to accept that the ability to observe such variables at relatively low cost precludes the presence of information asymmetry. First, health status is not necessarily observable with perfect accuracy (Crocker & Snow, 1986). Second, prohibiting the ability to use such variables in the determination of contribution rates is tantamount to placing a 'veil' in front of the 'eyes' of the medical scheme. Even though the medical scheme is fully aware of these risk-related characteristics, they are compelled through regulation to exclude them from the determination of the risk profile of each member. Hence, community rating creates an artificial dichotomy of information asymmetry between the medical scheme and the member. Thus, community rating enforces information asymmetry between the two parties.

Community rating is not a unique feature to South Africa's private health insurance sector; many countries have applied the principle in various lines of insurance. Neuhaus (1995) examines the salient features of community rating with respect to private health insurers in Australia. Australian insurers are required by law to apply community rating, whereby they are not permitted to differentiate their contribution rates by risk-related characteristics such as age, gender and health status (Neuhaus, 1995: 95). Insurers can only discriminate with respect to contribution rates in relation to a single person requiring cover versus family-related cover. As in South Africa's case, there is no explicit regulation governing contribution rate levels (Neuhaus, 1995).

MacIntyre (1962) suggests that the reason for imposing a mandatory community rating system is based upon social equity. Equitable outcomes can take the form of various guises. One might interpret equity to mean that all medical scheme members should pay identical contributions or receive identical benefits (claim payments to a

member) irrespective of their relative risk. A variation of this view is based upon the notion of a member's ability to pay. Hence, those members with greater resources would be expected to pay more for medical scheme cover than those with fewer resources. Alternatively, equity can be defined to mean that members should pay according to the benefits that they receive. This interpretation advocates that equity within the health insurance sector is achieved when members pay contributions that are commensurate with their relative risk.

There exists a trade-off between the first interpretation of equity and efficiency – equal contribution rates across members with different levels of risk will reduce efficiency. Low-risk individuals will be induced to purchase less medical scheme cover than required, whilst high-risk individuals are likely to purchase greater levels of medical scheme cover than their relative risk demands. In addition, members' incentives to mitigate losses will be distorted through the equalisation of contribution rates. Members who do not pay the full cost of their health insurance will have less incentive to reduce their risk since they are already paying effectively discounted contribution rates. With respect to the final interpretation of equity there is no trade-off between equity and efficiency. Aligning the full costs of medical scheme cover to the benefits received is consistent with maximising efficiency. Members are encouraged to reduce their risk (and/or magnitude of loss) if the ensuing contribution rate savings exceeds the costs of reducing their risk.

Unlike in South Africa, all Australian residents are also covered by an adequate public healthcare system (Medicare). Medicare has led to far less reliance upon private healthcare, thereby reducing the privately-insured population significantly. Neuhaus (1995) does indicate that it is debatable whether community rating can achieve its social objective of higher enrolment through universal affordability. Buchmueller and DiNardo (2002: 283) put forward that pure community rating:

“...is often depicted primarily as a mechanism to transfer wealth from consumers whose expected medical expenses are lower than [the] average to those whose expected expenses are higher than [the] average”.

Studies have argued that community rating leads to a reduction in societal insurance coverage through a mechanism described as the *adverse selection death spiral* (Gradison, 1995; Hartnedy, 1994). In the absence of community rating, contribution rates would be risk-rated, whereby the young and healthy members are charged lower rates due to their lower expected claims. In view of the fact that community rating requires medical schemes to charge identical contribution rates with respect to age, gender and health status; the rate charged to younger individuals is expected to rise in response to the higher costs associated with the older members of the pool. Depending upon the age-specific price elasticity of demand for medical scheme cover, younger members (lower risk) are likely to reduce their coverage. The reduction in coverage by lower-risks induces higher future expected claims for the pool. This inducement translates into a further increase in contribution rates, which once again impacts negatively upon coverage levels. Therefore, the risk pool could disintegrate into a single member with the highest expected claims (highest risk).

2.2.3 Risk-Related Cross-Subsidisation and the Risk Equalisation Fund

Government's future vision for the South African healthcare system was outlined in a report published in 2002 by the health sub-committee of the Committee of Inquiry into a Comprehensive System of Social Security – also referred to as the Taylor Committee (Department of Social Development, 2002). A follow-up and more comprehensive report discussing this vision was published by the Department of Health in that same year (Department of Health, 2002). The key recommendation that emanated from these publications was that South Africa should move towards a National Health Insurance system that would integrate both the public sector and private medical schemes within the context of a universal contributory system (McLeod et al., 2004). The central feature of government's plan would be to have a national health insurance system that provided universal coverage to all South Africans, whilst simultaneously ensuring accessibility, equity and efficiency in the delivery of healthcare outcomes (Theron & van Eeden, 2009). Even though the

National Health Act No. 61 of 2003 does not refer directly to mandatory health insurance, it is clear from the legislation that government's objective is to address the pervasive inequalities between the public and private healthcare sectors. And although the transformation of the national health system may have moral traction, the move towards a mandatory national health insurance system will have several consequences – intended and unintended – for both the public and private healthcare sectors, including the private financing of healthcare within the country.

By implementing the principles of open enrolment, community rating and the prescribed minimum benefits through the Medical Schemes Act No. 131 of 1998, government sought to force a cross-subsidisation to take place between those who represent lower risks (the young and healthy) to the medical scheme and the higher risk individuals, such as the elderly and unhealthy. McIntyre et al. (2007) notes that the open enrolment coupled with community-rated contributions have provided a degree of cross-subsidisation to occur within individual medical scheme's pools. Moreover, the introduction of the PMBs offered a guaranteed minimum package of healthcare benefits, which was a uniform standard across all medical scheme beneficiaries. Thus, the authors suggest the introduction of the PMBs encouraged further risk-related cross-subsidisation. Theron and van Eeden (2009) examine the extent of cross-subsidisation within the private health insurance sector and find that:

“...wealthier individuals, which tend to use private healthcare, receive proportionally more healthcare benefits than the rest of the population, [but] they also pay more for healthcare [cover] than the benefits they receive” (Theron & van Eeden, 2009: 4).

Thus, Theron and van Eeden (2009) conclude that there is significant cross-subsidisation between the wealthy and poor in South Africa. Although, Theron and van Eeden (2009) do not investigate the success of risk-related cross-subsidisation, that is the cross-subsidisation between the young and elderly, or between the healthy and unhealthy, it nevertheless suggests that government's objective to address some of the disparities between the public and private healthcare sectors are having some success. A key criticism of community rating is that it may lead to adverse selection given the artificial information asymmetry that is created between the member and the

medical scheme. But there appears to be little evidence that this is indeed taking place in South Africa (Theron & van Eeden, 2009). Moreover, Buchmueller and DiNardo (2002) argue that despite the prevailing belief that an insurance coverage reducing death spiral is a likely consequence of community rating, it is not necessarily the outcome that models of insurance markets would predict.³³

Under the current regulatory framework, McLeod et al. (2004) argues that medical schemes can still market themselves and design their benefit offerings in such a way as to attract younger and healthier enrolees, whilst discouraging the elderly and unhealthy from joining. McIntyre et al. (2007) suggest that a substantial amount of resources are allocated to the marketing activities of various open medical schemes so as to attract the young and healthy, which include as McIntyre et al. (2007: 66) put it “...the perverse incentivisation of intermediaries”. Hence, under the current regulatory environment the so-called *cherry-picking* (or *cream-skimming*) of low risk individuals by medical schemes could continue to flourish albeit in a more subtle manner. McLeod and Ramjee (2007) also mention that medical schemes are able to design their benefit options in such way so as to benefit from the requirement that each option be treated as a separate risk pool for community-rating purposes. Since each benefit option is required by law to be self-sustaining, it induces a risk-pooling to occur at an option level (McLeod & Ramjee, 2007). Whilst restricted schemes tend to offer very few benefit options, open schemes tend to offer a large variety of different options, partly for competitive reasons, but also as a subtle mechanism to classify individuals into appropriate risk pools.³⁴ Therefore, this together with the *cherry-picking* undertaken by medical schemes via their marketing activities would to some extent defeat the purpose of the regulatory reforms introduced by the government and it implies that one would simply return to the *status quo* prior to the introduction of the Medical Schemes Act No. 131 of 1998. As a result of the *cherry-picking*, some medical schemes may find themselves with a greater proportion of older and unhealthy members compared to the average ‘representative’ scheme, which would

³³ Refer to Chapter 3 for a more detailed discussion.

³⁴ The Registrar of Medical Schemes has become aware of this practice, whereby medical schemes attempt to construct multiple benefit options to take advantage of the separate risk pools that these benefit options can create. Thus, the Registrar is increasingly refusing to authorise minor variations in benefit packages as separate benefit options (McLeod & Ramjee, 2007).

translates into these schemes bearing a greater burden with respect to a higher community rate for the PMB package (McLeod et al., 2004). The practice of *cherry-picking* diverts healthcare resources away from those who require it the most and this can only be overcome if mechanisms of financial compensation between schemes with different risk profiles can be created (McIntyre et al., 2007).

McIntyre et al. (2007) identify the important feature that is absent from the regulatory structure; namely, a mechanism to facilitate risk cross-subsidisation across individual medical schemes and benefit options.³⁵ The proposed solution is the introduction of a system of risk equalisation. Indeed, McLeod et al. (2004) mention that South Africa is unusual in having applied the principles of open enrolment and community-rating without risk equalisation. Neuhaus (1995) asserts that risk equalisation is a theoretical solution to the problem of supporting community rating where one has, as in South Africa's case, different medical schemes that are offering a multitude of different benefit offerings. For the author, the crucial issue is to design an appropriate risk equalisation system based upon the criterion that any risk equalisation for deciding upon a suitable risk equalisation mechanism should be that it minimises the opportunities for arbitrage between medical schemes.

Under risk equalisation, a Risk Equalisation Fund (REF) is established, which in its simplest form, receives funds from medical schemes with lower risk profiles and pays out funds to medical schemes with a higher risk profile (McLeod et al., 2004). It is suggested that risk equalisation between medical schemes will remove the age and health status factors as competition parameters amongst the medical schemes (McIntyre et al., 2007). The REF serves as a mechanism that equalises contribution schedules (or tables) across medical schemes. The objective of the REF is to discourage medical schemes from *cherry-picking* the younger and healthier members of society, which simultaneously erodes the balanced age and health profiles of other

³⁵ McLeod et al. (2004) argue that the absence of a mechanism such as risk equalisation was not a policy oversight on behalf of the government, but rather a question of timing. It is suggested by the authors that it is only now possible to effectively to implement such a strategy to facilitate the risk cross-subsidisation across medical schemes and benefit options.

medical schemes (McLeod et al., 2004). The proposed REF has been designed in such a way that financial transfers across medical schemes will be based upon the industry community-rate for the PMBs (McIntyre et al., 2007). Therefore, the overall risk profile of all medical scheme members and their corresponding beneficiaries will be considered in calculating the actuarial costs that relate to the benefits provided under the PMB package (including the benefits provided under the Chronic Disease List). McLeod (2004: 9) views the risk equalisation system that has been proposed as:

“...a mechanism to ensure that everyone pays the same industry community rate for the common package of benefits, not the rate determined by the age and health profile of the medical scheme they have chosen to join”.

Another related policy issue that is raised by McLeod et al. (2004) refers to the subsidy framework for medical schemes. According to the Taylor Committee, the tax expenditure subsidy to medical schemes of R7.8 billion (at the time the Committee reported their findings in 2002), represented an effective tax subsidy of over R1000 per medical scheme beneficiary per month. This figure represents more than the total amount spent per individual who depends upon the public sector for the receipt of their healthcare services. In addition, the tax rebate that individuals receive in relation to medical scheme contributions, further rewards high income individuals and those that elect the more comprehensive (and expensive) medical scheme options. Thus, it has been suggested that this is perhaps part of the reason why the medical scheme contribution tax rebate is likely to be abandoned, once the National Health Insurance framework is implemented in South Africa (McLeod et al., 2004).

After the Taylor Committee's findings were presented in 2002, the Formula Consultative Task Team was established by the Department of Health in July 2003 to spearhead the design of an REF in South Africa. On behalf of the Formula Consultative Task Team, McLeod et al. (2004) published a report outlining a proposed structure for the design of the REF. An International Review Panel of experts from six different countries supported McLeod et al. (2004)'s findings and

further emphasised that the introduction of the REF in South Africa should be an urgent priority (McIntyre et al., 2007).

In 2005, a 'shadow process' for the REF was introduced with Cabinet approval, whereby all medical schemes were required to submit information regarding the age, gender and chronic disease profile of their beneficiaries. From this information, an industry community-rate for the PMB package was calculated to be R224.90 per beneficiary per month (McIntyre et al., 2007). Under the proposed REF framework, the financial compensatory transfers to and from the fund will be based upon this industry community-rate for the PMB package. Thus, medical schemes with a better-than-average risk profile will be required to make certain contributions towards the fund, whilst the medical schemes with a worse-than-average risk profile will receive payments from the REF (McIntyre et al., 2007). In addition, medical schemes will be informed in advance concerning the amount that they may need to pay towards the REF or the amount that they can expect to receive from the REF. Therefore, McIntyre et al. (2007) claim that the prospective nature of the REF will permit medical schemes to determine their expected revenue streams with greater certainty at the beginning of each operating period.

After a rigorous process of testing various factors, McLeod et al. (2004) suggest that the following factors should be included in the determination of the risk equalisation formula:

- Age;
- A pregnancy/maternity indicator; and
- Measures of the chronic disease burden of the medical scheme, which includes the number of medical scheme beneficiaries with:
 - one of the identifiable 25 PMB-CDL chronic conditions
 - multiple CDL diseases; and
 - HIV/AIDS members who are currently on Anti-Retroviral therapy.

Equation 2-1 describes the general formula for obtaining the subsidy per medical scheme beneficiary under the REF that was outlined in a Technical Report by Grobler, Theron and Cooper (2003). From Equation 2-1, it is evident that the subsidy per medical scheme will be dependent upon the average cost of treating the PMB-CDL conditions.

Equation 2-1: General Formula for the Determination of the Subsidy per Medical Scheme Beneficiary under the REF framework (Grobler, Theron & Cooper, 2003)

$$Sub_{ij} = \frac{\{AC \times \sum_i \sum_j M_{ij} \times [(W_{ij} / \sum_i \sum_j (W_{ij} + M_{ij})) \times \sum_i \sum_j M_{ij} - 1] + TS\}}{\sum_i \sum_j M_{ij}}$$

where Sub_{ij} represents the subsidy per annum for medical beneficiaries with disease i and age band j , AC is the average cost of the PMB-CDL condition for the total medical scheme population per beneficiary per annum, W_{ij} represents the cost weighting attached to medical scheme beneficiaries with disease i and age band j , M_{ij} is the total number of medical scheme beneficiaries with disease i and age band j across all medical schemes and TS is the total subsidy available for all medical scheme beneficiaries in all medical schemes.

McIntyre et al. (2007) suggest that the REF framework is merely the starting point in an attempt to achieve an optimal level of risk-related cross-subsidisation. They go further to point out the shortcomings with the REF framework. The proposed REF framework excludes any risks associated with the healthcare benefits that fall outside of the PMB package. This is essentially because the PMB package represents the only common set of healthcare services covered by all medical schemes (McIntyre et al., 2007). Therefore, the ability to adequately implement risk cross-subsidises through the REF is restricted. Moreover, the REF framework does not consider other risk factors outside of those outlined in the contribution tables, such as demographic profile of medical scheme members (McLeod et al., 2004; McIntyre et al., 2007). Hence, the REF compensatory transfers may not accurately reflect the differences in the risk profile of the medical schemes. However, McIntyre et al. (2007) does note that the risk factors included within the REF framework are far more extensive than those found in the risk equalisation systems utilised by many other countries.

After the “shadowing process” that took place during 2005, the medical scheme industry expected the full implementation of the REF by early 2007 (McLeod & Ramjee, 2007). However, the process of drafting the new legislation that would introduce the REF framework took much longer than anticipated. Ultimately, the efforts to incorporate the REF framework into the current regulatory environment culminated in the drafting of the Medical Scheme Amendment Bill No. 58 of 2008.

The Medical Scheme Amendment Bill No. 58 of 2008³⁶ proposes to amend the Medical Schemes Act No. 131 of 1998, to expand the role of the Council for Medical Schemes to take responsibility for the operation of a Risk Equalisation Fund (REF), which McLeod and Ramjee (2007) argue will enhance the risk-pooling function of medical schemes by creating an industry-wide risk pool and simultaneously, an industry-wide community rate for PMBs. The Medical Scheme Amendment Bill No. 58 of 2008 sets out clearly the purpose of the legislative changes; namely, to amend the Medical Schemes Act No. 131 of 1998 in such a way so as to allow for the establishment of a risk equalisation fund whereby medical schemes would be required to participate. The proposed reforms would extend the functions of the Council for Medical Schemes in relation to operating the REF and create a framework to facilitate the collection of information from medical schemes to the Council for Medical Schemes for purposes of risk equalisation. Part 4 of the Medical Scheme Amendment Bill No. 58 of 2008 sets out the methodology and procedures to be adopted to calculate the appropriate financial compensatory transfers that will ensure the sharing of expected costs (claims) across all medical schemes. Therefore, it is necessary to amend the provisions relating to benefits and contributions as set out in the Medical Schemes Act No. 131 of 1998. The Bill also seeks to amend the provisions relating to the composition of the Boards of Trustees of medical schemes and adjunct to this the eligibility of persons to serve as trustees or principal officers of medical schemes. The Medical Scheme Amendment Bill No. 58 of 2008 goes further to outline the respective duties and functions of the Board of Trustees and those of the principal officers.

³⁶ From here on also referred to as ‘the Bill’.

The promulgation of the Medical Scheme Amendment Bill No. 58 of 2008 has as yet to be completed, the Bill has been withdrawn from Parliament on a number of occasions and according to the Council for Medical Schemes (2010) it is only likely to be passed into law during 2011.³⁷ This is despite the International Review Panel's recommendation back in 2005 that the REF should be implemented as a matter of urgency. It appears that a delay of six years is sufficiently urgent. With the delay in the implementation of risk-adjusted transfers through the REF, McLeod and Ramjee (2007) argue that certain medical schemes, that would be net contributors under the REF, have taken the opportunity to become more vocal in their attempts to be excluded from REF. Legislators have been urged to ignore these special requests for exclusion from the REF, since it implies that the remaining schemes (and ultimately medical scheme beneficiaries) will end up bearing the additional cost (McLeod & Ramjee, 2007).

McLeod and Ramjee (2007) claim that one of the primary purpose of the REF is to discourage medical schemes from competing on the basis of risk-selection (*cherry-picking* the young and healthy), but rather encourage schemes to compete on the basis of ensuring that their members receive cost-effective healthcare delivery. Those medical schemes that reduce the cost of healthcare delivery for their members will be able to retain that benefit and pass this onto their members through lower contribution rates for the PMB package (McLeod & Ramjee, 2007). According to McIntyre et al. (2007), the REF is necessary to transform medical schemes into a vehicle that extends private healthcare coverage to a broader and more diverse public.

³⁷ Even though the Risk Equalisation Fund (REF) has not yet been implemented, the Council for Medical Schemes has continued to collect the necessary information from medical schemes that it would require to compute the financial transfers across medical schemes. The Council for Medical Schemes (2010) states that it is currently evaluating the information collected thus far.

2.2.4 Government's Initiatives to Extend Medical Scheme Cover to Low-Income Earners and the Proposed National Health Insurance System

An important set of recommendations that arose out of the health sub-committee of the Committee of Inquiry into a Comprehensive System of Social Security of 2002 (or Taylor Committee) related to a phased roll-out set of reforms in view of establishing a National Health Insurance (NHI) system. According to Magennis and van Zyl (2009), the sub-committee simply defines the NHI as a “universal contributory environment”. After the successful implementation of the voluntary contributory environment, the second phase would involve extending the framework into a mandatory environment (Magennis & van Zyl, 2009).

Following the publication of the findings of the Taylor Committee, the Department of Health established a Ministerial Task Team in 2004. The task team was asked to decide upon which of the Taylor committee's proposals to take forward. It was determined that at the time, the goal to pursue a NHI system would not be possible in the short term, but in the interim it would be beneficial to pursue certain elements of a social health insurance (SHI) system (Magennis & van Zyl, 2009). The Department of Health felt that the REF and the reform of the tax subsidy with respect to medical scheme contributions would be the precursors to implementing a mandatory health insurance system for all South Africans. In particular, through the consultative process of the Ministerial Task Team, it was determined that extending health insurance coverage to low-income individuals would be an important step in steering the healthcare sector towards universal access to affordable healthcare for all South Africans.

Low-Income Medical Scheme (LIMS) initiative

The Ministerial Task Team recognised that the income cross-subsidies envisaged as part of a social health insurance system may take some time to be realised (McLeod & Ramjee, 2007). Thus, the task team launched the Low-Income Medical Scheme

(LIMS) consultative process in 2005. It was hoped insights could be gleaned from a range of stakeholders to determine the best way to extend medical scheme coverage to low-income employees from the formal sector (Magennis & van Zyl, 2009).³⁸ The LIMS initiative garnered significant industry support, and McLeod and Ramjee (2007) claim that it contributed significantly to the market's understanding about the challenges in providing affordable medical scheme cover to the low-income segment. Several factors were identified as being significant obstacles in extending medical scheme coverage to the low-income segment. These include the high levels of unemployment in South Africa, the lack of affordability of current medical scheme contribution rates and the high actuarial cost of the PMB package (McLeod & Ramjee, 2007).

The LIMS process highlighted the absence of any affordable low-cost options that would directly cater for the low-income earners. In accordance with the risk profile of the current lower cost options that are available in the market, McLeod and Ramjee (2007) argue that they appear to be competitively priced offering reduced benefits and aggressive managed care interventions. If a common set of benefits were to be offered to the low-income segment and these members were pooled at the scheme level rather than at option level (which is currently the case), they suggest that the contributions would need to rise sharply to compensate for the lack of income cross-subsidies since a LIMS would have a largely homogenous group of low-income earners. Magennis and van Zyl (2009: 20-21) list the key recommendations that emerged from the LIMS consultative process as follows:

- LIMS should be open to any formal sector employee or self-employed person (and their dependents) who earns less than R6 500 per month (in 2005 terms).
- LIMS should be offered by new schemes and also under any new benefit options offered by existing schemes.
- Employers and employees should each make a 50 percent contribution towards the monthly premium. In addition, the employee's share of the premium should not exceed 8 percent of household income.

³⁸ Low-income was broadly defined as being between R2 000 and R6 000 per month (McLeod & Ramjee, 2007).

- Government should directly subsidise the monthly LIMS membership contributions.
- The common set of benefits offered under LIMS should include a package with certain acute and some chronic out-patient or ambulatory care. LIMS members would be expected to obtain in-patient care at public healthcare facilities at no cost.
- LIMS schemes would be ring-fenced from other medical schemes and thus, any cross-subsidisation across other medical schemes or even financial transfers under the REF would not be permitted.

The LIMS initiative proposes to extend risk-pooling to a larger proportion of the South African population, who often bear an excessive amount of out-of-pocket expenses with respect to their healthcare expenditure. Nevertheless, Magennis and van Zyl (2009) criticise the LIMS recommendations for a number of reasons. First, to make medical scheme coverage affordable for LIMS members, the recommendation is to simply offer LIMS members a restrictive benefit package. As the authors correctly argue the focus should rather be upon curtailing the high cost of medical scheme coverage by addressing the problem of supply-induced demand. Second, the LIMS process did not adequately consider the option of extending medical scheme membership to low-income individuals within existing medical scheme structures (Magennis & van Zyl, 2009). It does not preclude the possibility of some form of government subsidy or income cross-subsidisation to assist low-income earners with a lower contribution rate. Finally, Magennis and van Zyl (2009: 21) are concerned that the creation of LIMS schemes may "...compound inequities and entrench further fragmentation [within] South Africa's healthcare financing [sector]".

Whilst the LIMS proposals were well-received in 2006, there has been little further movement with respect to their implementation. In addition, before the proposals could be implemented enabling legislation would still be required. The Medical Scheme Amendment Bill No. 58 of 2008 does include a single enabling line, but it does appear that the LIMS initiative has lost impetus. This may be partly due to

government's determination to press forward with its plan to adopt a National Health Insurance (NHI) system.

National Health Insurance (NHI)

The government has publicly announced its intention to introduce a mandatory health insurance system with the view to provide universal health coverage for all South Africans. Until only recently, the details surrounding the National Health Insurance (NHI) system were somewhat sketchy and most of the debate tended to revolve around conjecture and speculation rather than factual discussion. Nevertheless, the Taylor Committee's findings that were published in 2002, Department of Social Development (2002) and Department of Health (2002) all provided some insight into the direction in which government was likely to proceed with its roll-out of a NHI system for the country. More details have since emerged about the proposed NHI system with the release of a series of documents during 2009 (Econex, 2009a).³⁹ Further confirmation that the government was determined to press ahead with its plan for a NHI system was evident when the Department of Health established a National Health Insurance Advisory Committee back in 2009. The National Health Insurance Advisory Committee is expected to advise the Minister of Health with respect to the development of policy and legislation relating to the phased roll-out of a NHI system in the country.

Under new proposals tabled by the African National Congress (ANC) in 2010, a National Health Insurance (NHI) scheme is to be implemented for all South Africans from 2012 onwards and is to be funded by dedicated new taxes. The scheme – to be implemented over 14 years – would guarantee a full range of healthcare services to

³⁹ Department of Health. (2009a). *National Health Insurance Plan for South Africa*. Pretoria: National Health Insurance Task Team – Department of Health.

African National Congress. (2009a). *National Health Insurance Policy Proposal*. ANC National General Council. Released 22 June 2009.

African National Congress. (2009b). National Health Insurance: A unified, equitable and integrated national health system that benefits all South Africans. *ANC Today*, 9(29) 24–30 July 2009.

African National Congress. (2009c) *National Health Insurance (NHI)*. ANC National General Council. Released September 2010.

Department of Health (2009b) Establishment of the National Health Insurance Advisory Committee. Government Gazette, 11 September 2009.

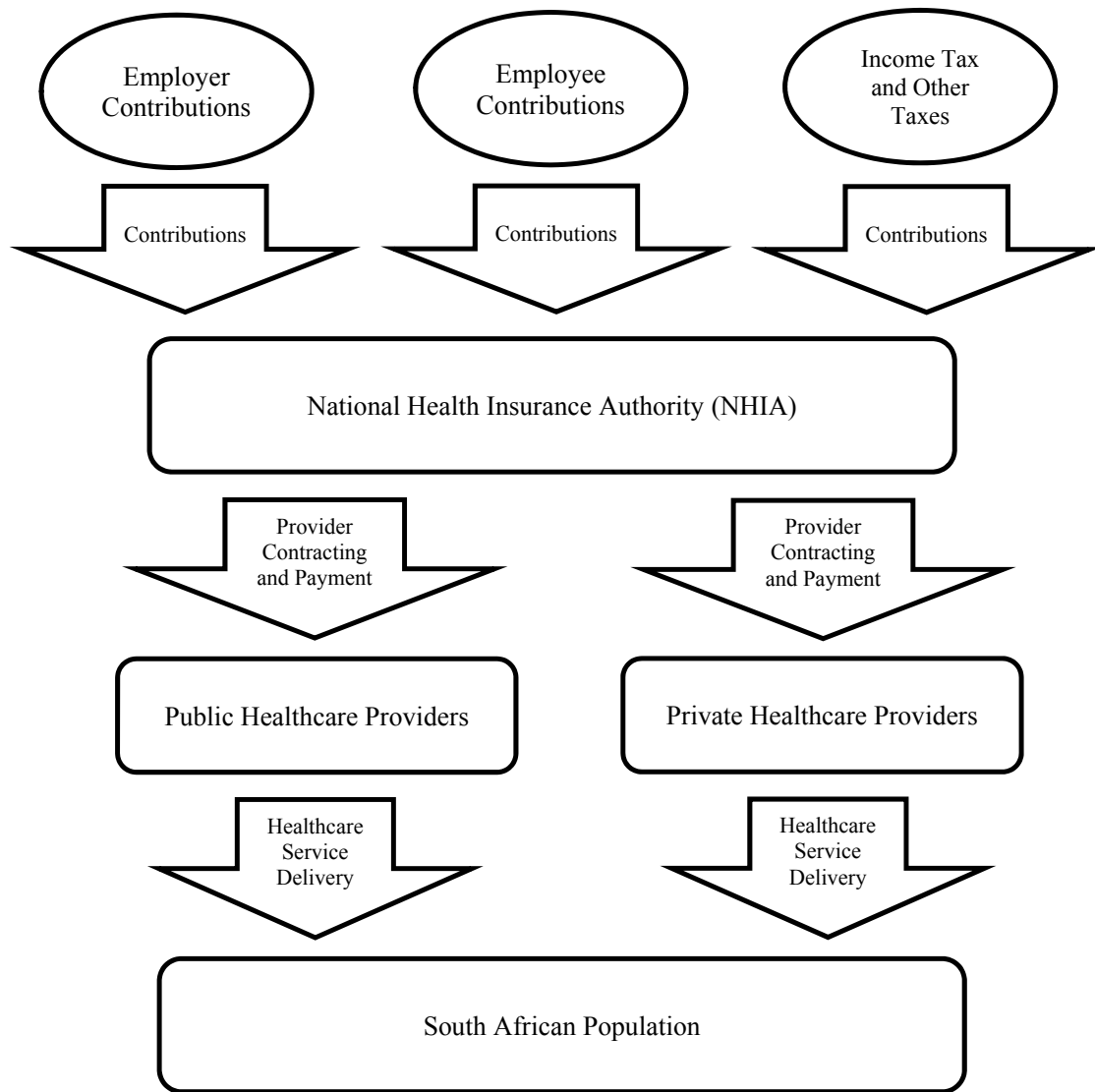
everyone, regardless of their ability to pay and it is intended to make some private sector facilities available to people currently denied access to private healthcare services (Department of Health, 2010). The forecasted initial increase in the South African government's health budget of R117 billion for 2012 will see health's share of government spending to increase by one fifth, from 12 percent to 14.5 percent. To pay for this increase in public healthcare expenditure South Africans can be expected to brace themselves for significant tax increases. Under the plan, all South Africans would have to be members of the scheme. Everyone earning an income would have to contribute, whilst no one would be obliged to use the services. Private hospitals would be encouraged to join the system but would not be obliged to do so. Individuals would be free to buy private medical insurance and to go to hospitals that did not participate in the scheme. NHI members will be free to decide which participating clinic or doctor in their area to register with, but will not have any choice with respect to specialists or hospitals that they will be referred to in the event that the primary healthcare facility or practitioner is unable to treat their illness or injury. Medical treatment will be free at the point of service. According to government's proposal the greatest portion of the cost for the new NHI will come out of the existing healthcare budget. The outstanding amount will be funded by one or more of the following: a ring-fenced VAT increase, a surcharge on personal income tax, a special payroll levy and the abolition of tax exemptions for medical costs.

In accordance with the information available thus far, a National Health Insurance Authority (NHIA) will be established to oversee and operate a single-payer fund (Econex, 2009a). The NHIA will pool resources and purchase healthcare services on behalf of the entire South African population. Econex (2009a) posit that funding will be sourced from a combination of employer and employee contributions, and existing fiscal funding from the national and provincial health budgets. The employer and employee contributions would immediately imply that a further payroll tax will be levied. Despite the African National Congress (2009b)'s assurances that the additional tax is likely to be modest, it does raise concerns about further increasing the burden of taxation. Moreover, the current proposals advocate the removal of the tax deductibility of medical scheme contributions, which will further increase the tax

burden for current medical scheme members. The current proposals also provide for some tax relief for certain categories of low-income employees, whereby they will be exempt from making contributions to the NHIA. Whilst the government argues that this funding arrangement should be sufficient to fund the NHIA, it is unclear whether the pool of resources will be sufficient to cover the healthcare needs of the entire South African population (Econex, 2009a).

Figure 2-1 illustrates the proposed funding arrangement under the South African NHI system. Contributions from the three sources of funding are pooled together to purchase healthcare services from both the public and private healthcare sectors (Econex, 2009a).

Figure 2-1: Proposed Funding Arrangement for the South African NHI System



Source: Econex (2009a: 4)

Government believes that through the NHI system significant cost savings will be realised. In particular, the creation of a single funder and purchaser of healthcare services will translate into considerable cost savings with respect to administration. Econex (2009a) mention that these cost savings are a result of the move towards a single purchaser of healthcare services as opposed to the multi-purchaser model, whereby medical schemes reimburse healthcare providers for services rendered to

medical scheme beneficiaries. It is put forward that under the NHI framework, healthcare services will be delivered by both public and private healthcare providers, and funding for these healthcare services will be forthcoming from the NHIA. Healthcare service providers would have to be accredited by the NHIA based upon certain predefined criteria (Econex, 2009a). The service delivery model will see public and private healthcare providers, including healthcare facilities, contracted by the NHIA (Econex, 2009a). Thus, the preferred model for the purchase of the healthcare services would be on a contracted basis, but the nature of the contracting relationship between the government and in particular, the private healthcare sector requires further clarification. According to Econex (2009a), contracted healthcare providers would render their services to both contributory and non-contributory households based upon the principle of universal access on a per capita basis.

From the African National Congress (2009a) NHI policy proposal document, Table 2-1 outlines the suggested list of healthcare services that would be covered under a proposed standard NHI benefit package. The proposed standard NHI benefit package represents a considerable improvement over what is currently being offered through the public healthcare sector (Econex, 2009a).

Table 2-1: Suggested List of Healthcare Services covered under the Proposed Standard NHI Benefit Package

Primary Care and Preventative Healthcare Services	Mental Healthcare Services
In-Patient Care	Dental Services (excluding cosmetic dentistry)
Out-Patient Care	Substance Abuse Treatment
Emergency and Trauma Care	Chiropractic Services
Prescription Medication	Rehabilitation Care
Appropriate Technologies for Diagnosis and Treatment	Hearing Care (including the supply of hearing-aids)
Basic Vision Care and Vision Correction Devices (excluding laser vision correction for cosmetic purposes)	

Source: African National Congress (2009a) - NHI Policy Proposal Document

The newly-introduced standard benefit package that would be offered under the NHI framework with zero co-payments is likely to induce a significant increase in the

demand for healthcare services and this would occur in the context of an already capacity-constrained healthcare sector. One need only consider the shortages currently being experienced with respect to generalist and specialist practitioners.⁴⁰ Econex (2009a: 5) argue that given the finite healthcare resources available in the country, “...it is inevitable that a major increase in demand for healthcare will require some form of rationing to match demand and supply”. Under the proposed framework, it is clear that private healthcare providers cannot be mandated to sell their services to the NHIA. Therefore, unless the capitation remuneration arrangements are sufficient to persuade private healthcare providers to contract with the NHIA, the public-private partnership to supply healthcare services will simply revert to a public healthcare delivery system. African National Congress (2009a) envisages no differentiation in the price paid to public and private healthcare providers, which implies that the capitation price will have to converge towards a price that private healthcare providers are willing to accept for their services. Econex (2009a) argues that this may lead to doubts about the financial sustainability of the NHI framework, if indeed; the NHIA decides to pursue contracting quality healthcare services from the private healthcare sector.

Another contentious issue relates to the role that current medical schemes will play under the NHI system. One model suggests that medical schemes could continue to operate as private health insurers offering additional top-up health insurance cover for those private healthcare services not covered under the standard NHI healthcare package. This would be similar to how private health insurance operates in the United Kingdom, whereby individuals can purchase top-up private health insurance over and above the healthcare benefits they receive under the National Health Service (NHS).⁴¹ In the United Kingdom, private health insurance policyholders also benefit by being able to secure private healthcare treatment timeously - thereby escaping the NHS waiting lists. An alternative idea could see medical schemes acting as intermediaries

⁴⁰ Refer to the further discussion in Subsection 2.3 relating to the regulation of healthcare provision in South Africa.

⁴¹ The National Health Service (NHS) was established in 1948 in the United Kingdom. It provides free healthcare for all United Kingdom citizens and permanent residents, regardless of their age, occupation or ability to pay. The NHS is funded through National Insurance contributions (a payroll tax) and general taxation via the national health budget.

between healthcare providers and the NHIA; however, it is unclear as to what would be their exact role or function under this model.

The central premise behind the introduction of the NHI system is to expand healthcare coverage to all South Africans. The principle of *solidarity* forms the backbone to the current set of proposals underlying the NHI system, whereby government advocates the notion that the delivery of healthcare services should be based upon need rather than upon the ability to pay. By implication, free access to healthcare services with zero co-payments, at the point-of-service, regardless of the person's ability to pay, will require a significant degree of income cross-subsidisation between those individuals who can pay for healthcare and those who are unable to do so (Econex, 2009a). The proposed NHI framework does pose several challenges with respect to implementation and operational procedures. Nevertheless, government is determined to proceed on this path to provide universal coverage within a fourteen-year timeframe. Econex (2009a) point out that this is indeed an extremely optimistic target, considering that only two countries in the world have managed to achieve universal coverage in less than 40 years from their initial enabling social health legislation – Japan (36 years) and South Korea (26 years).

Taxpayers are likely to see a rise of more than 10% in their total tax liability. There are approximately only 5.3 million taxpayers in South Africa and of these, 1.2m pay approximately three-quarters of all personal and company tax collected. South African taxpayers are going to be expected to fork out an additional R247 billion to finance the government's proposed social security fund (National Security Savings Fund), a basic income grant for the unemployed and the national healthcare insurance scheme (NHI). This will increase government expenditure from around 31 percent to more than 40 percent of Gross Domestic Product (GDP) and the tax burden from 27 percent to 37 percent.

Given that most of the burden of taxation falls upon only a fraction of the employable adult population, it raises the question whether or not South Africa's massive social and healthcare spend is sustainable in the medium term.⁴² At present South Africa is already one of the countries with the highest proportion of people depending upon social grants. Approximately 13 million (or 26 percent of the total population) depend upon these social grants. Therefore, it translates into the fact that each taxpayer is effectively supporting two people on social grants. Furthermore, government's plans for a basic income grant to the unemployed will add another 7 million to the list of dependents. A sharp rise in taxation and contributions to retirement and medical scheme funds is unavoidable if the plans for a social security fund and the NHI are to be realised.

From Pauly (1974)'s analysis one can infer that public intervention, as in the case of introducing mandatory health insurance via the NHI, may be a solution to improve the market outcome observed under imperfect information. He explains that compulsory insurance could ultimately assist insurers in collecting more accurate information about the purchasers of voluntary additional insurance cover. According to theory of adverse selection, high-risk individuals would demand more insurance coverage and hence, buy coverage over and above that of the mandatory NHI insurance. Therefore, health insurers, in this case medical schemes, can potentially identify more accurately the high-risk individuals from the low-risk individuals. Thus, by implication those who seek say, top-up medical scheme cover, will be revealing themselves as higher risks. However, it is not appropriate to extend the adverse selection theory predictions too far in relation to the South African private health insurance sector. Risk aversion, wealth and income, the actual (or even perceived) quality of healthcare treatment that one might expect to receive under the NHI, the ability to secure treatment timeously may all be factors that affect whether or not a person decides to purchase additional top-up health insurance cover (over and above the standard NHI benefit package). The principles of *open enrolment* and particularly, *community rating* would in any event hinder any advantage one could possibly hope to glean from such information. Thus,

⁴² All South Africans consumers do contribute to the state's coffers via indirect taxes, such as Value-Added-Tax (VAT), or through customs and excise duties, when goods are imported into the country. Therefore, it is perhaps presumptuous to suggest that only a fraction of the population bears most of the tax burden. It does imply that the tax base is broader than one might anticipate.

the application of Pauly (1974)'s analysis may not be amenable to South Africa's private health insurance sector.

2.2.5 Brief Assessment of the South African Private Health Insurance Industry

Since 1994, the South African private health insurance industry has undergone a substantial shift towards the principle of *solidarity* and away from the principle of *mutuality*. A key mechanism to achieve government's objective of extending coverage in terms of both the number of medical scheme members and the benefits that they (and their dependents) enjoy has been the rewriting of the legislation governing medical scheme business (Doherty & McLeod, 2003). The Medical Schemes Act No. 131 of 1998 reflects the underlying changes in government policy in favour of increased regulatory control of not only medical schemes but also those entities that are contracted out by medical schemes to perform various allied services. These entities include medical scheme administrators, managed care organisations and medical scheme intermediaries (brokers) (Pearmain, 2000).

Doherty and McLeod (2003) argue that the explicit tool utilised by the Medical Schemes Act No. 131 of 1998 to achieve the stated objective of increased medical scheme coverage is the combination of three key factors; mandatory open enrolment, community-rating and the Prescribed Minimum Benefits (PMBs) package. Although, it is acknowledged that community-rating and PMBs would likely lead to an increase the costs of cover for younger (and/or more healthy) individuals, it was hoped that the new legislation would, through the application of these three factors, open up medical schemes to a large and more diverse population of potential members. In particular, it was hoped that the low-income segment of the market would have greater access to affordable medical scheme coverage. If successful, it would create larger risk pools and thus, encourage medical schemes to compete on the basis of efficiency, which would ultimately lead to lower overall costs of health insurance cover for the majority of medical scheme members (Doherty & McLeod, 2003).

Considering the significant legislative changes that have taken place with respect to medical scheme business since the first democratically-elected government took office in 1994, it is necessary to evaluate the success that the regulatory reforms have had in achieving the stated-objective of increased medical scheme coverage.

Recent Trends in the Number of Registered Medical Schemes, Membership and Solvency

The number of medical schemes has steadily declined since the early 1990s. In fact, according to McLeod and Ramjee (2007), this downward trend has been observable since as far back as 1974. In 1974, there were a total of 305 medical schemes, of which, 252 were registered medical schemes and 53 so-called exempt medical schemes (McLeod & Ramjee, 2007).⁴³ Over time many of these exempt (or Bargaining Council) schemes have gained registered status and for those that remain they represent only a small fraction of the total number of medical scheme beneficiaries covered.⁴⁴ Doherty and McLeod (2003) indicate that of the Bargaining Council schemes that remain, they tend to offer only primary healthcare benefits by a salaried panel of healthcare providers.

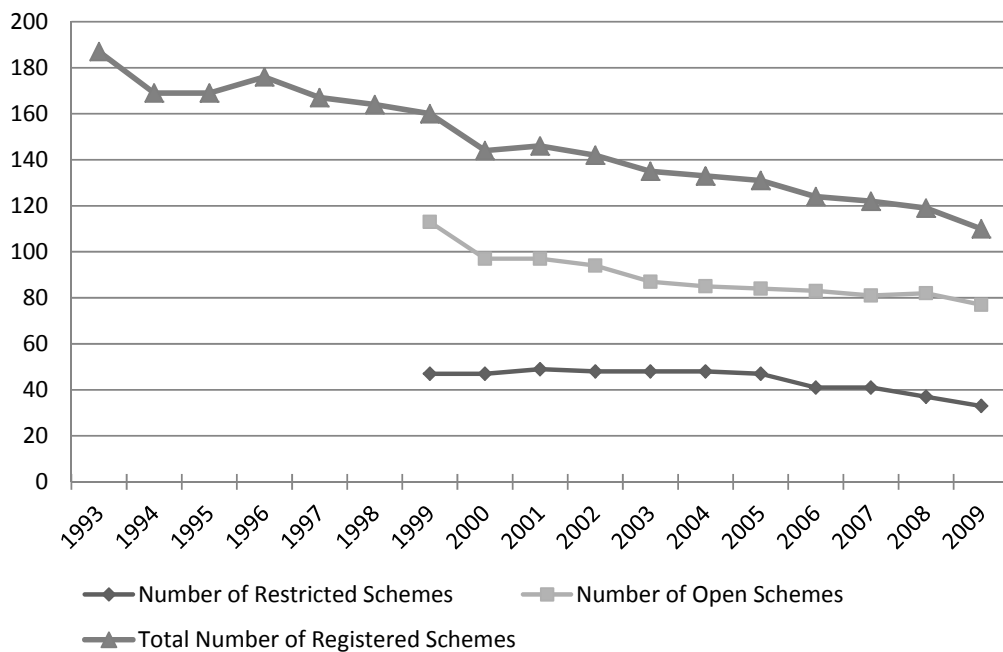
Excluding the Bargaining Council medical schemes, Figure 2-2 illustrates that the total number of registered medical schemes has declined steadily from 187 in 1993 to only 110 by 2009. The decline in the total number of registered medical schemes may be viewed as a period of consolidation as many amalgamations have occurred during this time. A more recent development is to classify registered medical schemes along

⁴³ Exempt medical schemes, later referred to as Bargaining Council schemes, were medical schemes that were granted certain exemptions from medical scheme legislation – originally, with respect to the Medical Schemes Act No. 72 of 1967 and later the Medical Schemes Act No. 131 of 1998. Historically, these schemes covered members of the police, defence force and correctional services (Doherty & McLeod, 2003).

⁴⁴ The Council for Medical Schemes has for the most part ceased reporting on Bargaining Council medical schemes within their Annual Reports (Council for Medical Schemes, 2004).

open versus restricted lines.⁴⁵ From Figure 2-2, it is clear that the number of both open and restricted medical schemes has steadily declined over the period 1999 to 2009. Since 1999, the number of medical scheme members in open and restricted schemes has diverged significantly, with open schemes benefitting from medical scheme intermediaries (brokers) aggressively targeting people in restricted schemes in order to move them over to open schemes. In addition, some restricted medical schemes have simply opted to open up their membership and thus, convert their restricted medical scheme status to that of an open medical scheme.

Figure 2-2: Trends in the Total Number of Registered Medical Schemes, Number of Open and Restricted Medical Schemes (1993 – 2009)



Source: South African Council for Medical Schemes Annual Reports (1993 – 2010)

⁴⁵ As note previously, open medical schemes permit new individuals from the general public to become members, subject to a mandatory waiting period and in certain cases late joiner penalties; whilst restricted membership schemes only permit individuals that meet the specified criteria of the scheme to become members. The basis of the restriction can only be employment or former employment in a profession, trade, industry or calling, or by a particular employer or class of employer.

As at the 31 December 1993, the total number of registered medical scheme members stood at 1,993,417 and the total number of registered medical scheme beneficiaries (principal members and dependents) was 5,399,886 individuals.⁴⁶ Over the next six-year period there was modest growth in the number of members purchasing cover from registered medical schemes – the average annual growth rate over this period was 2.21 percent over the period 1993 to 1999. By the end of 1999, the number of members covered by registered medical schemes stood at 2,276,495 and the total number of beneficiaries was 6,025,550. Therefore, the membership data refutes the common belief that the deregulation that took place towards the end of the 1980s and the early 1990s had led to a significant decline in medical scheme membership.

Once the Medical Schemes Act No. 131 of 1998 became effective from 1 January 2000, one can observe a sharp rise in the number of medical scheme members; however, rather than being the result of government's regulatory reforms, it can be explained by the reclassification of Polmed and Transnet from exempt to registered medical scheme status. Figure 2-3 confirms that after the initial increase in the number of medical scheme members during 2000, for the next six years the trend in medical scheme membership reverted to the prior growth rates. The total number of medical scheme members increased from 2,598,865 in 2000 to 2,985,350 individuals by the end of 2006 – an average annual growth rate of 2.31 percent over the period 2001 to 2006. Medical scheme beneficiaries increased from 6,729,551 in 2000 to 7,127,343 individuals by the end of 2006 – an average annual growth rate of only 0.96 percent over the same period. Thus, growth in medical scheme membership that followed the implementation of the Medical Schemes Act No. 131 of 1998 did not translate into a significant increase in medical scheme coverage.

⁴⁶ Exempt (or Bargaining Council) medical scheme members and their dependents were excluded from the assessment of membership, as the reporting of these schemes by the Council for Medical Schemes is inconsistent. In fact, the Council for Medical Schemes ceased reporting the detailed membership figures for these schemes from 2005 onwards. The analysis is further complicated by changes in the definition of registered schemes over the 1993 to 2009 period, since medical schemes such as Transnet and Polmed that were previously classified as 'exempt', were suddenly redefined as registered medical schemes from the year 2000 onwards (McLeod & Ramjee, 2007). Thus, it is important to acknowledge the once-off rise in membership figures due to the Transnet and Polmed reclassification.

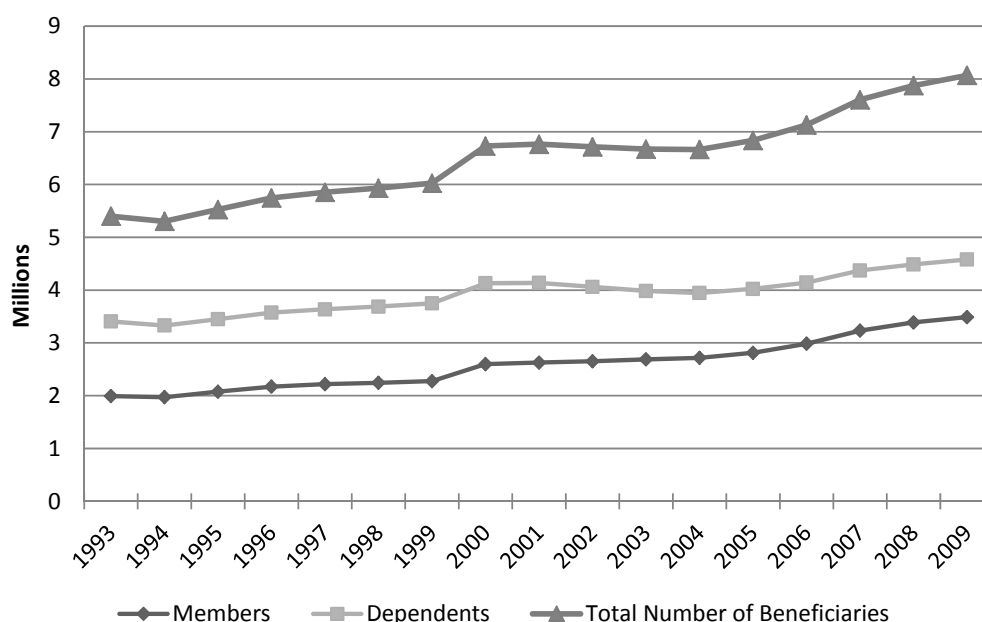
Since the end of 2006, the growth rate in medical scheme membership has improved marginally. By 2009, medical scheme membership had increased to 3,488,009 individuals - an average annual growth rate of only 5.19 percent over the period 2007 to 2009. Correspondingly, medical scheme beneficiaries have increased to 8,068,505 individuals by the end of 2009 – an average annual growth rate of only 4.13 percent over the period 2007 to 2009.⁴⁷ However, if one examines the data from the Council for Medical Schemes (2007)'s Annual Report, it is unclear whether suddenly the number of members and dependents are inflated by the reclassification of the remaining Bargaining Council medical schemes. In addition, the increase in medical scheme beneficiaries may in part be due to previously uncovered lives being incorporated into the Government Employee Medical Scheme (GEMS), which has been in operation since 2006 (McLeod & Ramjee, 2007).

As McLeod and Ramjee (2007) point out, perhaps a more meaningful measure of medical scheme coverage is to relate the number of beneficiaries to the total population. Based upon mid-year estimates of South Africa's total population from Statistics South Africa (StatsSA), the ratio of the number of medical scheme beneficiaries to the total population has marginally increased from 15.4 percent in 2000 to 16.36 percent in 2009.⁴⁸ However, McLeod and Ramjee (2007) note that in 1994, the ratio of the number of medical scheme beneficiaries to the total population was 17 percent. Thus, these findings are indicative of the failure of the new regulatory framework to broaden the scope of coverage to the population-at-large.

⁴⁷ The difference in the growth rates observed in relation to medical scheme members versus total medical scheme beneficiaries can be attributed to the problem of affordability rather than any real changes in family size (McLeod & Ramjee, 2007).

⁴⁸ Statistics South Africa. (2000). *Mid-year Population Estimates 2000*. Pretoria: Statistics South Africa. Retrieved February 13, 2010 from <http://196.25.65.5/publications/P0302/P03022000.pdf>.
Statistics South Africa. (2009). *Mid-year Population Estimates 2009*. Pretoria: Statistics South Africa. Retrieved February 13, 2010 from <http://www.statssa.gov.za/publications/P0302/P03022009.pdf>.

Figure 2-3: Total Number of Medical Scheme Beneficiaries – Members and Dependents (1993 – 2009)



Source: South African Council for Medical Schemes Annual Reports (1993– 2010)

The new regulatory framework that emerged with the enactment of the Medical Schemes Act No. 131 of 1998 did place considerable emphasis on ensuring that medical schemes maintain adequate solvency levels. Regulation 29 of the Medical Schemes Act No. 131 of 1998 prescribes the minimum accumulated funds that need to be maintained by registered medical schemes, where accumulated funds refers to net asset value of the scheme excluding funds set aside for specific purposes and unrealised non-distributable profits (Council for Medical Schemes, 2010). Accumulated funds expressed as percentage of gross contributions, or solvency ratio, must be maintained at a minimum level of 25 percent.⁴⁹ The solvency ratio is a useful measure of the financial soundness and sustainability of the medical scheme, and it further gives an indication of the ability of the medical scheme to withstand unforeseen (or unexpected) fluctuations in incurred claims (benefits paid out to healthcare providers for healthcare services rendered to medical scheme members and

⁴⁹ The minimum accumulated funds required to ensure that the solvency ratio is kept at the required level of 25 percent, is more commonly referred to as the *reserves* of the medical schemes (Council for Medical Schemes, 2010).

their dependents). Moreover, with improvements in the skills and capacity at the Council for Medical Schemes, monitoring the solvency of medical schemes has essentially become part of the ‘daily’ routine. Indeed, there is now detailed quarterly monitoring of all schemes and monthly reviews of medical schemes identified to be in financial distress (McLeod & Ramjee, 2007).

From Table 2-2, in 2000, approximately 3.22 million individuals were beneficiaries of registered medical schemes with a solvency protection of less than 10 percent of contributions (McLeod & Ramjee, 2007).⁵⁰ By December 2009, this number had been significantly reduced to only 2.25 million individuals who were beneficiaries of registered medical schemes with a solvency protection level of less than 25 percent of contributions – an improvement of 36.1 percent.

Table 2-2: Prescribed Solvency Levels and the Number of Medical Scheme Beneficiaries

Year	Number of Medical Scheme Beneficiaries within Medical Schemes that were Below the Prescribed Solvency Ratio	Number of Medical Scheme Beneficiaries within Medical Schemes that were Above the Prescribed Solvency Ratio	Total Number of Medical Scheme Beneficiaries
2000	3,224,080	3,505,460	6,729,540
2001	3,227,396	3,537,004	6,764,400
2002	3,770,379	2,943,755	6,714,134
2003	3,649,418	3,022,383	6,671,801
2004	2,614,433	4,048,130	6,662,563
2005	2,819,467	4,016,154	6,835,621
2006	3,363,751	3,763,592	7,127,343
2007	3,829,041	3,776,195	7,605,236
2008	2,058,427	5,816,399	7,874,826
2009	2,246,674	5,821,831	8,068,505

Source: South African Council for Medical Schemes Annual Reports (2000 – 2010)

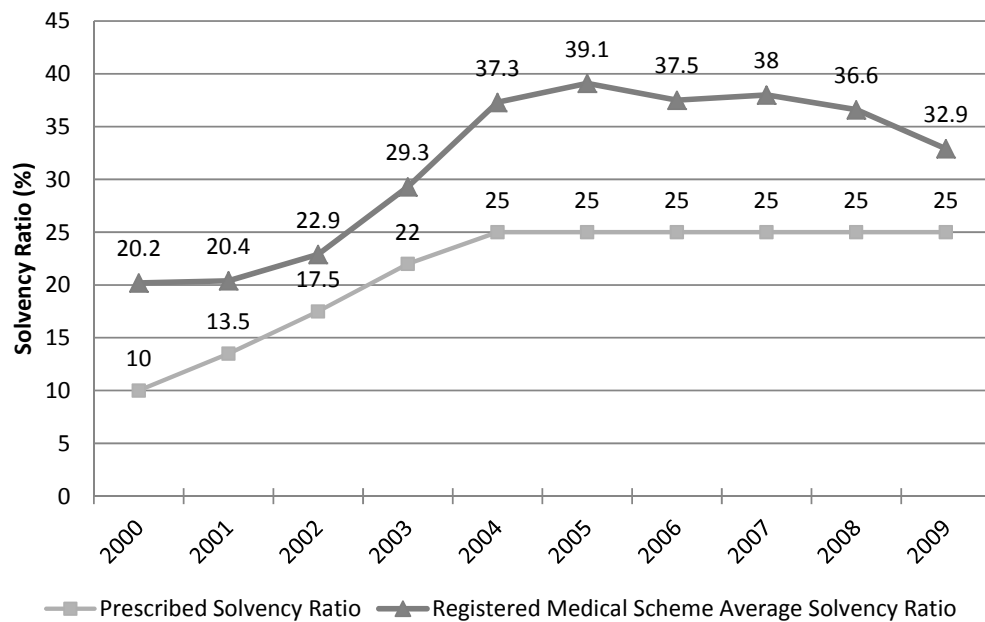
Since 2000, the average solvency ratio for all registered medical schemes has been consistently higher than the prescribed level. Figure 2-4 illustrates that in 2000, the average solvency ratio for all registered medical schemes was 20.2 percent, it rose

⁵⁰ The minimum solvency ratio initially started out at 10 percent in 2000 and through a transitional period, the solvency ratio was gradually increased to its current level of 25 percent.

steadily to reach a peak of 39.1 percent by 2005 and then it edged progressively lower to touch 32.9 percent by 2009 (Council for Medical Schemes, 2010).

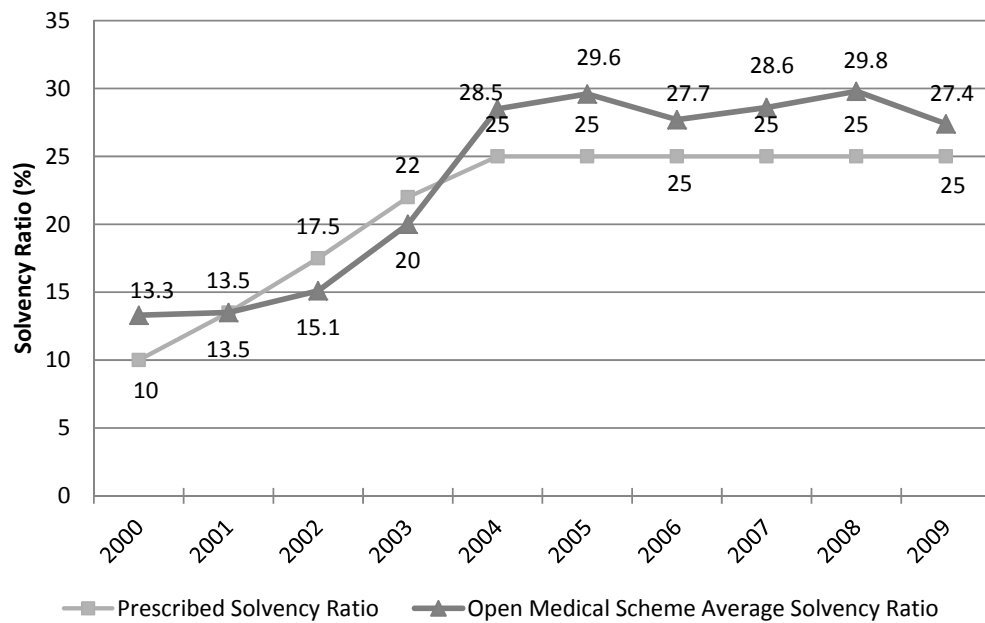
The average solvency ratio for open schemes was 27.4 percent in 2009 compared to 13.3 percent back in 2000. Although, from Figure 2-5 it is clear that the solvency ratio has dipped below the prescribed level on two occasions within the 2000 to 2009 period. It is however, currently above the prescribed level of 25 percent (Council for Medical Schemes, 2010). Open medical schemes do have greater difficulty in maintaining solvency levels when compared to restricted members schemes, since by definition, their pool of members can have a greater diversity of risks. In certain cases, open schemes may be reluctant to apply interim increases in members' contributions or indeed raise contributions too excessively as medical scheme brokers may persuade members to leave the scheme, thereby causing solvency ratios to deteriorate further (McLeod & Ramjee, 2007). Moreover, medical schemes that are growing their membership base rapidly, risk suffering strain with respect to having sufficient *reserves* to maintain appropriate solvency levels.

Figure 2-4: Industry Solvency Trends for all Registered Medical Schemes (2000 - 2009)



Source: South African Council for Medical Schemes Annual Report (2010: 199)

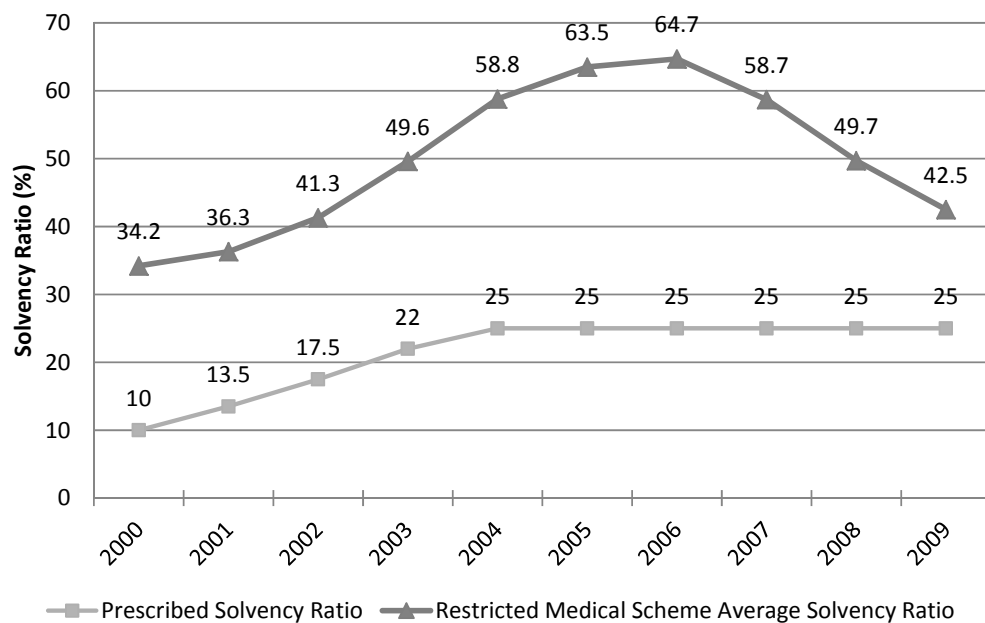
Figure 2-5: Industry Solvency Trends for Open Medical Schemes (2000 - 2009)



Source: South African Council for Medical Schemes Annual Report (2010: 199)

The average solvency ratio for restricted schemes was 42.5 percent in 2009 compared to 34.2 percent back in 2000. Although, 2009's figure of 42.5 percent is considerably higher than the prescribed level, it nevertheless represents a significant drop from the peak of 64.7 percent in 2006. Whilst restricted schemes tend to maintain higher solvency levels than their open scheme counterparts, Figure 2-6 illustrates that the average solvency ratio for restricted schemes has progressively deteriorated since 2006 (Council for Medical Schemes, 2010).

Figure 2-6: Industry Solvency Trends for Restricted Medical Schemes (2000 - 2009)



Source: South African Council for Medical Schemes Annual Report (2010: 200)

Measuring Competitiveness

It is necessary to prove that the South African medical scheme market is sufficiently competitive despite the legislation in order to apply Rothschild and Stiglitz (1976)'s predictions concerning the behaviour of insurance markets in the presence of imperfect information. Once the medical scheme market is shown to be sufficiently

competitive, it is then possible to test for the presence of adverse selection as described in the empirical section of this thesis.

Despite the decline in the number of registered medical schemes from 187 in 1993 to only 110 by 2009, South Africa's medical schemes market is characterised by few entry and exit barriers. Over the 1993 to 2009 period, many new entrants to the medical schemes market have been observed. Indeed, the current largest private open medical scheme by gross contribution income, Discovery Health Medical Scheme, is a relatively newcomer to the market, having only been established in March 1992. In addition, a significant number of registered medical schemes have also been liquidated, either voluntarily or involuntarily over this period. The number of new entrants in any given year has largely offset those medical schemes exiting the market through liquidation. Thus, the primary reason for the decline in the number of registered medical schemes between 1993 and 2009 has been due to a period of consolidation as evidenced by the large number of amalgamations. Hence, at first glance it does appear that South Africa's private health insurance sector is a vibrant and competitive market with few entry and exit barriers.

As a consequence of the extensive consolidation that has taken place within the medical schemes market; many observers have questioned whether the industry has become plagued by a high degree of concentration. Some may argue that as a direct result of the consolidation, the industry may have become far more concentrated in the hands of a few medical schemes (McIntyre et al., 2007).

To test whether the South African private health insurance sector is sufficiently competitive, the thesis adopts a simple index that will measure the size of medical schemes in relation to the industry as a whole; namely, the Herfindahl-Hirschman Index (HHI). The HHI may be a simple measure, but it is a useful indicator of the degree of competition (concentration) within a market (Hirschman, 1964; Brown & Warren-Boulton, 1988). The HHI is defined as the sum of the squares of the market

shares of the fifty largest firms within an industry, where the market share is expressed as a proportion of the total market share.⁵¹ The HHI can range between 0 and 1, with increases in the index indicating a decrease in competition, whilst the converse would be true if the index were to fall. Mathematically, the HHI can be described by Equation 2-2.

Equation 2-2: General Formula for the Herfindahl-Hirschman Index (HHI) (Hirschman, 1964; Brown & Warren-Boulton, 1988)

$$HHI = \sum_{i=1}^N s_i^2$$

where **HHI** represents the value of the Herfindahl-Hirschman Index, s_i is the market share of medical scheme i and N represents the number of registered medical schemes (in this case the fifty largest registered medical schemes by gross contribution income).

The HHI as described by Equation 2-2 can take on value of between $1/N$ and one. Brown and Warren-Boulton (1988) suggest that an HHI value of less than 0.1 indicates a sufficiently competitive market, whilst values between 0.1 and 0.18 indicate a certain degree of concentration and finally, a value in excess of 0.18 indicates a high degree of concentration.

In the case of South Africa's medical scheme industry, an HHI was calculated for each year corresponding to the period analysed in the empirical section of this thesis; namely, between 1993 and 2009. For each year, the fifty largest registered medical schemes by gross contribution income were considered, and the market shares utilised in the computation of the HHI corresponded to each of these medical scheme's share of the total gross contribution income of all registered medical schemes. The results of the HHI computation are summarised in Table 2-3. The results suggest the South African medical scheme industry is sufficiently competitive, but it is evident that the

⁵¹ If there are fewer than fifty firms making up an industry, it is appropriate to sum the squares of the market shares of all the firms within that specific industry (Hirschman, 1964).

degree of concentration has become progressively higher over the period 1993 to 2009.

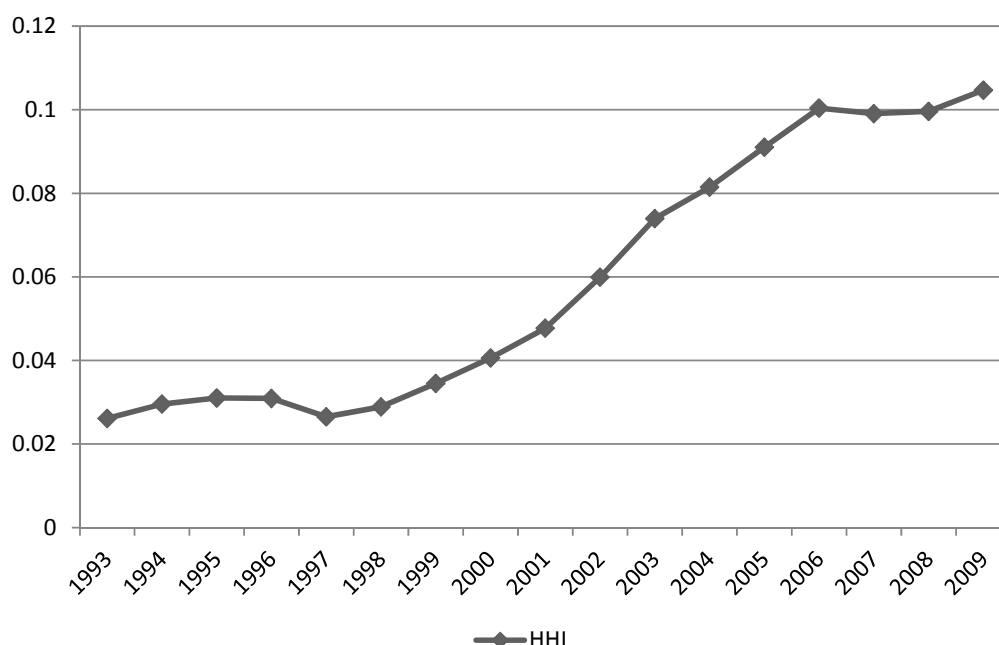
Table 2-3: Degree of Competition within South Africa's Private Health Insurance Sector (1993 - 2009)

Year	Herfindahl-Hirschman Index (HHI) Value
1993	0.0261
1994	0.0296
1995	0.0310
1996	0.0309
1997	0.0265
1998	0.0289
1999	0.0345
2000	0.0406
2001	0.0477
2002	0.0599
2003	0.0740
2004	0.0815
2005	0.0910
2006	0.1004
2007	0.0991
2008	0.0996
2009	0.1047

Source: Own calculations and the South African Council for Medical Schemes Annual Reports (1993 – 2010)

Figure 2-7 illustrates that there has been a noticeable upward trend in the HHI since 1998 for South Africa's medical scheme industry. In 1998, the HHI stood at 0.0289, but by 2009 it had reached 0.1047. Although, the degree of competition remains adequate, it is concerning that competition is becoming progressively weaker as measured by the HHI. In addition, it is noted that after 1993, the HHI does appear to trend marginally downwards, indicating a period of increased competition, which coincides exactly with Reekie (1999)'s view that the deregulation that took place towards the end of the 1980s and early 1990s, created a regulatory framework that promoted competition.

Figure 2-7: Herfindahl-Hirschman Index (HHI) for South Africa's Medical Scheme Industry (1993-2009)



Source: Own calculations and the South African Council for Medical Schemes Annual Reports (1993 – 2010)

2.3 Regulation relating to Healthcare Provision

At present total annual healthcare expenditure by both the public and private sectors is in excess of R220 billion and this represents approximately 8.5 percent of South Africa's Gross Domestic Product (GDP) (Erasmus & Fitchen, 2010). The revised national health budget for 2010 was R101.9bn⁵², which represents 3.9 percent of GDP (National Treasury, 2010). South Africa's healthcare expenditure relative to GDP exceeds that of many countries with a similar level of economic development and in some cases healthcare expenditure is at levels comparable with those observed in high-income countries such as the United Kingdom (McIntyre et al., 2007). Despite healthcare expenditure being relatively high by international standards, South Africa

⁵² This represents the amount allocated towards healthcare expenditure in the fiscus in real terms (2010 prices).

performs poorly with respect to many health status indicators, such as infant mortality and life expectancy. Thus, McIntyre et al. (2007) suggest that a strong argument can be made that South Africa rather than facing a lack of resources with respect to healthcare provision should utilise existing resources more efficiently and equitably.

McIntyre et al. (2007) points out that with respect to healthcare provision, South Africa has a two-tier healthcare system, which is delineated along socio-economic lines. A wealthy minority is serviced by the private healthcare sector and in this case, healthcare is typically financed through medical scheme cover, whilst the majority of South Africans rely upon a relatively under-resourced public healthcare sector (McIntyre et al., 2007). Although, individuals without medical scheme cover have access to free primary healthcare at public sector healthcare facilities, in many instances they still incur out-of-pocket expenditure for items such as non-prescription medication and general practitioner consultations (McIntyre et al., 2007). The authors emphasise the disparities in resources across the two-tiers. According to McIntyre et al. (2007), resources in the public healthcare sector have been relatively constant in real per capita terms, whereas there has been a significant rise in real capita spending per medical scheme beneficiary.

When one considers access to appropriate healthcare services or healthcare professionals within each sector, the disparities become quite palpable. Each public sector pharmacist serves between 12 and 30 times as many patients as their private sector counterparts (McIntyre et al., 2007). Each public sector generalist practitioner serves between 7 and 17 times more patients than private sector generalist practitioners. The figure for specialist practitioners is a shocking 23 times as many. Whilst for nurses the figure is 6 times as many patients. Furthermore, there are more than twice as many private hospital beds available to medical scheme beneficiaries than for those who depend upon the public healthcare sector (McIntyre et al., 2007).

From 1994 onwards, the first democratically-elected South African government made a concerted effort to abandon the principle of *mutuality* in favour of the principle of *solidarity* with respect to access to healthcare and the financing thereof. This paradigm shift was first articulated within the White Paper for the Transformation of the Health System in South Africa (Department of Health, 1997). McIntyre et al. (2007)⁵³ outlines the proposals from this white paper, some of the suggestions included:

- Primary healthcare should be the guiding approach that steers the transformation of the South African health system.
- Within the health system, equitable outcomes should be promoted and address past historical injustices. In particular, certain vulnerable groupings, such as women, children, the elderly, disabled and the poor should receive priority.
- Healthcare financing should serve as a redistributive mechanism with respect to resources.
- Mandatory social health insurance is an important tool in promoting equity within healthcare financing.

Many of the principles contained within the white paper were later reflected in the National Health Act No. 61 of 2003⁵⁴. The objective of the National Health Act No. 61 of 2003, as set out in Section 2 of the Act, refers to the establishment of a national health system that encompasses both the public and private sectors in healthcare delivery. In fact, Section 45 (1) of the Act empowers the Minister of Health to “prescribe mechanisms to enable a co-ordinated relationship between private and public health establishments in the delivery of health services”. The National Health

⁵³ McIntyre, D., Thiede, M., Nkosi, M., Mutyambizi, V., Castilo-Riquelme, M., Gilson, L., Erasmus, E., Goudge, J. (2007). *A Critical Analysis of the Current South African Health System*. Health Economics Unit, University of Cape Town and Centre for Health Policy, University of the Witwatersrand. This paper is also referred to as the SHIELD Report, as it is a report into the findings of the Strategies for Health Insurance for Equity in Less Developed Countries project. McIntyre et al. (2007) is merely the first report outlining the initial findings of the SHIELD project. The project’s objective is to investigate existing healthcare financing arrangements in three countries, namely; Ghana, Tanzania and South Africa – with a view to identify the major challenges for equitable outcomes within healthcare provision and delivery. The project has examined alternative approaches to health insurance to address equity challenges within national health systems.

⁵⁴ The National Health Act No. 61 of 2003 is also referred to as the National Health Act (2004), since it was only promulgated in July 2004.

Act No. 61 of 2003 goes further by actively promoting the rights of women, children, the elderly and disabled to have access to affordable healthcare. McIntyre et al. (2007) point out that the legislative changes, such as the National Health Act No. 61 of 2003, represent an explicit expression of government's intention to create a national health system that is far more inclusive whilst at the same time adopting redistributive mechanisms to correct the historical injustices of the previous *apartheid* government.

McIntyre et al. (2007) separate the regulation relating to healthcare provision into the following three categories:

- quantity and distribution of healthcare provision;
- quality of healthcare provision; and
- price of healthcare services and products.

Regulation with respect to the Quantity and Distribution of Healthcare Provision

In South Africa, legislation that directly influences the quantity and distribution of healthcare providers has been lacking (Cornell, Goudge, McIntyre & Mbatsha, 2001). Prior to 1994, private hospital groups were required to apply to the national Department of Health for permission to build new hospitals, increase the number of hospital beds in established hospitals or even obtain approval for the purchase of advanced equipment (McIntyre et al., 2007). After the first democratic elections in 1994, there was a considerable degree of confusion surrounding whether the national or provincial health authorities would be responsible for the licensing of private healthcare facilities. As a consequence and to allow for some time to develop appropriate policy guidelines surrounding the licensing of private healthcare facilities, the first democratically-elected government decided to impose a temporary moratorium on the building or expansion of private hospitals (Cornell et al., 2001).

Legislative clarification with respect to the responsibilities surrounding the licensing of private healthcare facilities was forthcoming with the promulgation of the National

Health Act No. 61 of 2003 (McIntyre et al., 2007). Under the new legislation, responsibility for the planning and development of new private healthcare institutions was devolved to the provincial health departments. Accordingly, a Provincial Health Council was established within each province to give advice concerning best practice and standards for healthcare institutions.⁵⁵ In addition, each council directs the development, procurement and the utilisation of healthcare technology within their respective province (McIntyre et al., 2007). The Director-General of the national Department of Health is tasked with the responsibility of issuing licenses (referred to as a ‘certificate of need’) for all private hospitals. According to Section 36(1) of the National Health Act No. 61 of 2003, a person may not:

“...establish, construct, modify or acquire a health establishment or health agency; increase the number of beds in or acquire prescribed health technology at a health establishment or health agency; or provide prescribed health services...without being in possession of a certificate of need”.

Before the Director-General issues (or renews) a certificate of need, he/she is required to consider the following criteria:

- “(a) the need to ensure consistency of health services development in terms of national, provincial and municipal planning;
- (b) the need to promote an equitable distribution and rationalisation of health services and health care resources, and the need to correct inequities based [up]on racial, gender, economic and geographical factors;
- (c) the need to promote an appropriate mix of public and private health services;
- (d) the demographics and epidemiological characteristics of the population to be served;
- (e) the potential advantages and disadvantages for existing public and private health services and for any affected communities;
- (f) the need to protect or advance persons or categories of persons designated in terms of the Employment Equity Act, 1998 (Act No. 55 of 1998), within the emerging small, medium and micro-enterprise sector;

⁵⁵ Each Provincial Health Council consists of provincial and local government officials, and healthcare officials.

- (g) the potential benefits of research and development with respect to the improvement of health service delivery;
- (h) the need to ensure that ownership of facilities does not create perverse incentives for health service providers and health workers;
- (i) if applicable, the quality of health services rendered by the applicant in the past;
- (j) the probability of the financial sustainability of the health establishment or health agency;
- (k) the need to ensure the availability and appropriate utilisation of human resources and health technology;
- (l) whether the private health establishment is for profit or not; and
- (m) if applicable, compliance with the requirements of a certificate of non-compliance”.

(Section 36(3) of the National Health Act No. 61 of 2003)

Therefore, the “certificate of need” is ostensibly a mechanism by which the South African Department of Health can control the distribution of healthcare services, influence the quantity and quality of healthcare services within the country (McIntyre et al., 2007). The National Health Act No. 61 of 2003 and in particular, the “certificate of need” have garnered considerable criticism from healthcare professionals and private hospital groups. Despite the vociferous opposition from these groups, the legislation was promulgated successfully (McIntyre et al., 2007). But it is only in 2006 that government initiated the implementation of the National Health Act No. 61 of 2003 and it has to be said that many of the provinces are still in the process of aligning their provincial legislation with the Act. Some of the criticism levelled against the new legislation stems from a view that government appears to be antagonistic towards for-profit private healthcare. It is regrettable that this is the perceived view of many private healthcare stakeholders. However, Biermann (2004) emphasises that it is important that in order to achieve a more equitable public-private healthcare mix that offers a sustainable (and adequate) healthcare delivery system for all South Africans, government needs to acknowledge that there is no such ideal as free healthcare – to coin the old adage – there is after all no free lunch. Healthcare services provided by the public sector are in any event funded by taxpayers, which include businesses and individuals. Without the profit motive, so much of what is

produced would simply cease to exist and this includes healthcare technologies and even medicines (Biermann, 2004).

As free market provocateurs would argue, in a free society, consumers reign supreme since profits accrue to those who serve the customer best (Biermann, 2004). The author points out those who do not meet customer's needs simply go out of business. Therefore, under the public provision of healthcare where there is no profit motive, it is difficult to ascertain whether or not patients' requirements are being met. Nevertheless, according to Biermann (2004) government appears to eschew the profit motive in the provision of healthcare, viewing it somehow as immoral. Without profits little or no progress would be made with respect to technological medical advancements, new medical treatments or pharmaceutical drugs. For example, the pharmaceutical industry devotes a considerable amount of resources and effort in the development of new medicines and by their very nature these ventures are fraught with uncertainty as there is no guarantee of success (Reekie, 1999). Arrow (1971: 223) states that:

“By definition, research is a venture into the unknown...The outcome of any research project is necessarily uncertain, and the most important results are likely to come from projects whose degree of uncertainty to begin with was greatest”.

Pharmaceutical companies rely entirely upon the efficacy of the cures that these medicines may offer to patients (Biermann, 2004). Therefore, it is the pursuit of profits that ultimately determine the path of new innovation with respect to medical advancements.

Biermann (2004) also considers the impact that the National Health Act No. 61 of 2003 will have with respect to healthcare professionals. Once again a free society advocates that healthcare providers are free to practise their trade where they wish and to provide their services to whom they choose – provided that these transactions are voluntary (Biermann, 2004). Contrast this with a system that dictates where healthcare providers may offer their services, the nature of what services and equipment they

may offer to patients, the fees that they can charge for their services, the type of medicines (generic or otherwise) they may prescribe, what prices pharmaceutical companies may charge for medicines, and even how many hospital beds there should be within healthcare facilities (Biermann, 2004). The author argues that these draconian measures to rebalance the public-private healthcare mix are a planning tool of a society that is not free. For Biermann (2004), the pre-eminent planning tool of a free society is the profit and loss system.

Some proponents of the new legislation, such as McIntyre et al. (2007) argue that the profit-loss system, leads to a misallocation of healthcare resources. This is perhaps clearly evident when one considers the over-concentration of capacity in terms of private healthcare facilities and providers within urban areas versus the under-provision experienced by rural communities (Biermann, 2004). Hence, the “certificate of need” is earmarked as the mechanism that can address these imbalances. However, as Biermann (2004) indicates that there is little evidence to suggest that attempts by government to control the distribution of healthcare services are indeed successful. He argues that after decades of government control in the provision and delivery of healthcare in countries such as Canada and the United Kingdom, the principle of equity continues to remain an elusive goal.

Biermann (2004) goes further to assert that government’s introduction of the “certificate of need” is a precursor to chaos. The National Health Act No. 61 of 2003 requires the Director-General of the national Department of Health to make accurate and informed decisions with respect to the healthcare requirements of the country. Moreover, the Director-General is tasked to align the current healthcare infrastructure and capacity with the healthcare requirements of the South African population (Biermann, 2004). In his seminal paper entitled “*Der Wettbewerb als Entdeckungsverfahren*”⁵⁶, Hayek (1968) asserts that entrepreneurs can discover the exact needs of consumers only through the trial and error process of the competitive market and its guiding device – the price system (Biermann, 2004). Thus, without the

⁵⁶ The English title of Hayek (1968)’s paper is “Competition as a discovery procedure”.

profit motive, Biermann (2004) argues that it is questionable to suggest that government officials will be capable of allocating healthcare resources efficiently and equitably.

The Pharmacy Act No. 53 of 1974 together with the Pharmacy Amendment Act No. 88 of 1997 governs the licensing of wholesale and retail pharmacies. A person or entity wishing to establish a pharmacy must apply to the Director-General of the national Department of Health for a license (McIntyre et al., 2007). The Pharmacy Amendment Act No. 88 of 1997 established the new South African Pharmacy Council and set out the council's functions under the legislation, including the investigative and disciplinary powers of the council.⁵⁷ The ambit of the council was also extended to include the public sector. Moreover, the Pharmacy Amendment Act No. 88 of 1997 amended the provisions relating to pharmacy education and training, registration requirements and the ownership of pharmacies (McIntyre et al., 2007).

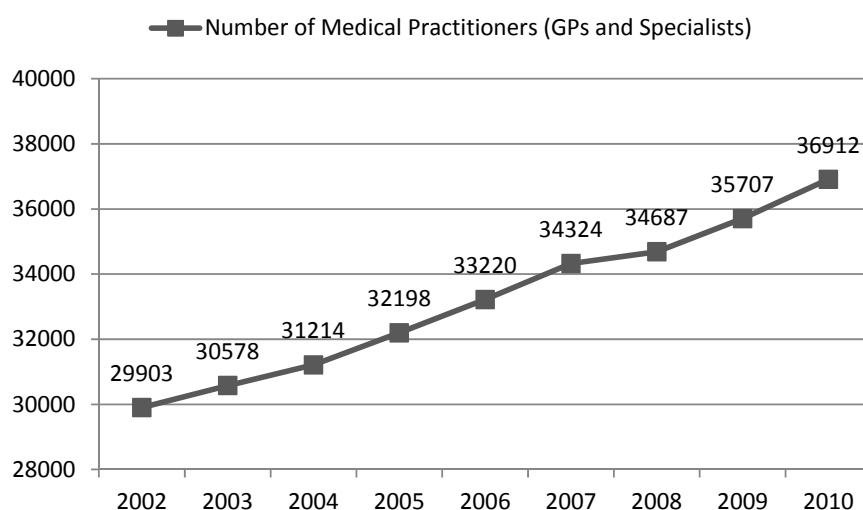
Whilst the “certificate of need” is an attempt to address the distribution of healthcare provision, McIntyre et al. (2007) maintain that it does not directly address the supply of healthcare professionals. South Africa is currently facing a severe shortage of healthcare professionals. Some estimates of the number of generalist and specialist (medical) practitioners and nurses appear to be woefully over-estimated (Econex, 2009b). Official estimates from the Health Professions Council of South Africa (HPCSA) indicate that at present the total number of generalist and specialist practitioners practicing in the country is equal to 36,912 (African National Congress, 2010). Furthermore, it is suggested that there has been a steady increase in the number of generalist and specialist practitioners since 2002.⁵⁸ Venter, Erasmus and Theron (2010) highlight the difficulties in obtaining accurate data regarding the number of healthcare professionals actively working in South Africa. The HPCSA figures include all medical practitioners who are registered with the council; however, some practitioners practice overseas and simply maintain their registration status with the

⁵⁷ Before a licence is granted to a new pharmacy, the South African Pharmacy Council is required to inspect the premises and report back to the Director-General of the Department of Health.

⁵⁸ Refer to Figure 2-8.

council (Venter et al., 2010). In addition, the HPCSA data include medical practitioners that no longer practice, for example, those individuals occupying managerial or other administrative positions.

Figure 2-8: Growth in Medical Practitioners (GPs and Specialists), 2002-2010



Source: HPCSA data from African National Congress (2010).

To obtain a more accurate assessment of the numbers of generalist and specialist practitioners practising within South Africa, Venter et al. (2010) explore alternate data sources. For the public sector, one possible source is to examine the government's personnel and salary administration system (PERSAL). In 2010, the number of generalist and specialist practitioners on the government's payroll was 11,309 and 4,442 respectively (Venter et al., 2010). Medical scheme data estimates that the total number of generalist practitioners to be 6,949 and 5,695 specialist practitioners operating within the private sector (Council for Medical Schemes, 2010).⁵⁹ To avoid double counting, Venter et al. (2010) point out that one should exclude from the public sector part-time and sessional medical practitioners that work primarily in the private sector and also exclude medical practitioners in the private sector that work primarily in the public sector. According to Venter et al. (2010), many industry

⁵⁹ The data was based upon the number of medical practitioners visited by medical scheme members and subsequently reimbursed by the schemes (Venter et al., 2010).

experts believe that the incidence of public sector medical practitioners working part-time in the private sector is far more prevalent than the reverse. Since neither dataset adequately captures the incidence of public and private sector cross-overs by medical practitioners, Venter et al. (2010) apply various possible scenarios to arrive at an estimate of the absolute number of generalist and specialist practitioners actively working in the country.⁶⁰

Venter et al. (2010) estimate that in 2010, the absolute number of generalist and specialist practitioners actively working in the country to be equal to 27,431. Of which the total number of general practitioners was 17,801, which equates to approximately 35.9 general practitioners per 100,000 people. The total number of specialist practitioners was equal to 9,630 or 19.4 per 100,000 people. Thus, the ratio of generalist and specialists practitioners per 100,000 people equates to 55.3. This is significantly lower than developed countries, such as Australia (247 per 100,000), United Kingdom (230 per 100,000) and the United States (256 per 100,000). It is even lower than other emerging economies, such as Brazil (185 per 100,000) and Mexico (198 per 100,000) (Venter et al., 2010). By 2020, based upon the current ageing profile of medical practitioners, training and attrition rates, Venter et al. (2010)'s analysis suggest that this ratio will deteriorate further to only 45.8 per 100,000.⁶¹

Various data sources can be combined to determine the total number of nurses practicing within South Africa. Once again utilising PERSAL, one is able to ascertain the number of nurses working within the public sector. Other alternative data sources include the All Media and Product Survey (AMPS), the Labour Force Survey (LFS) or registration information from the South African Nursing Council (SANC) (Econex, 2009b). In 2008, the AMPS data estimated the number of nurses to be 256,000, whilst LFS estimated a total of 189,000. According to SANC, there were 213,000 registered

⁶⁰ The scenario that Venter et al. (2010) finally decided upon as being the most appropriate was to adjust both the PERSAL and medical scheme industry figures in the following way: a 2.5 percent downwards adjustment was applied to the number of generalist practitioners in both the public and private sectors, whilst a further 5 percent was subtracted from the specialist figures in both sectors.

⁶¹ For low income countries, the mean value for the ratio of medical practitioners per 100,000 people is found to be in the region of 50 (Venter et al., 2010). Hence, by 2020, the ratio of medical practitioners per 100,000 people in South Africa is expected to fall well-below that level.

nurses in 2008 (Econex, 2009b). As Econex (2009b) note, both the AMPS and SANC data appear to overstate the true number of nurses in the country. In particular, SANC registration data includes nurses that could either not be working or working overseas, but are simply maintaining their registration status. The PERSAL data indicates that in 2008, 104,571 nurses were working in the public sector.⁶² Unfortunately, there is no reliable source of information relating to the number of nurses employed in the private sector. In 2008, SANC estimated there were 82,000 nurses employed in the private sector (Econex, 2009b). As noted already, this is likely to be overstated. The Hospital Association of South Africa (HASA) estimated only 40,000 nurses were employed in the private sector during 2008 (Econex, 2009b). Thus, based upon PERSAL and HASA's estimate, in 2008, a total of approximately 144,571 nurses were employed in the public and private healthcare sectors.

The common perception is that the quantity of healthcare professionals is significantly skewed in favour of the private sector. For example, Wade and Kahn (2007)'s analysis of the proportion of healthcare professionals across the public and private healthcare sectors, supports the view that the public healthcare sector appears to be woefully under-staffed in comparison to the private sector with respect to generalist and specialist practitioners. However, they acknowledge the paucity of the data that they utilised in arriving at their conclusion. It is therefore, not surprising that the paucity in the data has led many to believe that the disparity with respect to the quantity of healthcare professionals favours the private sector so greatly. Venter et al. (2010)'s analysis provides a more accurate assessment of the public-private mix with respect to the quantity of healthcare professionals employed within the South African healthcare sector. Contrary to popular perceptions, the authors find that the proportion of healthcare professionals across the public and private sectors are more equitable than previous studies suggested.⁶³ In particular, the majority (61.9 percent) of generalist practitioners are employed in the public sector. On the hand, the majority

⁶² Of the 104,571 nurses working in the public sector; 47,834 are professional/registered nurses (RNs), 22,702 are enrolled/staff nurses (ENs) and 34,030 are enrolled auxiliary/assistant nurses (ENAs) (Econex, 2009b).

⁶³ However, The number of individuals serviced by the public and private sectors are very different and therefore, despite the near equality with respect to the proportion of healthcare professionals across the two sectors, the reality is nevertheless

(56.2 percent) of specialist practitioners are employed within the private sector (Venter et al., 2010). Table 2-4 presents a summary of the results obtained by Wadee and Kahn (2007). Table 2-5 presents a summary of the results obtained from Econex (2009b) and Venter et al. (2010).

Table 2-4: Summary of Public versus Private Sector Mix - Healthcare Professionals (Wadee & Kahn, 2007) ⁶⁴

Healthcare Sector	Generalist Practitioners (percent)	Specialist Practitioners (percent)	Nurses (percent)
Public	27.4	24.8	58.9
Private	72.6	75.2	41.1

Source: Wadee and Kahn (2007: 143).

Table 2-5: Summary of Public versus Private Sector Mix - Healthcare Professionals (2010)

Healthcare Sector	Generalist Practitioners (percent)	Specialist Practitioners (percent)	Nurses (percent) ⁶⁵
Public	61.9	43.8	72.3
Private	38.1	56.2	27.7

Source: Econex (2009b), Venter et al. (2010) and own calculations.

McIntyre et al. (2007) argue that the main instrument available to government to influence the production of health professionals is to direct how many students are admitted for suitable training within the various education institutions, such as universities and nursing colleges. In 2006, the Department of Health initiated a programme to evaluate the human resource constraints within the healthcare sector and a plan (Human Resource Plan) was outlined to provide broad guidelines on how to address the staff shortages observed within the healthcare sector in the country (Department of Health, 2006). According to the Human Resource Plan, the Department of Health would be responsible for planning the production of healthcare professionals and establish certain targets for increases in the number of healthcare professional graduates relative to the total number of graduates produced in the country.

⁶⁴ Wadee and Kahn (2007)'s data is adapted from the work undertaken by Sanders and Lloyd (2005).

⁶⁵ Estimates for the number of nurses within each sector relates to the year 2008.

An adjunct to this mechanism would be to create an environment that is conducive to retain healthcare professionals trained within the country. One quarter of all medical practitioners trained in the country have left to pursue their careers elsewhere (Council for Medical Schemes, 2009). In the case of specialist practitioners, the country is currently unable to replace at a fast enough rate those specialists who are either emigrating or retiring. The African National Congress (2009: 20) state that:

“...it is estimated that about 67% of nurses who trained in the period 1997 to 2005 do not appear on the South African Nursing council register...Some leave the country to seek greener pastures in countries that pay them higher salaries such as Saudi Arabia, Oman, [United Kingdom], [United States], Canada and Australia.”

In addition, the public healthcare sector struggles to retain staff since private practice is always an attractive alternative to public service. In the public sector, just under 50 percent of general practitioner posts, 44 percent of specialist practitioner posts and 42% of all other healthcare professional posts are vacant (Council for Medical Schemes, 2010). As such there are no legal constraints applied to a healthcare professional wishing to enter into the private healthcare sector, with the exception of medical doctors and pharmacists who are now required to spend a year completing community service in a public healthcare institution upon graduation (McIntyre et al., 2007). The government has attempted to retain public healthcare professionals by implementing rural and ‘scarce skills’ allowances, which supplement their remuneration packages, for certain categories of healthcare professionals. Moreover, the South African government has entered into bilateral agreements with for example, the United Kingdom, to promote “ethical recruitment practices” (McIntyre et al., 2007: 38). In essence, these represent attempts by the government to restrict the ability of locally-trained healthcare professionals from seeking employment outside of South Africa. These measures are draconian and they may simply encourage school-leavers, interested in the healthcare professions, to pursue their studies elsewhere if at all possible. The notion that the government believes that it can (and should) dictate labour mobility with respect to the healthcare sector is unfortunate. The government should rather concentrate on creating a local environment that is conducive for individuals to remain in the country. Although, there are a myriad of factors that are causing many healthcare professionals to seek employment outside of South Africa,

addressing these common concerns should be a high priority for the government as they have deep and dire consequences for the wider population. Some of these factors include the excessive crime rate, deteriorating criminal justice system, corruption, mismanagement of government funds, lack of accountability, practices that promote mediocrity rather than excellence and the enrichment of a minority of elite, well-placed and well-connected individuals at the expense of the majority of South Africans. It is only through courageous leadership and moral fortitude that the South African government will be able to begin to place the country on a more prosperous and sustainable path that will ultimately benefit all South Africans.

Another policy instrument available to the government to improve the supply and distribution of healthcare professionals is to recruit healthcare professionals from other countries (McIntyre et al., 2007). The policy as discussed within Department of Health (2006) emphasises the objective of recruiting foreign healthcare professionals with the relevant skill set and competencies to work in particularly, under-serviced and remote areas of the country. Indeed, since 1994, South Africa has already recruited a number of general practitioners from countries such as Tunisia, Cuba and Iran. In 2004, the South African government also launched the Community Health Worker Programme to create community-based generalist healthcare workers that would assist with health promotion, disease prevention and primary healthcare (McIntyre et al., 2007).

Regulation with respect to the Quality of Healthcare Provision

The regulation governing the quality of healthcare provision in South Africa primarily involves imposing strict controls surrounding the production and sale of medicines, and safeguarding the high standard of education and training for healthcare professionals (McIntyre et al., 2007). Certain regulatory provisions permit the inspection of healthcare facilities; however, capacity is limited to carry-out such inspections or even to enforce compliance with directives to improve standards. Nevertheless, it is possible for a healthcare facility's licence to be revoked should it be

found that the quality of healthcare provided at a facility is deficient (McIntyre et al., 2007).

Significant emphasis has been placed upon regulating the quality of medicines supplied within the country. The Medicines and Related Substances Control Act No. 101 of 1965 (as amended)⁶⁶ specifies that only medicines that have been approved by the Medicines Control Council (MCC) and registered with the Medicine Regulatory Authority (MRA) may be distributed within South Africa (McIntyre et al., 2007). The primary function of the MCC is to evaluate applications for the registration of any new medicines. The MCC also applies a schedule to each medicine registered and these stipulate who may distribute these medicines and under what conditions they may do so. For example, Schedule 0 medicines may be sold by any retailer, whilst Schedule 1 to 6 medicines can only be sold by qualifying healthcare practitioners, subject to pre-specified conditions. Schedule 7 to 8 medicines may only be sold by a supplier who has the pre-requisite permit from the Director-General of the Department of Health. In addition, medical practitioners, nurses or dentists are required to attend a course and pass an examination before they can be licensed to dispense medication (McIntyre et al., 2007).

The MRA is responsible for licensing and inspecting the facilities of pharmaceutical manufacturers. Moreover, the MRA frequently tests locally-produced medicines and those medicines that are imported into the country to ensure that they adhere to specific quality and safety standards (McIntyre et al., 2007). A more recent phenomenon is the distribution and sale of counterfeit or unregistered medicines, the authors argue that the MRA has to a large extent been successful in curtailing these criminal practices.

⁶⁶ These amendments include the Medicines and Related Substances Control Amendment Act No. 90 of 1997, Medicines and Related Substances Amendment Act No. 59 of 2002 and Medicines and Related Substances Amendment Act No. 72 of 2008. Other relevant legislation includes the Medicines and Medical Devices Regulatory Authority Act No. 132 of 1998, Prevention of and Treatment for Substance Abuse Act No. 70 of 2008, Pharmacy Act No.53 of 1974 and Pharmacy Amendment Act No. 88 of 1997.

In South Africa, the responsibility with respect to the quality of training of healthcare professionals is largely delegated to professional councils (or associations). Thus, these councils are tasked to safeguard the level of competency for their respective healthcare professionals and to ensure that an appropriate standard of quality is maintained. The four primary professional councils in the country include the:

- Health Professions' Council;
- Nursing Council;
- Pharmacy Council; and
- Allied Health Professions' Council.

These bodies are also responsible for the registration of the qualifying healthcare professionals within their respective vocation. Although, these professional councils are responsible for maintaining a certain level of competency, the practicalities often translate into merely determining whether a candidate has obtained the necessary qualifications and completed any requisite practical training (including community service, residency or internship).⁶⁷ Therefore, McIntyre et al. (2007) state that the professional councils have very limited influence with respect to increasing the supply of healthcare professionals or for the matter directly influencing the level of competency of the healthcare professionals – it appears to be largely a ‘rubber-stamping’ exercise of certification rather than quality enhancement. Regulatory authority is devolved to these councils through primarily four pieces of legislation (as amended); namely, the Health Professions Act No. 56 of 1974⁶⁸, Pharmacy Act No. 53 of 1974⁶⁹, Allied Health Professions Act No. 63 of 1982⁷⁰ and Nursing Act No. 33 of 2005.

⁶⁷ The requisite training qualifications may be obtained either locally or internationally.

⁶⁸ As amended by the Health Professions Amendment Act No. 29 of 2007.

⁶⁹ As amended by the Pharmacy Amendment Act No. 6 of 1995, Pharmacy Amendment Act No. 88 of 1997 and Pharmacy Amendment Act No. 1 of 2000.

⁷⁰ As amended by the Allied Health Professions Amendment Act No. 6 of 2000.

At present, price regulation in the South African healthcare sector is dominated by legislation that relates to the prices of medicines. The Medicines and Related Substances Control Act No. 101 of 1965 (as amended⁷¹) makes provision for a Pricing Committee to formulate recommendations regarding medicine pricing regulation to the Minister of Health (McIntyre et al., 2007). McIntyre et al. (2007) summarises the significant elements of the current South African legislation and accompanying regulations with respect to the pricing of medicines. The current legislation and accompanying regulations, emphasise transparency with respect to the pricing of medicines to ensure that all stakeholders are aware of the manufacturer price for medicines, wholesaler or distributor fee and the dispensing fee.

A *single exit price* policy is in effect whereby all pharmaceutical manufacturers are required to sell a particular medicine at the identical price to all purchasers.⁷² As Davie and Urbach (2006) indicate, the single exit price requires pharmaceutical manufacturers and pharmaceutical importers to incorporate all their costs into the medicine's price. This includes the costs of distributing and delivering medicines to the final user (Davie & Urbach, 2006). Hence, this requirement does not permit the application of a different pricing structure to take into account factors such as the physical location of the patient, the cost to maintain medicines at appropriate temperatures or even consideration about the shelf life of certain medicines. In the past, pharmaceutical manufacturers simply applied price discrimination principles in response to varying price elasticities of demand.⁷³ Reekie (1999) argues that pharmaceutical manufacturers merely responded to observable price sensitivities by pricing at a lower level in the more price sensitive sector such as in the public healthcare sector. Therefore, through pursuing their own self-interest and attempting to maximise expected profits, pharmaceutical companies achieved a redistribution of income across the public and private sectors. However, by applying the single exit

⁷¹ As amended by the Medicines and Related Substances Control Amendment Act No. 90 of 1997, Medicines and Related Substances Control Amendment Act No. 59 of 2002 and Medicines and Related Substances Control Amendment Act No. 72 of 2008.

⁷² The *Single Exit Price* must be printed on the medicine packaging (McIntyre et al., 2007).

⁷³ Individuals serviced by the private sector are generally in full employment and are less price sensitive than poorer individuals serviced by the state (Reekie, 1999).

price policy it removes the possibility for this redistributive mechanism to flourish. In addition, Reekie (1999) contends that the enactment of the Pharmacy Amendment Act No. 88 of 1997, which repealed the ban on the corporate ownership of pharmacies, would have resulted in lower medicine prices at the point-of-service. Reekie (1999) claims that the establishment of larger retail outlets would encourage greater competition and at the same these outlets would be able to benefit from economies of scale.

Discounting of medicines is prohibited under the current legislation. McIntyre et al. (2007) state that prior to the current legislation, it was common practice for pharmaceutical manufacturers to offer private healthcare facilities and dispensing medical practitioners a substantial discount on medicines to ensure that these medicines were dispensed regularly and in the case of a private hospital included on the hospital's formulary.

Previously, significant mark-ups were applied to medicines throughout the supply chain. Thus, government sought to move retail pharmacies and dispensing medical practitioners away from the 'trade in medicine', and towards a model whereby they would simply receive a professional dispensing fee (McIntyre et al., 2007). However, these regulatory reforms have attracted significant opposition from wholesalers and retail pharmacy groups, and implementation has been plagued by several delays due to litigation. Therefore in 2010, the Minister of Health decided to propose a new dispensing fee structure. Under the current proposals, the dispensing fee structure would be applied across four categories of medicines – these categories would be determined by the medicine's *single exit price*.

The suggested four-tier structure would be as follows: medicines costing less than R75 would attract a dispensing fixed fee of R6 plus a dispensing variable portion of no more than 46 percent of the medicine's single exit price; medicines costing between R75 and R200 would attract a dispensing fixed fee of R51 plus a dispensing

variable fee of no more than 33 percent of the medicine's single exit price; medicines costing between R200 and R700 would attract a dispensing fixed fee of R15 plus a dispensing variable fee of no more than 15 percent of the medicine's single exit price; and medicines costing in excess of R700 would attract a dispensing fixed fee of R121 plus a dispensing variable fee of no more than 5 percent of the medicine's single exit price. If these adjustments to the dispensing fee structure are implemented, South African consumers can look forward to exorbitant increases in the price of medicines at the point-of-service as these dispensing fees are factored into prices.

Regulation exists enforcing a maximum that may be charged with respect to logistics, such as the transportation and distribution of medicines (McIntyre et al., 2007). Under the logistics fee cap, government's stated objective of improving access to affordable medicines for the poor will have the contrary effect. As Davie and Urbach (2006) argue the poor tend to reside far from major urban centres, thus the logistics fee cap will discourage distributors of medicines to supply these outlying areas. Moreover, the longer this policy is in force, it will ensconce a greater a misallocation of affordable medicines in favour of more wealthy consumers (Davie & Urbach, 2006). The logistics fee cap, as with any price control, is often promoted and devised as a mechanism to assist the poor and alleviate poverty. However, in almost every case, it achieves the exact opposite effect.

In South Africa, wealthier individuals tend to have access to high-volume, low mark-up retailers that are located in the major urban centres, whilst the poor tend to shop at low-volume, high mark-up retailers in a more rural setting. Indeed, Davie and Urbach (2006) point out that rural areas are typically not well-served by large high-volume retailers or pharmacies. Therefore, poorer patients purchase medicines from small-scale retailers that tend to have higher margins due to their low trade volume (Davie & Urbach, 2006). The fact that the poor tend to shop at these outlets is a factor of convenience, the travelling costs (and opportunity costs, such as time and effort) is likely to offset any discount received at the urban retailer. Hence, the authors suggest that poor consumers are merely behaving rationally by choosing to purchase their

medicines at these rural establishments. As with any form of price control, such as the logistics fee cap, it will not and cannot protect consumers or indeed induce lower medicine prices (Davie & Urbach, 2006). The authors claim that attempts by government to control prices simply interfere with the normal market process, reduces competition and ultimately harms the consumer.⁷⁴ Davie and Urbach (2006: 1) go further to explain that:

“...consumers are most effectively protected by open competition, which gives the maximum power to consumers to punish or reward [companies] based [up]on their performance”.

A policy whereby government endeavours to micro-manage the distribution of medicines through a logistics fee cap can, according to Davie and Urbach (2006), be expected to have a number of adverse effects and unexpected consequences. Hence, the bid to control economic outcomes through regulation is often based upon a narrow set of preferences, which risks alienating the general preferences of all market participants (Davie & Urbach, 2006).

Whilst the government recognises that modern medicines do require considerable research effort and cost to develop, and it further acknowledges that purchasers of these medicines should expect to contribute towards these costs, it believes that these costs⁷⁵ should be borne according to each country's ability to pay (Pillay, 2006). Pillay (2006) concedes that World Health Organization (WHO) endorses the view that the principle of differential pricing of essential medicines is an acceptable practice by certain pharmaceutical manufacturers. However, the Pricing Committee's view is that the purchase prices of medicines should relate to their therapeutic performance and take into account national socio-economic factors. According to Pillay (2006), many countries have managed to negotiate medicine prices with pharmaceutical manufacturers using evidence-based comparisons. It is therefore the desire of the Pricing Committee and the Department of Health to establish such a programme in

⁷⁴ Price controls essentially deviate from the notion that Posner (1974) espouses whereby government intervention should address market failures and enhance social welfare rather than attempt to tackle market 'problems', such as pricing behaviour.

⁷⁵ Government believes that these costs should first be accurately determined before the costs can be distributed amongst the purchasers of medicines.

South Africa (Pillay, 2006). At the outset, the Pricing Committee sought to ensure that South African purchasers of pharmaceuticals were not paying higher prices than their counterparts in other countries. Thus, the Pricing Committee proposed the implementation of a method for international benchmarking of manufacturer prices. Countries were selected on the basis of their stringent regulation with respect to the quality and price of medicines, and the easy accessibility of pricing data. Under the regulations, each pharmaceutical manufacturer of *originator medicines*⁷⁶ will be required to compare the price they charge for their product in South Africa with that charged in Australia, Canada, New Zealand and Spain (McIntyre et al., 2007). If the price is lower in any of the identified jurisdictions, they will be required to lower that price accordingly. McIntyre et al. (2007) state that in the case of generics, the price must be set at a level that is at least 40 percent below the price of the originator equivalent.

The Medicines and Related Substances Control Act No. 101 of 1965 (as amended) also promotes generic substitution. Pharmacists are required to inform patients about generic equivalents and to discuss with the patient the benefits of using the generic alternatives. McIntyre et al. (2007) actually stipulates that the legislation requires pharmacists to dispense the generic equivalent unless specifically instructed to by the patient; or expressly stipulated within the prescription or if the generic equivalent happens to be more expensive than the originator medicine. The legislation also permits compulsory licensing and parallel importation. In other words, a judicial authority such as the Minister of Health or Competition Board, may issue a license for the domestic manufacture of a medicine still under patent without the agreement of the patent holder (McIntyre et al., 2007).⁷⁷ This represents a significant move away from the protection of intellectual property; however, the government claims that this would only be undertaken under circumstances where the availability of the medicine

⁷⁶ Medicines can be classified into either *originator medicines* or *independent multi-source medicines* (generics). Originator medicines refer to medicines registered in South Africa that are currently protected by a patent or had been previously protected by a patent. Generics refer to medicines registered in South Africa that have never been protected by a patent. Generics are manufactured by companies other than the company that originally held the patent.

⁷⁷ An acceptable royalty payment would be paid to the patent holder. The question immediately raised would be acceptable to whom? It is unclear who will determine the “acceptability” of such a compensatory payment.

is critical to public health or as recourse to anti-competitive practices, such as excessive prices or the failure to supply sufficient quantities. Parallel importation refers to the practice whereby an entity, such as government, imports a medicine from sources other than the original manufacturer. Typically, where pharmaceuticals are sold at lower prices in other countries, it may be possible to take advantage of the price differential offered to purchasers in alternate jurisdictions (McIntyre et al., 2007).

The opposition to many of the provisions within the legislation and accompanying regulations has been vociferous, which has resulted in a plethora of legal challenges testing the validity of such provisions (McIntyre et al., 2007). Despite government's success in promulgating the relevant legislation, in particular, the Medicines and Related Substances Control Amendment Act No. 59 of 2002, it is yet to make use of the provisions relating to either compulsory licensing or parallel importation (McIntyre & Doherty, 2004).

The National Health Act No. 61 of 2003 also makes provision for a National Health Reference Price List (NHRPL), which recommends the fees that should be charged for different healthcare services provided by private healthcare professionals or private healthcare facilities (McIntyre et al., 2007). The NHRPL outlines a reference price list for healthcare services rendered, diagnostic and treatment procedures performed, and consumable items utilised by healthcare establishments, healthcare providers or allied healthcare workers within the private healthcare sector. Remarkably, the National Health Act No. 61 of 2003 explicitly states that medical schemes are permitted to use the NHRPL to determine their own benefit offerings and in the case of private sector healthcare providers including healthcare practitioners, the Act permits the use of the NHRPL to construct their fee schedules (McIntyre et al., 2007).

As the NHRPL serves merely as a guideline for prices, McIntyre et al. (2007) argues that within the South African private healthcare sector there is no regulatory control of prices, with the exception of medicines and associated dispensing services. However, in 2010, the Department of Health released a discussion document proposing the creation of a statutory price determination authority.⁷⁸ The purpose of the draft policy framework is set out as follows:

“...for achieving stability in the determination of prices in the South African health private health system...The policy framework seeks to achieve price stability through the establishment of an enabling mechanism for the rational and fair determination of final prices used by suppliers of health services for billing patients and medical schemes and the tariffs used by schemes as a basis for determining their levels of reimbursement” (Department of Health, 2010: 10).

The proposal seeks to bring back multilateral tariff negotiations under the auspices of the statutory price determination authority. Mandatory arbitration will be utilised to break deadlocks, and steer medical schemes and healthcare providers towards equitable and affordable outcomes. The Department of Health (2010) points out that medical schemes have traditionally operated according to a fee-for-service reimbursement system, whereby the healthcare expenditure incurred by a medical scheme member or beneficiary is simply paid to the relevant healthcare provider by the medical scheme.

Considering that medical scheme members typically select their healthcare service provider, the medical scheme ends up acting merely as an indirect and third-party payer with little influence over the tariff negotiation between the medical scheme member (or beneficiary) and healthcare provider (Department of Health, 2010). Furthermore, since the healthcare provider relies almost exclusively upon the reimbursements from medical schemes as the predominant source of their income, and as the medical scheme member is essentially not paying in full for the healthcare services received at the point-of-service, there is a tendency for over-utilisation. Moreover, the medical scheme member or beneficiary is typically at an informational

⁷⁸ Department of Health (2010) Discussion Document – The Determination of Health Prices in the Private Sector. Pretoria: National Department of Health.

disadvantage with respect to the diagnosis and treatment of any medical condition or healthcare incident. Therefore, the medical scheme member or beneficiary is inclined to accept any diagnosis or treatment prescribed by their healthcare provider. Hence, the Department of Health (2010: 10) argues that the medical schemes “...will face an infinite liability” if they do not negotiate appropriate reimbursement prices with healthcare service providers. It is important to note that these “reimbursement prices” do not necessarily need to equate to the actual prices charged by the healthcare providers. Thus, by implication the difference between the reimbursement and actual price charged by the healthcare provider would represent an out-of-pocket expense for the medical scheme member.

The Department of Health (2010) draft policy framework represents probably the most significant reform to South Africa’s private health insurance sector since the enactment of the Medical Schemes Act No. 131 of 1998. If the draft policy is implemented and if it works as intended, the Department of Health (2010) argues it will arrest the oligopoly power that effectively resides with private healthcare institutions and private healthcare professionals. Therefore, the proposed reforms relating to the creation of a statutory price determination authority would directly address the high medical inflation that has plagued the private healthcare sector for more than two decades.

2.4 Cost Containment

Contracted periods of high medical inflation have been associated with the ability of healthcare providers to ‘capture’ the regulatory environment partly as a result of the fee-for service reimbursement structure for healthcare services, which creates incentives for over-supply. Doherty and McLeod (2003) review the strategies adopted by medical schemes to contain costs. These strategies have been successful in controlling utilisation, but have had several adverse consequences with respect to equity (Doherty & McLeod, 2003). Unless the strategies are able to address the problem of the over-supply of healthcare services in the private sector, there is little

optimism that cost escalation can be arrested. The authors concur with Reekie (1999) that only once managed care is implemented properly can cost containment be achieved through the extensive use of healthcare provider networks that are reimbursed through risk-sharing arrangements. Another factor which has played a role in the observed cost escalation is the flight of skills to other countries and the scarcity of quality healthcare facilities. McIntyre et al. (2007) argue that the key weakness with the current regulatory environment governing medical scheme business in the country is that it does not address the:

“...vicious cycle between supply (quantity and distribution) and price issues within the private [healthcare] sector” (McIntyre et al., 2007: 48).

They describe the exodus of generalist and specialist practitioners, and other allied healthcare professionals from the public to the private healthcare sectors as deeply concerning. In the early 1980s, approximately 40 percent of generalist and specialist practitioners were employed in the private healthcare sectors, but by 1990 this had increased to well over 60 percent and to over 70 percent by the late 1990s (McIntyre et al., 2007). The authors note that this trend is mirrored in the number of private (for-profit) hospital beds over a similar period. In the early 1980s, there were about 8,000 private (for-profit) hospital beds, which increased to over 11,000 by the late 1980s (McIntyre et al., 2007). Significant rises were recorded thereafter to 16,415 by 1994 and 23,706 by 1999. Since then the increase has been modest and by 2007, the number of private (for-profit) hospital beds had levelled off to approximately 25,000 (Cornell et al., 2001; van den Heever, 2007; McIntyre et al., 2007).

Despite the significant rise in the supply of healthcare providers, McIntyre et al. (2007) assert that the number of individuals able to afford these private healthcare services, in particular medical scheme members, had not increased at the same pace. According to the authors, the end result was an excess supply of private healthcare providers relative to their designated clientele that they serve – medical scheme members and their dependents. Interestingly, they raise the question, how could the private healthcare professionals and the private healthcare facilities expect to generate sufficient income in the face of what was an apparent over-supply of their services to

the private healthcare sector. McIntyre et al. (2007) suggest that the primary response of private healthcare providers was to promote increased utilisation of healthcare services. As most healthcare expenditure that is incurred by consumers tends to be non-discretionary, funding healthcare expenditure via a risk pool is more appealing to insurers since healthcare services are demanded due to medical necessity rather than personal preference. However, medical scheme beneficiaries are at an informational disadvantage when compared to the healthcare professionals that advise them with respect to the appropriate diagnosis and treatment of medical conditions or incidents. Furthermore, since the fee-for-service reimbursement model does not require medical scheme beneficiaries to pay for the full cost of the healthcare services they receive at the point-of-service, the incentive to reduce utilisation is diminished considerably. The authors suggest that as a result over time the balance of power with respect to influencing prices has shifted away from medical schemes towards private healthcare providers. McIntyre et al. (2007: 50) goes further to argue that:

“...these factors have contributed to unconstrained cost spiral in the private [healthcare] sector, which has resulted in a rapidly widening public-private mix gap.”

2.4.1 Is Medical Scheme Coverage becoming more affordable?

Medical scheme contributions have increased significantly over the last decade. This is despite the implementation of the Medical Schemes Act No. 131 of 1998 and the government's stated objective of providing universal access to affordable medical scheme coverage. Specifically, the corresponding principles of open enrolment, community-rating and the Prescribed Minimum Benefits (PMBs) package were meant to directly address the cost escalation in medical scheme contributions and deterioration in medical scheme benefits. Many commentators wish to attribute the blame for the inordinate increases in medical scheme contributions to the actual medical schemes; however, as McIntyre et al. (2007) point out a considerable portion of that blame should rest with the private healthcare providers. Private healthcare providers have been allowed to wield undue power with respect to the quantity, distribution and pricing of private healthcare services (McIntyre et al., 2007; van den

Heever, 2007). Thus, by further extension, the government should also bear some of the blame for allowing private healthcare providers to “capture” the regulatory environment.

To determine whether or not medical scheme members are being charged monthly contributions that are aligned with the benefits that they are receiving per month, it is necessary to examine the medical scheme data from the Annual Reports published by the Council for Medical Schemes over the period 1993 to 2010.⁷⁹ The methodology to analyse the trends in medical scheme contributions and net claims incurred require a number of adjustments before any appropriate comparisons can be drawn.

The inconsistent reporting of accurate information regarding exempt or Bargaining Council schemes over between 1993 and 2009 may contaminate any inferences that can be made if one were to include these schemes in the dataset. Moreover, the fact that the Bargaining Council schemes are exempt from certain provisions of the legislation may lead to further inconsistencies. Once Polmed and Transnet medical schemes were reclassified as registered schemes rather than exempt schemes in 2000, the relative size of the Bargaining Council medical schemes segment within the overall medical scheme sector shrank considerably. In fact, the Council for Medical Schemes stopped reporting the details about Bargaining Council medical schemes from their 2005 Annual Report onwards. In 2004, members of the Bargaining Council schemes constituted only 4.1 percent of the total number of both registered and Bargaining Council medical scheme members. Therefore, for the purposes of consistency, it was necessary to exclude the so-called exempt or Bargaining Council schemes from the dataset and thus, only registered medical schemes are considered in this thesis. The registered medical scheme data provided in the Council for Medical Schemes Annual Reports are of an annual frequency.

⁷⁹ The nature of the reporting cycle requires one to examine the subsequent year’s publication to obtain the relevant data for medical schemes in that particular year. For example, to obtain the medical scheme data for the year ending 2009, it is necessary to examine the Council for Medical Schemes Annual Report published in 2010.

Second, the data series that were required for these initial computations included the following:

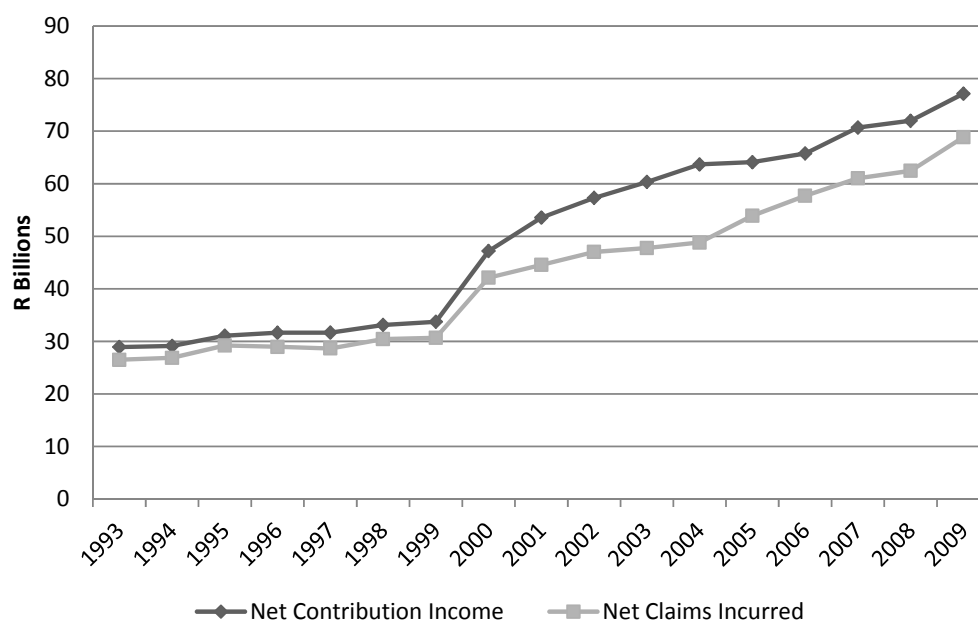
- ***South African Consumer Price Index (excluding interest rates on mortgage bonds - CPIX)***: can be used to derive a measure of the annual consumer inflation rate. The CPIX was utilised to convert the other monetary data series from nominal into real terms (relative to 2009 prices).
- ***Total number of medical scheme beneficiaries***: represent the total number of registered medical scheme members and their dependents.
- ***Net Annual Contribution Income (or Gross Annual Risk Contribution Income)***: refers to gross annual contribution income less the contributions made towards medical savings accounts during that corresponding year. Gross annual contribution income simply refers to the contributions (premiums) paid annually by medical scheme members to registered medical schemes in return for medical scheme cover. These contributions include the non-risk contributions (savings contributions) paid into medical savings accounts on behalf of the medical scheme member. Since the savings contributions represent contributions made by medical scheme members to pay out-of-pocket healthcare expenses, they essentially do not represent a benefit that the medical scheme beneficiary receives as part of the “insurance” portion of their medical scheme cover.
- ***Net annual claims incurred***: Net annual claims incurred are the annual claims paid out to healthcare providers for services rendered to the medical scheme member or their dependents. Net annual claims incurred do not include out-of-pocket expenses paid to healthcare providers from members’ medical savings accounts.
- ***Gross annual administration expenses***: refer to the annual administration expenses of operating the medical scheme, but they do not include broker fees, managed healthcare expenditure or the net income / (expense) on commercial reinsurance.

In real terms, the total net contribution income collected by registered medical schemes from medical scheme members has risen sharply from R28.9 billion in 1993 to over R77.1 billion by 2009. This is despite only modest increases in the total

number of medical scheme beneficiaries over the same period (from approximately 5.4 million individuals in 1993 to just over 8 million by 2009). As illustrated in Figure 2-9, total net claims incurred have followed a similar trend. In 1993, total net claims incurred stood at R26.5 billion versus R68.9 billion in 2009. What is perhaps more interesting is the fact the gap between total net contribution income and total net claims incurred has widened considerably since 2000. By implication, it suggests that medical scheme members are paying considerably more for the benefits that they receive under medical scheme coverage. The divergence between total net contribution income and total net claims incurred suggests that medical schemes are increasingly charging over-and-above the actuarially fair premium for medical scheme coverage. Therefore, the argument that the increases in medical scheme contributions can be explained by the higher fees charged by healthcare providers for services rendered is thrown into doubt by these results. It can be argued that the difference may be a direct result of the new regulatory reforms, whereby in particular, community-rating force medical schemes to apply widespread income cross-subsidisation (between the wealthy and poor) and risk-related cross-subsidisation (between the healthy and unhealthy), which has induced an additional redistributive ‘tax’.⁸⁰

⁸⁰ This is may be indicative of a similar situation in Dahlby (1983)’s analysis of the Canadian automobile insurance market where statistical discrimination was prohibited with respect to an individual’s gender. Dahlby (1983) finds that enforced risk-related cross-subsidisation between females and males, leads to society paying more for automobile insurance.

Figure 2-9: Total Net Contribution Income versus Total Net Claims Incurred (1993 – 2009) (in 2009 prices)

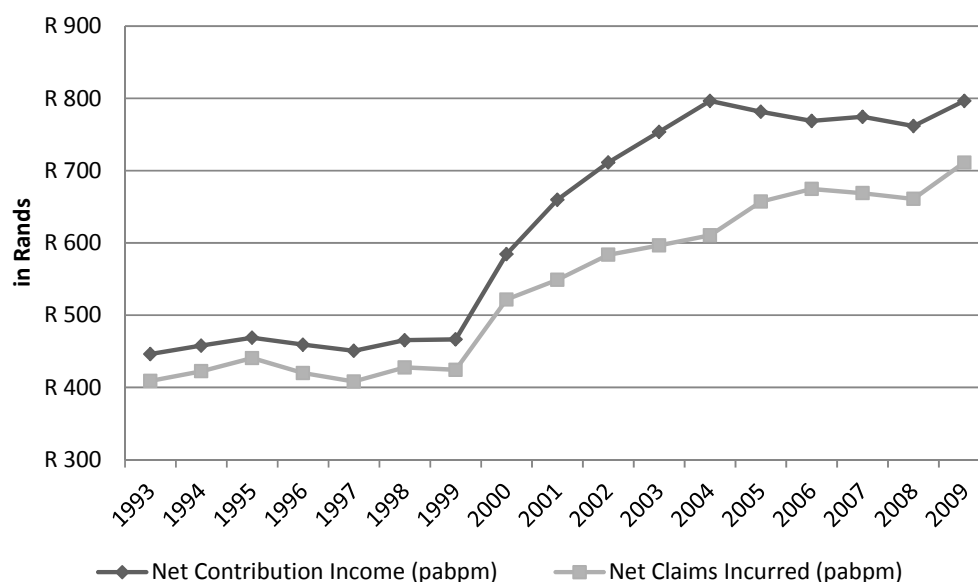


Source: Own calculations; South African Council for Medical Schemes Annual Reports (1993 – 2010); and values are adjusted by CPIX to reflect 2009 prices.

In 2009, R796.63 (in 2009 prices) was paid per month by the average registered medical scheme beneficiary (excluding the contributions paid towards medical savings accounts) for medical scheme cover. In real terms, this represents almost double the amount paid by a medical scheme beneficiary in 1993 (R446.23). Interestingly, 2009's figure is more than 30 percent higher than the amount paid by the average medical scheme beneficiary per month (R584.45) at the time when the Medical Schemes Act No. 131 of 1998 became effective in 2000. The average net claims incurred per beneficiary per month in 2009 was R711.19, which is higher than the corresponding values observed in 2000 (R521.63) and 1993 (R408.97).⁸¹ From Figure 2-10, it is evident that both the average net contribution and net claims incurred per beneficiary per month has increased consistently between 1993 and 2009.

⁸¹ These values have been adjusted by CPIX to reflect 2009 prices.

Figure 2-10: Net Contribution Income versus Net Claims Incurred (1993 – 2009) (pabpm) (in 2009 prices)



Source: Own calculations; South African Council for Medical Schemes Annual Reports (1993 – 2010); pabpm – per average beneficiary per month; and values are adjusted by CPIX to reflect 2009 prices.

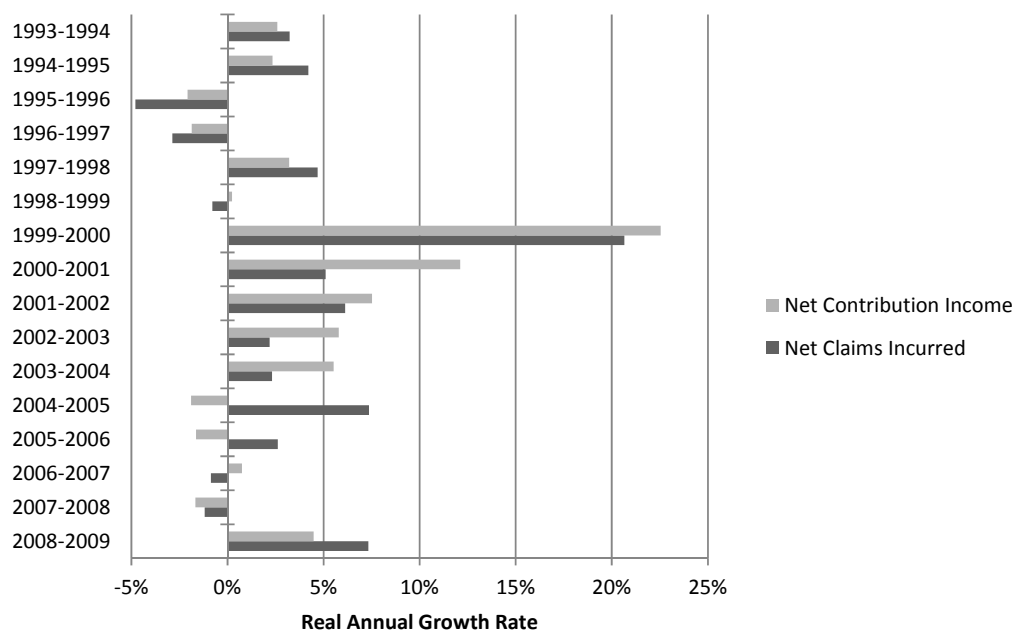
Figure 2-11 demonstrates the annual real increases in net contribution income and net claims incurred, per registered medical scheme beneficiary, between 1993 and 2009. The annual growth rate in both net contribution income and net claims incurred per average medical scheme beneficiary has not always outpaced consumer inflation since 1993. For the period just after the deregulation, net contribution income per average medical scheme beneficiary declined in real terms, it fell by 2.08 percent over the year 1995-1996 and further 1.86 percent during 1996-1997. Over the same two periods, net claims incurred per average medical scheme beneficiary, fell by 4.79 percent and 2.87 percent respectively. Therefore, Reekie (1999: 4)’s argument that the deregulation would address the healthcare “cost explosion” is supported by these results.

The reclassification of Polmed and Transmed from exempt to registered medical scheme status in 2000 can partially explain the significant rise in both net

contributions (22.54 percent) and net claims (20.66 percent) per average beneficiary over the 2000-2001 period. However, absolute increases in net contributions and net claims incurred as a result of Polmed and Transmed's inclusion would have been partly offset by the increase in the total number of medical scheme beneficiaries. Therefore, the implementation of the Medical Schemes Act No. 131 of 1998 from the year 2000, appears to have had a significant effect in increasing both net contribution income and net claims incurred per average beneficiary. Indeed, for the next four years after the implementation of the new regulatory framework, net contributions and net claims per average beneficiary recorded considerable increases above inflation.

In 2009, the annual increase in net contribution income per average medical scheme beneficiary was 4.48 percent above inflation. The annual increase in net claims incurred per average medical scheme beneficiary was 6.82 percent above inflation over the same period.

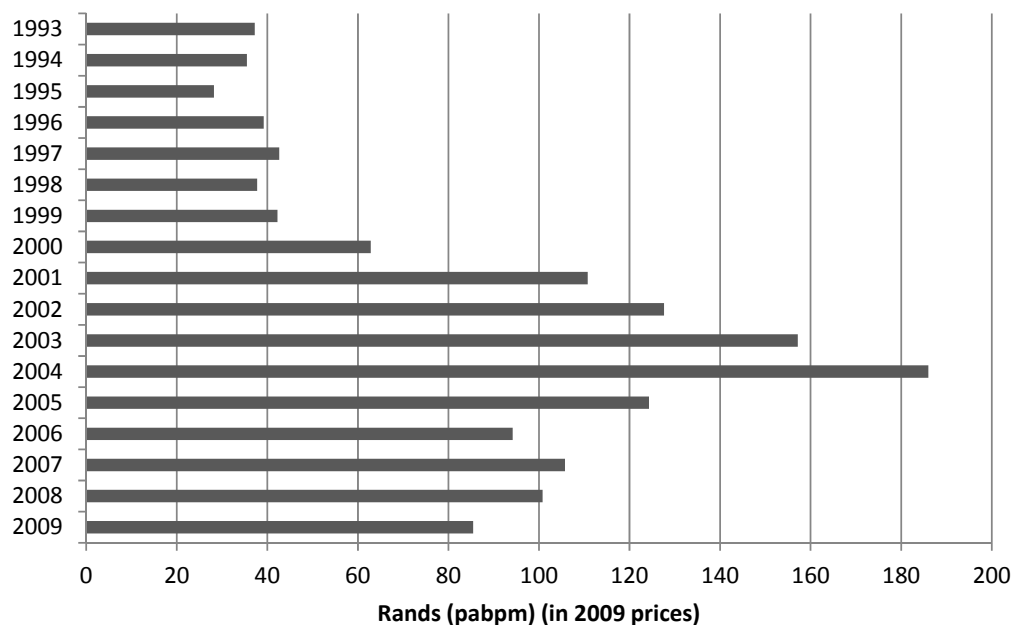
Figure 2-11: Annual Real Growth Rate in Net Contribution Income and Net Claims Incurred (pabpm) (1993 - 2009)



Source: Own calculations; South African Council for Medical Schemes Annual Reports (1993 – 2010); pabpm – per average beneficiary per month; and values are adjusted by CPIX to obtain real changes.

From 2000 onwards, the gap between the average net contribution and net claims incurred per beneficiary per month has widened significantly. Figure 2-12 displays the difference between the average risk contributions paid over to registered medical schemes per average beneficiary per month and the net claims incurred per average beneficiary per month (benefits received by the average registered medical scheme beneficiary per month). The difference has increased dramatically from R37.25 in 1993 to R85.44 in 2009 (in real terms).

Figure 2-12: Difference between the Average Real Monthly Net Contributions and Net Claims Incurred (pabpm) (1993 - 2009) (in 2009 prices)

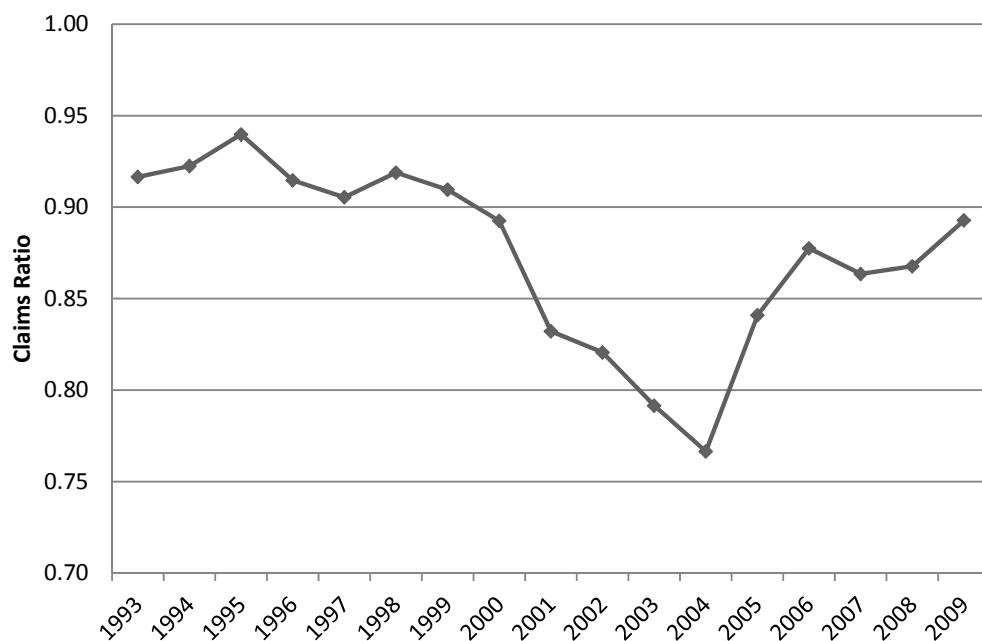


Source: Own calculations; South African Council for Medical Schemes Annual Reports (1993 – 2010); pabpm – per average beneficiary per month; and values are adjusted by CPIX to reflect 2009 prices.

The claims ratio, similar to the market loss ratio under traditional insurance, measures the ratio of total net claims incurred over total net contribution income. Therefore, the claims ratio can be interpreted as the Rand amount of protection medical scheme beneficiaries receive for every Rand paid in contributions. The claims ratio has started to once again rise from its low of 79.15 percent in 2004 to 89.28 percent by

2009. The general upward trend in net claims incurred versus net contribution suggests that it may be premature to conclude that there has been no discernable deterioration in the risk profile of registered medical scheme beneficiaries. However, the increase in net claims incurred versus net contributions may simply be further evidence that healthcare providers are continuing to drive-up healthcare costs unilaterally (Council for Medical Schemes, 2010).

Figure 2-13: Claims Ratio of South Africa's Private Health Insurance Sector (1993 - 2009)



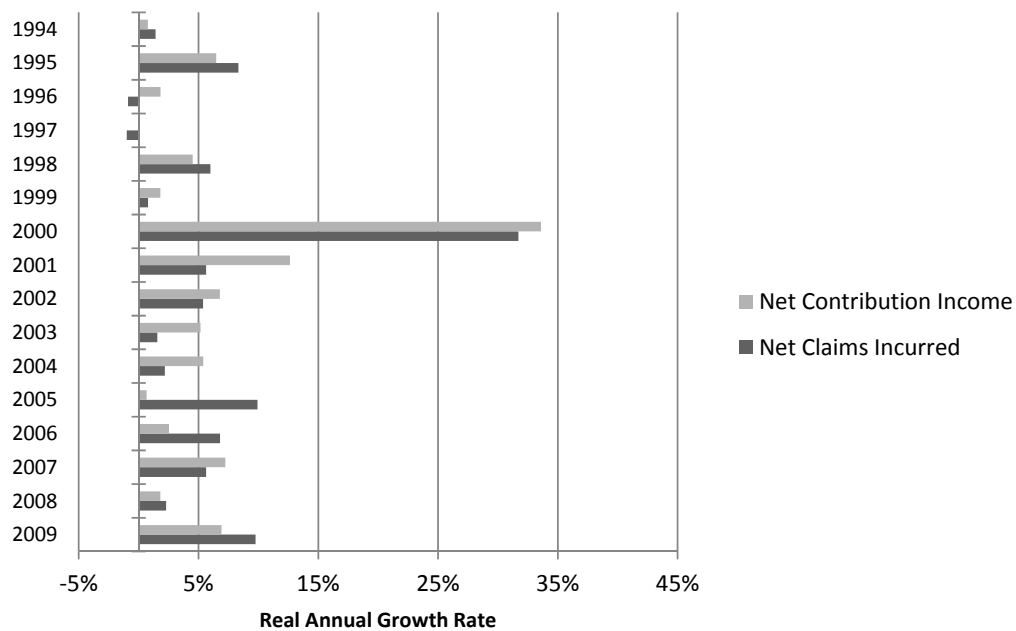
Source: Own calculations and South African Council for Medical Schemes Annual Reports (1993 – 2010).

From the above results it appears that the claims ratio is rising, in other words the ratio of benefits received versus contributions is increasing, and therefore over the long term medical scheme membership will become more costly.

2.4.2 Medical Scheme Cost Drivers

If one considers the annual real growth rate in net contribution income (and not per average medical scheme beneficiary), then it is apparent that each year medical schemes raise their contributions rates faster than household income and consumer inflation.⁸² Even though the annual real growth rate in net claims incurred also demonstrates higher-than-inflation increases, it is argued that medical schemes offer the same, if not fewer, benefits. The reason for this is that healthcare providers are simply charging much higher fees each year for the equivalent quantity of healthcare services rendered.

Figure 2-14: Annual Real Growth Rate in Net Contribution Income and Net Claims Incurred (1993 - 2009)



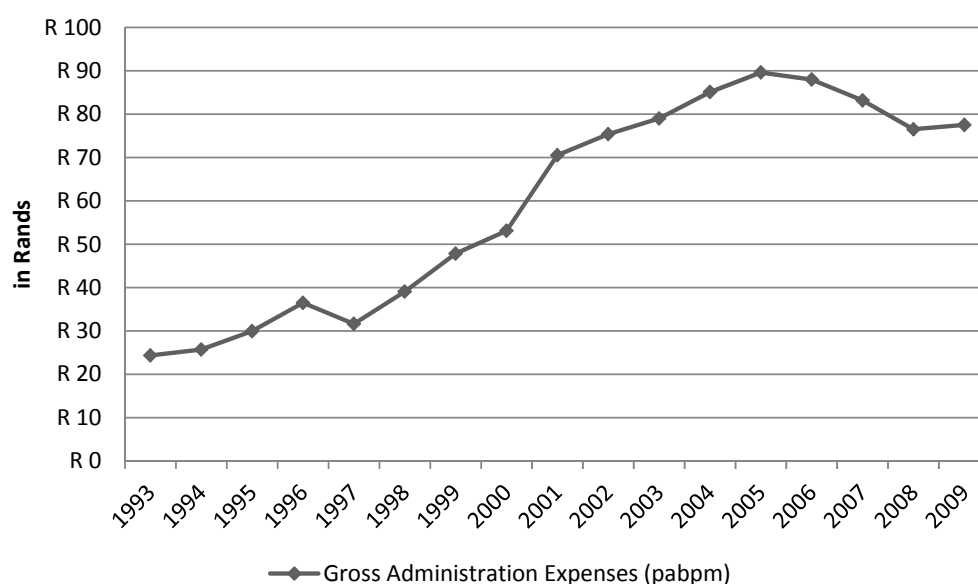
Source: Own calculations; South African Council for Medical Schemes Annual Reports (1993 – 2010); values are adjusted by CPIX to obtain real changes.

⁸² Refer to Figure 2-14

Many studies have examined the various causes of South Africa's high and rising private healthcare financing costs (Doherty & McLeod, 2003; McIntyre et al., 2007; van den Heever, 2007; Magennis & van Zyl, 2009). Van den Heever (2007) suggests that non-healthcare costs such as administration expenses, broker fees and reinsurance were considerable cost contributors during the 1990s, but according to the author these have for the most part levelled off. However, if one examines gross administration expenses per average medical scheme beneficiary per month between 1993 and 2009, there has been a significant escalation in the general administration costs of a medical scheme.⁸³ Although, it is fair to note that since 2005, gross administration expenses per average medical scheme beneficiary per month has fallen marginally. In 1993, R24.35 (in 2009 prices) was spent per medical scheme beneficiary per month on general administration (excluding managed care fees, broker fees, net reinsurance profit / loss and bad debts). But by 2009, this figure had more than tripled to R77.54 per beneficiary per month. Medical scheme administrators have argued that due to the complex nature of community-rating and the mandatory PMBs, there has been a corresponding increase in the costs of administering a medical scheme (Doherty & McLeod, 2003).

⁸³ Refer to Figure 2-15

Figure 2-15: Gross Administration Expenditure (pabpm) (1993-2009) (in 2009 prices)



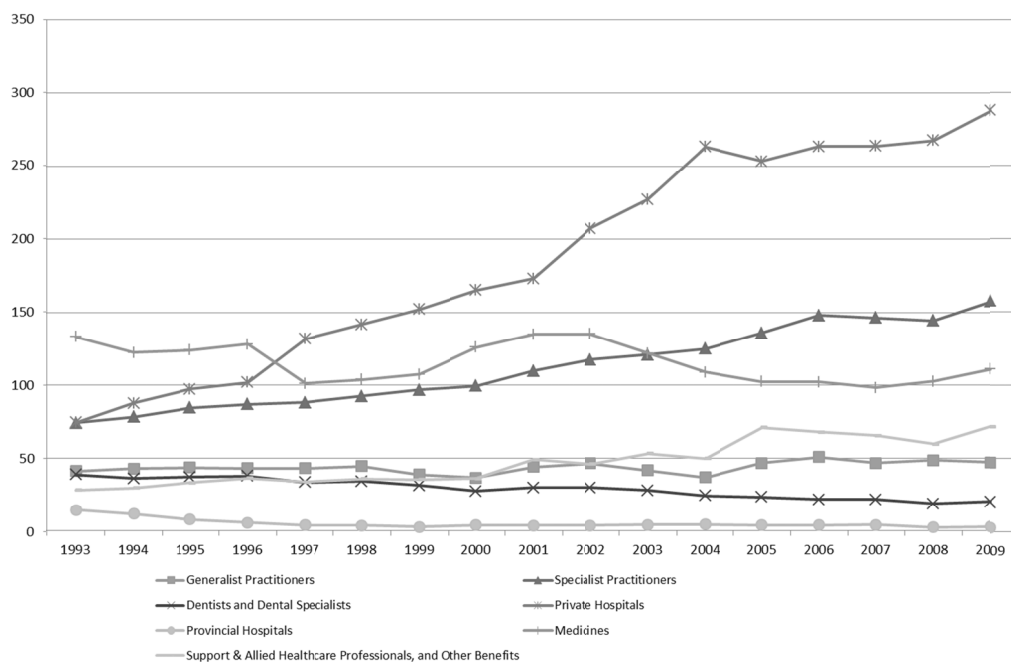
Source: Own calculations; South African Council for Medical Schemes Annual Reports (1993 – 2010); pabpm – per average beneficiary per month; and values are adjusted by CPIX to reflect 2009 prices.

Figure 2-16 depicts the distribution of healthcare risk benefits paid to the various categories of healthcare providers over the period 1993 to 2009 per average beneficiary per month.⁸⁴ Since as far back as 1997, private hospitals have progressively devoured a larger slice of the private healthcare expenditure “pie”. The Council for Medical Schemes has stated categorically that private hospitals and specialist practitioners are the biggest culprits with respect to the surges in medical scheme costs over the last decade (Council for Medical Schemes, 2010). In 2000, the average medical scheme beneficiary paid R165.32 (in 2009 prices) over to private hospitals per month and this compares to R287.88 by 2009. As a portion of the total risk benefits paid to healthcare providers, private hospitals received a 40.44 percent share of all risk benefits during 2009. This compares to 17.99 percent in 1993 and 32.38 percent in 2000. According to Magennis and van Zyl (2009), the increased billings to medical schemes by private hospitals are driven by increases in patient days (greater admission rates and longer duration stays at hospital) and the cost per

⁸⁴ Benefits paid out of the medical scheme’s risk pool are only considered. Thus, payments made to healthcare providers out of members’ medical savings accounts are excluded.

patient day (which is a function of efficiency, price of technology and resources utilised). Another key factor, which has led to private hospitals becoming the primary contributor to medical scheme's costs, is that private hospital market is dominated by three major hospital groups – Netcare, Medi-Clinic and Life Healthcare. At present only around 5 percent of all private hospital beds are located outside of these three hospital groups, which compares to more than 50 percent just 15 years ago (van den Heever, 2007; Magennis & van Zyl, 2009). Therefore, this oligopolistic situation has permitted these private hospital groups to become price setters and thus, as a consequence they have ratcheted up their prices (van den Heever, 2007).

Figure 2-16: Distribution of Healthcare Risk Benefits paid to Healthcare Providers (1993-2009) (pabpm) (in 2009 prices)



Source: Own calculations; South African Council for Medical Schemes Annual Reports (1993 – 2010); pabpm – per average beneficiary per month; and values are adjusted by CPIX to reflect 2009 prices.

The Hospital Association of South Africa (HASA) denies such claims and offers an alternative reason for the observed cost escalation (van den Heever, 2007; Magennis & van Zyl, 2009). HASA's view is that increased utilisation of private hospital care

by medical scheme beneficiaries is a direct result of the increase in the number of medical scheme beneficiaries, and the deteriorating risk profile of medical scheme pools. Young and healthy medical scheme members are allegedly exiting the market and thereby leaving behind an ageing medical scheme population, who have a rising disease burden that ultimately translates into greater utilisation of private hospital healthcare. Indeed, van den Heever (2007) nor the Council for Medical Schemes (2009) are convinced by HASA's explanation, and they point out that at present most new entrants to the medical scheme market tend to be exactly the young and healthy, who have cost profiles below the existing market. Hence, as a consequence private hospital margins should be heading lower; however, South Africa's private hospitals continue to maintain the highest margins of any private hospital group in the world.

Another significant contributor to higher medical scheme costs is specialist practitioners. In 1993, the average medical scheme beneficiary paid R74.18 per month over to specialist practitioners. But by 2009, this had more than doubled to R157.57 (in 2009 prices). Part of the reason for the increase in risk benefits paid to medical specialists can be attributed to the shortages in the supply of qualified specialist practitioners. Therefore, medical specialists enjoy a certain degree of market power and in the absence of regulation; they can charge patients up to 300 percent above medical scheme tariffs without any penalty.⁸⁵ An adjunct to this may involve the relationships that exist between specialists and private hospitals. Private hospitals tend to compete amongst themselves for specialist practitioners rather than for medical scheme business. The belief is that patients tend to follow specialist practitioners. Hence, private hospitals in their attempts to secure the services of medical specialists, they are often prepared to fall into line with the specialist's wishes (and / or demands). An example of this is the purchase and utilisation of new medical technology. New advances in medical technology often attract widespread attention as they offer significant benefits for patients in the diagnosis and treatment of many health-related conditions (Aaron, 1991; Newhouse, 1992; Fuchs, 1996). Private hospitals use new technological innovations, such as advanced medical equipment, to attract specialist practitioners to their facilities. The high costs of new medical technologies does not

⁸⁵ If a specialist practitioner charges a patient above the reimbursement price of the medical scheme, it simply forces the medical scheme member to pay the difference from out of their own pockets.

prove to be an obstacle for private hospitals, since the fee-for-service reimbursement system pays healthcare providers graciously for utilising new innovations (Weisbrod, 1991). Private hospitals can simply recover the cost of these purchases from medical schemes by raising their facility fees.

The high costs of medicines, which tended to keep pace with hospital expenditure during the 1980s and early 1990s, began to moderate from the mid-1990s. In particular, from 2002, with the establishment of a Pricing Committee to determine a Single Exit Price (SEP) for all medicines, van den Heever (2007) argues the regulatory reforms helped to negate any significant cost increases with respect to medicines. In 1993, the average medical scheme beneficiary paid R133.22 per month for medicines. By 2002, it stood at R135.24 (in 2009 prices). However, as a direct result of the introduction of the SEP, medicines started to have less of an impact in real (inflation-adjusted) terms. In 2009, the average medical scheme beneficiary paid R111.10 per month for medicines. Although, the Minister of Health has announced that for 2011, there will be a zero percent increase in the SEP, the new dispensing fee arrangements for pharmacists may add as much as 4 percent to medical scheme contributions for the year ahead.

To explain the observable cost escalation in healthcare costs, healthcare providers often "...cite a number of unavoidable factors" and these include the following: the rising cost of medical technology and innovations, the ageing profile of the population requires ever more frequent and expensive healthcare interventions and finally, the emergence of new diseases or diseases that become increasingly difficult to treat, such as Acquired Immune Deficiency Syndrome (AIDS) which is caused by the Human Immunodeficiency Virus (HIV), and multi-drug resistant tuberculosis (Doherty and McLeod (2003: 50). Although, all these factors certainly play a role in increasing healthcare costs; Reekie (1999), Doherty and McLeod (2003), and Erasmus and Fitchen (2010), all concur that the primary reason for the healthcare cost explosion is the persistence of the fee-for-service reimbursement system. The fee-for-service

reimbursement system creates incentives for healthcare providers to over-supply healthcare services.

The Council for Medical Schemes (2010) is particularly scathing in its criticism of the role private hospitals and specialist practitioners play in driving healthcare costs upwards. The argument centres around the absence of normal competitive market pressures in the private healthcare sector, whereby healthcare costs rise due to both the prevalence of supernormal profits and endemic inefficiencies (van den Heever, 2007; Council for Medical Schemes, 2010). Inefficiencies arise due to the over-supply of healthcare services including medicines and consumables, and the overcapitalisation within private healthcare institutions.

2.4.3 Strategies to Contain Healthcare Costs

Medical schemes employ a variety of mechanisms in their attempt to control healthcare costs. Doherty and McLeod (2003) discuss the techniques available to medical schemes and some of these include:

- Rationing cover, imposing levies or co-payments and utilising medical savings accounts;
- Attracting the young and healthy;
- Formularies for Acute and Chronic Medication;
- Hospital Pre-Authorisation and Case Management;
- Disease Management Programmes; and
- Negotiated Tariffs and Preferred Provider Networks.

Rationing Cover, Levies or Co-Payments and Medical Savings Accounts (MSAs)

The central tenet behind these mechanisms is that by shifting some of the risk back onto the medical scheme member it is likely to induce a reduction in moral hazard and discourage over-utilisation of healthcare services. Rationing cover requires medical schemes to set annual limits on the amount that an individual member and their

dependents may claim for certain categories of healthcare within a given timeframe, typically one year (Doherty & McLeod, 2003). By imposing these limits upon benefits, the expectation is that it will deter medical scheme beneficiaries from utilising healthcare services unnecessarily. The authors contend that it may allow medical schemes to avoid paying for legitimate healthcare expenses.

Levies (excess or deductibles) and co-payments relate to individual claims as opposed to annual healthcare expenditure. A levy excludes a fixed portion of any claim from coverage. In other words, a medical scheme member is required, in the event of a claim against the medical scheme, to pay first out of his or her own pocket up until the levy amount is reached and only then does the medical scheme step in to cover any additional amount over and above the levy amount per claim made. Co-payments (or co-insurance) require the medical scheme member to pay a pre-defined percentage of any claim incurred by the member with respect to healthcare expenditure.

A member of a medical scheme plan with a medical savings account (MSA) pays a monthly contribution to the medical scheme that consists of both a traditional health insurance (risk) premium and a fixed amount that is set aside 'within' the member's medical savings account. Funds that accumulate within the MSA are designed to cover day-to-day healthcare expenses, which includes general practitioner consultations, acute treatment and out-patient hospital care. MSAs were devised to encourage medical scheme members to control their own utilisation of healthcare services (Doherty & McLeod, 2003).

Attracting the Young and Healthy

Despite the Medical Schemes Act No. 131 specifically prohibiting medical schemes from discriminating persons with respect to their age, prior and current health status, it does not preclude medical schemes from marketing (and tailoring where possible)

their products (benefit options) for a specific target audience.⁸⁶ Thus, medical schemes can surreptitiously attract the young and healthy rather than accept the overall age and risk profile of the market. In a competitive market, it is beneficial to keep annual contributions as low as possible relative to one's competitors so that one can sustain and possibly augment your market share (Doherty & McLeod, 2003). In addition, low contributions ensure that you have a further competitive edge in attracting low risk individuals such as the young and healthy.

Formularies for Acute and Chronic Medicine

Doherty and McLeod (2003: 58) explain that:

“...formularies are lists of drugs that a managed care organisation determines to be clinically appropriate and cost effective. Reimbursement by schemes is then restricted to items on the formulary, although members can obtain other products if they are prepared to pay the difference”.

Formularies often include generics as these are typically, more cost-effective whilst at the same being clinically as effective as medicines under patent. Most medical schemes adopted formularies for chronic medication back in the 1990s and more recently, medical schemes require medical scheme beneficiaries to register on a chronic medicine management programme (Doherty & McLeod, 2003).

Hospital Pre-Authorisation and Case Management

Hospital pre-authorisation is straightforward whereby medical scheme members or their dependents are required to obtain authorisation before admission to a hospital. If pre-authorisation is not obtained, claims arising from the hospital stay will not be reimbursed by the medical scheme (Doherty & McLeod, 2003). The intention is to prevent unnecessary treatment in what is an expensive treatment setting. Case management refers to the active monitoring of patients once in a private healthcare facility. The purpose is to ensure that the treatment delivered to the patient is taking

⁸⁶ Doherty and McLeod (2003) mention the so-called “Swiss-Cheese” medical scheme benefit design, which intentionally leaves gaps in benefit cover so as to discourage higher risk individuals.

place in the correct care setting and at the same time is clinically appropriate. For example, once a patient's health condition permits, moving the patient to a ward with a lower intensity of care can save inordinate sums of money. Increasingly, medical schemes are requiring members to move from private hospitals to so-called "step-down" facilities once their condition permits. Step-down facilities do not provide the same intensity of care as private hospitals and thus, the treatment (and rehabilitation) of patients is at a much lower cost to the scheme than private hospitals.

Disease Management Programmes

Disease management programmes involve the active management by medical scheme administrators with respect to the prevention, diagnosis and treatment of pre-defined diseases such as, diabetes, pregnancy or asthma (Doherty & McLeod, 2003). The authors suggest that these programmes typically allow the more accurate collection of information about the member's condition, but also the dissemination of best standards and practice to healthcare providers in relation to the treatment of the disease (Doherty & McLeod, 2003).

Negotiated Tariffs and Preferred Provider Networks

Medical scheme tariffs are negotiated between healthcare providers and medical schemes to determine the reimbursement prices that medical schemes are prepared to pay with respect to the rendering of healthcare services to their beneficiaries. The Board of Healthcare Funders (BHF), who represent most medical schemes, negotiates on behalf of medical schemes. Doherty and McLeod (2003) suggest that these negotiations should place downward pressure on the price of healthcare services. However, because of the market power that healthcare providers wield and their ability to manipulate the supply of healthcare services, they essentially end up dictating the prices that they are willing to accept for their services. The authors point out that medical schemes have been reluctant to exercise collective bargaining power and thus, they have failed to contain the healthcare cost escalation. One option available to medical schemes is to opt to negotiate directly with individual healthcare

providers (or networks of providers) in order to achieve cost savings; however, unless they represent a sufficient number of members it is difficult to envisage how they can influence pricing significantly.

Government has realised that the “free-for-all” that allows private healthcare providers to dictate pricing and thereby drive up healthcare costs unilaterally cannot continue unchallenged any longer (Department of Health, 2010). Therefore, this is providing impetus for the creation of the statutory price determination authority.. sAt present, tariff negotiations between healthcare providers and medical schemes often end in angry standoffs; this typically arises due to the inevitable dichotomy between what medical schemes believe they can afford to pay for healthcare services versus what healthcare providers believe that they are entitled to. Therefore, the proposal is to introduce a mandatory dispute resolution mechanism by means of arbitration, whereby the arbitrator may only choose between the initial offerings of the two parties (Department of Health, 2010). In other words, the arbitrator cannot construct a compromise bid between the two positions. Furthermore, no additional evidence may be introduced. Nor are any appeals permitted under the proposal. Thus, the arbitration procedure ensures that no party presents unrealistic bids since they would risk losing out to the other party’s bid via arbitration. The arbitration mechanism will require healthcare providers to justify their prices for healthcare services and medical schemes will need to justify why they cannot afford to pay more (van den Heever, 2007; Council for Medical Schemes, 2010).

2.4.4 Principle of Managed Care

Reekie (1999) argues that the application of sound insurance and risk management principles only became a reality once managed care became a legal possibility within the private health insurance. Under managed care, the delivery of healthcare services (from primary healthcare to complex surgical procedures) to a defined group of individuals (enrolees or members) is provided by a network of organisations such as hospitals, clinics, physicians and other healthcare providers. In return for a fixed (or

capitation) fee, enrollees of a managed care organisation receive healthcare services where required from a predefined network of healthcare providers. Traditionally, healthcare providers are either employed by the managed care organisation on a salaried basis or they receive a fixed fee for supplying healthcare services to members. A managed care organisation typically relies upon a general practitioner to act as the gatekeeper to ensure that the appropriate level of care is provided to the enrollee and to prevent over-utilisation. The general practitioner will assess whether an enrollee requires further specialised diagnosis and/or treatment, such as recommending a referral to specialists, surgery or physiotherapy. Thus, it is thought that the managed care organisation would be expected to control healthcare expenditure by removing the financial incentives for healthcare providers to overprescribe, over-treat and over-hospitalise their patients. In essence, managed care creates incentives to keep enrollees healthy by emphasising preventative care and health promotion activities. In the event that enrollees do become ill, managed care attempts to treat these individual in the most cost-effective manner whilst simultaneously ensuring that the quality of care provided is adequate.

While it is instructive to provide a general description of managed care organisations, it is important to recognise that the concept of managed care is undergoing continual evolution. As Reekie (1999) points out South Africa took direction from developments in the United States to create a legislative framework that permitted managed care to become a reality. This was achieved through the enactment of the Medical Schemes Amendment Act No. 23 of 1993. Shortell, Gillies and Anderson (1994) outline the emergence of an organised model for health care delivery in the United States:

“...the organizational landscape of health care delivery is being rearranged. The nation’s hospitals are merging, consolidating, and aligning at a dizzying pace; more than 40 percent of physicians are practicing in groups, and existing physician groups are beginning to consolidate; physicians and hospitals are forging new strategic alliances; and formerly hospital-based systems are reinventing themselves as broad-based, vertically integrated delivery systems” (Shortell et al., 1994: 47).

Shortell et al. (1994) go further to provide an early definition of an organised delivery network that can easily be interpreted as the definition for a managed care organisation:

“...a network of organizations that provides or arranges to provide a coordinated continuum of services to a defined population and is willing to be held clinically and fiscally accountable for the outcomes and the health status of the population served. Most such systems own or are closely aligned with an insurance product.” (Shortell et al., 1994: 47).

Two key features characterise the contemporary managed care organisation. The first being that managed care organisations tend to demonstrate extensive reliance upon healthcare (clinical) information systems. At the outset, healthcare information systems were created as part of the administration or clerical function of the health insurer – the objective being to assist in the billing and record-keeping processes. In the United States for example, Shortell et al. (1994) discuss how during the early 1990s, many large health facilities budgeted millions of US dollars to integrate information systems. A significant challenge facing managed care organisations is the design of information systems that provide direct clinical support to improve the quality of healthcare delivered to enrolees. The second feature of managed care organisations is their disinclination towards acute hospital care. As most hospital procedures incur significant costs, acute hospital care represents an expensive method of healthcare delivery. As noted earlier, general practitioners often serve as the gatekeepers within the managed care environment. Their role is to direct patients towards the most cost-effective treatment setting.

Types of Managed Care Organisations

Broadly speaking managed care organisations can be classified into three main types:

- *Health Maintenance Organisations (HMOs)* - provides enrolees with relatively comprehensive healthcare services, rather than reimbursements for healthcare

expenditure.⁸⁷ Therefore, for a fixed (or capitation) fee, enrollees face few out-of-pocket expenses and are covered for most inpatient and outpatient treatment, which includes medical consultation (diagnosis), surgery and prescribed medication. Upon joining the HMO, each enrollee is assigned a primary general practitioner (“gatekeeper”). HMOs generally require that all healthcare services be delivered through the plan’s network of healthcare providers and the primary general practitioner should authorise any healthcare services rendered.⁸⁸ HMOs that directly employ healthcare providers within their network – including the primary general practitioner – are referred to as staff HMO plans. In general, healthcare providers would be paid salaries by the HMO. Alternative remuneration schemes do exist, for example, greater compensation to reflect a larger patient load. HMOs that establish their healthcare network by contracting with healthcare providers are called an Independent Practice Association (IPA). They provide healthcare services to IPA enrollees at a negotiated capitation or retainer fee, or even on a fee-for-service basis. Typically, an IPA may contract out healthcare services to a wide-spectrum of healthcare professionals such as general practitioners, specialists; physical, speech and occupational therapists. In some cases, an IPA may only offer primary healthcare services or a single specialisation.

- *Preferred Provider Organisations (PPOs)* – also referred to as *Participating Provider Organisations*. PPOs offer enrollees two distinct tiers of coverage. Enrollees, who opt to receive healthcare services from the PPO’s network of preferred healthcare providers, would benefit from lower deductibles or coinsurance payments than if they chose to receive healthcare from non-network providers.⁸⁹ Unlike HMOs, PPOs do not appoint a primary general practitioner to serve as the “gatekeeper”. By agreeing to charge reduced (or discounted) rates to PPO enrollees, healthcare facilities and providers become part of the PPO’s

⁸⁷ Much of the scholarly discussion and healthcare policy research surrounding managed care organisations focus upon HMOs due to their pioneering role in combining health insurance and healthcare provision.

⁸⁸ The HMO enrollee is usually liable for any unauthorised healthcare services.

⁸⁹ PPOs simply create financial incentives for enrollees to opt for the PPOs’ own network of healthcare providers.

preferred network.⁹⁰ Healthcare providers that join the PPO's network typically submit themselves for utilisation reviews as part of the contractual terms. In particular, an utilisation review is required for any hospital stay that an enrollee may need. Many PPOs require that enrollees obtain a pre-admission certificate for any hospital stay and a second opinion in the event that a surgical procedure is necessary.

- *Point-of-Service (POS) plans* – are effectively a hybrid form of an HMO and PPO. Similar to an HMO, a POS plan would assign a primary general practitioner who would authorize within-network healthcare that would be covered by the POS plan under more generous coverage terms. The second dimension of a POS plan involves the PPO element, whereby the plan offers two tiers of insurance cover. An enrollee that receives healthcare services from a network provider benefits from coverage levels that are far more generous when compared to receiving treatment from outside the network.⁹¹

Conventional health insurance coverage tends to lead to an over-consumption of healthcare by an individual insured (*moral hazard*) and the insured is generally only concerned with the out-of-pocket cost of healthcare rather than the full cost at the point of service. Under fee-for-service health insurance plans, the healthcare provider advises the medical scheme member on the frequency, type and magnitude of healthcare services required. Since the consumer is at an informational disadvantage with respect to medical knowledge when compared to the healthcare professional and together with the structure of the fee-for-service remuneration scheme, the incentives created for the over-consumption of healthcare are significant. Managed care organisations appear to address the information asymmetries that arise within fee-for-service arrangements and in particular, the over-consumption of healthcare exemplified by traditional health insurance cover.

⁹⁰ Healthcare providers contracted by the PPO do not necessarily guarantee that they will treat every enrollee, but because of the favourable cost-sharing arrangements that enrollees enjoy, healthcare providers are more likely to take advantage of the greater volume of potential patients. Healthcare providers that are part of the PPO network are also likely to benefit from prompt payment for services rendered.

⁹¹ Under a POS plan, out-of-pocket costs are lower (greater coverage) when enrollees utilise network providers for healthcare services and higher (less coverage) when they do not. Most POS plans do not require pre-authorisation for enrollees to obtain healthcare services from providers outside of the network.

Merging healthcare provision and health insurance allows managed care organisations to utilise financial incentives that encourage healthcare providers to limit utilisation, restrict the services that they provide through command-and-control methods, and to bargain with healthcare provider networks to obtain lower prices (Cutler, McClellan & Newhouse, 2000). Managed care encourages healthcare providers to keep enrollees healthy by emphasising preventive and health promotion practices. When enrollees do become ill, the managed care system will strive to treat them in the most cost-effective way possible. In effect, managed care facilitates the process of risk-sharing between medical schemes and healthcare providers.

Managed Care in Practice

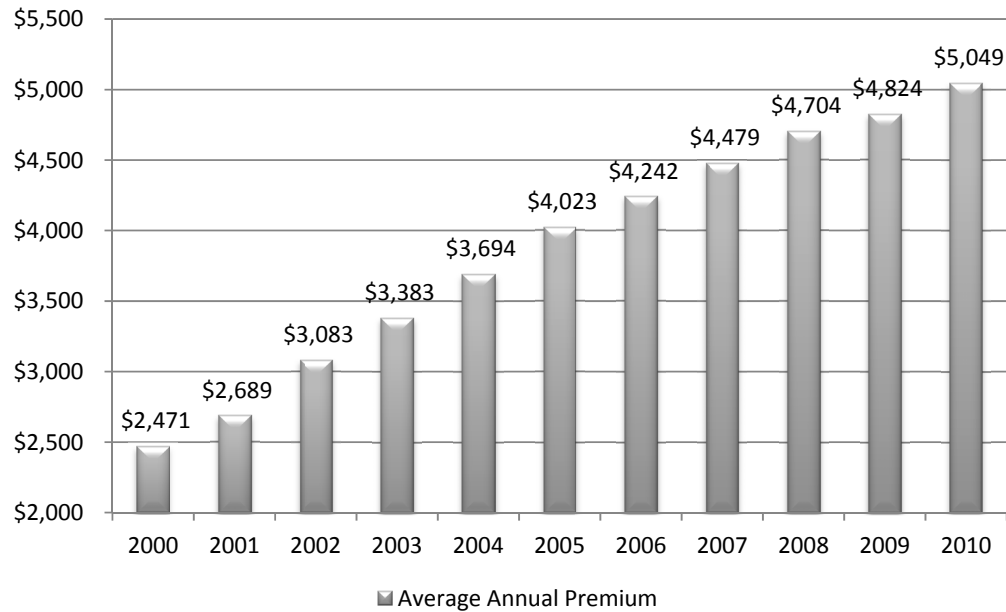
To gauge the success of managed care versus traditional fee-for-service health insurance plans, it is useful to examine a market where managed care is a well-established phenomenon such as employer-sponsored health insurance in the United States. In the United States, employer-sponsored health insurance covers approximately 157 million non-elderly individuals, which represents more than half of the total population (Kaiser Commission on Medicaid and the Uninsured, 2010). The Kaiser Family Foundation⁹² together with the Health Research and Educational Trust (HRET) conduct an annual national survey of employer health insurance benefits in the United States. In the Kaiser/HRET Survey of Employer-Sponsored Health Benefits 2010, 2,046 public and private employers were interviewed over the period January to May 2010 (Claxton, DiJulio, Whitmore, Pickreign, McHugh, Osei-Anto & Finder, 2010). In most cases, employers offer only one type of health insurance plan to employees. However, some employers do offer a variety of health insurance plans from which employees select their preferred plan. The Kaiser Family Foundation/HRET (2010) survey questioned firms about how many plans of each given type they offer to employees. Interestingly, employers may differentiate their offerings according to the employee grade or location. For example, executive or

⁹² The Kaiser Family Foundation is a non-profit private organisation, based in California, United States. The Foundation produces research and analysis on health-related issues.

senior-level management may be offered a very different selection of healthcare plan types compared to intermediate-level management.

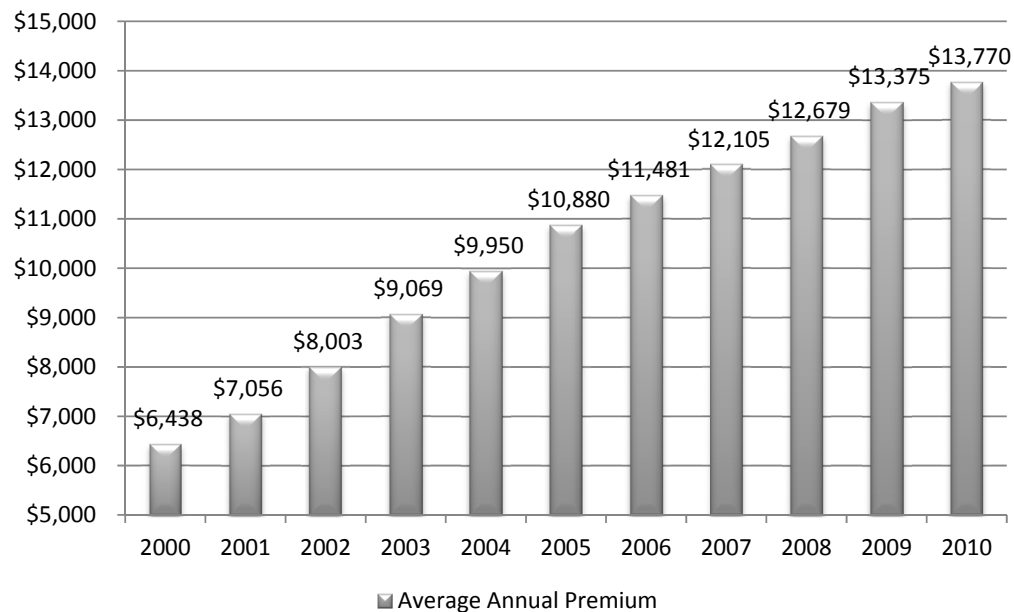
The results indicate that in 2010, the average annual premium for single health insurance coverage was \$5,049 (compared to \$4,842 in 2009) and \$13,770 (compared to \$13,375 in 2009) for family coverage. For single health insurance coverage this represents an increase of 4.28 percent and for family coverage, an increase of 2.95 percent. Over the same period, general inflation increased by 2.2 percent and wages rose by 2.3 percent (Claxton et al., 2010). Thus, for employer-sponsored health insurance the annual cost of coverage in the United States has only marginally outpaced general and wage inflation. If managed care is gaining market share versus conventional fee-for-service health insurance plans, it may be indicative that managed care indeed has positive implications with respect to containing healthcare costs as one might expect. Comparing the annual cost of coverage over a ten-year period (2000-2010), Figure 2-17 illustrates that for single coverage, annual premiums have been steadily rising from \$2,471 in 2000 to \$5,049 – representing a rise of 104.33 percent. In the case of family coverage, the corresponding rise is 113.89 percent. From Figure 2-18, the average annual premium has increased from \$6,438 in 2000 to \$13,770 in 2010 (Kaiser Family Foundation/HRET, 2010).

Figure 2-17: United States Employer-Sponsored Health Insurance - Average Annual Health Insurance Premium - Single Coverage (2000-2010)



Source: Kaiser Family Foundation/Health Research and Education Trust, Employer Health Benefits 2000-2010 Annual Surveys.

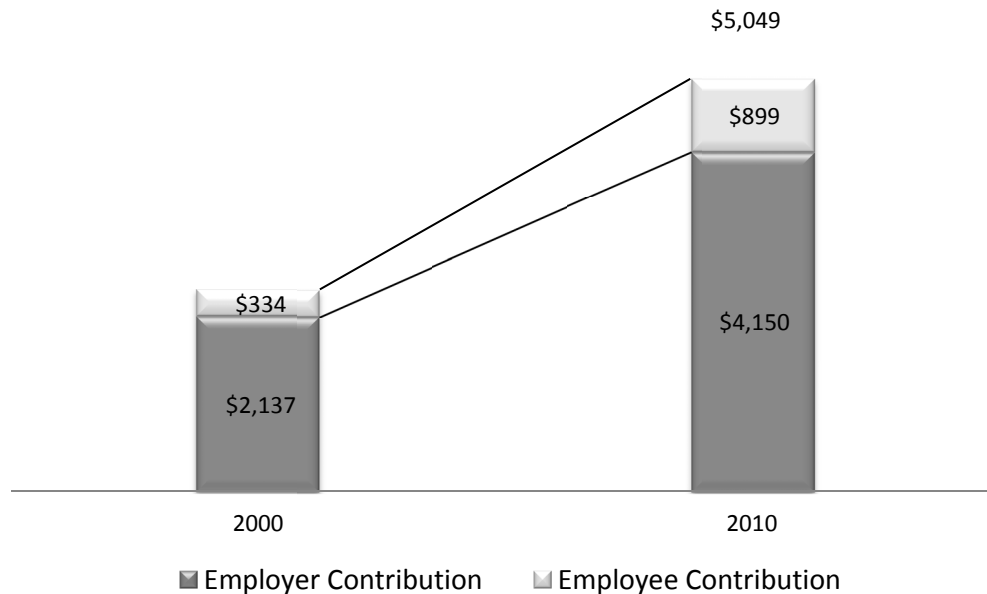
Figure 2-18: United States Employer-Sponsored Health Insurance - Average Annual Health Insurance Premium - Family Coverage (2000-2010)



Source: Kaiser Family Foundation/Health Research and Education Trust, Employer Health Benefits 2000-2010 Annual Surveys.

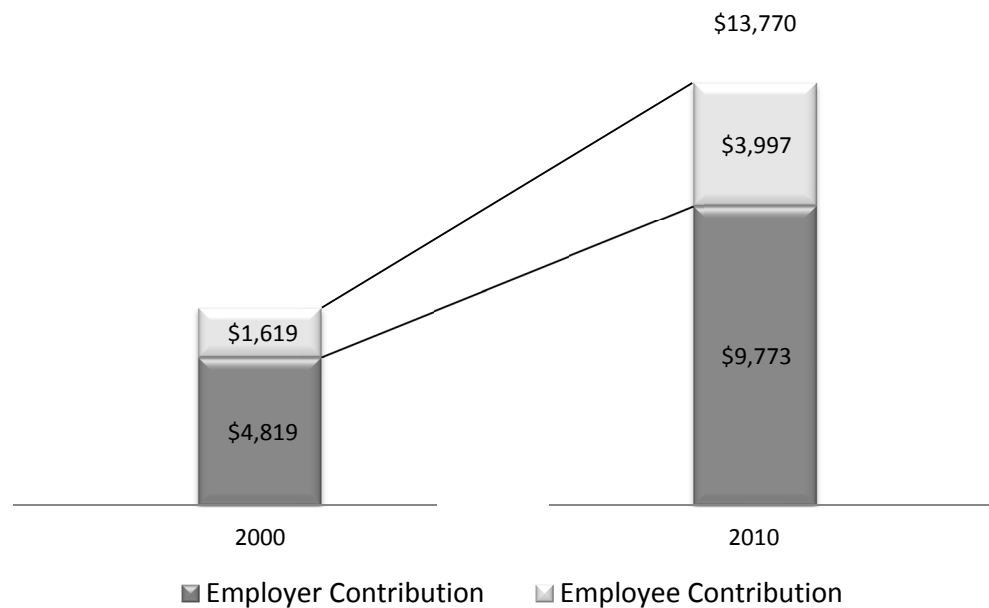
Under employer-sponsored health insurance, the employer typically makes a substantial contribution towards the cost of health insurance coverage for their employees. As employer-sponsored health insurance in the United States does involve contributions from both the employer and employee to cover the costs of insurance coverage, it is useful to examine the extent to which employees contribute towards their own health insurance coverage. Figure 2-19 demonstrates that the employee contribution towards single coverage has increased on average from \$334 per annum in 2000 to \$899 per annum in 2010 (Kaiser Family Foundation/HRET, 2010). Therefore, on a monthly basis an employee pays on average \$74.92 for single health insurance cover in 2010. In Figure 2-20, employees contributed on average \$3,997 per annum for family coverage during 2010 – up from \$1,619 per annum ten years earlier (Kaiser Family Foundation/HRET, 2010). Thus, employees with family coverage contribute on average \$333.08 per month towards their health insurance cover.

Figure 2-19: United States Employer-Sponsored Health Insurance - Average Annual Health Insurance Premium - Employer-Employee Contributions for Single Coverage, 2000 versus 2010



Source: Kaiser Family Foundation/Health Research and Education Trust, Employer Health Benefits 2000, 2010 Annual Survey.

Figure 2-20: United States Employer-Sponsored Health Insurance - Average Annual Health Insurance Premium - Employer-Employee Contributions for Family Coverage, 2000 versus 2010



Source: Kaiser Family Foundation/Health Research and Education Trust, Employer Health Benefits 2000, 2010 Annual Survey.

Table 2-6 presents the employees' share of the cost of coverage for both single and family coverage. Employees are paying considerably more towards their health insurance cover; the increase in the employee-share is most notable over the period 2009 to 2010. For most of the decade, the employee-share hovered around 15 percent and 25 percent for single and family coverage respectively. During 2010, the share rose sharply to 18 percent and 29 percent for single and family coverage respectively, which is a notable change from the steady share that employees were accustomed to paying (Claxton et al., 2010). The authors report that more than 30 percent of firms in the United States either curtailed the scope of healthcare insurance benefits offered to employees or simply increased the cost sharing in response to the recent economic downturn.

Table 2-6: United States Employer-Sponsored Health Insurance - Average Employee Contribution towards the Cost of Coverage

Year	Employee-Share of the Total Average Premium (Single Coverage)	Employee-Share of the Total Average Premium (Family Coverage)
2000	13.52%	25.15%
2001	13.20%	25.33%
2002	15.12%	26.70%
2003	15.02%	26.60%
2004	15.11%	26.74%
2005	15.16%	24.94%
2006	14.78%	25.89%
2007	15.49%	27.10%
2008	15.33%	26.45%
2009	16.15%	26.28%
2010	17.81%	29.03%

Source: Kaiser Family Foundation/Health Research and Education Trust, Employer Health Benefits 2000-2010 Annual Surveys.

Within the United States employer-sponsored health insurance sector there has been a considerable shift towards managed care plans as opposed to conventional fee-for-service plans.

“In response to high healthcare inflation, HMO enrolment grew rapidly [during] the late 1980s and early-to-mid 1990s in the [employer-sponsored health insurance] market” (Kaiser Family Foundation/HRET, 2000: 66).

This together with the trends in the cost of coverage creates a suitable laboratory to explore whether or not managed care does indeed lead to lower healthcare costs, which Cutler et al. (2000) allude to.

Towards the end of the 1990s the popularity of managed care in the employer-sponsored health insurance market did wane as the cost of health insurance remained relatively stable (Kaiser Family Foundation/HRET, 2000). At the time, it appeared that there had been a consumer backlash against managed care, particularly against HMO plans; this was largely attributed to the restrictive conditions applicable under such arrangements. Since 2000, however, the trend once again reversed in favour of managed care with PPOs faring particular well as they offered enrolees the additional

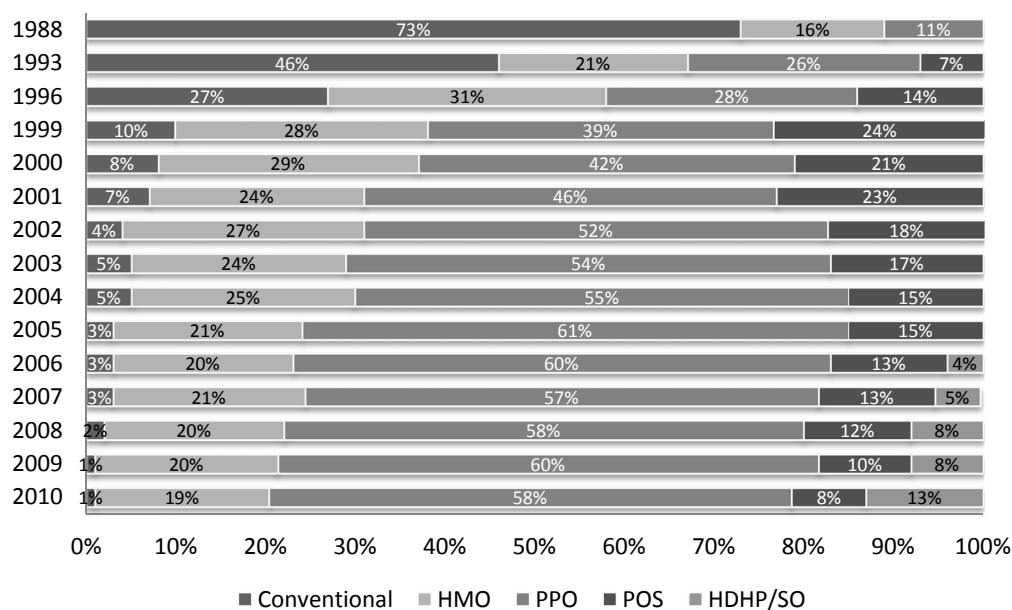
feature of permitting treatment from non-network healthcare providers, albeit on less generous terms.

Figure 2-21 illustrates the distribution of health plan enrolment for covered employees in the United States for selected years from 1988 until 2010. The market share of conventional fee-for-service plans within the employer-sponsored health insurance sector has shrunk from 73 percent in 1988 to a paltry 1 percent by 2010. PPO plans currently dominate the marketplace, enrolling 58 percent of covered employees. HMOs enrol 19 percent, whilst POS plans had a market share of approximately 8 percent. A relatively new entrant to the marketplace, high-deductible⁹³ health plans with a savings option (HDHP/SOs), has seen their market share grow rapidly in the United States from just 4 percent in 2006 to 13 percent by 2010.⁹⁴

⁹³ A general annual deductible is the amount an enrollee must pay before further healthcare expenditure is covered (partially or fully) by the health insurance plan.

⁹⁴ Only after legislative changes in the United States, did the establishment of savings arrangements for healthcare become a reality within employer-sponsored health insurance. Thus, prior to 2006 HDHP/SOs were excluded from The Kaiser Family Foundation/HRET Survey of Employer-Sponsored Health Benefits.

Figure 2-21: Distribution of Health Care Plan Enrolment for Covered Employees in the United States, 1988-2010



Source: Kaiser Family Foundation/Health Research and Education Trust, Employer Health Benefits 1999-2010; KPMG Survey of Employer-Sponsored Health Benefits, 1993, 1996; the Health Insurance Association of America (HIAA), 1988.

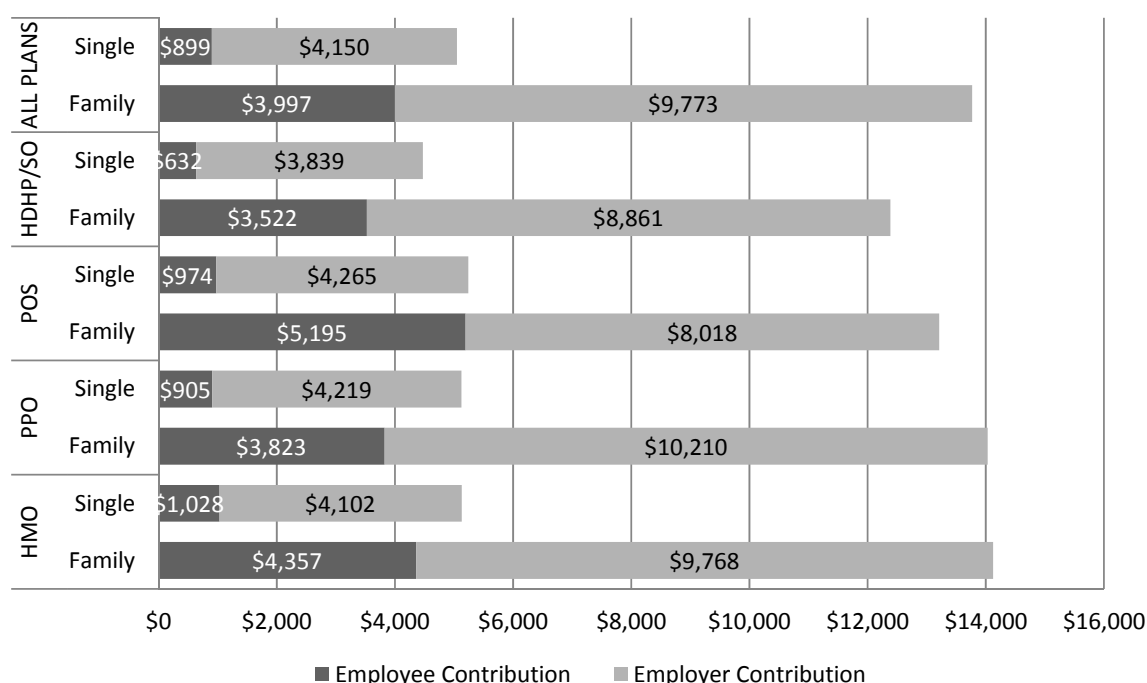
A general trend within the United States employer-sponsored health insurance sector has been observed where there has been a shift towards plan types with higher deductibles. By 2010, 27 percent of employees covered by employer-sponsored health insurance in the United States opted for a plan with a deductible of \$1,000 or more. This compares to 22 percent in 2009 and only 10 percent in 2006.⁹⁵ Furthermore, one would expect that employees enrolled in an HMO would be less likely to opt for a deductible, but the authors note that the percentage of employees enrolled in an HMO that have opted for a deductible increased from 16 percent in 2009 to 28 percent in 2010. Trends favouring higher deductible plans may indicate a change in preferences by both employers and employees towards plans that reduce benefits to provide more affordable cover. The ability of employees to directly affect preferences in favour of high-deductible health insurance plans is somewhat limited as employers determine

⁹⁵ This data relates to health plans that offer single coverage. Similar trends are observed under family coverage (Claxton et al., 2010).

the selection of healthcare plans that they make available to their employees. However, employees can influence the choice of health insurance plan(s) that is made available by employees through consultation with employers, contract negotiation and trade union bargaining. This is perhaps not surprising given the difficult economic conditions that arose as a result of the 2007-2010 global financial crisis (Claxton et al., 2010).

If one examines the average annual contribution (premium) for covered employees by managed care health plan, Figure 2-22 indicates that HMOs and PPOs charge the highest annual contributions for family coverage, \$14,125 and \$14,033 respectively. HDHP/SOs charge \$12,383 per annum for family. Thus, under family coverage, the difference between the most and least expensive plans (on average) is only \$1,742. Whilst one would expect HDHP/SO plans to be more affordable due to the high deductible feature, but as noted previously 28 percent of HMO enrollees are now opting for a deductible as well. Interestingly, the extent to which employers contribute towards the cost of coverage varies across both managed care plan type and single versus family coverage. At the one extreme, under POS plans that offer family coverage, employers subsidise the cost of coverage on average by 60.68 percent. At the other extreme, under HDHP/SOs that offer single coverage, one observes that employers contribute on average 85.86 percent towards the cost of coverage.

Figure 2-22: Average Annual Contribution (Premium) for Covered Employees by Managed Health Care Plan Type in the United States for Single and Family Coverage, 2010.



Source: Kaiser Family Foundation/Health Research and Education Trust, *Employer Health Benefits 2010*.

Managed Care, Utilisation and the Quality of Healthcare Delivered

Some detractors of the managed care framework have argued that managed care may lead to an under-provision of healthcare services as cost containment is paramount. It is often argued by critics that managed care sacrifices high-quality healthcare for cost savings (Gottfried & Sloan, 2002). Thus, managed care is thought to provide inferior health insurance coverage when compared to fee-for-service plans. This is however tempered by the fact that under-provision may lead to more severe financial cost implications in the future as the health status of enrollees deteriorates. The deterioration in health status would dictate a greater utilisation of healthcare services by the enrollee going forward.

Miller and Luft (1994) employed a rigorous review of peer-reviewed literature to compare the performance of HMOs to that of fee-for-service plans. Their review suggests that compared to fee-for-service plans, HMO plans exhibit significantly lower utilisation with respect to hospital services, expensive or discretionary procedures and tests. However, HMO enrollees when compared to their fee-for-service counterparts received more preventative tests, procedures and examinations such as cancer screening, pelvic, rectal and general physical examinations. HMO enrollees were also found to be more likely to visit general practitioners. When comparing the quality of healthcare received (health outcomes) under HMO versus fee-for-service plans, the results appear mixed (Miller & Luft, 1994). In a subsequent study, Miller and Luft (1997) also considered Preferred Provider Organisations (PPOs), which for the most part they grouped with fee-for-service (FFS) plans.⁹⁶ They concluded that the quality of care between managed care and fee-for-service plans were largely comparable.

Dudley, Miller, Korenbrot and Luft (1998) examine how financial incentives may impact upon the quality of healthcare enrollees receive. According to the authors because of the difficulty in measuring healthcare quality and the absence of appropriate mechanism to correct for risk, financial incentives are unlikely to managed care health plans to maximise healthcare quality. Nevertheless, they also find that the quality of clinical outcomes between FFS plans and HMOs are similar. Thus, any discernible differences in quality between FFS and HMOs can be explained by benefits design and coverage decisions rather than an inherent preference to deliver high or low-quality healthcare.

It is further argued that managed care may discourage healthcare providers from adopting new technological improvements through the purchase (lease or utilisation) of new equipment or other treatment tools, as these may translate into an increase in costs that cannot directly be passed onto patients. Hence, the diminution of financial

⁹⁶ Although, often categorised as a managed care organisation, a PPO is strictly not a managed care organisation, but rather enters into a relationship (partnership) with health insurers (or third-party administrators).

incentives to adopt new technological improvements is expected to hamper innovation. An appropriate question therefore to ask is: whether a higher penetration of managed care within the medical scheme business could be expected to lead to a deceleration in the growth in availability of new medical treatment technologies? From empirical evidence the answer appears to be ambiguous – in some cases “yes” and in other cases “no” (Friedman & Steiner, 1999; Baker, 2001; Baker & Phibbs, 2002).

Baker (2001) investigates how managed care may influence technological diffusion by examining the relationship between Health Maintenance Organisations’ (HMOs) market share and the diffusion of magnetic resonance imaging (MRI) equipment. The results indicate that across various states, increases in HMO market share are associated with the slower diffusion of MRI into hospitals over a period between 1983 and 1993, with substantially lower overall MRI availability through until the end of the 1990s. The author also finds that areas dominated by managed organisations display markedly lower rates of MRI procedure utilisation. Baker and Phibbs (2002) empirically examine the relationship between HMO market share and the diffusion of neonatal intensive care units. Once again higher HMO market share is associated with the slower adoption of intermediate neonatal intensive care units, but it is not the case with respect to the adoption of advanced neonatal intensive care units (Baker & Phibbs, 2002). Moreover, they infer that contrary to the common supposition that slower technological diffusion is harmful to patients; critically-ill new-borns may experience better health outcomes in areas where HMOs have a greater market share. The reason offered is that since health outcomes for seriously-ill new-borns are better when treated within an advanced neonatal intensive care unit, under an HMO-dominated area the restricted availability of intermediate neonatal units may increase the chance of receiving care within a healthcare facility that has an advanced neonatal intensive care unit. Therefore, the slower diffusion of intermediate intensive care units within HMO-dominated markets may benefit patients.

Friedman and Steiner (1999) analyse the availability of hospital intensive care units by adult patients who are under age 65 and not covered by *Medicaid*⁹⁷ in the United States. Utilising data for the states of Massachusetts and Florida during 1992, the authors consider factors such as the patient's condition, severity of illness, type of admission, emergency status, and the extent to which the total intensive care unit of the hospital is utilised. In both states there is no discernible difference in admission rates under managed care versus fee-for-service arrangements. In Massachusetts, they find that the length of stay in an intensive care unit is less for managed care and uninsured patients than for other privately-insured patients.

These empirical results of Friedman and Steiner (1999), Baker (2001), and Baker and Phibbs (2002) all suggest, as Baker (2001) asserts, that technology diffusion in healthcare can respond (positively or negatively) to changes in financial and other incentives, which may have further implications for both healthcare costs and the welfare of patients.

Cutler et al. (2000) examine how Health Maintenance Organisations (HMOs) differ from traditional health insurers with respect to the treatment of heart disease and they find that HMOs experience between 30 percent and 40 percent lower healthcare expenditures than traditional health insurance plans. Lower expenditures may translate into lower quality of healthcare provision. The authors do not find any discernible difference in the quality of care (actual treatment and health outcomes) provided by HMOs versus traditional health insurance plans.

⁹⁷ In the United States, *Medicaid* is a means-tested social health insurance programme funded by both the federal and state government, and is managed at state-level. *Medicaid* provides health coverage or nursing home coverage to eligible individuals and families with low incomes and financial resources. Having limited financial resources is the primary determinant for eligibility, but poverty alone is not the only criteria for eligibility. Other criteria include age, pregnancy, having a disability and immigration status. *Medicaid* does not pay money directly to the *Medicaid* enrollee, but rather reimburses healthcare providers for healthcare services rendered. Depending upon the state's law, enrollees may also be asked to pay a small portion of the cost (co-payment) for certain healthcare services.

In the United States, at the very least, managed care refers to the utilisation of selective networks of contracted healthcare providers, the implementation of mechanisms to incentivise enrolees to utilise these networks and some form of risk sharing with the networks (Doherty & McLeod, 2003). Doherty and McLeod (2003) contends that medical schemes in South Africa have until recently not taken advantage of establishing provider networks and negotiating risk-sharing arrangements with these providers. The authors claim that the reason may once again stem from the market power that these private healthcare provider groups wield.

2.5 Conclusion

The Medical Schemes Act No. 131 of 1998 has not had the desired effect of lowering healthcare financing costs. On the contrary, it has led to a rapid increase in medical scheme contributions coupled with a widening gap between contributions and benefits. The upward trend in medical scheme contributions is of great concern from the perspective of equity let alone efficiency (Doherty & McLeod, 2003). The authors go further to suggest that as incomes tend to be inflation-linked, it is difficult to envisage how people can keep up with the present escalation in medical scheme contributions. Moreover, employers who subsidise their employee's medical scheme contributions are also struggling to keep pace with the increasing cost of financing private health insurance. They note that increasingly, employers are opting to structure employee's remuneration packages in such a way that it leaves the decision of whether or not to purchase medical scheme cover up to the employee. In addition, fewer employees are willing to offer new employees subsidised medical scheme cover that extends beyond retirement (Doherty & McLeod, 2003).

Private hospitals have progressively consumed a larger slice of the private healthcare expenditure "pie". Three major hospital groups dominate the private hospital

landscape in South Africa and this has permitted oligopolistic practices to flourish within private healthcare provision. Currently, approximately 5 percent of all private hospital beds are located outside of these three hospital groups and this compares to more than 50 percent just 15 years ago (van den Heever, 2007; Magennis & van Zyl, 2009). Hence, these private hospital groups have become price setters rather than price takers. Another contributing factor to higher medical scheme costs is the shortages observed in the supply of healthcare professionals and in particular, specialist practitioners. These shortages have allowed specialist practitioners to enjoy a certain degree of market power and in the absence of regulation; they are able to charge patients well-above medical scheme tariffs without any penalty.

Healthcare providers often cite various factors to explain the observable cost escalation in healthcare costs. These include the rising cost of medical technology and innovations; the ageing profile of the population and the emergence of new diseases or diseases that become increasingly difficult to treat, such as AIDS or multi-drug resistant tuberculosis (Doherty and McLeod (2003). Whilst these factors may explain a portion of the cost escalation in healthcare costs; Reekie (1999), Doherty and McLeod (2003), and Erasmus and Fitchen (2010), claim that the primary reason for the healthcare cost explosion is the persistence of the fee-for-service reimbursement system. These authors maintain that the medical scheme industry should move away from the fee-for-service reimbursement system and towards more risk-sharing arrangements with healthcare providers.

Medical schemes have historically only concentrated upon implementing managed care elements that influence utilisation, such as rationing and co-insurance. Therefore, medical schemes have not attempted to tackle the incentives to over-supply healthcare services in the context of a fee-for-service reimbursement system (Reekie, 1999; Doherty & McLeod, 2003). Government's proposed new statutory price determination authority that will bring back multilateral tariff negotiation between healthcare providers and medical schemes with a mandatory dispute resolution mechanism

through arbitration, should also be beneficial in arresting the cost escalation observed within healthcare provision.

3 THEORY OF ADVERSE SELECTION

The objective of Chapter Three is to provide a detailed account of adverse selection theory. The chapter discusses the seminal literature in this area with the view to describe the response of insurance companies operating in a competitive insurance market to the problem of information asymmetry. Moreover, a suitable empirical framework based upon the theoretical literature is presented to directly test for the presence of adverse selection in insurance markets. The empirical evidence of numerous studies is also discussed in this chapter.

The term adverse selection itself originated in the context of insurance and Arrow (1963) is one of the early sources analysing its welfare implications within the health insurance market. . Siegelman (2004) describes adverse selection as:

“...the process by which insureds utilize private knowledge of their own riskiness when deciding to buy or forego insurance” Siegelman (2004: 1223).

For example, consider two individuals; *A* and *B*. Assume further that individual *A* is aware that they have a much greater likelihood of dying within the next week than the insurance company would anticipate. For individual *A*, life insurance that is priced according to the average likelihood of death for the overall population will appear to be a very attractive purchase (Siegelman, 2004). On the other hand, individual *B* is aware that they are more likely to live much longer than the average person with the identical observable characteristics, such as age, gender and health status. Hence, individual *B* will find that life insurance that is priced according to the average risk would not be an attractive purchase. By implication *A* is likely to demand more life insurance and *B* would curtail their purchase of life insurance. Therefore, insurers end up charging an average premium rate to the population that consists of increasingly higher risk individuals. Unless insurers respond by increasing premiums they would incur significant losses (Siegelman, 2004).

The seminal literature developing the theory of adverse selection includes Akerlof (1970)'s "Market for Lemons", Spence (1973)'s "Job Market Signaling", Spence (1974)'s book "Market Signalling", and particularly in the insurance context, Rothschild and Stiglitz (1976)'s "Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information". Indeed, Löfgren, Persson and Weibull (2002) present a useful review of the contributions that this seminal literature has made in the understanding of adverse selection and particularly, in explaining how insurance markets respond to imperfect information.

Akerlof (1970)'s classic paper examines the interaction between quality and uncertainty. Akerlof (1970) focuses upon the situation when one party to a transaction has more information about the characteristics of commodities exchanged than the other party to that transaction. More specifically, Akerlof (1970) demonstrates that in the used car market, where the seller has superior information about the car that they are selling than the buyer, it leads to a situation whereby only "lemons" (or poor quality used cars) are traded.⁹⁸

Spence (1973, 1974) demonstrates how signalling can be utilised to counteract the effects of adverse selection.⁹⁹ More specifically, Spence (1973, 1974) considers how education serves as a signalling mechanism within the labour market. For example, where employers cannot distinguish between prospective high- and low-productivity employees, Akerlof (1970)'s *Lemons Principle* suggests that the labour market would revert into a market for only low-productivity employees. Education is measured on a continuous scale, where ever-increasing values indicate higher levels of education. Spence (1973) assumes that the necessary cost (time, effort or expense) to acquire levels of education is lower for high-productivity individuals at each level of education. Note he assumes that education provides no improvement to productivity nor do individuals undertake education for any other reasons besides its value in

⁹⁸ Refer to Sub-Section 3.1.1 for a more detailed discussion of Akerlof (1970)'s analysis.

⁹⁹ In this context, signalling refers to the observable actions taken by agents to persuade opposing parties about the value and quality of their products or services (Löfgren et al., 2002).

providing signalling capacity.¹⁰⁰ Thus, individuals would choose the least amount of education that would be sufficient to serve as an appropriate signal to potential employers for their associated level of productivity. Assume further that new labour market entrants acquire education before entering the labour market. Under perfect information, perfect information and constant returns to scale, Spence (1973, 1974) demonstrates that individuals would elect to acquire no education before entering the labour market and each new employee would be paid according to their level of productivity (Löfgren et al., 2002). Under asymmetric information, high-productivity would acquire education as a signal of their productivity (ability) (Löfgren et al., 2002).

In Rothschild and Stiglitz (1976)'s seminal article, they investigate how the insurer (who is uninformed about the risk profile of the insured) can respond to improve the market outcome in the presence of information asymmetry. Rothschild and Stiglitz (1976) present a model to describe the presence of adverse selection within insurance markets under imperfect information. An important result emanates from this work whereby a market for insurance may fail to form, or as Rothschild and Stiglitz (1976: 629) declare "...not only may a competitive equilibrium not exist, but when equilibria do exist, they may have strange properties". The authors offer the separating equilibrium as a possible solution to address the observed market failure, whereby insurers offer separate contracts designed in such a way so as to attract the appropriate risk (either high or low risk individuals). This is achieved by adjusting effectively the level of cover offered and thereby the premium charged for insurance – the so-called self-selection mechanism of Salop and Salop (1976).¹⁰¹

Belli (2001) describes adverse selection as strategic behaviour by the more informed party against the less informed party. The insured is often seen to have superior information about their risk profile than the insurance company. In the health insurance context, each individual will select amongst a set of contracts based upon

¹⁰⁰ It is possible to extend the analysis whereby education does enhance productivity and this would merely affirm the conclusion

¹⁰¹ Refer to Sub-Section 3.1.3 for a more detailed discussion of Rothschild and Stiglitz (1976)'s paper.

their expected probability of utilising health care services. Thus, those individuals who anticipate requiring medical care regularly choose more generous plans compared to those requiring limited use. Cutler (1996: 30) describes the pervasiveness of adverse selection within health insurance markets as:

“Almost all health insurance systems where individuals are allowed choice of insurance have experienced adverse selection. Medicare [enrolees] who choose managed care¹⁰² are healthier than...[those] who do not. The Federal Employees Health Benefits Program...has adverse selection between more and less generous policies. The spread in premiums between more and less generous policies is 68 percent greater than benefits alone would dictate...And almost every large firm that has encouraged employee choice has found the cost of the most generous policies increases sufficiently rapidly than these policies are no longer viable”.

Cutler (1996)’s comments suggest that when offered a choice between generous and moderate healthcare indemnity plans, individuals who are more certain as to their need for coverage will seek out the more generous plans.

Insurers attempt to overcome the problem of adverse selection by employing a *screening device*. Within a labour market context, Salop and Salop (1976) describe the screening device as a *rule of thumb* whereby a certain set of observable characteristics correlated with the parameter of interest are used to determine the anticipated job performance of a new applicant. In terms of insurance, information gathering and subsequent risk classification serve as the tools to appropriately assess the risk profile of an individual or homogenous group of individuals. Thus, allowing the insurance company to charge premiums that reflect the inherent risk profile of the insured. Through risk classification, insurers utilise certain individual characteristics, such as age, gender and past claim experience, in an attempt to classify individuals into homogenous risk pools.¹⁰³ The premise is that the insurer can simply charge all members within the same pool an identical premium as their risk should be indistinguishable. Thus, the insurer is indirectly utilising the individual’s risk characteristics to charge the person a premium that is commensurate with their risk

¹⁰² Belli (2001) note that managed care plans impose stricter controls and restrictions over the utilisation of health services than traditional indemnity plans.

¹⁰³ These risk characteristics are generally based upon the historical claims experience of the insurer.

status. Premiums based upon this risk classification may be erroneous for two primary reasons:

- The characteristics utilised in the classification procedure may not fully describe the risk profile of the individual; and/or
- Information gleaned from the insured regarding their particular risk characteristics may be inaccurate.

In the absence of accurate information concerning these characteristics, the insurance companies are unable to differentiate between various risks within a particular risk pool. Under the standard adverse selection setting, since insurers are unable to distinguish between the different risk classes within the same pool, all members must be offered the identical price for insurance coverage. If one assumes that there are only two distinct risk classes, namely, high risk and low risk individuals. The insurance company would charge a weighted-average premium based upon the proportion of high risk and low risk individuals within each pool. Once the premium is determined, a situation arises whereby the low risk individuals are being charged a premium that overestimates their true underlying risk. Whilst on the other hand, the high risk individuals are charged a premium that underestimates their true underlying risk. Thus, the low risk members of the pool perceive accurately that they are effectively subsidising the high risk members within their pool. In this case, these low risk individuals are more likely to opt out of the insurance market as the premium they are charged is not commensurate with their risk level. Whereas the higher risk members of the pool are likely to demand more insurance because they are being charged a discounted price for insurance. As the low risk members exit the pool, the risk profile of the pool deteriorates further, and consequently the premium is adjusted upwards to reflect the overall higher risk profile of the pool.¹⁰⁴ The situation that arises is akin to that described by Akerlof (1970)'s 'lemons' principle whereby the lemons (high risks) drive out the low risks. In the case of two risk classes, namely, high and low risk, one can observe that each distinct risk profile causes individuals to behave very differently. When insurance companies offer a menu of policies, a

¹⁰⁴ As the low risk members of the pool exit the insurance market, the risk profile of the pool deteriorates and this would be observed through higher actual claims experienced by the insurer.

coverage-risk correlation can be expected to manifest itself in a tendency among high-risk individuals to select policies that offer more comprehensive coverage.¹⁰⁵ The asymmetry of information between the purchaser and seller of insurance is what the theory suggests causes the adverse selection phenomenon to occur within insurance markets. In the health insurance sector, it is important because each individual chooses among the set of contracts offered by medical aid schemes according to their expected probability of utilising healthcare services. Hence, those who foresee an extensive use of healthcare will tend to choose plans that are more generous while those who expect to have limited use opt for more moderate plans.

Individuals purchase insurance to protect themselves against possible adverse outcomes that can result in a loss. A key premise behind insurance is the principle of indemnification whereby individuals seek protection so that they can be placed in the identical financial position prior to the insured event occurring. If the insured believes that the premium charged by the insurer underestimates their underlying risk, it is possible that they may perceive that insurance offers them an expected monetary gain. This ability to profit from insurance coverage would violate the principle of indemnification. This, however, is unsustainable from the insurer's perspective, as observed claims would exceed expected claims. The insurer would need to adjust premiums upwards to reflect the higher associated risk and offset future underwriting losses. Traditionally, risk aversion has been seen as the key determinant of the demand for insurance by individuals. The decision to purchase insurance by individuals is however driven by several factors. Adverse selection appears to be housed within a behavioural context whereby the decision to purchase insurance (and the choice of appropriate coverage levels) is determined by an individual's relative risk status to that of their pool. The intrinsic mechanism used by medical schemes to overcome the problem of adverse selection, is through offering alternative contracts, where higher deductibles can be traded-off against lower contribution rates. Thus, members through their own self-selection effectively separate into distinct risk classes (Salop & Salop, 1976).

¹⁰⁵ When insurers offer a single insurance product, the coverage-risk correlation can be expected to manifest itself through a greater tendency by high-risk individuals to purchase insurance.

Belli (2001) describe this particular self-selection *screening strategy* as even more critical to the sustainability of the insurance market when regulation is in place that does not allow premiums to reflect individual risk (premium rate restrictions) or as in South Africa's case does not permit medical schemes to utilise certain pertinent information relating to a members' health status.

Our traditional view of adverse selection is from the supply side or from the insurers' perspective. Insurers require accurate information concerning the risk characteristics of the insured so as to make optimal decisions about policy offerings and the appropriate pricing of these contracts. As discussed previously, adverse selection exists when the buyer has superior (more accurate) information than the seller. This implies heterogeneity of insureds with respect to their utility functions, risk profiles and hence their associated premiums (Valsamakis, Vivian & du Toit, 1999). If the insurer knew the true nature of risk (and was permitted to act upon that information) for each insured then the premium charged would accurately reflect the inherent risk of the individual, thereby averting the theoretical adverse selection *death spiral*.

It is important to note that adverse selection can in fact affect both sides of the insurance market. Susceptibility to information asymmetry is a feature that is not only privy to the insurer. The insurer may possess superior information over the insured. Therefore, on the demand side (insured's perspective), information asymmetry can occur when an individual is misinformed about the quality of products and services that an insurer offers as well as the appropriateness of coverage. The buyer of insurance is often faced with problems in acquiring the information they require to make informed purchasing decisions, these problems include difficulty in obtaining accurate information with respect to the:

- appropriate price for insurance coverage (actuarially fair premium and risk premium);
- interpretation of insurance contract provisions;

- quality of service offered by different insurers;
- financial strength and soundness of insurers; and
- ability of insurers to meet their contractual obligations.

Due to this information asymmetry the insurance buyer will be at a disadvantage and might purchase a policy, which is unnecessarily expensive or perhaps contains inadequate coverage. Owing to the lack of technical knowledge and expertise on the part of the consumer, an insurance company may exploit their comparative advantage by selling the consumer sub-optimal insurance policies. Furthermore, the insured may suffer the consequences of being unaware (or at best partially informed) of the technical provisions contained within the policy documentation due to their lack of expertise and technical knowledge. Therefore, on the demand side of the insurance market, the consumer is susceptible to abuse by the insurance company. The insurer can exploit their comparative advantage because of the inherent information asymmetry.

For example, an individual, say Mr X, who has just moved into a new property in a different geographical area, may approach an insurance company ABC to purchase a comprehensive homeowners insurance policy. Despite this particular area having infrequent natural disasters such as hurricanes or earthquakes, company ABC recommends a policy which will compensate Mr X for damage arising from such natural disasters. As Mr X is unaware of the weather patterns and geomorphology of this area, Mr X pays an additional premium to protect his property against hurricane and earthquake damage. Mr X's perception of the risk of such events is misplaced, which creates an opportunity for the insurance company to charge an additional premium for the hurricane and earthquake protection that is not commensurate with the risk. Company ABC is exploiting their informational advantages by encouraging Mr X to insure against an event that is unlikely to occur. In addition, Mr X may have an inherent fear of natural disasters, which would predispose him to this form of manipulation by the insurance company and ultimately alter his demand in favour of this type of insurance coverage.

The true extent of the insured's risk with respect to a loss-causing event(s) may indeed be known to the insurer rather than the insured. Thus, the insured's perception of their true risk may be erroneous, which once again may cause the insured to make sub-optimal insurance purchasing decisions. The insurance company can take advantage of their superior knowledge about the insured's risk to influence demand decisions through contract offerings and pricing. Although, competitive forces should mitigate the extent to which the insurer can exploit the insured, it nevertheless has the potential to contaminate (or at least offset) our traditional view of how adverse selection may affect the insurance market. Informational asymmetries may operate in the opposite direction favouring insurance companies over the insured.

The seminal papers by Akerlof (1970), Spence (1973, 1974), and Rothschild and Stiglitz (1976) have stimulated considerable discussion into the theory of adverse selection. Other papers, such as Pauly (1974), argue that in the absence of perfect information, competitive outcomes in insurance markets may be non-optimal. He proves that these outcomes are not only suboptimal when compared to the infeasible optimum obtained under perfect information, but also compared to the set of feasible optima. Spence (1976) elaborates further about signalling and screening with particular focus upon contingent contracts. The author concludes that exogenously costly signals can transmit information, but it is often at some discernible real social cost. Wilson (1977) also presents a model of insurance markets where one has incomplete information. By suggesting a less stringent definition of equilibrium than the one initially described by Rothschild and Stiglitz (1976), he is able to demonstrate that by permanently removing insurance contracts that make negative expected profits, equilibrium is far more likely to exist.¹⁰⁶ Löfgren et al. (2002)'s account of the analytical contributions of George Akerlof, Michael Spence, Michael Rothschild and Joseph Stiglitz, do provide a useful backdrop to explaining the functioning of markets in the presence of asymmetric information. In addition, various surveys of empirical studies have specifically examined whether or not adverse selection is indeed present

¹⁰⁶ Wilson (1977) was first elucidated in Wilson (1976)'s Ph.D. thesis entitled "Equilibrium in a Class of Self-Selection Models" and Rothschild and Stiglitz (1976) also included Wilson (1976)'s notion of equilibrium within their analysis.

within different insurance segments, lines and markets. The surveys of empirical studies include Dionne and Doherty (1992), Cutler and Zeckhauser (2000), Chiappori (1999, 2000), Dionne, Doherty, and Fombaron (2001), Chiappori and Salanié (2003), and Hall (2006).

More recent contributions, such as Fang, Keane and Silverman (2008), Koufopoulos (2007), Seog (2009) and Wambach (2000) have extended the theory. Seog (2009) investigates the outcomes when each party to an insurance contract simultaneously has superior information. In the case of medical schemes, one would assume that medical scheme members have superior information about specific risks while medical schemes have superior information about general risks (Seog, 2009: 279). Under standard adverse selection theory, the insured is considered to possess private information about their specific risk.¹⁰⁷ However, the author asserts that it is also conceivable for the insurer to possess superior information about certain risks. For instance, individuals would know the extent of their smoking habit (on average how many cigarettes they smoke per day), whilst health insurers may have superior information about how smoking affects the health condition of an individual (Seog, 2009: 279-280). The author, utilising a similar methodology found in Chassagnon and Villeneuve (2005) and Villeneuve (2003), proves that under these assumptions, outcomes could well deviate from traditional adverse selection theoretical predictions.

3.1. Imperfect Information in Competitive Insurance Markets

Before presenting a model to describe the functioning of a competitive insurance market under imperfect information, it is beneficial to provide an account of the *Lemons Principle* as described by Akerlof (1970) to illustrate how information asymmetry through adverse selection can cause the failure of the insurance market to form.

¹⁰⁷ Dionne, Doherty and Fombaron (2000) and Chiappori (2000) provide a detailed review of the standard adverse selection models.

3.1.1. Lemons Principle

In Akerlof (1970)'s investigation into the relationship between quality and uncertainty, he identifies many applications whereby the interaction between quality differences and uncertainty can explain various observable features within several markets. He states that the '*Lemons Principle*' may help to explain the reluctance of employers to employ minorities, to determine the costs of dishonesty, to understand the operation of credit markets within underdeveloped economies and to explain why individuals aged over 65 find it difficult to obtain insurance coverage against healthcare expenditure, disability or death. Information asymmetry develops between the buyer and seller of commodities because the buyer is uncertain as to the quality of the good being purchased. The seller on the other hand has superior information regarding the quality of the item that he/she is trying to sell to the buyer. The buyer generally uses a static market statistic, such as a price, to act as a measure for the quality of an item (Akerlof, 1970). Akerlof (1970: 488) goes further by stating:

“...there is incentive for sellers to market poor quality merchandise, since the returns for good quality accrue mainly to the entire group who statistic is affected rather than to the individual seller. As a result there tends to be a reduction in the average quality of goods and also in the size of the market”.

This feature is illustrated precisely by Akerlof (1970)'s examination of the automobile market. The author was perplexed by the discrepancy observed between the price of a new car and the price of a used car. It is often the case that the price of a new car is significantly higher than that of a comparable used car. Surely, as he argued the fact of merely owning a new car cannot attract such a premium. He introduces a simple example to illustrate. Assume that one has four types of cars; new, used, good and bad (“lemons”) cars. Table 3-1 describes the four possible outcomes when purchasing a motor vehicle.

Table 3-1: Possible Outcomes when Purchasing a Motor Vehicle

	Good	Bad
New	(New, Good)	(New, Bad)
Used	(Used, Good)	(Used, Bad)

Assume further that the proportion of cars produced that are good is equal to p and thus the corresponding proportion of cars produced that are lemons is equal to $1-p$. If new and used cars were priced identically, the following scenario would conceivably transpire; a buyer, say Mr X, of a new car would assign a probability of p that the car is a good and a corresponding probability $1-p$ that the car is a lemon. Once Mr X has an opportunity to drive the vehicle, he assigns a new probability that the car is good, say p^* , and the corresponding probability $1-p^*$ that the car is a lemon. Through driving the vehicle, Mr X is able to glean more accurate information concerning the quality of the car. For example, Mr X may become aware of a noise whilst driving the vehicle and this could be indicative of an underlying mechanical flaw, which would detract from the quality of the vehicle. In this case, p^* would be lower than his original p . Importantly, Mr X now has superior information relating to the quality of the vehicle than at the time he initially purchased the vehicle. If $p^* < p$, Mr X would choose to sell his “used” car and using those proceeds he could purchase a new car with a higher probability of it being good. As Akerlof (1970) notes the purchaser of Mr X’s used car is at an informational disadvantage when compared to Mr X and as such can only assign a probability of p that the used car he is purchasing is a good car. If on the other hand, Mr X’s assessment of the original car that he purchased is $p^* > p$, then Mr X would be unwise to sell his used car and use the proceeds to purchase a new car. Under these circumstances he would be purchasing a new car with a lower probability of it being good. The result of this analysis is that if new and used cars were priced identically, then ultimately only lemons would be traded.

Here as Lofgren, Persson and Weibull (2002:198) suggests Akerlof (1970)’s analysis illustrates “how private information may lead to the malfunctioning of markets”. The

seriousness of this exemplification cannot be underestimated. It is not difficult to extend this example into a context whereby the traded commodities with varying degrees of quality no longer need to be identically priced, and yet the outcome will be as serious. Consider only the used automobile market, where once again one has bad (“lemons”) and good cars in fixed proportions as before. Assuming that the proportion of cars available that are good is equal to p and thus the corresponding proportion of cars produced that are lemons is equal to $1-p$. Each buyer seeks only to purchase one vehicle, but at the time of purchase the buyer cannot distinguish between the low and high quality cars. Suppose that buyers assign a value of V_B to a bad car and a value of V_G to a good car, where $V_B < V_G$. Sellers assign a value of V_B^* to a bad car and a value of V_G^* to a good car. Here the value that the seller assigns to good and bad cars is lower than the values assigned by buyers. Mathematically, $V_B^* < V_B$ and $V_G^* < V_G$.

Following a similar analysis as outlined by Lofgren et al. (2002), with separate markets for both bad and good used cars, it is conceivable that within the low quality used car market every price between V_B^* and V_B would support beneficial transactions (Lofgren et al, 2002). Similarly, for the high quality used car market prices between V_G^* and V_G would also support beneficial transactions between the two parties (that is between the sellers and buyers)¹⁰⁸. Since buyers are unable to distinguish between good and bad used cars at the time of purchase, the resultant information asymmetry actively encourages sellers to market low-quality used cars within the high-quality used car marketplace. In the absence of suitable mechanisms to mitigate the information asymmetry, the low-quality and high-quality markets would merge into a single market for all used cars (Lofgren et al, 2002). The buyers’ valuation of used cars would thus converge towards a single price measure ($\bar{V}$) for all used cars, which can be described by Equation 3-1.

¹⁰⁸ The beneficial transactions that would conclude between sellers and buyers would represent socially efficient outcomes as all gains from trade would be realised (Lofgren et al, 2002).

Equation 3-1: Buyers' Weighted-Average Valuation of Used Cars once a Single Market Emerges

$$\bar{V} = pV_G + (1 - p)V_B$$

where $\bar{V}$ represents the weighted-average price of a used car that a buyer is willing to pay, p represents the proportion of cars available that are good and $1 - p$ represents the corresponding proportion of cars that are bad. V_G and V_B are the buyers' assigned values for a good and bad car respectively.

Thus, if both good and bad used cars are sold, in the presence of information asymmetry buyers would be willing to pay at most $\bar{V}$ for a used car. However, if sellers' valuation of a good used car fell below the weighted-average price buyers were willing to pay, that is $V_G^* < \bar{V}$, then sellers would refrain from selling any good used cars. Ultimately, a market for only bad used cars (lemons) would emerge and thus buyers' weighted-average valuation of a used car would decline accordingly. The buyers' weighted-average price that they would be willing to pay for a used car would decline towards V_B . Therefore, as long as $V_B^* < V_B$, sellers would continue to offer only lemons. The result is a market for only lemons (Akerlof, 1970). Akerlof (1970: 490) explains that the bad cars drive out the good cars "...because they sell at the same price as good cars since it is impossible for a buyer to tell the difference between a good and a bad car; only the seller knows". A useful extension to this analysis is to consider varying degrees of quality and as Akerlof (1970) points out that without loss of any generality, it is quite conceivable that the very bad would drive out the bad who drive out the not-so-bad who drive out the average who drive out the not-so-good who drive out the good who drive out the very good. Ultimately, at the extreme one would be left with a market with only the lowest quality item, which clearly would be no market at all.

Akerlof (1970)'s presents an example that provides a mathematical confirmation of the assertion that information asymmetry with respect to quality may in the case of the automobile market lead to an inability of the market to form. The author assumes that the demand for automobiles is a function of two variables, namely, the price of the automobile (p) and the average quality of the used cars traded denoted by μ . Thus,

the quantity of automobiles demanded is given by $Q^d = D(p, \mu)$. The supply of automobiles is dependent upon price, $S = S(p)$ and the average quality is also a function of the price such that $\mu = \mu(p)$. It is understandable to believe that an individual's assessment of automobile's quality will be influenced by the price of the car. Often people perceive that if an item is expensive it somehow possesses a higher quality compared to an item that attracts a lower price. In equilibrium, demand would equal supply at a given level of average quality (Akerlof, 1970).

Using utility theory, Akerlof (1970) constructs a hypothetical example to illustrate the effect of quality uncertainty. Assume further that there are two groups of car traders, the first group who possess all the cars and the second group who has no cars. A total of N cars are owned by the individuals making up Group One. As the cars are owned by Group One, they are effectively pre-owned (used) and thus the market that he is considering is a market for used cars only. The utility function of Group One is given by Equation 3-2 and Group Two's utility function is given by Equation 3-3. Both utility functions describe a linear function with respect to utility. As the author notes assuming that individuals are risk neutral does not distract from the results of the analysis. Non-linear utility would essentially render the analysis unnecessarily complex. Further, by assuming linear utility functions, the author removes potential risk aversion affects that may contaminate the results. The nature of the utility functions also suggests that for an individual in either group, the addition of a second car, or indeed the k th car, will add the exact same amount of utility as the first car (Akerlof, 1970).

Equation 3-2: Utility Function of Group One (Akerlof, 1970)

$$U_1 = M + \sum_{i=1}^n x_i$$

where U_1 represents the utility function of Group One, M represents the monetary value the consumption of other goods besides cars, x_i the quality of the i th car and n the number of cars.

Equation 3-3: Utility Function of Group Two (Akerlof, 1970)

$$U_2 = M + \sum_{i=1}^n 3/2 x_i$$

where U_2 represents the utility function of Group Two, M represents the monetary value of consumption of other goods besides cars, x_i the quality of the i th car and n the number of cars.

Akerlof (1970: 491) makes the following further assumptions:

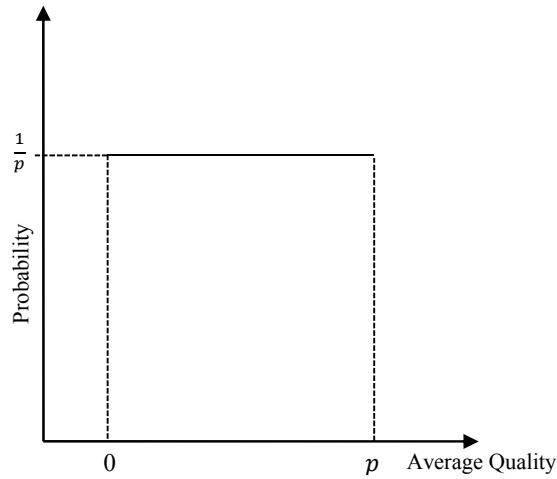
1. Both groups of individuals are von Neumann-Morgenstern utility maximisers.
2. The N cars have uniformly distributed quality x , with $0 \leq x \leq 2$. x_i is the quality derived from “consuming” the i th car. As x is uniformly distributed between 0 and 2, the expected value of x would be given by $E[x] = 1$.
3. The price of all “other goods” (everything except cars) M is unity. For each group, individuals derive one *util* of utility for every one dollar worth of every other good consumed.
4. Group One has income (wealth) Y_1 and Group Two has income (wealth) Y_2 .

The average quality, μ , is itself assumed to be distributed uniformly between 0 and p (that is $0 \leq \mu \leq p$).¹⁰⁹ Here by implication p represents the full price one would pay for a car essentially with “perfect” quality and at the other extreme, a price of 0 would be paid for a car with “zero” quality. The crucial element underpinning Akerlof (1970)’s analysis is the fact that the average quality, μ , is uncertain. He does not adequately emphasise that it is precisely for this reason that information asymmetry surfaces. By describing μ as a random variable following a particular distribution, it subjugates average quality to take on any possible value within a pre-defined

¹⁰⁹ The continuous uniform distribution is used for illustrative purposes, other distribution may be used but these would unnecessarily complicate the analysis. For a random variable X that follows a continuous uniform distribution, all possible values within an interval of say $[a, b]$ have an equal likelihood of occurring. The distribution would be defined by the two parameters, a and b , which represent the minimum and maximum values that the random variable can take. In summary $X \sim U(a, b)$, with probability density function $f(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b \\ 0, & x < a \text{ or } x > b \end{cases}$ and first moment of the distribution is given by $E[X] = \frac{a+b}{2}$.

continuous interval with equal probability. It is thus left to the buyer to make an assessment of the likely quality of any car being purchased. Whilst at the same time it is understood that the seller views the quality of the car being sold as a certainty. Therefore, the buyer can only make an informed expectation of the likely outcome for the quality of the car and in this case this would translate into the expected value of μ . Given that the average quality can take on any value between 0 and p , and $\mu \sim U(0, p)$, the probability density function would be represented by Figure 3-1.

Figure 3-1: Probability Density Function for the Average Quality μ



“The [total] demand for used cars will be the sum of the demands by both groups” (Akerlof, 1970: 491). From the author’s model it is trivial to establish the demand equations for each group. Ignoring indivisibilities, Group One’s demand for used cars is described within Equation 3-4 and Group Two’s demand function for used cars is given by Equation 3-5.

Equation 3-4: Demand for used cars by Group One (Akerlof, 1970)

$$D_1 = \begin{cases} \frac{Y_1}{p}, & \mu/p > 1 \\ 0, & \mu/p < 1 \end{cases}$$

where Y_1 represents the income of Group One, p is the price of a car and μ represents the average quality of a car.

Equation 3-5: Demand for used cars by Group Two (Akerlof, 1970)

$$D_2 = \begin{cases} \frac{Y_2}{p}, & 3\mu/2 > p \\ 0, & 3\mu/2 < p \end{cases}$$

where Y_2 represents the income of Group Two, p is the price of a car and μ represents the average quality of a car.

To explain how the demand described in Equation 3-4 is derived, one notes that if the ratio of the average quality to the price of a car is greater than unity, it implies that an individual would be able to purchase a car at a “bargain” rate – the car bought would be worth more than the price paid for it. Effectively, individuals are purchasing cars with an average quality translated into a value that exceeds the underlying price paid. Hence, Group One would justifiably utilise all their income to purchase cars, since for every dollar spent on a car you are receiving in excess of a “dollar worth of quality”.¹¹⁰ Further, as individuals within Group One derive one *util* worth of utility for every dollar worth of quality “consumed” with respect to a car owned, it translates into the observation that individuals purchasing a car are in fact receiving more than a dollar worth of quality. This means that they receive more than one *util* for every dollar spent on a used car. Now because individuals only receive one *util* for every dollar spent on “other goods”, it is clear individuals would therefore rather allocate all their income to the purchase of used cars.

If on the other hand, the ratio of the average quality to the price of a car is less than unity, it implies that individuals within Group One that purchase a used car would receive a car that is worth less than the price paid for it. Thus, they receive less than one *util* for every dollar spent on a used car and thereby individuals within Group One would allocate all their income to the purchase of “other goods” instead of used cars – the demand for used cars by Group One would be zero.

¹¹⁰ From the utility functions described in Equation 3-2 and Equation 3-3, it is apparent that individuals’ utility are a function of the quality they derive from owning a used car.

Similarly from Equation 3-5, it is possible to describe how the demand for used cars is derived for Group Two. From Equation 3-5, it is clear that individuals within Group Two derive one-and-a-half *utils* worth of utility for every dollar worth of quality “consumed” with respect to a car owned, which compares favourably against spending on “other goods” because individuals only receive one *util* for every one dollar spent. Now as long as the price of a car does not exceed one-and-half times the average quality of the car, individuals within Group Two would allocate all their income in the pursuit of purchasing cars. In other words, even if they receive less than one dollar worth of quality from a purchased car, the utility they derive from the car purchased exceeds one *util* thereby making it an attractive purchase over “other goods”. Notably, as long as they receive at least two-thirds of one dollar worth of quality for every dollar spent on a car, the utility derived from “consuming” the purchased car will exceed one *util* and therefore, individuals within Group Two would allocate all their income to purchasing used cars.

When the price of a car exceeds one-and-half times the average quality of the car one receives, Group Two’s individuals even though they value cars more than “other goods” in terms of utility, would refrain from purchasing cars in favour of “other goods”. Thereby, their demand for used cars will revert to zero.

The total demand $D(p, \mu)$ for used cars by both groups would simply be the summation of the individual demands (Akerlof, 1970). Equation 3-6 describes the total demand for used cars by both groups of individuals.

Equation 3-6: Total Demand for used cars by both Groups (Akerlof, 1970)

$$D(p, \mu) = \begin{cases} \frac{Y_1 + Y_2}{p}, & p < \mu \\ \frac{Y_2}{p}, & \mu < p < 3\mu/2 \\ 0, & p > 3\mu/2 \end{cases}$$

where Y_1 represents the income of Group One, Y_2 represents the income of Group Two, p is the price of a car and μ represents the average quality of a car.

From Equation 3-6, three possible demand states occur depending upon the price of a used car relative to the average quality derived from a used car. Under the first scenario, if the price paid for a used is less than the average quality one derives from a used car, both groups would obtain in excess of one dollar worth of quality for every dollar spent on a car and hence this would translate into more than one *util* of utility for every dollar spent. Thus, both groups would allocate all their income in favour of purchasing used cars rather than “other goods”. For the second possibility, for every dollar spent on a used car, the buyer would only receive between 66.6 cents and one dollar worth of quality. In terms of utility, for Group One the utility derived would be less than one *util* for every dollar spent and for Group Two the utility would still exceed one util for every dollar spent. Therefore, Group One would demand no cars whilst Group Two would still demand cars in favour of “other goods”. The final scenario occurs when for every dollar spent on a used car; the buyer would receive less than 66.6 cents worth of quality. This translates into less than one *util* for every dollar spent on a used car for both groups and therefore every individual would allocate all their income in favour of “other goods” rather than used cars.

Recall information asymmetry is present because buyers are uncertain as to the quality of any used car they purchase. Sellers on the other hand are understood to possess superior knowledge concerning the quality of the used car that they are selling. As a consequence buyers can at best make an informed assessment of the

average quality of used cars $\mu \sim U(0, p)$. The information asymmetry is captured by assuming that the average quality is a random variable following a continuous uniform distribution. Out of the class of estimates for the average quality, the first moment of the distribution would be the best unbiased and consistent estimate of the average quality of a used car. Hence, a buyer's assessment of the average quality of a used car would equate to $E[\mu] = p/2$ (the expected value of μ equals the price of a used car divided by 2).

Since the expected value of μ is equal to $p/2$, it is now possible to substitute the expected value of μ for each μ under the various constraints within Equation 3-6. From this one obtains Equation 3-7.

Equation 3-7: Total Demand for used cars by both Groups substituting for the $E[\mu]$

$$D(p, \mu) = \begin{cases} \frac{Y_1 + Y_2}{p}, & p < \frac{p}{2} \\ \frac{Y_2}{p}, & \frac{p}{2} < p < 3p/4 \\ 0, & p > 3p/4 \end{cases}$$

where Y_1 represents the income of Group One, Y_2 represents the income of Group Two, p is the price of a car and μ represents the average quality of a car.

It is immediately noticeable that after the substitution for the expected average quality, the constraint under the first scenario stipulates that $p < \frac{p}{2}$, and since p , the price of a used car, can only be greater than or equal to 0 ($p \geq 0$) – this constraint is untenable. Thus, the first scenario is not possible. Under the second scenario, whilst $p > \frac{p}{2}$ is true, the second part of the constraint is once again false p cannot be less than three-quarters of p ($p \nless 3p/4$) – the second scenario is not feasible. For the

final scenario, the constraint $p > 3p/4$ is indeed true and thus the only possible scenario is that the demand for used cars is zero ($D(p, \mu) = 0$). Therefore, from this mathematical exercise, Akerlof (1970) proves that the demand for used cars would revert to zero – no market for used cars would form.

The *Lemons Principle* as advocated by Akerlof (1970) can easily be related to the health insurance context. The author indeed relates it to the medical insurance market where individuals over 65-year old often have difficulty in obtaining medical cover. Individuals who are over 65 generally have a higher propensity to require health care services compared to younger individuals holding all other factors constant. Claims experience suggests that senior citizens not only claim more frequently but the size of their claims tends to be larger. Akerlof (1970: 492) asks the question: “...why doesn’t the price rise to match the risk?” He argues that if the price were to rise to reflect the greater risk, increasingly those over-65s who are more certain that they will require medical cover would seek or at least maintain their current level of cover. Over-65s with a lower risk profile are more likely to opt out as they would not be prepared to pay a premium that effectively subsidises their “high” risk counterparts.¹¹¹ As the risk of the pool deteriorates, the premiums would need to rise once more to reflect the higher claims experience and the so-called adverse selection *death spiral* would manifest itself. It is clear that Akerlof (1970)’s analogy of bad cars (or lemons) driving out good cars can easily be translated into the notion that higher risk over-65s drive out lower risk over-65s. Ultimately, as with the case with the varying degrees of quality, which in this case would translate into the varying degrees of risk within the over-65 cohort; it is conceivable that no market for medical insurance for over-65s will form.

¹¹¹ The premium they would pay far exceeds their expected claims experience and hence the risk premium – the amount which is over-and-above their actuarially fair premium – is substantially relative to the expected cost of claims.

Salop and Salop (1976: 619) highlight Akerlof (1970)'s contribution as:

“...[Akerlof (1970)] showed that the inability of one side of the market to convey its information about the product convincingly to the other side of the market prevented the consummation of transactions, which would occur if the information transfer could be accommodated”.

Akerlof (1970)'s analysis does accentuate the far-reaching consequences of information asymmetries within the automobile market and the difficulties that senior citizens experience in purchasing medical insurance.

Lofgren et al. (2002) acknowledge that Akerlof (1970) outlines an important response by economic agents to address the consequences of adverse selection through the creation of institutions that mitigate information asymmetry. Guarantees, brand-name goods and chains, and licensing (certification) practices all reduce quality uncertainty. Many consumer purchases, particular large purchases such as motor vehicles or electrical appliances, all carry guarantees or at least some form of warranty to protect the consumer from quality uncertainty. Should a purchased item fail or malfunction, the consumer has recourse as he or she can return the item to be repaired or replaced (generally at no cost to the consumer) during the warranty or guaranty period. To some extent the risk of poor quality is shifted back onto the seller of the goods. Consumers have the prerogative to curtail purchases of brand-name goods should the quality of a purchased item not live up to their expectations (Akerlof, 1970). He states that restaurant or hotel chains alleviate quality uncertainty as consumers can once again refrain from utilising these services should the quality of service fall short of what they would expect to receive at any other member of that respective chain. Certification and licencing of skilled labour (and professions) indicate a reliable level of proficiency (Akerlof, 1970). Even qualifications from academic or vocational institutions serve as “brand names” that reflect either positively or negatively concerning the ability of a particular individual to perform a specific function.

3.1.2. Gresham's Law

An alternative exposition of this analysis is articulated within Gresham's Law.¹¹² Gresham's Law describes how bad money drives out good money if their exchange rate is set by law. "Bad" money refers to coinage that has a commodity value (the actual value of the metallic constituents of the coins) which is considerably less than its nominal (or face) value. While "good" money is money that has value which is indistinguishable between its commodity and nominal value. Gresham's Law arises when both these forms of money are in circulation and have to be accepted as legal tender with identical face values for economic transactions. Non-trivial sanctions are imposed upon persons who attempt to discriminate against bad coins in favour of good coins. This is achieved by simply requiring, through legal-tender laws, that all coins must be accepted and accorded their nominal value. A so-called "Prisoners' Dilemma" game forms, whereby only bad coins are employed for economic transactions between buyers and sellers (Selgin, 1996). Since sellers must accept either good or bad coins at their official face value, buyers offer only inferior coins for purchases. Sellers on the other hand, anticipate this behaviour by buyers and thereby price their wares accordingly (Selgin, 1996).

A practical example of Gresham's Law in practice occurred in Canada, where up until 1968 silver coins were in wide-circulation. Once the market value of silver exceeded the nominal value of the coinage, Canada debased their coinage by switching to cheaper metal alternatives. Subsequently, the silver coins disappeared from circulation as individuals chose to retain these in anticipation of realising a greater future return from the silver content of the coins. Fetter (1932: 495-6) argued that the tendency for good coins to actually leave a country was the result, not of debasement *per se*, but rather the result of an adjustment in the prices of the associated metals. Prices rise in response to an excess supply of coins. This occurs since official debasement, and even the unofficial reduction of good coins into their metal constituents, increases the nominal stock of money to be derived from any given quantity of metal (Fetter, 1932).

¹¹² Sir Thomas Gresham (c. 1519 – 1579) was an English merchant and financier who served in the court of King Edward VI, Queen Mary I and Queen Elizabeth I of England.

The expression “Gresham’s Law” was devised by Henry Dunning Macleod, a Scottish economist, who in his published work *Elements of Political Economy* decided to name the tendency for bad money to drive out good money after Sir Thomas Gresham (Macleod, 1858: 476). Sir Thomas Gresham was associated with the concept as he wrote a letter to Queen Elizabeth I on the occasion of her accession in 1558 urging Queen Elizabeth I to restore the debased currency of England. Gresham argued that bad and good coinage should not be in circulation simultaneously. He was concerned that after the debasements of English silver coins that took place during the reigns of King Henry VIII and King Edward VI, there was a discernible exodus from the country of not only the coinage with a higher silver content, but also the exit of gold coins (Burgon, 1965). Giffen (1891)’s claim that Gresham was only responsible for the suggestion that bad coinage drives good coinage of the same metal out of circulation was mistaken. Gresham’s letter to Queen Elizabeth I explicitly express Gresham’s concern that the debasement of silver coins has led to the disappearance of gold from circulation (Fetter 1932: 490-1).

At the time Macleod (1858) articulated Gresham’s Law, the notion that bad money drives out good money was not a new phenomenon and in fact it is mentioned as far back as 405 BC in a play written by Aristophanes entitled “The Frogs”. Within this play, Aristophanes describes the observation of how bad money replaced good money in respect to day-to-day transactions (Aristophanes, Barrett and Dutta, 2007). Further references to this tendency occur in medieval writings; most notably in the works of Nicole Oresme and Nicolaus Copernicus.¹¹³ In 1357, Oresme published a treatise on the ethics and economics of money production, which examined the causes and effects of inflation.¹¹⁴ The insights gleaned from Oresme’s treatise concerning inflation still have applicability in contemporary contexts. Copernicus’ contribution

¹¹³ Nicole Oresme (c. 1323 – 1382) was a theologian and philosopher who published works on various subjects including economics, mathematics, physics, astronomy, philosophy and theology. In addition, he served as a counsellor in the court of King Charles V of France. Nicolaus Copernicus (1473 – 1543) was a scientist, who most famously disputed the view that the Earth was the centre of the universe.

¹¹⁴ The treatise was published in Latin as the *Tractatus de Origine, Natura, Jure et Mutationibus Monetarum* or translated into English as the *Treatise on the Origin, Nature, Law and Alteration of Money*.

occurred after he had advised the *Prussian Diet*¹¹⁵ about monetary reform (Volckart, 1997). It was in 1526 when Copernicus published a study, which formulated an early iteration of the theory underlying Gresham's Law.¹¹⁶ He concluded that bad coinage drives good coinage out of circulation as individuals exchange bad coins for good coins, and subsequently melt down the good coins to sell the metallic constituents (Volckart, 1997).¹¹⁷

Although Gresham's Law was originally conceived within the context of bad coinage driving out good coinage of the same metal, since both forms of coinage have equal legal tender status, it is not difficult to extend the law's application to other situations. Selgin (1996) suggests that legal-tender laws may compel people to treat coins and paper notes, as equivalents, thereby forcing the more esteemed form of money out of circulation. Some writers have applied Gresham's Law more generally thereby deviating from the original premise of the law. Mundell (1998)'s assertion that the tendency of redeemable banknotes, which lacked legal tender status, to replace gold or silver coins is yet another application of Gresham's Law is in fact erroneous.¹¹⁸

Rolnick and Weber (1986) question the validity of Gresham's Law as they argue that bad money will drive good money to a premium rather than driving it out of circulation. In particular, Rolnick and Weber (1986) discuss situations involving *bimetallic* legislation, the interaction between metallic money and the introduction of

¹¹⁵ The *Diet* of Prussia or *Landtag*, took place in the Duchy of Prussia, a vassal state of Poland. The term *diet* is derived from Medieval Latin *dieta*, meaning parliamentary assembly.

¹¹⁶ The study focused upon the value of money, and was entitled *Monetae cudendae ratio*.

¹¹⁷ Copernicus also formulated a version of the quantity theory of money and his recommendations concerning monetary reform were actively implemented by the leaders of both Prussia and Poland to achieve a stable currency (Armitage, 1952 and Volckart, 1997).

¹¹⁸ Selgin (1996) points out that in so far as the redeemable notes are regarded by their holders and by others who accept them, not as "bad" money, but rather as money that was just as "good" as the coins into which they can be readily converted.

paper substitutes for gold or silver coinage.¹¹⁹ An important omission from Rolnick and Weber (1986)'s discussion is any reference to the debasement of coinage. It is precisely debasements that are discussed in all early iterations of Gresham's Law (Selgin, 1996). Moreover, Rolnick and Weber (1986) do not consider the influence of legal-tender legislation, which requires individuals to accept both good and bad money at par for economic transactions (Selgin, 1996). Although Gresham's Law is often extended to argue against bimetallism, Giffen (1891) points out that Gresham made no direct reference to bimetallism or in fact to:

“the analogous case of inconvertible paper when the paper drives the metal out of circulation” (Giffen 1891: 304).

Selgin (1996) comments that one should err on the side of caution when applying Gresham's Law outside of the law's original context. Akerlof (1970: 490) himself recognises that the analogy of bad cars (lemons) driving out good cars cannot be blanketly accepted as yet another example of Gresham's Law without certain caveats. Under Gresham's Law, both buyers and sellers can supposedly distinguish between bad and good money. Hence, the absence of information asymmetry is a serious omission for the general application of Gresham's Law within the adverse selection context of insurance.

3.1.3. Rothschild and Stiglitz (1976)'s Model

Rothschild and Stiglitz (1976) are critical of theorists who banish discussions surrounding imperfect information to footnotes. In this seminal work, Rothschild and Stiglitz (1976: 629) state the purpose of their study is to:

¹¹⁹ Bimetallism refers to a monetary standard where the country's monetary unit is defined according to either a certain quantity of gold or a certain quantity of silver. For example, for most of the 19th century in the United States, one dollar was defined as consisting either of 22.5 grains (1.4625 grams) of gold or 371 grains (24.115 grams) of silver. Individuals were permitted to exchange quantities of either metal into gold or silver dollar coins. Effectively, bimetallism creates a fixed rate of exchange for the two metals. Giffen (1892) presents a detailed argument against bimetallism and he emphasises that if market forces lead to the value of the metal exceeding its nominal currency value, then the most likely outcome would once again be disappearance of coinage from circulation.

“... [analyse] competitive markets in which the characteristics of the commodities exchanged are not fully known to at least one of the parties to the transaction”.

The authors investigate the effects of imperfect information in competitive insurance markets and their examination has proved to be a most valuable conceptual framework to assess the impact of information asymmetry within a competitive insurance context. Rothschild and Stiglitz (1976) study the response by insurers to information asymmetry and they explore whether or not equilibrium does exist. In their paper, they consider the particular situation where insurance companies do not have accurate information concerning the risk characteristics of a prospective policyholder. The key conclusion that emanates from this classical paper can be summarised as follows:

“[Rothschild and Stiglitz (1976)] are able to show that not only may a competitive equilibrium not exist, but when equilibria do exist, they may have strange properties... if individuals were willing or able to reveal their information, everybody could be made better off. By their very being, high-risk individuals cause an externality: the low-risk individuals are worse off than they would be in the absence of the high-risk individuals. However, the high-risk individuals are no better off than they would be in the absence of the low-risk individuals” (Rothschild & Stiglitz, 1976: 629).

In a perfect world, the insurer would know the probability of loss for all individuals and thus insurance markets would be characterised by symmetric information. Therefore, both the policyholder and the insurance company would have perfect information regarding the probability of suffering a loss and the potential magnitude of the loss or economic consequence of suffering a loss. Hence, insurance markets would be viable since insurers would be able to charge premiums that were commensurate with the associated risk that they face and individuals would be prepared to pay an actuarially fair premium (pure premium) for protection against the loss-causing event(s). In fact, since individuals are risk averse, individuals would be prepared to pay in excess of the pure premium to cover the administration, claims handling and loading fees of the insurer.

However, Rothschild and Stiglitz (1976) consider the case that one party to the insurance transaction, namely the policyholder, has more information regarding their probability of suffering a loss than the other party, namely the insurer. For example, in the case of genetic testing, an individual would undergo genetic screening whereby so-called genetic-markers would be detected that indicate a predisposition to certain congenital diseases. Presumably, the individual would be eager to keep this information private when applying for health insurance. This situation of asymmetric information results in a problem of adverse selection. Hoy, Orsi, Eisinger and Moatti (2003) discuss the potential impact of the diffusion of genetic testing on health insurance markets and the authors utilise the theoretical approach underlying Rothschild and Stiglitz (1976)'s model to obtain their findings. Despite the presence of information asymmetry, Hoy et al. (2003) demonstrate that a market equilibrium can be reached either on the basis of separating contracts, which discriminate between risk levels or through offering pooling contracts that are based upon the average risk of the cohort. The choice between separating and pooling contracts depends upon the relative proportion of high-risk individuals within each cohort. The authors examine specifically the hereditary risk associated with a predisposition to breast cancer. Breast cancer and the corresponding genetic research confirm that due to a limited incidence of gene mutations in the general female population, they argue that a pooling equilibrium is the likely outcome. Interestingly, Hoy et al. (2003) argue that the claim that the diffusion of genetic tests will inevitably lead to adverse selection within the health insurance market, if the utilisation of this information is prohibited, is unduly exaggerated.

A useful starting point before explaining Rothschild and Stiglitz (1976)'s model is to outline the basic framework to an example presented by the authors. Assume that there exists only two states of nature; the first being the situation where no loss event occurs (no accident state) and the second when a loss event takes place (accident state). Consider an individual who has wealth (income) denoted by W should no accident take place. This would represent the situation whereby the individual is in the no accident state. However, should an accident occur, the individual's wealth would

be given by $W - d$. Here d refers to the size of the financial loss suffered by the individual. Now suppose the individual has an opportunity to purchase insurance, which would indemnify him or her against the loss should the accident occur.¹²⁰ In the event that the person decides to purchase insurance, a premium of α_1 would be paid by the individual to the insurance company to gain protection against the loss-causing event. Hoy et al. (2003) defines this premium as $\alpha_1 = prd$, where p is the probability of an accident occurring ($(1 - p)$ represents the probability of no accident taking place), r is the amount of cover afforded by the insurance contract in the case of a loss occurring. Hence, α_1 represents the pure premium (actuarially fair premium). r is bound by the interval $[0, 1]$, if $r = 1$ then full cover is provided and thus the individual will pay a premium corresponding to full indemnification and if $r = 0$ then no insurance cover is provided, which implies that the premium charged would be zero ($\alpha_1 = 0$).

If an accident does indeed occur, the individual can expect to receive a reciprocal payment of $\hat{\alpha}_2$ from the insurer – this would represent the payment received by the insured once a claim has been successfully lodged (and approved) with the insurer. Hence, the insured would receive a payout that would represent the claim payment net of the premium paid ($\alpha_2 = \hat{\alpha}_2 - \alpha_1$). According to Hoy et al. (2003), the claim payment received by the insured would be a function of both the fractional level of cover afforded by the insurance contract (r) and the magnitude of the loss (d). Therefore, the claim payment would simply equate to the fractional level of cover afforded by the insurance contract multiplied by the magnitude of the loss ($\hat{\alpha}_2 = rd$).

The different possible wealth combinations in the no accident and accident states would equate to $(W, W - d)$ without insurance and with insurance it would be $(W - \alpha_1, W - d + \alpha_2)$. Let the individual's wealth in the no accident state be W_1 and let W_2 be the individual's wealth in the accident state – this is irrespective of whether or not the person has purchased insurance. The various wealth combinations (with and

¹²⁰ Indemnification against the insurable event may be full or partial.

without insurance) under the two possible states of nature are summarised within Table 3-2. Rothschild and Stiglitz (1976) use the vector $\alpha = (\alpha_1, \alpha_2)$ to fully describe the insurance contract.¹²¹

Table 3-2: Wealth-Combinations with and without Insurance

	Without Insurance	With Insurance
No Accident State	$W_1 = W$	$W_1 = W - \alpha_1$
Accident State	$W_2 = W - d$	$W_2 = W - d + \alpha_2$

According to the authors, these insurance contracts (α 's) are traded on the insurance market so the next step in the analysis is to present the framework that describes both the demand and supply functions of the participants in the insurance market. Assume further that there are only two types of entities in the insurance market, individuals who purchase insurance contracts and insurance companies that sell insurance contracts (Rothschild & Stiglitz, 1976).

Demand for Insurance Contracts

Individuals purchase insurance so as to alter their pattern of wealth across the two states of nature; namely, no accident and accident states (Rothschild & Stiglitz, 1976). Hence, individuals have pre-defined preferences between their wealth in the no accident and accident states. As individuals do not *per se* have control over which

¹²¹ Rothschild and Stiglitz (1976) indicate that an actual insurance contract may be far more complicated than their outlined framework dictates. An insurance contract may offer coverage against a vast array of potential losses. As the authors explain a formal generalization of the scheme can be achieved if one assumes that an individual will, in the absence of insurance, have an income of W_i if state i occurs. An insurance contract is thus an array $(\alpha_1, \dots, \alpha_n)$ whose i -th coordinate describes the net payment of the individual to the insurance company if state i occurs (Rothschild & Stiglitz, 1976). Hoy et al. (2003) attempt to extend the analysis by allowing the claim payment to reflect different coverage levels. But in fact Rothschild and Stiglitz (1976) do not exclude this possibility within the outline of their framework and in any event their graphical analysis incorporates this feature.

state of nature that they will find themselves in¹²², they have to take a decision as to the level of wealth they prefer in either state. There exists a trade-off between the wealth in the no accident and accident states, which occurs as the person decides to purchase insurance (and at what coverage level) or not. To determine the demand for insurance contracts by individuals, Rothschild and Stiglitz (1976) employ the numerical utility of the type developed by von Neumann and Morgenstern (2007).¹²³

Nash (1950) uses of the concept of *anticipation* within an example to succinctly describe the idea of expected utility. In Nash's example, an individual, Mr. Smith, anticipates an equal chance that he will receive a *Buick* or *Cadillac* motor vehicle the following day, which Nash (1950) describes as *Buick anticipation* or *Cadillac anticipation*. The outcome of which is to be decided by a simple coin toss. Nash (1950: 156) describes the individual's anticipation "...as a state of expectation which may involve the certainty of some contingencies and various probabilities of other contingencies". According to Nash (1950), the important property is that if p is the probability of receiving the *Buick* (anticipation A) and $(1 - p)$ is the reciprocal probability of receiving the *Cadillac* (anticipation B) on that following day (where $0 < p < 1$), then the combination of both individual anticipations is itself an anticipation represented by $E[A \cup B] = pA + (1 - p)B$.

The decision to purchase insurance by an individual is determined by considering the expected utility individuals derive from having insurance cover versus the expected utility they obtain by not purchasing insurance. Since one can assume that individuals attempt to maximise their expected utility, it follows that they will demand insurance only if their expected utility from purchasing an insurance contract exceeds the expected utility from not purchasing an insurance contract. The expected utility

¹²² Moral hazard is excluded from the analysis as it is a completely separate feature of insurance markets. If considered it would potentially contaminate the results of the analysis by preventing one from adequately separating the effects of adverse selection and moral hazard. In addition, one assumes that the principle of indemnification is applicable, as one does not assume individuals will have an opportunity to profit from insurance (speculate).

¹²³ Von Neumann and Morgenstern first published this applied utility framework in their book entitled the *Theory of Games and Economic Behavior* (Von Neumann & Morgenstern, 1944).

theorem states under a few mild assumptions that an individual's preferences for wealth in either state of nature can be described by a function of the form:

Equation 3-8: Individual's Expected Utility Function (Rothschild & Stiglitz, 1976)

$$\hat{V}(p, W_1, W_2) = (1 - p)U(W_1) + pU(W_2)$$

where $\hat{V}$ represents the expected utility function of an individual, p represents the probability of an accident occurring and thus $(1 - p)$ represents the probability of no accident occurring, $U(\cdot)$ represents the utility function of the individual with respect to wealth, W_1 is the individual's wealth in the no accident state and W_2 is the individual's wealth in the accident state.

When an individual purchases insurance, their expected utility function can be described by Equation 3-9 and when they choose not to purchase insurance, their expected utility function by Equation 3-10.

Equation 3-9: Individual's Expected Utility Function under Insurance (Rothschild & Stiglitz, 1976)

$$\hat{V}(p, \alpha) = \hat{V}(p, W - \alpha_1, W - d + \alpha_2)$$

where $\hat{V}$ represents the expected utility function of an individual, p represents the probability of an accident occurring, W is the individual's initial wealth, d is the loss suffered by the insured should the insured event occur, α_1 is the premium paid by the insured for insurance cover and α_2 is the payout received by the insured in the event of a claim net of the premium paid by the insured.

Equation 3-10: Individual's Expected Utility Function under No Insurance (Rothschild & Stiglitz, 1976)

$$\hat{V}(p, 0) = \hat{V}(p, W, W - d)$$

where $\hat{V}$ represents the expected utility function of an individual, p represents the probability of an accident occurring, W is the individual's initial wealth and d is the loss suffered by the insured should the insured event occur.

From all the insurance contracts that are available to purchase, an individual will select the contract (α) that maximises his or her expected utility. Note the model assumes that individuals can only purchase one insurance contract whereby the insurer specifies both the price and the quantity of insurance offered. A person does have the option not to purchase insurance. Therefore, an individual will only purchase insurance if $\hat{V}(p, \alpha) \geq \hat{V}(p, 0)$.

Rothschild and Stiglitz (1976) assume that all individuals are identical except with respect to the propensity of having an accident. In addition, individuals are assumed to be risk averse, their utility function with respect to wealth is concave and thus $\frac{\partial^2 U}{\partial W^2} = U''(W) < 0$.¹²⁴ It is mentioned by the authors that $\hat{V}(p, \alpha)$ would be a quasi-concave function in this case.¹²⁵

If one assumes that the insurance market can be segmented into high-risk and low-risk individuals. High-risk individuals have a probability of having an accident (p_H), and low-risk individuals have a probability of having an accident (p_L), where $p_H < p_L$. λ represents the proportion of the total population that is high risk and $(1 - \lambda)$

¹²⁴ A real-valued function f defined on an interval is defined as being concave if, for any two points x_1 and x_2 in its domain D and any h in $[0,1]$, the following holds:

$$f(hx_1 + (1 - h)x_2) \geq hf(x_1) + (1 - h)f(x_2).$$

The function would be strictly concave if:

$$f(hx_1 + (1 - h)x_2) > hf(x_1) + (1 - h)f(x_2) \text{ for any } h \text{ in } (0,1) \text{ and } x_1 \neq x_2.$$

¹²⁵ A function is quasi-concave if the function is concave and has either no local maxima or at most one local maximum.

represents the proportion of the total population that is low risk, where $0 < \lambda < 1$.¹²⁶ Another assumption by Rothschild and Stiglitz (1976) is that individuals know their respective probabilities of having an accident, whilst insurers do not know the associated accident probability attaching to each individual. Under these circumstances, one reverts to the traditional adverse selection setting, where if everyone only revealed their risk status to the insurer, both groups of individuals would be better off. High-risk individuals impose an externality upon the low-risk individuals and this in turn makes low-risk individuals worse off. For instance, high risks would be eager to conceal their risk status in the hope that they will attract lower premiums. Now due to the information asymmetry, insurers are unable to distinguish between the high and low-risk individuals and therefore one option would be to charge a premium based upon a weighted average probability, $\bar{p} = \lambda p_H + (1 - \lambda)p_L$, or a so-called pooling contract. But since the premium is not commensurate with each respective risk, low risks are likely to opt out of insurance as they would not be prepared to subsidise the high risks. To accommodate the two distinct risk classes, the expected utility functions described in Equation 3-9 and Equation 3-10 can be re-written as follows:

Equation 3-11: Individual's Expected Utility Function under Insurance (with two risk classes)

$$\widehat{V}_i(p_i, \alpha) = (1 - p_i)U(W - \alpha_1) + p_iU(W - d + \alpha_2), i = H, L$$

where $\widehat{V}_i$ represents the expected utility function of an individual who is either a high-individual denoted by a subscript H or a low-risk individual denoted by a subscript L , p represents the probability of having an accident for an individual who is either a high-individual denoted by a subscript H or a low-risk individual denoted by a subscript L , W is the individual's initial wealth, d is the loss suffered by the insured should the insured event occur, α_1 is the premium paid by the insured for insurance cover and α_2 is the payout received by the insured in the event of a claim net of the premium paid by the insured.

¹²⁶ It is not appropriate to allow $\lambda = 0$ or $\lambda = 1$, as this would imply that our cohort is either completely made up of only low-risk individuals or high-risk individuals respectively. Effectively, this would simply revert to the situation where one has a homogenous group with probability p of having an accident.

Equation 3-12: Individual's Expected Utility Function under No Insurance (with two risk classes)

$$\hat{V}_i(p_i, 0) = (1 - p_i)U(W) + p_iU(W - d), i = H, L$$

where $\hat{V}_i$ represents the expected utility function of an individual who is either a high-individual denoted by a subscript H or a low-risk individual denoted by a subscript L , p represents the probability of having an accident for an individual who is either a high-individual denoted by a subscript H or a low-risk individual denoted by a subscript L , W is the individual's initial wealth and d is the loss suffered by the insured should the insured event occur.

Supply of Insurance Contracts

In terms of the supply side, Rothschild and Stiglitz (1976) note that the supply for insurance contracts is less straightforward to describe than the demand side. The key question is how do insurance companies decide exactly which insurance contracts to offer to individuals. Moreover, the return of an insurance contract is in fact a random variable, as an insurer can only make an assessment as to the expected cost of claims and as such objective risk is always present.¹²⁷ The authors assume that insurance companies are risk-neutral, are only concerned with maximizing expected profits and are able (due to their financial resources) to sell any number of contracts that they think will make an expected profit. Therefore, to the insurer any insurance contract α sold to an individual would be worth the following:

Equation 3-13: Insurer's Expected Profit Function for each α sold (Rothschild & Stiglitz, 1976)

$$\pi(p, \alpha) = (1 - p)\alpha_1 - p\alpha_2$$

where π represents the expected profit function for each insurance contract α that is sold, p represents the probability of an accident occurring, α_1 is the premium received by the insurer who provides insurance coverage to the insured under the each insurance contract α and α_2 is the payout to the insured in the event of a claim net of the premium received from the insured.

¹²⁷ Objective risk refers to the difference between actual claims versus expected claims.

The insurance market is assumed to be competitive where free entry is evident. With these set of assumptions, it is apparent that any contract which is demanded and that is expected to be profitable will be supplied (Rothschild & Stiglitz, 1976).

Further Assumptions

Rothschild and Stiglitz (1976) also outline a set of assumption regarding how individuals and insurance companies determine the accident probability, which is an essential element in both the demand and supply functions, described in Equation 3-8 and Equation 3-13 respectively. A bold assumption is that individuals know precisely their accident probabilities, while insurers do not know the associated accident probability attaching to each individual. Hence, as the authors point out that since individuals are identical in all respects save their propensity to have accidents, it implies that insurance companies cannot discriminate among potential customers on the basis of their risk characteristics. Further, if one assumes that individuals can be segmented into only two risk classes; namely, high and low risks, it is important to note that insurers are believed to know the exact probability values for high and low-risk individuals. To reiterate high-risk individuals have a probability of having an accident (p_H), and low-risk individuals have a probability of having an accident (p_L), where $p_H < p_L$. Moreover, insurers even know the proportion of high-risk to low-risk within the population. Where λ represents the proportion of the total population that is high risk and $(1 - \lambda)$ represents the proportion of the total population that is low risk, and $0 < \lambda < 1$. The key element of information that insurers do not possess is the knowledge of exactly which individuals are high or low-risk. Insurers are able to generate a weighted-average probability of having an accident for the pool of individuals as a whole, $\bar{p} = \lambda p_H + (1 - \lambda)p_L$, and insurers can then calculate a premium for the pooling contract based upon the weighted-average probability of having an accident. As already noted the premium computed in this way would not be commensurate with each respective risk, and thus low risks are likely to opt out of insurance as they would not be prepared to subsidise the high risks. With the low-risk individuals leaving the insurance market, the weighted-average probability of having an accident would increase since the proportion of high-risk to low-risk individuals rises concurrently. It is therefore conceivable that this would create a self-enforcing

cycle that continually discourages lower risk individuals to maintain their insurance coverage.

The alternative is to offer a so-called separating set of contracts, one set for high-risk individuals and another for low-risk individuals. Seeing that insurers cannot distinguish between high and low-risk individuals, the idea is to offer contracts that through the selection by individuals end up revealing their true risk status. Rothschild and Stiglitz (1976) refer to this feature as the self-selection mechanism, as proposed by Salop and Salop (1976). The insurer is able to make inferences about an individual's accident probability through the individual's purchasing decisions. Holding all other factors constant, individuals with high accident probabilities will demand more insurance than those who have low accident probabilities (Rothschild & Stiglitz, 1976). Importantly, the self-selection mechanism is not a profitable method to determine an individual's risk characteristics, because insurers have to offer less than full cover to low risk individuals to ensure that these contracts do not attract high-risk buyers, thereby foregoing potential premium income. Insurers utilise the self-selection mechanism in response to the failure of the pooling contract to provide cover for both high and low-risk individuals within a stable insurance market.

Insurance companies would prefer to know their customers' individual risk characteristics to determine the appropriate terms under which they should offer insurance cover. Rothschild and Stiglitz (1976: 632) indicate that:

“Information that accrues after purchase may be used only to lock the barn after the horse has been stolen. It is often possible to force customers-to make market choices in such a way that they both reveal their characteristics and make the choices the firm would have wanted them to make had their characteristics been publicly known”.

A further assumption made by Rothschild and Stiglitz (1976) considers that individuals can only purchase one insurance contract. In effect, it implies that insurers

specify both the price and quantity of insurance to be purchased (Rothschild & Stiglitz, 1976). This is perhaps a surprising assumption given that in most competitive markets, suppliers determine only price and have no control over the quantity of good that a customer wishes to purchase. It is equivalent to a person going to a local supermarket and wishing to buy twenty pints of milk, and the shop owner insisting that they can only purchase say five pints of milk. Naturally, this is an unlikely event, with the possible exception that a shop owner may want to limit quantities that can be purchased by each customer during a sale. By limiting the quantities of an item that a customer may purchase, it effectively reduces revenue and thereby potential profits for the shop owner. Insurance is by its nature quite distinct from other traditional goods and services. Insurance is, as Arrow (1971) pointed out, an exchange of money now for money later contingent upon the occurrence of a certain event or events. Individuals do not demand insurance *per se* to meet an immediate specific need. Rothschild and Stiglitz (1976) defend their proposition of what they consider price and quantity competition in the insurance context as opposed to traditional price competition.

Equilibrium Conditions

The key feature of Rothschild and Stiglitz (1976) is to examine whether or not a stable insurance market can form in the presence of information asymmetry (or imperfect information). To test whether this is indeed the case, the authors draw upon a methodology to formally test whether or not equilibrium in a competitive insurance market exists in the presence of imperfect information. Perfect competition is an ideal construct as many industries are characterised by a limited number of companies, entry barriers and other structural impediments including information asymmetries. Martin (1993) argues that perfect competition requires complete and perfect knowledge. Scherer and Ross (1990) discuss the concept of *workable competition* as a practical standard to evaluate the structure and performance of industries.¹²⁸ Rothschild and Stiglitz (1976)'s analysis holds within the context of workable competition and the characteristics of the insurance market include free entry, a large

¹²⁸ It is accepted that workable competition exists when the structural characteristics of a market reasonably approximate the conditions for perfect competition (Scherer & Ross, 1990).

number of firms selling similar insurance contracts and no firm being able to affect the market price.

Rothschild and Stiglitz (1976) define equilibrium in a competitive insurance market, as a set of insurance contracts: such that when individuals:

“...choose a contract to maximise their expected utility the following two conditions must hold (i) no contract in the equilibrium set makes negative expected profits; and (ii) there is no contract outside the equilibrium set that, if offered, will make a nonnegative profit” (Rothschild and Stiglitz, 1976: 633).

The nature of the equilibrium proposed by Rothschild and Stiglitz (1976) is of a Cournot-Nash type, where insurers are assumed to be myopic.¹²⁹ In other words, each insurer assumes that their actions are independent of their competitor's actions (Rothschild & Stiglitz, 1976; Hoy et al., 2003). Nash (1950) outlines the classical economic problem of a two-person bargaining or exchange situation that can be applied to other contexts, such as bilateral monopoly. Nash (1950: 155) sketches the situation where “...no action taken by one of the individuals without the consent of the other can affect the well-being of the other one”. He assumes that the two individuals are rational, they can accurately compare their satisfaction from different strategic decisions and they have perfect information concerning the tastes and preferences of the other individual. After a theoretical discourse, Nash (1950)

¹²⁹ The Cournot-Nash Equilibrium is named after Antoine Augustin Cournot (1801 – 1887) and John Forbes Nash, Jr. (1928 –). Cournot was a French economist and mathematician. He was one of first individuals to successfully apply mathematics to economic questions. In 1838, Cournot proposed a version of the Nash equilibrium concept in his *Researches into the Mathematical Principles of the Theory of Wealth* (Schelling, 1960). But Nash can be ascribed with the formalisation of the properties that define an equilibrium in a bargaining context with two or more entities.

John Nash is an American mathematician and 1994 winner of the Nobel Memorial Prize in Economic Sciences. His contribution within the areas of game theory, differential geometry and partial differential equations has been significant. In particular, Nash's insights into the forces that govern randomness and the strategies that are employed within specific event spaces have drawn much acclaim. In 1950, Nash wrote a 28-page doctoral dissertation on non-cooperative games, which formed the basis for a further four articles, and outlined the definition and properties of what is now referred to as a Nash equilibrium (Nash, Kuhn & Nasar, 2002).

Cournot's version of equilibrium considered the premise that firms determine their own output decisions based upon the expected profit maximisation motive. However, a firm's optimal output decision depends upon the output decisions of competing firms (Schelling, 1960).

provides a solution to bargaining problems where “...two individuals who have the opportunity to collaborate for mutual benefit in more than one way” undertake some form of exchange (Nash, 1950: 155).¹³⁰ In a game involving two or more players, where each player is believed to know the strategies of all the other players and if no player participating in the game has anything to gain by altering their own strategy unilaterally then the current set of strategic choices constitute a Nash equilibrium. Therefore, each strategy under a Nash equilibrium represents the best response to all other possible strategies within that equilibrium (Nash, 1950; Nash, 1953; Nash, Kuhn & Nasar, 2002). These ideas underpin the Cournot-Nash equilibrium proposed by Rothschild and Stiglitz (1976).

Rothschild and Stiglitz (1976)’s first equilibrium condition stipulates that contracts in equilibrium have to generate positive expected profits, even if these profits are not actually realised. The second equilibrium condition considers insurance contracts outside of the equilibrium contract (or equilibrium set of contracts), if a contract outside of equilibrium exists that generates positive expected profits or at least breaks even, then this would represent a violation of equilibrium.

Equilibrium with Perfect Information

A useful starting point is to examine the situation where one has perfect information. This serves as a benchmark against which one can compare the situation of information symmetry against the situation where one introduces information asymmetry. Rothschild and Stiglitz (1976) initially make the bold assumption that all individuals are identical even with respect to their probability of having an accident. By doing so, insurance companies are assumed to have perfect information about their policyholders’ probability of loss. Using historical claims experience and since all policyholders are identical, insurers can easily determine the probability of having an

¹³⁰ According to Nash (1950: 155) a solution “...means a determination of the amount of satisfaction each individual should expect to get from the situation, or, rather, a determination of how much it should be worth to each of these individuals to have this opportunity to bargain”.

accident for the homogenous risk class. As the authors contend only when policyholders have different accident probabilities will imperfect information arise.

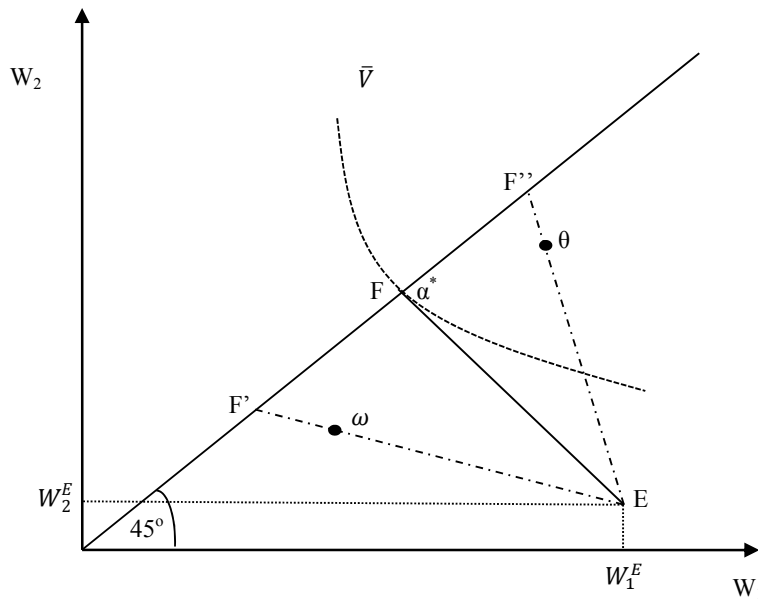
The situation with perfect information is equivalent to the case where all individuals have different accident probabilities, but they reveal their true risk to the insurer. In this case, the insurer is able to classify all individuals according to their accident probability and thereby create homogenous risk classes. Each homogenous risk class would contain members with identical accident probabilities, which reverts to the case where one is considering equilibrium for a single risk pool with members having equal accident probabilities. Essentially, the insurance market splits into several sub-markets for each different accident probability (Rothschild & Stiglitz, 1976).

Utilising Rothschild and Stiglitz (1976)'s graphical representation one is able to test whether or not a Cournot-Nash type equilibrium exists under perfect information. Figure 3-2 depicts the situation where equilibrium is considered for a homogenous risk pool with members having identical probabilities of loss (information symmetry).

Before explaining Figure 3-2, a number of points need to be noted. The horizontal and vertical axes represent states of nature defined previously – wealth in the no accident state (W_1) and wealth in the accident state (W_2) respectively. Any point within the Cartesian plane represents different combinations of wealth in the two states of nature – no accident and accident. In effect, it characterises a possible insurance contract with certain restrictions. The 45-degree line is important as it specifies the situation where the insured has equal wealth in both states of nature. It is precisely this feature of insurance that risk averse individuals seek – the principle of full indemnification. The principle of full indemnification states that after the loss-causing event has taken place, insurance coverage should place the insured in the identical financial position

that the insured was in prior to the loss.¹³¹ By implication, it also excludes the possibility of the insured profiting from insurance – for example, through speculation or over-insurance. Thus, all points along the 45-degree line represent the situation where the insured would have full insurance cover (complete coverage). Purchasing full insurance cover would place an individual on the 45-degree line thereby removing the uncertainty with respect to the pattern of wealth under either state of nature as the insured's wealth is the same irrespective of whether or not an accident (loss event) takes place.

Figure 3-2: Equilibrium with Perfect Information (Rothschild & Stiglitz, 1976)



The supply of insurance contracts is represented by the fair-odds line (EF). Insurance companies attempt to maximise Equation 3-13 – the expected profit equation. But under Rothschild and Stiglitz (1976)'s competitive market assumptions, in particular free entry, competitive forces within the insurance market ensure that policies bought in competitive equilibrium will have expected profits converging towards zero. Hence, Equation 3-13 can be re-written as follows:

¹³¹ The principle of full indemnification is a subset of the principle of indemnification. The principle of indemnification is defined as “the insured is restored to his or her approximate financial position prior to the occurrence of the loss” (Rejda, 1995: 22). With partial cover (incomplete coverage) the insured would not be fully restored to his or her financial position prior to the loss-causing event.

Equation 3-14: Insurer's Expected Profit Function under Competitive Conditions – Expected Profit converges towards zero (Rothschild & Stiglitz, 1976)

$$\pi(p, \alpha) = (1 - p)\alpha_1 - p\alpha_2 = 0$$

$$\therefore \frac{\alpha_2}{\alpha_1} = \frac{(1 - p)}{p}$$

where π represents the expected profit function for each insurance contract α that is sold, p represents the probability of an accident occurring, α_1 is the premium received by the insurer who provides insurance coverage to the insured under the each insurance contract α and α_2 is the payout to the insured in the event of a claim net of the premium received from the insured.

Equation 3-14 suggests that the ratio of net pay-out on an insurance contract to the actuarially fair premium (pure premium) equates to the ratio of the probability of no accident to the probability of an accident taking place. Therefore, it follows that the fair-odds line describes all insurance contracts that if offered to individuals with an accident probability of p would break even. Further from Equation 3-14, the slope of the fair-odds line would be given by $-\frac{(1-p)}{p}$,¹³² that is it is equal to the ratio of the probability of not having an accident to the probability of having an accident.

The graphical representation of the fair-odds line hides further insights, which Rothschild and Stiglitz (1976) fail to adequately describe. Along the line EF , insurers offer insurance contracts that reflect the actuarially-fair premium for individuals with an associated accident probability of p . Thus, the slope of the fair-odds line reflects the inherent risk of the individual purchasing the contract if it is correctly specified, that is if the associated accident probability of the individual equals the probability used to determine EF .

¹³² By definition, the negative sign attaching to the slope of EF is traditionally ignored without any loss of generality.

For example, if $p = 0.5$ then the slope of EF can be computed as $\frac{\partial(EF)}{\partial W_1} = -\frac{(1-p)}{p} = -\frac{1-0.5}{0.5} = -1$. Now if individuals became more risky, assume say that their accident probability increases from 0.5 to 0.9 ($p = 0.5 \rightarrow p^* = 0.9$). Then the slope of the new fair-odds line would equate to $\frac{\partial(EF)}{\partial W_1} = -\frac{(1-p^*)}{p^*} = -\frac{1-0.9}{0.9} = -0.111$. Hence, the slope of the fair-odds line has become less negative (in essence the slope has become larger). This change would cause the fair-odds line to become flatter relative to the horizontal axis (W_1). In Figure 3-2, the fair-odds line EF would swivel towards the horizontal axis at the point E . Therefore, the line EF' would better reflect the fair-odds line with new associated accident probability of p^* . Importantly, this highlights a key feature of the graphical representation of the fair-odds line, in that the flatter the fair-odds line the higher the risk of the individuals it is based upon. Conversely, if individuals became less risky, then the slope of the fair-odds line would become smaller (more negative), swivelling upwards at point E . This would translate into a steeper fair-odds line, say EF'' shown in Figure 3-2. Thus, a steeper fair-odds line implies a lower risk.

Now as the slope of the fair-odds line reflects the risk of those individuals whose accident probability it is derived from, it follows that the actuarially premium is also reflected by the same fair-odds line with one important caveat. In fact, the flatter the fair-odds line, the higher the associated accident probability and the higher the premium charged to the insured per unit of coverage. Equally, the steeper the fair-odds line, the lower the corresponding accident probability and the lower the premium charged per unit of coverage. Although, the premium charged under the insurance contract is reflected in the gradient of the fair-odds line, it also depends upon the degree of coverage afforded by the insurance contract. The coverage itself is intrinsically reflected in the relative position of the insurance contract to the 45-degree line and the initial uninsured state. In summary, all insurance contracts constructed along the fair-odds line EF , that is for individuals with an accident probability of p , charge individuals an actuarially fair premium based upon this precise accident probability for varying levels of insurance cover. Furthermore, these insurance

contracts are expected to break even as long as those individual purchasing the policies have the corresponding accident probability of p .

In Figure 3-2, Point E represents the initial endowment point or uninsured state. Thus, at point E the person has no insurance cover. At E individuals have wealth combination across the states of nature given by (W_1^E, W_2^E) . Notice that the wealth in the no accident is significantly larger than the wealth in the accident state. As the individual is uninsured their wealth in the no accident state would simply be equal to W , but if they had an accident their wealth is reduced by d , which is the financial loss suffered by the insured when an accident occurs. Therefore, $(W_1^E = W, W_2^E = W - d)$. Rothschild and Stiglitz (1976) goes further to explain that as an individual purchases an insurance policy $(\alpha = (\alpha_1, \alpha_2))$ along the fair-odds line EF , wealth in either state of nature changes. Purchasing insurance requires the payment of the premium α_1 , the premium payment occurs irrespective of the state of nature (no accident or accident) and thus, wealth in both states of nature would be reduced. Movement away from E and along the fair-odds line EF , reduces the individual's wealth in the no accident state $(W_1 = W - \alpha_1)$.

In the accident state, despite paying the insurance premium and suffering the financial loss due to the accident, an individual's wealth increases relative to the initial uninsured state of E . This occurs because an individual can expect to receive a claim payment $\hat{\alpha}_2$ from the insurance company once the insured-event takes place. Thus, the wealth combinations under either state of nature once insurance is purchased is given by $(W_1 = W - \alpha_1, W_2 = W - d - \alpha_1 + \hat{\alpha}_2)$. Recall the premium is a function of three variables; namely, the financial loss suffered by the insured when an accident, the accident probability and the proportion of cover afforded by the insurance policy. Therefore, $\alpha_1 = f(d, p, r)$, where $\alpha_1 = prd$. r is the proportion of cover afforded by the insurance contract in the case of a loss occurring and $0 \leq r \leq 1$ (Hoy et al., 2003). When $r = 1$ the insurance policy affords full coverage and when $r = 0$ the insurance policy affords zero coverage. For a value between zero and one, this indicates that insurance policy affords the insured partial or incomplete coverage.

The claim payment itself is a function of the level of coverage purchased under the insurance contract. If full cover is purchased the claim payout under the policy should equate to the financial loss suffered due to the accident ($\hat{\alpha}_2 = d$). Full coverage would occur where the person purchases a contract that lies at the intersection of the fair-odds line EF and the 45-degree line (at point F). Recall at this point wealth in the no accident state and wealth in the accident state would be equal – adhering to the principle of full indemnification. When a person with an accident probability p purchases an insurance contract at point F , the wealth combination under each state of nature would be equivalent ($W_1^F = W - \alpha_1, W_2^F = W - d - \alpha_1 + \hat{\alpha}_2 = W - \alpha_1$) since ($\hat{\alpha}_2 = d$). Note the fair-odds line cannot extend beyond the point F as it would violate the insurance principle of indemnification. If one extended the fair-odds line beyond F , it implies that individuals who purchase insurance contracts beyond F are able to profit from insurance.¹³³ Under these circumstances, the person's wealth in the accident state would exceed their wealth in the no accident state. Therefore, an individual's motive to prevent loss is diminished to the point where it is conceivable that they would in fact seek to cause an accident so as to profit from their insurance protection. Moral hazard would indeed be evident.

In Figure 3-2, given perfect information insurers offer a set of insurance contracts along EF , as they attempt to maximise the expected profit equation (Equation 3-13). All insurance contracts along this fair-odds line (EF) if purchased would break even. In terms of the demand for insurance contracts, as previously discussed it is apparent that an individual's decision to purchase insurance is driven by the motive to maximise their expected utility – as defined by Equation 3-8. Since the insurer determines both the price and quantity (level of coverage) the individual can purchase, the key decision left to the individual is to decide whether or not to purchase insurance. If the individual decides to purchase insurance, it is then possible to make a further decision by selecting the particular policy that maximises their expected utility out of the set of insurance contracts on offer. As already noted in Figure 3-2, the set of

¹³³ This includes all insurance contracts that lie to the left of the 45-degree line.

contracts available to a person includes an infinite set of contracts along the fair-odds line EF , which includes point E (the uninsured state – where the individual would choose not to purchase insurance) and point F (full indemnification – the individual obtains full cover).

To capture an individual's motive to maximise their expected utility, Rothschild and Stiglitz (1976) evaluate an individual's preferences between wealth in the no accident state versus wealth in the accident state. In Figure 3-2, this is captured by the indifference curve $\bar{V}$, which slope is given by the marginal rate of substitution between wealth in the no accident state and wealth in the accident state (Rothschild & Stiglitz, 1976). Thus, the slope of $\bar{V}$ equals $U'(W_1)(1 - p)/U'(W_2)(p)$. Importantly, the shape of the indifference curve $\bar{V}$ has meaning by reflecting the relative preferences between wealth in either state of nature. There is an indifference map with an infinite set of indifference curves that have the same shape as $\bar{V}$ – they are essentially “parallel” to $\bar{V}$ in either direction – each reflecting different combinations of wealth in the no accident state and wealth in the accident state.

Along each indifference curve, the utility derived from each combination of wealth in the two states of nature offer the exact same amount of utility to the individual. Each successive indifference curve that is higher than $\bar{V}$ generates ever increasing utility than the wealth combination points along $\bar{V}$. Conversely, indifference curves that are lower than $\bar{V}$, attract lower amounts of utility for the individual. Considering that all individuals are identical including with respect to their propensity to have accidents, individuals have identical preferences with respect to wealth in the no accident state and wealth in the accident state. Hence, with perfect information all individuals' indifference maps are identical.

Rothschild and Stiglitz (1976) indicate that α^* in Figure 3-2 represents the equilibrium contract under perfect information. α^* is the proposed equilibrium contract under perfect information, if one considers that a risk averse individual seeks to maximise

their expected utility and this can only be achieved by starting at point E (the initial endowment point and uninsured state) and moving along the fair-odds line towards F by purchasing ever increasing amounts of insurance cover. Each successive insurance contract that affords greater cover allows an individual to move to a higher indifference curve reflecting greater expected utility. Therefore, the insurance contract at F (α^*) allows the individual to reach their highest indifference curve given the available insurance contracts on offer (and in addition α^* is preferred to the uninsured state at point E).

Mathematically, the slope of the fair-odds line EF is equal to the ratio of the probability of not having an accident to the probability of having an accident $\left(\frac{1-p}{p}\right)$, while the slope of the indifference curve $\bar{V}$ is equal to $U'(W_1)(1-p)/U'(W_2)(p)$. But by purchasing α^* at point F , an individual obtains full cover whereby their wealth in the no accident state equals their wealth in the accident state, implying that the respective utility derived from this wealth is identical $\Rightarrow U(W_1) = U(W_2) \Rightarrow U'(W_1) = U'(W_2)$. Hence, the slope of the indifference curve $\bar{V}$ would revert to $\left(\frac{1-p}{p}\right)$, which is precisely the slope of the fair-odds line EF . By implication this can only occur where the two curves, the fair-odds line EF and the indifference curve $\bar{V}$, are tangential. This takes place at the point F , that is at the insurance contract α^* .

It is not sufficient to accept that the insurance contract α^* is indeed the equilibrium contract without adequately testing the Cournot-Nash equilibrium conditions. The first condition requires that the equilibrium contract makes non-negative expected profits. Since α^* lies on the fair-odds line EF , it is based upon the accident probability p and thus the premium charged at α^* is the actuarially fair premium for individuals with an associated accident probability p . Further, α^* offers full coverage and therefore the premium attaching to the contracts reflects the “full” premium that an individual would pay to obtain full cover where their risk is reflected in the accident probability p . Now under perfect information, all individuals have the identical accident probability of p and it is precisely an individual with the accident probability

p that would be purchasing such an insurance policy. Therefore, the insurance contract α^* , which is offered by insurers and purchased by individuals will break even. The first condition is met.

The second Cournot-Nash equilibrium condition requires that contracts outside of equilibrium do not generate non-negative expected profits. To test this condition requires slightly more thought. From Figure 3-2, consider two contracts outside of equilibrium; specifically, ω and θ . Looking at these two contracts it is possible to test whether or not these contracts would possibly disrupt our equilibrium contract α^* . In the case of ω , if individuals were to purchase this contract, insurers could expect to make positive expected profits and thus violate the second equilibrium condition. Insurers would make positive expected profits at ω if individuals purchased this contract because an individual purchasing this contract would have an accident probability p , but the policy ω would in fact be based upon a higher accident probability than p . Consequently, the computed premium rate attaching to ω would be considerably higher than the premium rate attaching to contract α^* as the fair-odds line EF' passing through ω is much flatter than the fair-odds line EF passing through α^* . Thus, the insurance contract ω reflects a higher corresponding risk than the contract α^* dictates.

The premium rate refers to the premium charged per unit of coverage. Whilst the accident probability is a key determinant in the premium calculation, the level of coverage is another important element. For example, based upon a particular accident probability, assume that an individual's annual premium rate is equal to \$0.05 per \$1 of insurance cover provided. Thus, for an insured item valued at \$10000, an individual can expect to pay \$500 per annum to have complete (full) coverage. However, if a deductible or co-insurance feature attached to the insurance policy, one would expect to pay a premium that was less than \$500 per annum to reflect the incomplete (partial) cover being provided by the insurer. In the hypothetical example, it is important to note that the premium rate was unchanged, signifying that there is no

change in the risk profile of the individual purchasing the insurance policy simply because they are moving from complete coverage to incomplete coverage.

In the case, of insurance policies ω and α^* , it is evident that ω provides incomplete cover (ω lies to the south-east of the 45-degree line) and α^* provides complete cover (α^* lies on the 45-degree line). Thus, one might expect that the premium charged at ω would be less than the premium charged at α^* . However, this inference would be misplaced, because the two contracts have very different associated risks. The respective accident probabilities corresponding to either contract are quite different and therefore, the premium rates are also dissimilar. Accordingly, it is quite conceivable that the overall premium charged at ω would be larger than the premium charged at α^* . When compared to EF , EF' reflects a higher associated accident probability than p , a higher corresponding premium rate and therefore a higher risk. An individual opting to purchase ω over α^* would effectively be over-charged by the insurer as the individual's true risk is not being reflected in fair-odds line EF' .

For the insurer, individuals purchasing ω over α^* would create an opportunity to generate positive expected profits, thereby violating the second equilibrium condition. Before arriving at that conclusion, it is necessary to ask the simple question: will individuals want to purchase ω over α^* ? The answer is categorically “no”. By opting for ω over α^* , an individual would in fact be moving to a lower indifference curve, that is lower expected utility and this would contradict an individual's motive to maximise expected utility. The reason why ω attracts a lower expected utility than at α^* is two-fold; first ω offers less than full cover and second the premium rate is significantly higher than at α^* . An alternative way of looking at this can be obtained by carefully examining Figure 3-2. Movement from ω to α^* reduces the wealth in the no accident state as the overall premium charged by the insurer is significantly larger than at α^* (reflecting the higher associated risk). Moreover, the wealth in the accident state is also reduced at ω when compared to α^* . This is due to the larger overall

premium that would still be paid at ω in the accident state and the drop in coverage experienced caused by moving from α^* to ω .¹³⁴

Now consider the second contract outside of equilibrium (θ). Individuals would prefer θ to α^* , as they would be moving to a higher indifference curve and therefore experience greater expected utility. The reason for this would be because individuals despite experiencing a drop in coverage would pay a much lower premium rate (and overall premium). The fair-odds line EF'' , which passes through θ is considerably steeper than the fair-odds line EF . Accordingly, the fair-odds line EF'' reflects a much lower risk than the fair-odds line EF , and hence a lower probability of having an accident. From Figure 3-2, one can observe that the wealth in the no accident state at θ is larger than the wealth in the no accident state at α^* , this is owing to the fact that the premium rate (and overall premium) being charged at θ is lower than the premium rate (and overall premium) at α^* . In addition, the wealth in the accident state at θ is likewise larger than the wealth in the accident state at α^* . Given the reduced coverage level at θ when compared to α^* , which would translate into a lower payment from the insurer in the event of a claim, one might expect that under θ , the wealth in the accident would be less than the wealth in the accident state at α^* . However, this is clearly not the case. The only possible explanation is given that an accident has taken place, the insured receives the lower claim payment (compared to the situation should an accident occur under α^*) but at the same time pays a much lower overall premium under θ when compared to the situation at α^* . In essence, the lower premium ‘offsets’ the lower pay-out received under the insurance policy θ in the event of an accident. Therefore, at θ an individual experiences greater quantities of wealth in both the no accident and accident states than at α^* . Undoubtedly, greater wealth in either state of nature would obviously translate into greater expected utility.

Having established that individuals would opt for θ over α^* , the key question to consider is whether or not θ will make positive expected profits or at least break even

¹³⁴ The drop in coverage experienced by the individual moving from α^* to ω would translate into a lower claim payout under the policy should an accident occur.

as this would represent a violation of the second equilibrium condition. To answer this question one must consider the situation from the insurers' perspective when individuals opt to purchase θ over α^* . By selling θ , insurers would effectively open themselves up to making negative expected profits. The fair-odds line EF'' , which passes through θ , reflects a significantly lower risk than the risk attaching to individuals that would purchase θ . As a result the insurer would suffer greater than expected claims under θ , since the underlying probability of having an accident underestimates the true accident probability of those purchasing θ .

In conclusion, if ω or θ were offered to consumers, neither insurance contract would lead to non-negative expected profits. In the case of ω , individuals would continue to prefer α^* to ω , so no one would purchase ω and accordingly, it would not break even or generate any sort of expected profit. With respect to θ , individuals would opt for θ over α^* as it provides greater 'consumption' of wealth in both states of nature. But if insurers were to market θ to individuals with accident probabilities equal to p , θ would generate negative expected profits for the insurer. As such, ω nor θ would violate the second equilibrium condition.

From the analysis of Figure 3-2, the contract α^* maximises the individual's expected utility and breaks even (Rothschild & Stiglitz, 1976). Furthermore, under perfect information, insurance contracts outside of equilibrium (selling any other contracts preferred to α^*) do not violate the second equilibrium condition. Therefore, one can be satisfied that under perfect information equilibrium does exist – the insurance market will form. The authors summarise the 'ideal' situation where one has perfect information, by stating that in equilibrium each individual who purchases insurance will do so at actuarially fair odds and obtain complete cover.

The assumption that all individuals have identical risk profiles is unrealistic since no two individuals are identical and therefore individuals will have varying probabilities of loss. Hence, it is prudent to examine the more realistic situation where individuals face varying degrees of risk. Rothschild and Stiglitz (1976) segment the insurance market into two risk classes; namely, high and low risks. Low-risk individuals have a lower probability of loss than high-risk individuals, and accordingly should be charged a lower premium rate for insurance coverage than their high-risk counterparts. High-risk individuals are assumed to have a probability of having an accident (p_H), and low-risk individuals have a probability of having an accident (p_L),

where $p_H < p_L$.

Under these circumstances, Rothschild and Stiglitz (1976) argue that only types of equilibria can potentially exist: *pooling equilibria* where both high and low-risk individuals purchase the same contract and *separating equilibria* where high and low-risk individuals purchase different contracts. The authors first consider the case for a *pooling equilibrium*.

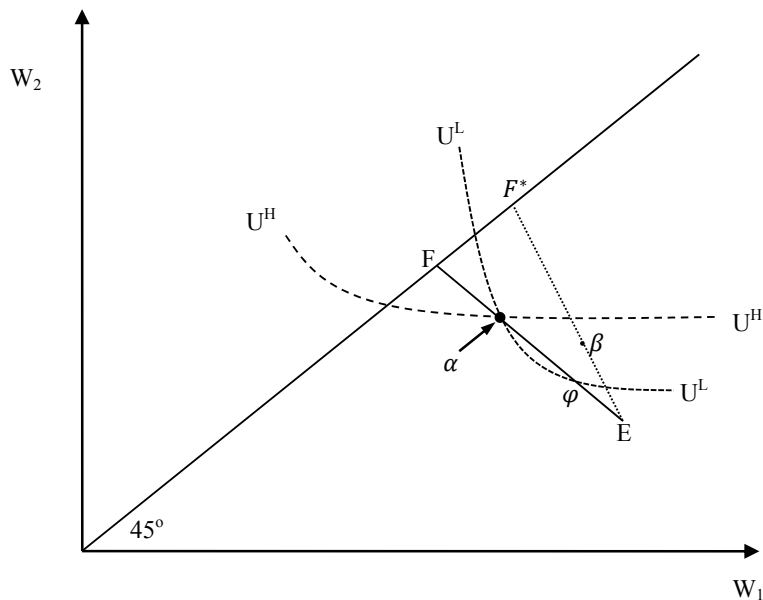
It is important to note that insurers are believed to know the exact probability values for high and low-risk individuals (Rothschild & Stiglitz, 1976). In addition, insurers are assumed to know the exact proportion of high-risk to low-risk individuals within the population. Here λ represents the proportion of the total population that is high risk and $(1 - \lambda)$ represents the proportion of the total population that is low risk, where $0 < \lambda < 1$. Absent from the insurers' information set is the fact that insurers cannot identify which exact individuals are high or low-risk. Thus, when individuals present themselves to insurance companies in order to purchase insurance, insurers cannot distinguish between individuals belonging to either of these two risk classes.

Insurers are unable to distinguish between individuals with a high probability of having an accident (high-risk) and those with a lower probability of having an accident (low-risk). At the same time, individuals are thought to know their respective probabilities of suffering a loss. Under the *pooling equilibria*, the option available to insurers is to charge policyholders an average premium for insurance cover, where the average premium is determined using the weighted-average probability of loss and the proportions of high-risk and low-risk individuals within the total population. The result is that high-risk individuals end up paying a less than actuarially fair premium rate for insurance cover whilst low-risk individuals pay well-above their respective expected claims cost for the same level of cover.

Insurers would compute the weighted-average probability of having an accident as follows: $\bar{p} = \lambda p_H + (1 - \lambda)p_L$, and subsequently insurers would utilise the weighted-average probability to determine the appropriate premium to charge the entire group. Considering that the premium calculated in this way would not be commensurate with the relative risk of either low-risk or high-risk individuals, one is likely to observe the low risks opting out of insurance as they would not be prepared to subsidise the high risks. With the low-risk individuals leaving the insurance market, the weighted-average probability of having an accident would increase since the proportion of high-risk to low-risk individuals rises concomitantly.

The option to offer a single set of contracts to both high and low-risk individuals (pooling equilibrium) is considered by analysing Figure 3-3.

Figure 3-3: Pooling Equilibrium with Imperfect Information (Rothschild & Stiglitz, 1976)

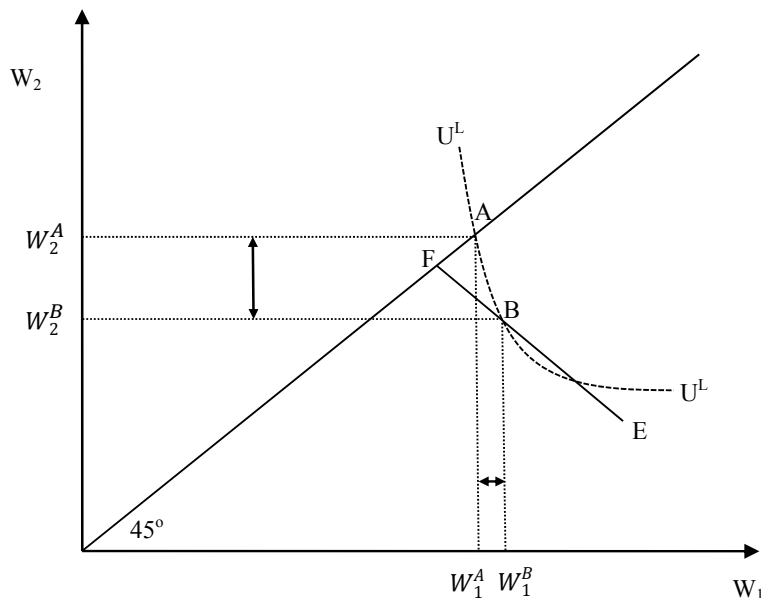


Given the fact that one now has a segmented market with both high and low-risk individuals, the key difference between Figure 3-2 and Figure 3-3 is the observation that no longer does one indifference map fully describe the preferences between wealth in the no accident state and wealth in the accident state. It is necessary to describe a separate indifference map for each market segment that is a set of indifference curves for high-risk individuals and a separate set of indifference curves for low-risk individuals. In Figure 3-3, U^H is a single indifference curve from the indifference map for high-risk individuals and U^L is a single indifference curve from the indifference map for low-risk individuals.

The shape of each indifference curve has important implications for the nature of the risk of the group of individuals that it describes. In Figure 3-4, if one considers only low-risk individuals, the indifference curve U^L demonstrates quite clearly the trade-off that exists between the wealth in the no accident state versus the wealth in the accident state. All points along the indifference curve U^L show wealth combinations in either state of nature that have exactly the same expected utility. The indifference curve U^L graphically summarises the preferences that low-risk individuals have between wealth in the no accident state and wealth in the accident state.

Movement from point A to point B in Figure 3-4 illustrates an important feature not discussed at length in Rothschild and Stiglitz (1976). A low-risk individual is willing to sacrifice a significant amount of wealth in the accident state just to obtain a small amount of wealth in the no accident state. Analytically, in Figure 3-4, $|W_2^A - W_2^B| \gg |W_1^A - W_1^B|$. This may appear intuitively obvious as low-risk individuals are more likely to be in the no-accident state of nature than in the accident state of nature because of their lower probability of having an accident (p_L), and thus low-risk individuals would prefer to have most of their wealth concentrated in the no-accident state versus the accident state.

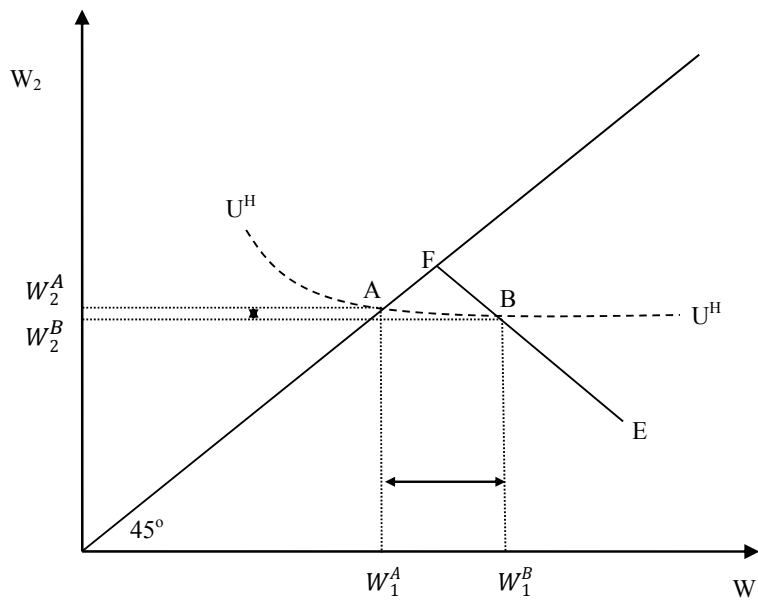
Figure 3-4: Trade-Off between Wealth in the No Accident State and Wealth in the Accident State for a Low-Risk Individual



To compare one can examine the case of high-risk individuals in Figure 3-5. Here high-risk individuals have a much flatter indifference curve (relative to the horizontal axis) than the low-risk indifference curve. Consider once again the movement from point A to point B. High-risk individuals would only be prepared to sacrifice very

little wealth in the accident state to obtain a great deal of wealth in the no accident state of nature. In this case, $|W_2^A - W_2^B| \ll |W_1^A - W_1^B|$. Intuitively, high-risk individuals are more likely to be in the accident state of nature than in the no-accident state of nature because of their higher probability of having an accident (p_H), and hence high-risk individuals would prefer to have most of their wealth concentrated in the accident state versus the no-accident state.

Figure 3-5: Trade-Off between Wealth in the No Accident State and Wealth in the Accident State for a High-Risk Individual



Emanating from this simple examination of the two distinct indifference curves for both low and high-risk individuals, one can conclude that the shape of the indifference curve captures the risk profile of individuals as they make choices between wealth in either state of nature. However, another feature inherent in the shape of the indifference curves that may be overlooked is the notion of risk aversion.¹³⁵ In essence, if the indifference curves are meant to describe an individual's preferences between wealth in the no accident and wealth in the accident, it surely also captures

¹³⁵ Risk aversion may be described as the unwillingness of an individual to take on risk when faced with uncertain outcomes. Thus, the degree to which someone is risk averse translates into the degree to which they attempt to avoid risk.

the nature of the risk aversion of the individual as well. The difficulty occurs when one attempts to separate the effects of risk *per se* and risk aversion in determining the shape of the indifference curve. Some may argue that they are intrinsically intertwined.

It is plausible to suggest that an individual who is extremely risk averse would display tendencies to protect wealth in either state of nature. Therefore, even though they might be low risk, their innate risk aversion may encourage them to trade larger amounts of wealth in the no accident state for ever smaller amounts of wealth in the accident state. Thus, the shape of the indifference curve for low-risk individuals could conceivably begin to mirror the shape of the indifference curve for high-risk individuals.

Returning to Figure 3-3, recall because the insurance market is segmented into two risk classes; namely, high and low risks, one has two separate indifference maps for each risk class. U^H is an indifference curve for high-risk individuals and U^L is an indifference curve for low-risk individuals. Furthermore, insurance companies as they are offering a single set of contracts for both risk classes, construct the pooling set of contracts based upon the weighted-average (or simply referred to as the average) probability of having an accident for the two risk classes combined. Therefore, the slope of the fair-odds line EF in Figure 3-3 is equal to $\frac{(1-\bar{p})}{\bar{p}}$, which is the ratio of the probability of not having an accident for the entire risk pool to the probability of having an accident for the entire risk pool. Note since insurers decide to offer only a single set of contracts to both market segments (pooling equilibrium), a single fair-odds line represents the insurance contracts offered to both high and low-risk individuals and is referred to as the true average fair-odds line.

Rothschild and Stiglitz (1976) argue it is simple to establish that there cannot be a pooling equilibrium where both high-risk and low-risk individuals are offered the same contract. If α is designated as the pooling equilibrium in Figure 3-3. The argument put

forward by Rothschild and Stiglitz (1976: 634 – 635) commences with the following statement:

“If $\pi(\bar{p}, \alpha) < 0$, then firms offering α lose money, contradicting the definition of equilibrium. If $\pi(\bar{p}, \alpha) > 0$, then there is a contract that offers slightly more consumption in each state of nature, which still will make a profit when all individuals buy it. All will prefer this contract α , so α cannot be an equilibrium. Thus, $\pi(\bar{p}, \alpha) = 0$, and α lies on the [fair-odds] line EF (with slope $(1 - \bar{p})/\bar{p}$)”.

Rothschild and Stiglitz (1976: 635) go further to explain that:

“...that at α the slope of the high-risk indifference curve through α , U^H , is $(p_L/1 - p_L)(1 - p_H/p_H)$ times the slope of U^L , the low-risk indifference curve through α ...The [indifference] curves intersect at α ; thus there is a contract β ...near α , which low-risk types prefer to α . The high risk prefer α to β . Since β is near α , it makes a profit when the less risky buy it, $(\pi(p_L, \beta) \approx \pi(p_L, \alpha) > \pi(\bar{p}, \alpha) = 0)$. The existence of β contradicts the second part of the definition of equilibrium; α cannot be an equilibrium”.

Rothschild and Stiglitz (1976)’s argument to explain the absence of a pooling equilibrium whilst accurate does not adequately utilise the graphical representation presented in Figure 3-3. First it is necessary to establish why α is proposed as the pooling equilibrium contract. Individuals aim to maximise their expected utility from purchasing insurance and as they are risk averse they seek protection from financial loss. In fact, both high and low-risk individuals seek full insurance cover as this would remove uncertainty from the two possible state outcomes (no accident and accident states).

Under imperfect information, it is now assumed that insurers do not have access to the exact risk status for each individual and therefore they can only determine the average premium for the entire population. Hence, insurers offer a set of contracts along the

fair-odds line EF with a slope equal to $\frac{(1-\bar{p})}{\bar{p}}$, which is based upon the average probability of having an accident $\bar{p}$. Thus, all available insurance contracts lie on the fair-odds line EF . Now it is not possible for both high and low-risk individuals to purchase an insurance contract that maximises their expected utility whilst simultaneously offering full coverage at F .

If one considers low-risk individuals, who attempt to maximise their expected utility by moving from point E (the initial endowment point – uninsured state) upwards along the fair-odds line EF , they would not maximise their expected utility at point F . By purchasing an insurance contract at point F , low-risk individuals do not reach their highest indifference curve possible – there exist other points along EF that would result in low risk reaching a higher indifference curve than the low-risk indifference curve that is tangential at point F . This can easily be observed if one considers low risk purchasing an insurance policy with less than full coverage, that is simply moving back downwards from F towards φ . As low risks select policies with ever decreasing levels of cover they would reach successively higher indifference curves until they reach point α . In fact at some point between insurance contracts α and φ , do low risks reach their highest indifference curve possible amongst the set of possible insurance contracts. It is only once a low-risk individual opts for insurance contracts below point φ (and towards point E) do they start to move onto lower indifference curves (compared to the indifference curves intersecting contracts α and φ) reflecting lower expected utility. The reason that point F does not afford the low risk individual, the highest expected utility is quite straightforward. By purchasing full cover at F , low risk individuals end up paying the ‘full’ premium (for full cover) but at a premium rate, which is significantly higher than their risk dictates. Hence, low risks suffer the greatest injustice at the point where they purchase full cover. Moreover, given that low risks would effectively subsidise the premium rate charged to high risk individuals¹³⁶ through the pooling mechanism, they prefer to opt for an insurance policy with lower levels of coverage given that they are paying an inflated premium.

¹³⁶ The premium rate charged to high-risk individuals is identical to the premium rate charged to low-risk individuals under the pooling equilibrium.

At the same time high-risk individuals would most definitely prefer to purchase an insurance contract at F out of all possible insurance contracts along the fair-odds line EF . At F , high risks receive full cover but pay only a premium that is commensurate with the average risk of the population. Hence, the average premium rate that they are charged is low relative to their inherent probability of having an accident. By purchasing an insurance policy at F out of all possible insurance contracts that are available, high risks would reach their highest indifference curve possible and therefore maximise their expected utility. Every successive cut in cover that a contract below F would afford to a high-risk individual would place a high risk individual on a lower indifference curve.

However, since only a single contract is to be offered to both high and low-risk individuals under a pooling equilibrium, a compromise contract has to be arrived at whereby high and low-risk individuals are able to reach their best possible indifference curve simultaneously. This would occur at the intersection of the indifference curves U^H and U^L in Figure 3-3. Therefore, the contract α is proposed as the compromise contract that affords both segments of the markets the highest expected utility.

To establish whether or not a pooling equilibrium exists, it is necessary to once again test if the equilibrium conditions hold. The first condition requires that the equilibrium contract makes non-negative expected profits. Since α lies on the fair-odds line EF , it is based upon the average accident probability $\bar{p}$ and thus the premium charged at α is the actuarially fair premium for individuals with an associated accident probability $\bar{p}$. α does not offer full coverage, the contract only offers partial (or incomplete) cover. α is not on the 45-degree line and from Figure 3-3 one can observe that an individual's wealth in the no accident state is larger than their wealth in the accident state.

If at the outset both high and low-risk individuals purchase α , the accident probability that reflects the combined risk classes would be the average accident probability $\bar{p}$. The true average fair-odds line EF is derived from precisely this average accident probability $\bar{p}$. Insurance companies would break even as long as the individuals purchasing contracts along the true average fair-odds line EF have the combined average accident probability $\bar{p}$. Therefore, the insurance contract α , appears to break even as both high and low-risk individuals purchase it. Initially, the first equilibrium condition holds.

The second equilibrium condition requires that contracts outside of equilibrium do not generate non-negative expected profits. To test if this condition holds, one can consider another insurance company introducing contract β to the market. Contract β represents a contract outside of equilibrium. According to the assumptions underlying Rothschild and Stiglitz (1976)'s analysis, the insurance market is considered to be competitive with free entry. Thus, the introduction of a new contract by another insurer would not be unexpected. The underlying reason why another insurer would introduce contract β to the market will become apparent momentarily.

Low-risk individuals would prefer contract β to α , as they would move to a higher indifference curve and therefore, higher expected utility. The reason why low-risks would favour β over α is because at β low-risks would be charged at a premium rate that is much lower than the premium rate at α . This can be seen in Figure 3-3 if one considers that there is a corresponding fair-odds line EF^* that passes through β . Assume that the associated accident probability used to construct the fair-odds line EF^* is given by p^* . Therefore, the slope of the fair-odds line EF^* would be given by $\frac{1-p^*}{p^*}$. However, as the fair-odds line EF^* is much steeper than the fair-odds line EF , it implies that the slope of EF^* is smaller than the slope of EF . This would only be possible if $p^* < \bar{p}$. Hence, the fair-odds line EF^* reflects a considerably lower risk than the fair-odds line EF and thus, contracts along the fair-odds line EF^* would charge a lower premium rate when compared to contracts along the fair-odds line EF .

Low-risk individuals would move to β as it charges a lower premium rate, which is more in-line with their risk status and this is despite of the cut in coverage that would be experienced by moving from α to β . Lower premiums translate into higher expected utility as more wealth would be available in both states of nature. In other words, for low risks the cut in premium more than compensates them for the cut in coverage.

High risks would remain at α as the cut in premium would not be sufficient to offset the cut in coverage experienced. For a high risk individual that is more likely to find themselves in the accident state of nature when compared to a low risk individual, it is precisely the wealth in the accident state that they are perhaps more eager to protect.

An important factor now concerns whether or not β violates the second equilibrium condition. If one considers that the accident probability attaching to β (p^*) is higher than the accident probability for a low-risk individual (p_L), that is $p^* > p_L$, then it is understandable why another insurance company would be interested in offering β . The insurance company would recognise that contract β would attract only the low-risk individuals and since the premium at β is based upon the accident probability p^* , it is essentially would be charging low-risk individuals too much relative to their risk. Hence, the insurance company would generate positive expected profits at β . A contract outside of equilibrium is making non-negative expected profits, which violates the second equilibrium condition.

In addition, as low-risk individuals move to β , the high-risk individuals that remain at α would cause the insurer to suffer negative expected profits, which would violate the first equilibrium condition – the equilibrium contract is making expected losses. This transpires because as the low-risk individuals leave α , the true accident probability should rise to reflect the higher risk of the pool of individuals that remain at α . However, α 's premium rate (and subsequent premium) is determined by the average accident probability and in this case the individuals that are purchasing α have an

accident probability $p_H > \bar{p}$. Thus, α would start to generate losses for the insurer. It can therefore be seen that the presence of contract β violates the second condition of equilibrium and further investigation reveals that the first condition of equilibrium is also violated as the proposed pooling equilibrium contract α generates losses. Therefore, a pooling equilibrium as outlined in Figure 3-3 is unattainable.

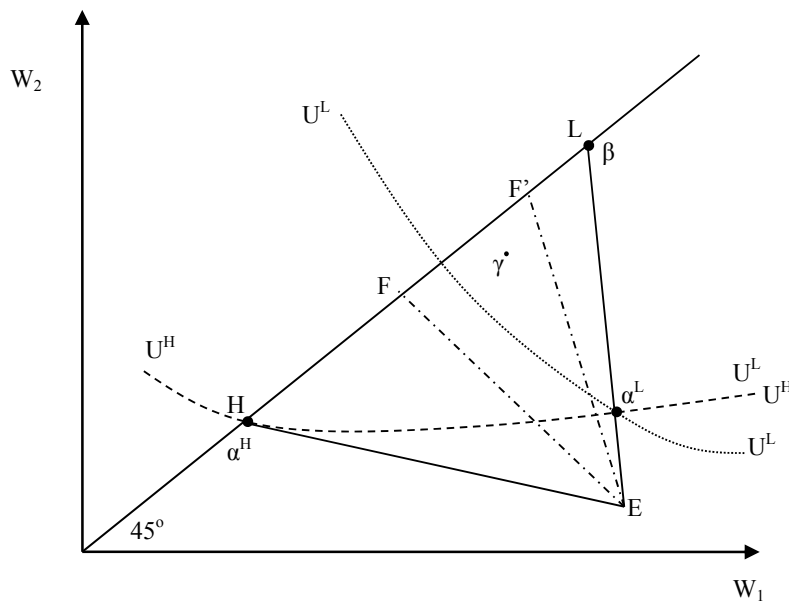
Equilibrium with Imperfect Information – Separating Equilibrium

Considering that the insurance market is characterised by two risk classes; namely, high and low risks, insurers can choose to respond by rather marketing two separate sets of insurance contracts, one set for the high-risk individuals and the other for low-risk individuals. Rothschild and Stiglitz (1976) posit that if there is to be an equilibrium, each risk type must purchase a separate insurance contract. Therefore, insurers would offer a set of contracts specifically for high-risk individuals, based upon their true probability of having an accident p_H . Thus, utilising Equation 3-14, the slope of EH would be given by $\left(\frac{1-p_H}{p_H}\right)$, which is the high-risk probability of not having accident divided by the high-risk probability of having an accident. In Figure 3-6, the fair-odds line EH describes a set of insurance contracts offering varying degrees of coverage, which will break even if and only if individuals that purchase policies along EH have accident probabilities p_H . Point E represents once again the initial endowment point with no insurance coverage and H represents full cover.

Similarly, insurers would offer a set of insurance contracts specifically for low-risk individuals, based upon the low-risk accident probability p_L (where $p_H > p_L$). Thus, correspondingly, there would be a fair-odds line EL , with a slope of $\left(\frac{1-p_L}{p_L}\right)$, that describes all possible insurance contracts based upon the accident probability p_L and as long as only low-risk individuals purchase insurance contracts along this line then insurance companies will break even.

Once again U^H represents an indifference curve from the indifference map for high-risk individuals that describes their preferences between their wealth in the no accident state (W_1) versus their wealth in the accident state (W_2). U^L is the corresponding indifference curve for low-risk individuals that describes their preferences between their wealth in the no accident state (W_1) versus their wealth in the accident state (W_2).

Figure 3-6: Separating Equilibrium with Imperfect Information (Rothschild & Stiglitz, 1976)



Ignoring all other contracts, besides those that lie on the fair odds line EH , high-risk individuals starting out at point E with no insurance would maximise their expected utility by moving towards H . Hence, at point H , α^H represents the insurance contract that maximises their expected utility and offers them full cover as it lies on the 45° line. In addition, insurance companies would break even as individuals purchasing the contract have an associated accident probability p_H , which is exactly commensurate with their risk. Likewise if one considers only contracts along the fair odds line EL , low-risk individuals starting out at point E with no insurance would maximise their expected utility by moving towards L . Hence, at point L , β represents the insurance contract that maximises their expected utility and offers them full cover as it lies on

the 45° line. Insurers selling β to low-risk individuals would break even as the contract lies on a fair-odds line that is based upon the accident probability p_L .

Therefore, at the outset one might believe that the separating equilibrium pair of insurance contracts that would be offered to high and low-risk individuals would be (α^H, β) . However, if these two contracts were marketed to both high and low-risk individuals, both groups of individuals would opt for contract β , as high risk individuals would reach a higher indifference curve at β than at α^H . The reason being they would still have full cover at β , but the premium that they would be charged would be much lower than the corresponding premium charged at α^H – this is reflected in the fact the wealth in both states at β is much greater than the wealth in both states at α^H . The premium charged at β is based upon the slope of the fair-odds line with a slope of $\left(\frac{1-p_L}{p_L}\right)$, which reflects the risk for a low-risk individual and thus insurance companies would be charging too little to the high-risk individuals. In fact, as both high and low-risks are purchasing β , the insurance company should be charging according to the average fair-odds line, with a slope of $\frac{(1-\bar{p})}{\bar{p}}$. Hence, as insurers cannot distinguish between high and low-risk individuals (information asymmetry), if they marketed both contracts to high and low-risk individuals, there would be no way to prevent high risks from purchasing β and therefore, the insurance companies would be making negative expected losses. Thus, equilibrium condition 1 would be violated, since no contract in equilibrium should make negative expected losses. Hence, the separating pair (α^H, β) cannot be the separating equilibrium.

Insurance companies need to offer an insurance contract to low-risk individuals that will not attract the high risk individuals and thus as Rothschild and Stiglitz (1976: 636) indicate:

“...[an] equilibrium contract for low-risk types must not be more attractive to high-risk types than α^H ; it must lie on the southeast side of U^H , the high-risk indifference curve through α^H ”.

Therefore, contract α^L meets that criterion, as it lies on the same high-risk indifference curve as α^H and thus, high risks are essentially “indifferent” between the insurance contract combination (α^H, α^L) . It is important to note, that the low-risk individuals now have to settle upon a contract that no longer offers them full cover even though they are still charged a premium rate that reflects their risk, α^L will have a lower premium than β simply because at α^L they are only receiving partial cover. Moreover, low-risks have had to opt for a lower-indifference curve than at β , the reason being that the reduction in premium due to partial cover is not sufficient to offset the reduction in cover that low-risks have to endure.¹³⁷ This establishes the proposed separating equilibrium pair (α^H, α^L) , whereby initially, at least, high-risk individuals would select contract α^H and low-risk would opt for α^L . The choice is indicative of Salop and Salop (1976)’s self-selection mechanism, where individuals through their choice of contract reveal the nature of their risk status.

To establish whether or not the separating pair (α^H, α^L) is in equilibrium, it is necessary to once again test if the equilibrium conditions hold. The first condition requires that both equilibrium contracts makes non-negative expected profits. Since α^H lies on the fair-odds line EH , it is based upon the accident probability for high-risk individuals p_H and the premium charged at α^H is the actuarially fair premium for individuals with an associated accident probability p_H . Thus, as only high risks are purchasing α^H , the insurance contract would break even. Low-risks would purchase α^L and as this lies on the fair-odds line EL with a slope of $\left(\frac{1-p_L}{p_L}\right)$, it is based upon the accident probability for low-risk individuals p_L and thus, the premium charged will be commensurate with that of a low-risk individual. Hence, α^L would also break-even. The first Cournot-Nash equilibrium condition holds.

The second equilibrium condition requires that contracts outside of equilibrium do not generate non-negative expected profits. To test if this condition holds, one can

¹³⁷ Note the premium rate per unit of cover is identical at both α_L and β , since both lie on the same fair odds line and therefore, both are based upon the probability p_L - the accident probability for a low-risk individual.

consider another insurance company introducing contract γ to the market as depicted in Figure 3-6. With contract γ available, high risks would prefer γ over α^H as they would reach a higher indifference curve implying greater expected utility. High risks move to a higher indifference curve at γ , because even they suffer a small decline in coverage (they would move away from the 45° line), the reduction in premium would more than compensate for the cut in cover. The slope of the fair-odds line that goes through γ is steeper than the slope of EH , which implies a lower accident probability say p_γ (where $p_\gamma < p_H$) and lower premium rate per unit of cover. Low risks would also choose γ over α^L . As they opt for γ over α^L , they would experience an increase in their insurance coverage – γ is closer to the 45° line than their original choice α^L . But they would be charged a much higher premium for two reasons. First, the increase in cover, *ceteris paribus*, automatically translates into a higher premium as they are effectively purchasing more cover. However, most importantly, the premium rate per unit of cover is higher at γ than at α^L and this is because the fair-odds line through γ is flatter than the fair-odds line through α^L , implying that the accident probability is higher ($p_\gamma > p_L$). Therefore, low-risks opt for γ over α^L , since the increase in cover more than compensates them for the increase in premium. This indicated in Figure 3-6 by the move to a higher indifference curve at γ .

Does γ violate the second equilibrium condition? To answer this question one needs to consider where the true average fair-odds line is situated. Since both high and low-risk individuals are purchasing γ , effectively you have the “average” or pooling contract, therefore, to break-even insurance companies would need to charge according to the true average accident probability $\bar{p}$. Now if the true average fair-odds line, the line based upon the true average accident probability $\bar{p}$ with a slope of $\frac{(1-\bar{p})}{\bar{p}}$, is given by EF as illustrated in Figure 3-6, that is below γ , then γ would make negative expected profits. Insurers would be charging both the high and low risk individuals (the average) too little for the risk profile they are representing. The fair-odds line through γ is steeper than EF , implying a lower accident probability and thus, a lower premium rate per unit of cover than the true average fair-odds line EF . Hence, insurers would suffer losses at γ and the second condition would not be

violated – a contract outside of equilibrium must not make non-negative expected losses. Therefore, the pair of insurance contracts (α^H, α^L) would represent a separating equilibrium.

However, if the true average fair-odds line is given by EF' as illustrated in Figure 3-6, above γ , then γ would make positive expected profits. Insurers would be charging both the high and low risk individuals (the average) too much relative to their combined risk. The fair-odds line through γ would be flatter than EF' , implying a higher accident probability and thus, a higher premium rate per unit of cover than the true average fair-odds line EF' . Hence, insurers would make profits at γ and the second condition would be violated, and (α^H, α^L) would not represent a separating equilibrium and thus, in this case the competitive insurance market will have no equilibrium, which implies no insurance market would form.

Rothschild and Stiglitz (1976: 637) state the conclusion of their analysis as follows:

“This establishes that a competitive insurance market may have no equilibrium”.

In addition, Rothschild and Stiglitz (1976: 638) go further to suggest that:

“...[one] of the interesting properties of the equilibrium is that the presence of the high-risk individuals exerts a negative externality on the low-risk individuals. The externality is completely dissipative; there are losses to the low-risk individuals, but the high-risk individuals are no better off than they would be in isolation”.

The results of this analysis are disturbing because they imply a great deal of ambiguity. Depending upon where the true average fair-odds line is situated relative to the contract γ , will determine whether or not one will have a separating equilibrium and therefore, whether or not an insurance market will form. Furthermore, the analysis suggests that the greater the proportion of low-risk individuals to high-risk individuals, in other words the smaller the value of λ within the equation $\bar{p} = \lambda p_H +$

$(1 - \lambda)p_L$, the steeper would be the true average fair-odds line and therefore, the greater the likelihood that one would not have a separating equilibrium.

Rothschild and Stiglitz (1976)'s realise the serious implications of their analysis and thus they also test the robustness of the analysis with respect to a number of the assumptions. By simply modifying the Cournot-Nash equilibrium conditions to reflect Wilson (1976)'s notion of equilibrium, the ambiguity is removed and one does indeed arrive at a separating equilibrium.¹³⁸

Community Rating and Open Enrolment

Buchmueller and DiNardo (2002) argue that despite the prevailing belief that an insurance coverage reducing death spiral is a likely consequence of community rating, it is not necessarily the outcome that models of insurance markets would predict. Consider the two-state model described by Rothschild and Stiglitz (1976), when insurers are permitted to charge risk-based premiums, in equilibrium even though high-risk individuals (old and unhealthy) pay higher premiums, both high and low-risk individuals (young and healthy) opt for full cover. Buchmueller and DiNardo (2002) indicate that the theory predicts two possible responses to the imposition of community rating.

First, the market response would be a separating equilibrium in which lower-risk individuals purchase a lower level of cover (less than full cover). In this situation, low-risk individuals are worse off, but the number of individuals purchasing insurance remains unchanged. Therefore, there is no apparent death spiral in coverage. If the lower-coverage option is unavailable or unsustainable for a separating equilibrium, a second outcome would occur where low-risk individuals would exit the market implying the occurrence of the death spiral. An example of when this might occur

¹³⁸ The Wilson (1976) equilibrium modifies the Cournot-Nash equilibrium conditions, by simply adding to the original conditions with the following proviso that if an insurance contract makes negative expected profits (losses) it is removed from the market never to return.

would be if regulations enforced a prescribed minimum benefit constraint.¹³⁹ Likewise if young and healthy individuals view the reduced-coverage option as having no value, it could induce low-risk individuals to opt out of health insurance altogether. This may occur if they believe that they have access to *free* coverage through the requirement that public hospitals provide emergency care in the absence of payment (Buchmueller & DiNardo, 2002).

Hence, Buchmueller & DiNardo (2002)'s analysis asserts that the introduction of mandatory community rating, such as that brought about with the promulgation of the Medical Schemes Act No. 131 of 1998, generate two possible outcomes that can be empirically tested:

- *No death spiral* – The number of individuals purchasing medical scheme cover does not decline and the age composition of covered individuals does not change, but health coverage is effectively reduced.¹⁴⁰ Under this scenario, medical schemes offering more comprehensive medical scheme coverage will observe younger persons opting for lower coverage plans. Low-risk individuals do not abandon coverage altogether but rather purchase (cheaper) less complete coverage (Buchmueller & DiNardo, 2002).
- *A market-wide adverse selection death spiral* – The number of individuals purchasing medical scheme cover would decline and there is an observable increase in the age of those individuals who have medical aid cover. At the same time the average benefits paid out per member would be expected to rise as the risk profile of members deteriorates.

¹³⁹ This is indeed the case with medical schemes in South Africa where they are required by law to offer a prescribed minimum benefit (PMB) package to members. Recall PMBs entail a minimum package that must be offered by all schemes where beneficiaries are covered in full for a set of specified conditions with no limits or co-payments. Medical schemes may insist on the use of a contracted network of healthcare providers and formularies of drugs to manage the healthcare of members with these conditions.

¹⁴⁰ The reduction in health coverage would be detected by a greater utilisation of medical savings accounts, where individuals bear a greater proportion of their own risk, or simply a decrease in benefits (claims) paid out to medical aid members.

Neuhaus (1995) asserts that mandatory community rating in a competitive insurance market can lead to market instability and adverse selection against insurers. Therefore, the implementation of the Medical Schemes Act No. 131 of 1998 through open enrolment and community-rating is likely to induce adverse selection, unless medical schemes are able to respond by structuring benefit options that suitably discriminate against high and low risks. Medical schemes with a mature membership profile are particularly disadvantaged under community rating, since the cost of providing health insurance increases rapidly with a person's age. Therefore, Neuhaus (1995)'s analysis does support the necessity to introduce risk equalisation transfers amongst medical schemes through the Risk Equalisation Fund (REF). Under mandatory community rating, it is therefore urgent to facilitate risk equalisation transfers across medical schemes in order to address the *cherry-picking* of lower risks through careful manipulation of marketing activities and benefit options.

3.2. Testing for the Coverage-Risk Correlation

Key to investigating the prediction of adverse selection within the medical scheme industry involves analysing the recent performance of South Africa's medical schemes and to this end, it is apparent from the second chapter that contrary to the 'doomsday' prediction, the new regulatory environment as epitomised by the Medical Schemes Act No. 131 of 1998, has not led to the contraction of the medical scheme industry. Interestingly, the industry that has benefitted the most by the current regulatory environment appears to be the suppliers of private healthcare services. In particular, private healthcare providers have managed to take advantage of the fee-for-service reimbursement system in the absence of adequate risk-sharing arrangements.

Specifically, Spence (1978) alludes to the fact that not only can individuals differ in their expected cost that they impose upon the insurer, but they may also differ in their respective preferences for certain levels of coverage. According to the author, high risk individuals are more likely to place a higher value upon insurance coverage than

low risk individuals. Therefore, as Rothschild and Stiglitz (1976)'s analysis suggests, in the presence of imperfect information, insurers can offer policies that vary in the amount of coverage and thereby, individuals can choose their level of coverage at different premium rates per unit of coverage. Spence (1978) explains that:

“Thus individuals in an insurance market can differ in expected cost or risk to the insurer, and they can differ in their valuations of coverage. These two dimensions are often correlated. For example, with attitudes toward risk held constant, high risk people will have a higher expected cost to the insurer and place a higher value on insurance coverage than low risk people. We may then see a positive correlation between benefits and costs of coverage” (Spence, 1978: 427).

Therefore, a primary indicator often utilised empirically to detect the presence of adverse selection concerns the correlation between insurance coverage and risk. Adverse selection theory predicts that members who are known to themselves (but not to their medical scheme) to be high risk will tend to select higher levels of health insurance coverage (lower deductibles). Therefore, one would anticipate a positive correlation between health insurance coverage and risk.

A positive coverage-risk correlation is not sufficient to confirm the presence of adverse selection as it is also a feature when moral hazard is present. Cohen and Siegelman (2010) emphasize that the coverage-risk correlation is not an exclusive predictor of adverse selection, but also an indicator within moral hazard theory. For example, under automobile insurance, an individual who purchases greater amounts of insurance coverage may be encouraged to be more reckless in their driving behaviour, which increases the likelihood of having an accident and claiming under the insurance policy. Hence, the insurance policy itself has altered the motive to prevent loss and thereby the probabilities (risk) upon which the insurance company has relied (Shavell, 1979).

In spite of this, Chiappori and Salanié (2000) argue that the mere existence of the positive coverage-risk relationship is a necessary condition for adverse selection to be

present. Thus, the null hypothesis is that there is a positive coverage-risk correlation present within the South African private health insurance market. The absence of a positive correlation between coverage and risk is a sufficient condition for rejecting the presence of adverse selection.

Insurance markets offer a relatively amenable setting for the empirical testing of adverse selection theory, due to the availability of adequate data (Cohen & Siegelman, 2010). In many other contexts, where adverse selection occurs data, is often secondary and unverifiable. For example, Cohen and Siegelman (2010) mention that within the employment arena, an employee may have superior knowledge about their ability than the employer. Here even where a researcher may have full access to the employer's employee performance reports, it may prove difficult to observe the quality of the employee's work because their work is often combined with that of other employees to produce outcomes (Cohen & Siegelman, 2010). In contrast, they claim that under insurance if a researcher is granted access to the insurer's records, it is possible to assess the policyholder's risk *ex ante* (when purchasing the insurance policy) and the realisation of the policyholder's risk *ex post* (once a loss-causing event occurs). The insurer may not be privy to all of the information that is available to the policyholder at the time of the loss, such as the state of mind of the insured when an accident occurred. Nevertheless, it is possible to directly link the choice of coverage level to observable outcomes, such as the claims experience of the insured (Cohen & Siegelman, 2010).

Empirical testing of the theoretical predictions under imperfect information has been carried out in various insurance contexts (Altman, Cutler & Zeckhauser, 1998; Chiappori & Salanié, 2000; Cardon & Hendel, 2001; Fang, Keane & Silverman, 2008; Finkelstein & Poterba, 2004; Puelz & Snow, 1994; Saito, 2006). The central approach to test the presence of adverse selection within these empirical studies was to examine the correlation between insurance coverage and risk (Cohen & Siegelman, 2010).

Löfgren et al. (2002) also review a number of empirical studies that have tested for the presence of a coverage-risk correlation and find that the evidence for a positive coverage-risk correlation is mixed at best. Cohen and Siegelman (2010) argue that the one should not expect the question of whether a coverage-risk relationship exists to be answered in exactly the same manner (or to apply the identical methodology even to answer the question) in all insurance segments, lines or markets. Therefore, empirical studies that arrive at opposing conclusions should not be viewed as necessarily contradictory (Cohen & Siegelman, 2010). According to the authors, empirical studies that have been informed by theoretical reasoning, provide ample reasons to expect the existence of adverse selection to vary across segments, lines and insurance markets. Factors such as the type of insurance product, individual purchaser's characteristics, institutional and regulatory factors all effect whether adverse selection is likely to be present and if present, the extent to which adverse selection is pervasive within the specified insurance segment, line or market.

3.2.1. Formulation of the Coverage-Risk Specification

As discussed previously, the key prediction of adverse selection theory is to observe a discernible positive correlation between insurance coverage and risk. Cohen and Siegelman (2010: 40) reiterate the common prediction as follows:

“...policyholders who are known to themselves (but not to their insurer) to be high risk will tend to choose higher insurance coverage (lower deductibles); thus, coverage and risk are expected to be positively correlated”.

The authors suggest that when insurers offer a menu of insurance policies, the coverage-risk correlation is expected to manifest itself in the tendency of high risk individuals opting for policies with more comprehensive cover (lower deductibles). If, however, insurers offer only a single policy, then the coverage-risk correlation would suggest that high risks are more likely to purchase insurance than low risks (Cohen & Siegelman, 2010). Furthermore, they emphasise that:

“...the prediction of a coverage–risk correlation implies that coverage will be correlated with risk, controlling for all relevant policyholder characteristics that are observable to the insurer” (Cohen & Siegelman, 2010: 44).

According to Chiappori, Jullien, B. Salanié and F. Salanié (2006), the coverage-risk correlation is a fairly robust conjecture within the context of adverse selection and it does appear in a wide variety of insurance lines and markets. Therefore, throughout the many empirical studies that investigate the presence of adverse selection, the coverage–risk correlation has been the primary estimation mechanism within many of the studies’ empirical specification.

If one has access to insurers’ individual policyholder data, the conventional way to test for the coverage-risk correlation is to run a regression with the following specification:

Equation 3-15: Regression Specification for Testing the Relationship between Risk and Coverage (Cohen & Siegelman, 2010)

$$r_i = \alpha + \beta c_i + \gamma_{i,k} X_i + \varepsilon_i$$

where r_i is a variable representing the *ex post* realisation of the i^{th} policyholder’s risk, c_i represents the i^{th} policyholder’s choice of coverage, and X_i represents a $k \times 1$ column vector of k characteristics for the i^{th} policyholder that are known to the insurer and potentially relevant for classifying the i^{th} policyholder’s risk. ε_i represents the residual.

Equation 3-15 can be extended to include a measure for expected coverage as Dionne, Gouriéroux and Vanasse (2001) recommend. The inclusion of expected coverage can address the problems of non-linearities or misspecifications. It is possible to obtain expected coverage by estimating the choice of coverage equation. In Equation 3-15, the left-hand side of the equation could be a continuous variable, such as the total cost of claims (benefits paid out); a dichotomous variable, such as the number of claims; or simply a dummy variable indicating whether or not a claim was submitted (Cohen & Siegelman, 2010).

An alternative approach suggested by Chiappori and Salanié (1997, 2000) is the bivariate model described in Equation 3-16 and Equation 3-17, where the two equations are estimated either simultaneously or sequentially depending upon the estimated correlation between the residuals from each equation (Cohen & Siegelman, 2010). If the correlation is statistically significantly different from zero at a predetermined confidence level, then the two regressions should be estimated simultaneously. Otherwise, the two regression equations from the bivariate model should be estimated sequentially. If a positive correlation is found between the two residuals, this would be consistent with a positive coverage-risk correlation (Cohen & Siegelman, 2010).

Equation 3-16: Coverage Specification under Chiappori and Salanié (1997, 2000)'s Bivariate Regression Model (Cohen & Siegelman, 2010)

$$c_i = f(X_i) + \varepsilon_i$$

where c_i represents the i^{th} policyholder's choice of coverage, and X_i represents a $k \times 1$ column vector of k characteristics for the i^{th} policyholder that are known to the insurer and potentially relevant for classifying the i^{th} policyholder's risk. ε_i represents the coverage specification residual.

Equation 3-17: Risk Specification under Chiappori and Salanié (1997, 2000)'s Bivariate Regression Model (Cohen & Siegelman, 2010)

$$r_i = g(X_i) + \eta_i$$

where r_i is a variable representing the *ex post* realisation of the i^{th} policyholder's risk, and X_i represents a $k \times 1$ column vector of k characteristics for the i^{th} policyholder that are known to the insurer and potentially relevant for classifying the i^{th} policyholder's risk. η_i represents the risk specification residual.

A further non-parametric test that Chiappori and Salanié (2000) suggest may overcome the relatively restrictive functional forms described in Equation 3-15, Equation 3-16 and Equation 3-17, is the non-parametric procedure based upon the χ^2 tests for independence. Cohen and Siegelman (2010) explain Chiappori and Salanié (2000)'s methodology, whereby they create 2^m cells from m exogenous dummy

variables. Under each cell, a 2×2 matrix is generated from two dummy variables. The first being *coverage*, which is equal to 1 if high and 0 otherwise. The second being *risk*, which is equal to 1 if the policyholder has had at least one loss-causing event and 0 otherwise. The authors then test for independence of coverage from risk, conditional upon being in a particular cell (Cohen & Siegelman, 2010).

A limitation of the empirical investigation undertaken by this thesis is that only aggregated medical scheme data is available. Thus, it is not possible to utilise Chiappori and Salanié (1997, 2000)'s methodology. Equation 3-15, Equation 3-16 and Equation 3-17 are all specifications that are applied to individual data in order to test for the coverage-risk relationship. Although, Cohen and Siegelman (2010) criticise Dahlby (1983) and Dahlby (1992), for her utilisation of aggregated data to test for the coverage-risk correlation, it can be argued that without any loss of generality, it is still valid to employ aggregated data to ascertain the presence (or lack) of a coverage-risk correlation within insurance markets.

Yes, it would certainly be beneficial to have full access to the individual data of all medical schemes within South Africa. In particular, data relating to the claims incurred (benefits paid out), coverage levels and contributions at the medical scheme member level. However, most importantly, the current regulatory framework that applies community-rating and open enrolment, immediately excludes any possibility of including characteristics that would be used to assess the risk of a medical scheme member or their beneficiaries. Therefore from Equation 3-15, Equation 3-16 and Equation 3-17 perspective, X_i , which represents a $k \times 1$ column vector of k characteristics for the i^{th} policyholder that are known to the insurer and are potentially relevant for classifying the i^{th} policyholder's risk become untenable for this study. In South Africa's case, after the year 2000, the only two relevant classification variables would essentially revert to income and the number of dependents.

Moreover, medical schemes have been vehemently against disclosing any of their individual member data for fear of potential exploitation by other competing medical schemes. It is only since the proposed implementation of the Risk Equalisation Fund (REF) that individual medical scheme beneficiary data has been shared by medical schemes with the Council for Medical Schemes. Since the enabling legislation for the REF has not been enacted by parliament yet, the Council for Medical Schemes is still in the process of assessing the latest data that they have received from medical schemes. In any event, the Council for Medical Schemes has no intention of releasing the individual medical scheme data into the public domain and therefore, it is not possible at this stage to obtain individual medical scheme data within South Africa. Nevertheless, a test for the coverage-risk correlation is indeed possible within South Africa's private health insurance sector; it merely requires some innovative reformulation of the empirical specification. Refer to Chapter Four for a detailed discussion relating to the empirical investigation undertaken by this thesis.

3.2.2. Contradictory Evidence

Despite the acceptance of the theoretical findings emanating from the seminal literature by Akerlof (1970), Spence (1973, 1974) and Rothschild and Stiglitz (1976) during the 1970s, it is only until the 1980s did the empirical testing of these theoretical findings get underway. Since then numerous empirical studies have tested for the presence of adverse selection in a variety of insurance segments, lines and markets. These studies include Puelz and Snow (1994), Altman et al. (1998), Chiappori and Salanié (2000), Cardon and Hendel (2001), Fang, Keane and Silverman (2008), Finkelstein and Poterba, (2004), and Saito (2006). Since the empirical studies have provided contradictory evidence as to the presence of the coverage-risk correlation, it is useful to examine a selection of these findings.

Cohen and Siegelman (2010) note that in the automobile insurance market, three early studies suggested the existence of a positive coverage–risk correlation (Dahlby, 1983; Dahlby, 1992; Puelz & Snow, 1994), but these findings were challenged by subsequent research. Dahlby (1983) examined the impact of prohibiting statistical discrimination in the Canadian automobile insurance market. By simulating the impact of prohibiting statistical discrimination, in this case with respect to the gender, She finds that not only does society end up purchasing less insurance, but society pays more for insurance as well (Dahlby, 1983). Furthermore, Dahlby (1983, 1992) finds that females are not prepared to subsidise males (insurers can no longer charge males and females differential premiums *ceteris paribus*) and therefore, a significant portion of females are predicted to leave the voluntary collision insurance market. Dahlby (1983, 1992)’s empirical analysis suffered from aggregated data with respect to coverage (Cohen & Siegelman, 2010). Puelz and Snow (1994) utilised individual data and established the existence of adverse selection within the automobile insurance market. Dionne, Gouriéroux, and Vanasse (2001) criticised Puelz and Snow (1994) for failing to take nonlinear effects into account. Once nonlinear effects were considered, Dionne et al. (2001) found that there was no residual adverse selection in each risk class (Cohen & Siegelman, 2010).¹⁴¹

Chiappori and Salanié (2000)’s study of the French automobile insurance market failed to detect a positive correlation between coverage and risk. The study focused upon a relatively homogeneous group of novice drivers. The drivers within the study had between one and three years driving experience. In France, automobile insurance is sold in two broad categories:

1. a mandatory contract covering only third-party liability; and
2. a broader optional contract that also covers first-party losses to the policyholder and the policyholder’s vehicle, even if the policyholder is at fault.

¹⁴¹ Dionne et al. (2001)’s utilised a different dataset in their empirical analysis to the data utilised in Puelz and Snow (1994)’s study.

Although optional contracts come in a variety of deductible “flavours”, the authors managed to simplify the analysis by looking only at the choice between the minimum mandatory coverage and any type of expanded coverage (Cohen & Siegelman, 2010).

Most of the data utilised by the insurer to set premiums was available to undertake the empirical analysis. In addition, the choice of contract by each policyholder and the nature of the accident(s) for which the policyholder claimed coverage were also available. Cohen and Siegelman (2010) highlight the several variants of the coverage–risk correlation estimation procedure that was undertaken by Chiappori and Salanié (2000), which included:

1. estimating two independent *probit* specifications for the type of contract purchased and the probability of an accident, and then testing whether the residuals are correlated;
2. estimating the two *probit* specifications as a bivariate *probit* and directly testing whether the estimated correlation parameter is zero; and
3. utilising a variety of discrete non-parametric methods, based upon testing whether, conditional upon the values of the key explanatory variables, loss probability and type of coverage are independent of each other in contingency tables.

Chiappori and Salanié (2000)’s results indicate that individuals that are more likely to submit claims (higher probability of having accidents) do not purchase more insurance (Cohen & Siegelman, 2010).

Richaudeau (1999) also examined the French automobile insurance market where the focus was upon the choice between basic third-party coverage and comprehensive insurance (Cohen & Siegelman, 2010).¹⁴² The data included in Richaudeau (1999)’s

¹⁴² Comprehensive insurance includes cover for third-party liability plus damage to the policyholder’s own vehicle in at-fault situations.

study, consisted of detailed risk-classification variables¹⁴³ for individual policyholders and their motor vehicles. There was also information concerning the total number of reported (and unreported) accidents for each policyholder. A probit regression was estimated to test for a positive correlation concerning the decision to purchase basic third-party coverage versus comprehensive insurance (Richaudeau, 1999). After controlling for the independent variables that explain the decision choice, the error term would be indicative of the individual's riskiness. The subsequent step included using the risk measure as an explanatory variable in a specification seeking to explain the number of accidents in which the policyholder was involved in during a given year (Cohen & Siegelman, 2010). Richaudeau (1999)'s results suggest that individuals who drive regularly are more likely to purchase comprehensive insurance even though they are not essentially higher risk. Even though this does not adhere to the traditional adverse selection definition, it should not be misinterpreted as insurance-induced risky behaviour - in other words the presence of moral hazard (Cohen & Siegelman, 2010).

Cohen (2005) tests the theoretical predictions of adverse selection models using data from the Israeli automobile insurance market. In contrast with Chiappori and Salanié (2000)'s study, the author finds evidence that is consistent with the presence of adverse selection within this market. Utilising a similar methodology to Chiappori and Salanié (2000), the author finds no correlation between coverage and risk for novice drivers (those with fewer than three years' driving experience) but does find a positive and statistically significant correlation for drivers with more than three years' experience. The reasoning suggested is that as individuals gain driving experience they acquire knowledge of their risk type, which induces a positive coverage-risk correlation (Cohen & Siegelman, 2010).

New policyholders with driving experience also tend to possess superior knowledge over the insurer. Among these individuals, the average number of claims was 36

¹⁴³ The number of miles driven per annum is an example of a risk-related variable that was included in the analysis. Refer to Richaudeau (1999) for a detailed list of the risk-related variables included in the study.

percent higher for those who select a low deductible as opposed to those that select the mean deductible (Cohen, 2005). It is argued that individuals that switch to a new insurer are disproportionately people with a poor claims history and they are inclined to under-report their past claims history when joining a new insurer (Cohen, 2005). This may suggest that new policyholders actively seek to select against their new insurer by pooling with policyholders that had better claims histories (Cohen & Siegelman, 2010). Cohen (2005)'s empirical results with respect to more experienced drivers do not suggest that one should expect to detect a positive coverage-risk correlation within the French automobile insurance market as studied by Chiappori and Salanié (2000). The author argues that the reason stems from the fact that French insurers share information about policyholders' risks, while this is not the case in Israel.

Another study by Saito (2006), this time analysing the Japanese automobile insurance market, concluded that there was only a weak and statistically insignificant positive relationship between the purchase of *own-vehicle* coverage and accident risk for both novice and experienced drivers. Utilising the bivariate *probit* estimation methodology as outlined by Chiappori and Salanié (2000), he uncovers a negative and statistically significant relationship between risk and the purchase of a zero-deductible policy (Cohen & Siegelman, 2010). Moreover, the author finds that drivers in high-risk prefectures in Japan were not more likely to opt for zero-deductible or *own-vehicle* policies compared to those drivers in low-risk prefectures (Cohen & Siegelman, 2010).

Saito (2006) does identify an important caveat in relation to the findings of his empirical study. Although, he affirms that the robustness of his results have been carefully scrutinised, the empirical methodology is unable to address the situation where an individual has an accident, but simply does not claim under the insurance policy. The so-called "accident $\neq$ claim" phenomenon is likely to distort the accurate assessment of an individual's risk. For example, an individual who may have a number of accidents and yet does not claim under their insurance policy, would

essentially appear to be a low-risk individual despite their obvious inherent riskiness. Therefore, one may be misled by the inaccurate assessment of risk that leads to a negative relationship between risk and the purchase of a zero-deductible policy. The author stresses that the “accident $\neq$ claim” issue cannot simply be disregarded, since data from the Japanese police department imply that this problem is not trivial. He highlights that the statistics surrounding the theft of motor vehicles in Japan indicate that, over four prefectures, the average percentage of cars stolen equated to 0.10 percent during 1999 (11,411 cars were stolen out of a total of 11,880,389 cars), whereas the insurance data from Saito (2006)’s study demonstrated a much higher theft rate amongst insured drivers (0.22 percent or 67 cars were stolen out of a total of 30,000 private-use cars). Hence, this implies that insured drivers are more likely to have their cars stolen, which is consistent with a positive coverage-risk correlation and thus, the predictions of adverse selection theory.

Adverse Selection in the Health Insurance Market

Adverse selection in the health insurance market has undergone extensive empirical investigation. Cutler and Zeckhauser (2000) reviews fourteen separate studies that analyse the tendency of higher risk individuals to select more generous benefit plans within the health insurance context. The empirical results from these studies all suggest the presence of some form of adverse selection (Cohen & Siegelman, 2010). However, Browne and Doerphinghaus (1993), Buchmueller, Couffinahl, Grignon and Perronnin (2004), and Ettner (1997) do not find evidence of a positive correlation between coverage and risk.

In their analysis of the premium differential across health-insurance plans; Altman, Cutler and Zeckhauser (1998) find that adverse selection occurs within the health insurance sector as a result of low-risk individuals exiting from generous plans rather than high-risk individuals opting into the more generous plans. Altman et al. (1998) also coin an interesting term, namely, *adverse retention*:

“Adverse retention is the tendency for people who stay put to magnify cost differentials between plans, as they will if they differ in age and costs are more than linear with age” (Altman et al., 1998: 122).

In Cutler and Zeckhauser (1998)’s examination of data from the Massachusetts Group Insurance Commission, they provide a compelling argument of *intramarket* selection. The authors examined the differences amongst employees, in particular, those employees that select more generous fee-for-service plans versus those that elect to enrollee in an HMO. They find that employees opting for the fee-for-service arrangements were more likely to suffer a medical event, such as having a heart-attack or a stroke, than their HMO counterparts.

Cutler and Reber (1998) investigate the private employee health insurance provided by Harvard University. Harvard University elected to alter the structure of their healthcare plans; this was achieved by moving away from subsidising only their most generous plans and towards a fixed-dollar subsidy irrespective of the generosity of the plan selected by employees (Cohen and Siegelman, 2010). This translated into an annual increase of approximately \$500 to the cost of the most generous plans. Cohen and Siegelman (2010) suggest the positive coverage-risk correlation hypothesis was dramatically borne out. Cutler and Reber (1998) found that lower-risk employees abandoned the most generous plan. This was most evident in employees that left the more generous plans in favour of opting rather for the HMO option, as they were on average 4 to 5 years younger and had lower healthcare expenses than those who remained (Cohen and Siegelman, 2010).

According to Saito (2006), the empirical findings against a coverage-risk correlation may be specific to the automobile insurance market. Indeed, he mentions that the empirical findings differ across various studies undertaken in other insurance markets. For example, Ausubel (1999) found evidence of the presence of adverse selection in the credit card market¹⁴⁴, Finkelstein and Poterba (2004) in the annuity market, and

¹⁴⁴ Ausubel (1999: 1) examine “...the results of large-scale randomized trials in preapproved credit card solicitations for direct evidence of adverse selection”.

Barrett and Conlon (2003) in the health insurance market. Cawley and Philipson (1999) findings coincide with Saito (2006)'s results, where they found evidence supporting the absence of adverse selection in the life insurance market.

Another important aspect that may influence whether or not adverse selection is present within insurance markets relates to the effect of rate regulation. Harrington and Danzon (2000) demonstrate that rate suppression in workers' compensation insurance results in an increase in the frequency and / or severity of employee injuries claims (Saito, 2006). As discussed previously, Buchmueller and DiNardo (2002) argue that community-rating does not induce an adverse selection *death spiral* in the health insurance market. Therefore, according to Saito (2006) any investigation that attempts to test for the presence of a positive coverage-risk correlation should be undertaken with due regard for the similarities, nuances and differences across insurance lines and markets.

3.3. Conclusion

Developments within South Africa's medical schemes market do provide a unique opportunity to test the underlying implications of Rothschild and Stiglitz (1976)'s analysis whereby in a competitive market, the consequences of pooling heterogeneous risks classes can be explored. Community rating and open enrolment provide an opportunity to test the theoretical predictions emanating from the seminal literature by Akerlof (1970), Rothschild and Stiglitz (1976), Spence (1973, 1974) and Wilson (1977), whereby information asymmetry can lead to adverse selection. The seminal literature predicts that where risk classification is restricted and price controls are enforced, adverse selection would cause the market to contract and potentially collapse (Rothschild & Stiglitz, 1976).

A suitable procedure to detect whether adverse selection is present within the South African private health insurance sector is to determine if there is a positive coverage-risk correlation. The existence of such a correlation is a necessary condition to detect the presence of adverse selection.

4 EMPIRICAL RESEARCH METHODOLOGY AND EMPIRICAL RESULTS

The chapter discusses the empirical methodology adopted by this thesis to test for the presence of adverse selection in the South African private health insurance sector. To determine whether a positive coverage-risk correlation does exist, a panel data estimation procedure is undertaken. In this chapter, the method of study and empirical findings are reported. Also, several reasons why one may find the absence of adverse selection in insurance markets are outlined.

After the seminal work by Akerlof (1970), Spence (1973, 1974), and Rothschild and Stiglitz (1976), considerable research has been devoted to creating models to test for the theoretical predictions emanating from adverse selection theory (Cohen, 2005). Rothschild and Stiglitz (1976)'s analysis as discussed in Chapter 3, demonstrate that in competitive insurance markets adverse selection may lead to the failure of insurance pools to form, as was the case in the United States where insurers did not offer insurance cover in so-called *red-lined* areas.

It was in fact the perceived presence of adverse selection within the South African private health insurance sector after the deregulation that took place during the late 1980s and early 1990s that was identified by the South African government as the key determinant in reformulating the regulation governing medical scheme business in the country (Reekie, 1999). Prior to the current regulatory environment, Doherty and McLeod (2003) argued that the application of the concept of risk-rating had a negative impact upon medical scheme coverage. According to Doherty and McLeod (2003), high risk individuals, such as the elderly and unhealthy, had their monthly contributions loaded as they represented a greater risk to the medical scheme. Moreover, medical schemes imposed life-long exclusions for pre-existing conditions and in certain cases, the elderly and/or unhealthy were simply excluded from coverage altogether.

Therefore, government believed that the deregulation of medical schemes had led to a deterioration in medical scheme coverage for precisely those individuals who required medical scheme cover the most – the elderly and unhealthy. As a consequence, the first-democratically elected government sought to correct, what in its point of view, was a manifestation of the negative effects of the adverse selection phenomenon within the South African private health insurance sector (Reekie, 1999). The reformulation of the regulation culminated in the implementation of the Medical Schemes Act No. 131 of 1998 and the accompanying regulations. The stated reforms, more especially open enrolment and community rating, were touted by the government as necessary to address the undesirable effects of adverse selection. As discussed previously, despite government's misinterpretation of the concept of adverse selection, the question was never raised whether or not in fact adverse selection is indeed a feature of the South African private health insurance market.

To apply the theoretical literature of Rothschild and Stiglitz (1976) in constructing a suitable empirical specification to test for the presence of adverse selection, it is first necessary to establish that South Africa's private health insurance sector is indeed competitive. This issue was addressed in Chapter 2 (sub-section 2.2.5) whereby the Herfindahl-Hirschman Index (HHI) was computed over the period 1993 to 2009. The results suggest that the medical schemes sector is sufficiently competitive.

To empirically test for the presence of adverse selection within the South African private health insurance market, the specifications described in Chapter Three; Equation 3-15, Equation 3-16 and Equation 3-17, may all be desirable methodologies. However, all of these specifications represent methodologies that can be applied only to individual insured data in order to test for the positive coverage-risk correlation. But as the medical schemes data available for this study is aggregated, the empirical methodology undertaken by this study will draw upon the approach outlined by Dahlby (1983; 1992).

4.1. Data and Panel Structural Design

To test for the presence of adverse selection, it is necessary to detect a positive coverage-risk correlation (Chiappori & Salanié, 2000). While not a sufficient condition to confirm the presence of adverse selection, it is nevertheless a necessary condition. Hence, the absence of a positive coverage-risk correlation is adequate to confirm the absence of adverse selection. Alternatively, it is possible to detect the presence of adverse selection by employing Akerlof (1970)'s *Lemons Principle* that implies that as if low-risks are driven out of the insurance market by the high risks, one would expect to observe a deterioration in the risk profile of those that remain behind in the insured pool. It is precisely this observation that will direct the empirical specification of this study.

Data

The data utilised for the empirical analysis was collected from the Council for Medical Schemes' Annual Reports over the period 1993 to 2010. Due to the reporting cycle of the Council for Medical Schemes, the sample covers the period 1993 until 2009 (inclusive). Each Annual Report published by the Council for Medical Schemes provides annual data concerning the operations of registered medical schemes within South Africa over a particular year. The data is disaggregated by registered medical scheme, which includes both open and restricted medical schemes. Some of the variables published in these reports include": the total number of medical scheme beneficiaries, gross contribution income (risk and savings-related contributions), gross administration expenses, gross relevant healthcare expenditure (healthcare benefits paid to healthcare providers on behalf of medical scheme beneficiaries), year-end reserve position and the solvency ratio of for each individual medical scheme.

Exempt (or Bargaining Council) medical schemes were excluded from the dataset, as the reporting of these schemes by the Council for Medical Schemes has been

inconsistent. In fact, the Council for Medical Schemes ceased reporting the detailed membership figures for these schemes from 2005 onwards. By the time the Council for Medical Schemes ceased reporting detailed membership data in relation to the exempt or Bargaining Council medical schemes, they only constituted just over 4 percent of all medical scheme members. Therefore, they represent only a small fraction of the medical scheme population. In addition, since exempt or Bargaining Council schemes are exempt from certain provisions under the prior and current legislation governing medical scheme business, it is not appropriate to include them in the analysis as one would ostensibly be comparing “apples” with “pears”.

With respect to the data contained within the Annual Reports published by the Council for Medical Schemes, it is important to note that there were changes in the registration status of certain medical schemes over the sample period – from exempt or Bargaining medical scheme to registered medical scheme. Therefore, medical schemes that gained registered status were included as part of the panel from the date (year) their registration was accepted by the Council for Medical Schemes. For example, medical schemes such as Transnet and Polmed that were previously classified as ‘exempt’, were given registered status from the year 2000 onwards. The inclusion of these relatively large medical schemes coincided with the effective implementation date of the Medical Schemes Act No. 131 of 1998 (McLeod & Ramjee, 2007). Therefore, it is important to acknowledge the once-off rise in membership figures that occur in 2000 due to this reclassification.

In each year of the sample period, new medical schemes were registered, some were liquidated and yet others amalgamated with one another. Thus, it was necessary to correct the panel for possible ‘double-counting’. Over the sample period there were a total of 234 individual medical schemes that were considered ($N = 234$), but it must be noted that not every medical scheme had the full complement of observations for each variable over the entire seventeen-year sample period (1993-2009) ($T = 17$). In some cases, this was due to the entry, exit and consolidation of medical schemes; but in others it was simply a case of the failure by the medical scheme to report the

necessary data to the Council for Medical Schemes. Therefore, for estimation purposes the panel dataset is unbalanced.

The following variables were utilised to construct the variables required for the panel estimation specification:

- **ID**: represents a unique identifier for each medical scheme, which is required to adequately capture the cross-sectional individual-specific (medical scheme) aspects within the empirical estimation.
- **South African Consumer Price Index (CPIX) (excluding interest rates on mortgage bonds - CPIX)**: can be used to derive a measure of the annual consumer inflation rate. The CPIX was utilised to convert all the monetary series from nominal into real terms (relative to 2009 prices).
- **Total number of medical scheme beneficiaries (BENEF)**: represent the total number of registered medical scheme members and their dependents. This variable was utilised as a scale parameter, by which certain variables could be transformed to reflect the values per medical scheme beneficiary.
- **Total number of medical scheme members (MEM)**: represent the total number of registered medical scheme members.
- **Net Annual Contribution Income (or Gross Annual Risk Contribution Income) (NCONT)**: refers to gross annual contribution income less the contributions made towards medical savings accounts during that corresponding year. Gross annual contribution income simply refers to the contributions (premiums) paid annually by medical scheme members to registered medical schemes in return for medical scheme cover. These contributions include the non-risk contributions (savings contributions) paid into medical savings accounts on behalf of the medical scheme member. Since the savings contributions represent contributions made by medical scheme members to pay out-of-pocket healthcare expenses, they essentially do not represent a benefit that the medical scheme beneficiary receives as part of the “insurance” portion of their medical scheme cover.
- **Net annual claims incurred (NCLMS)**: Net annual claims incurred are the annual claims paid out to healthcare providers for services rendered to the medical scheme member or their dependents. Net annual claims incurred do not include

out-of-pocket expenses paid to healthcare providers from members' medical savings accounts.

- **Gross annual administration expenses (ADMIN):** refer to the annual administration expenses of operating the medical scheme, but they do not include broker fees, managed healthcare expenditure or the net income / (expense) on commercial reinsurance.
- **Solvency ratio (SOLVR):** refers to the solvency ratio of a medical scheme. Regulation 29 of the Medical Schemes Act No. 131 of 1998 prescribes the minimum accumulated funds that need to be maintained by medical schemes (Council for Medical Schemes, 2010). Accumulated funds refer to the net asset value¹⁴⁵ of the medical scheme excluding *inter alia* funds set aside for specific purposes and unrealised non-distributable profits (Council for Medical Schemes, 2010). Accumulated funds expressed as percentage of gross contributions, or solvency ratio, must be maintained at a minimum level of 25 percent.¹⁴⁶ (Council for Medical Schemes, 2010). The solvency ratio provides an indication of the financial soundness and sustainability of a medical scheme (Council for Medical Schemes, 2010).

To obtain the specific variables utilised in the empirical specification as described in Equation 4-8 and Equation 4-9, a number of adjustments were necessary. First, the variables *NCLMS*, *NCONT* and *ADMIN* series were deflated using the CPIX variable, so that the variables were inflation-adjusted, and thus, reflected real values as opposed to nominal values – relative to 2009 prices. To prevent an inherent bias appearing within the empirical model, specifically from the relationship between net contribution income and the total number of members, that purely stems from the fact that as more people join a medical scheme (an increase in the total number of medical scheme members) it will automatically translate into an increase in the value of net contribution income (new member's contributions would augment current net contribution income) Therefore, it was necessary to modify the inflation-adjusted

¹⁴⁵ Net assets refer to total assets of the medical scheme less total liabilities of the medical scheme.

¹⁴⁶ The minimum accumulated funds required to ensure that the solvency ratio is kept at the required level of 25 percent, is more commonly referred to as the *reserves* of the medical scheme (Council for Medical Schemes, 2010).

variables ($NCLMS^*$, $NCONT^*$ and $ADMIN^*$), so that they reflected the average value per medical scheme beneficiary per month. Therefore, the transformations are described in Equation 4-1, Equation 4-2 and Equation 4-3.

Equation 4-1: Bias Adjustment to Inflation-Adjusted Net Annual Claims Incurred

$$NCLMSA_{it} = \frac{NCLMS^*_{it}}{BENEF_{it}} \times 1/12 \quad \text{for } i = 1, \dots, N \text{ and } t = 1, \dots, T$$

where $NCLMSA_{it}$ refers to the net claims incurred (in real terms), per average medical scheme beneficiary per month (pabpm), for medical scheme i and observed at time t (1993 – 2009). $NCLMS^*_{it}$ refers to the net annual claims incurred for medical scheme i observed at time t (in real terms – 2009 prices) and $BENEF_{it}$ denotes the total number of medical scheme beneficiaries in medical scheme i observed at time t .

Equation 4-2: Bias Adjustment to Inflation-Adjusted Net Annual Contribution Income

$$NCONTA_{it} = \frac{NCONT^*_{it}}{BENEF_{it}} \times 1/12 \quad \text{for } i = 1, \dots, N \text{ and } t = 1, \dots, T$$

where $NCONTA_{it}$ refers to the net contribution income (in real terms), per average medical scheme beneficiary per month (pabpm), for medical scheme i and observed at time t (1993 – 2009). $NCONT^*_{it}$ refers to the net annual contribution income of medical scheme i observed at time t (in real terms – 2009 prices) and $BENEF_{it}$ denotes the total number of medical scheme beneficiaries in medical scheme i observed at time t .

Equation 4-3: Bias Adjustment to Inflation-Adjusted Gross Annual Administration Expenses

$$ADMINA_{it} = \frac{ADMIN^*_{it}}{BENEF_{it}} \times 1/12 \quad \text{for } i = 1, \dots, N \text{ and } t = 1, \dots, T$$

where $ADMINA_{it}$ refers to the gross administration expenses (in real terms), per average medical scheme beneficiary per month (pabpm), for medical scheme i and observed at time t (1993 – 2009). $ADMIN^*_{it}$ refers to the gross annual administration expenses of medical scheme i observed at time t (in real terms – 2009 prices) and $BENEF_{it}$ denotes the total number of medical scheme beneficiaries in medical scheme i observed at time t .

Given the three distinct regulatory regimes governing medical scheme business in South Africa it is appropriate to consider the inclusion of two separate structural breaks within the random-effects empirical specification (Equation 4-9). The first structural break (D_1) is defined to occur in 1994, to distinguish between the pre-deregulation regime, where medical schemes did not apply appropriate insurance and risk management principles, and thus, suffered consecutive underwriting losses and displayed poor solvency positions; and the Reekie (1999) inspired deregulation of medical scheme business covering the period from 1994 until 2000. Therefore, the second structural break (D_2) is defined to occur in 2000, so as to capture the effect of the implementation of the Medical Schemes Act No. 131 of 1998.¹⁴⁷ Both structural breaks are believed to be fundamental, mutually exclusive, changes in the structure of the medical schemes industry. Thus, for the first structural break, D_1 is assumed to be zero prior to 1994, one from 1994 until 1999, and then reverts back to zero from 2000 onwards. The second structural break (D_2) is defined to be zero prior to 2000 and then one from 2000 onwards.

Panel Data Methodology

Cameron and Trivedi (2009) refer to panel data as repeated measurements at different points in time on the identical individual unit, such as an individual or medical scheme. Panel data regressions attempt to capture both the variation over units, as in the case of conventional regressions examining cross-sectional data, and variation over time (Cameron & Trivedi, 2009). The standard errors of panel data estimators need to be adjusted since each time period of data is not independent with respect to the other time periods in the panel.

Hsiao (2003) identifies a key benefit in employing a panel data estimation procedure; namely, it controls for individual (or unit) *heterogeneity*. By construction, a panel data

¹⁴⁷ The Medical Schemes Act No. 131 of 1998 was only came into full force after January 2000.

specification implies that units (medical schemes) are heterogeneous (Baltagi, 2008). The standard specification for a panel data regression can be described as follows:

Equation 4-4: Standard Specification under a Panel Data Estimation Framework (Baltagi, 2008)

$$y_{it} = \alpha + X'_{it}\beta + u_{it} \quad \text{for } i = 1, \dots, N \text{ and } t = 1, \dots, T$$

where i refers to medical schemes and t denotes time. Therefore, the subscript i denotes the cross-sectional dimension and t denotes time-series dimension. α is a scalar or intercept term, β denotes a $K \times 1$ vector of *slope* coefficients and X_{it} is the i^{th} observation of K explanatory variables. u_{it} denotes the disturbance term (Baltagi, 2008).

The most common panel data specifications use a one-way error component specification model for the disturbance term u_{it} and for this study it will be no different. Equation 4-5 describes the one-way error component specification for the disturbance term u_{it} .

Equation 4-5: One-Way Error Component Model for the Disturbance Term (Baltagi, 2008)

$$u_{it} = \mu_i + v_{it} \quad \text{for } i = 1, \dots, N \text{ and } t = 1, \dots, T$$

where μ_i describes the unobservable individual-specific (medical scheme) effect and v_{it} denotes the remainder disturbance term (Baltagi, 2008).

The distinction between fixed-effects (FE) and random-effects (RE) panel data models is often misleading because under both types of models, the μ_i term that describes the unobservable individual-specific effects is viewed to be stochastic in nature (Cameron & Trivedi, 2009). However, under fixed-effects models the μ_i are assumed to be fixed parameters to be estimated whereas the remainder disturbances v_{it} are independent and identically distributed ($v_{it} \sim IID(0, \sigma_v^2)$) (Baltagi, 2008). Under FE models, the μ_i 's are permitted to be correlated with the regressors X_{it} , which allows a limited form of endogeneity (Cameron & Trivedi, 2009). Moreover, the X_{it} are assumed to be independent of the $v_{it} \forall i$ and t (Baltagi, 2008). Baltagi (2008) argues that fixed-effects panel data models are therefore; more suitable when one is examining a specific set of individuals or firms, which is indeed the case in this

study where one is investigating all registered medical schemes in South Africa over the period 1993 and 2009.

Under the FE framework, when N is large, there tends to be too many parameters to estimate and as a consequence one loses a significant number of degrees of freedom (Baltagi, 2008). If one is able to assume that μ_i is random, then the loss of degrees of freedom can be avoided. Under the random-effects panel data model framework, it is assumed that $\mu_i \sim IID(0, \sigma_\mu^2)$, $v_{it} \sim IID(0, \sigma_v^2)$ and the μ_i are independent of the v_{it} . Furthermore, X_{it} are assumed to be independent of both the μ_i and $v_{it} \forall i$ and t (Baltagi, 2008). He suggests that the RE model is an appropriate specification when one is drawing N individuals (medical schemes) randomly from a large population. An example of this would be when one is conducting household panel studies, where N is typically large and therefore and a FE model would lead to a significant loss of degrees of freedom (Baltagi, 2008).

The rationale behind a RE model is that, unlike the FE model, the variation across medical schemes is assumed to be random and uncorrelated with the explanatory variables (or regressors) included within the model. Greene (2008) refers to this as:

“...the crucial distinction between fixed and random effects is whether the unobserved individual effect embodies elements that are correlated with the regressors in the model, not whether these effects are stochastic or not” Greene (2008: 183).

Therefore, if one believes that the differences across medical schemes have some influence upon the net claims incurred (pabpm) then a RE would be more appropriate.

Testing for unit roots in conventional time-series studies is now a relatively straightforward exercise, but testing for stationarity in panel data models does require additional considerations (Baltagi, 2008). Harris and Tzavalis (1998, 1999), Maddala and Wu (1999), Hadri (2000), Choi (2001), Levin, Lin and Chu (2002), Im, Pesaran and Shin (2003) and Breitung and Das (2008) have all made significant contributions in outlining appropriate methodologies to test for unit roots within panel data. Many of the tests for unit roots require a strongly balanced panel. Maddala and Wu (1999) and Choi (2001) propose an alternative Fisher-type test that does not require the panel dataset to be balanced, which precisely describes the nature of the panel dataset employed by this study.

Levin et al. (2002) argue that individual unit root tests have limited power against alternative hypotheses and this is particularly the case in small samples (Baltagi, 2008). Therefore, as cited by Baltagi (2008), Levin et al. (2002) propose that a more powerful unit root, where the null hypothesis for each individual time series within the panel dataset contains a unit root versus the alternative hypothesis that each time series is stationary. The suggested starting point for testing the hypothesis is given by the specification described in Equation 4-6, from which Augmented-Dickey-Fuller (ADF) tests can be undertaken:

Equation 4-6: Levin et al. (2002)'s Initial Specification to Test for Unit Roots in Panel Data Studies (Baltagi, 2008)

$$\Delta y_{it} = \rho y_{i,t-1} + \sum_{L=1}^{\rho_i} \theta_{iL} \Delta y_{it-L} + \alpha_{mi} d_{mt} + \varepsilon_{it} \quad \text{for } m = 1, 2, 3.$$

where y_{it} describes the designated data series that one is examining for the presence of a unit root. ρ_i describes the lag order under conventional ADF regressions (Baltagi, 2008). d_{mt} is the vector of deterministic variables and α_{mi} is the corresponding vector of coefficients for model $m = 1, 2, 3$. The models are defined by $d_{1t} = \{\emptyset\}$, $d_{2t} = \{1\}$ and $d_{3t} = \{1, t\}$ (Baltagi, 2008).

Since the lag order of ρ_i is unknown, Levin et al. (2002) propose a three-step procedure to test their hypothesis (Baltagi, 2008). Maddala and Wu (1999), and Choi (2001) propose an alternative Fisher-type test for unit roots. Following Baltagi (2008)'s exposition, let G_{iT_i} be a unit root test statistic for the i^{th} group as defined within Equation 4-6, and assume that as the time series observations for the i^{th} group tend towards infinity ($T_i \rightarrow \infty$), then $G_{iT_i} \Rightarrow G_i$, where G_i is a non-degenerate random variable (Baltagi, 2008).¹⁴⁸

Baltagi (2008) then define p_i to be the asymptotic probability value of a unit root test for the cross-section i , that is $p_i = F(G_{iT_i})$, where $F(\cdot)$ is the distribution function of the random variable G_i . From these assumptions, Maddala and Wu (1999) and Choi (2001) construct their Fisher-type test, which is given by the following specification:

Equation 4-7: Fisher's Inverse Chi-Squared Test Statistic to Test for Unit Roots in Panel Data (Maddala & Wu, 1999; Choi, 2001; Baltagi, 2008)

$$P = -2 \sum_{i=1}^N \ln(p_i) \quad \text{for } i = 1, \dots, N$$

where P is Fisher's inverse chi-square test statistic for unit roots and p_i is the asymptotic probability value of a unit root test for the cross-section i (Baltagi, 2008).

Equation 4-7 combines the p -values from all unit root tests of each cross-section i to obtain the test for the unit root of the panel data. Baltagi (2008) note that $-2\ln(p_i) \sim \chi_{(2)}^2$, therefore as $T_i \rightarrow \infty$ for finite N , it implies that $P \sim \chi_{(2N)}^2$. Besides having the advantage of applying the Fisher test to unbalanced panel datasets, the Fisher test can utilise different lag lengths for in the individual ADF regressions

¹⁴⁸ A non-degenerate random variable is a random variable that cannot be constant.

(Baltagi, 2008). A disadvantage of the Fisher test is that it does require Monte Carlo simulations to derive the corresponding p -values.¹⁴⁹

Hausman Test

From the initial observations outlining the FE and RE panel data models, it is suggestive that a FE model is an appropriate specification for this thesis' medical scheme panel dataset. The reason for this conjecture is that this study is considering a predefined set of medical schemes (all registered medical schemes in South Africa over the period 1993 and 2009) rather than a random sample from a large population of medical schemes. Nevertheless, there is a formal test that can be applied; namely, the Hausman test. The test is based upon Hausman (1978) and looks at the differences between the FE and RE estimators (Baltagi, 2008).

Baltagi (2008) stresses that the interpretation of the Hausman test is far too often elegantly characterised by the rejection of the null hypothesis implying an acceptance of the FE specification over the RE specification. However, the author argues that one needs to also consider the testable restrictions implied by the FE specification as outlined by Chamberlain (1984) before one can accept the outcome of the Hausman test.

Panel Structural Design

To test for the presence of adverse selection within the South African medical scheme industry over the period 1993 until 2009, it will be necessary to employ Akerlof (1970)'s *Lemons Principle* that implies that if low-risks are driven out of the insurance market by the high risks, one would expect to observe a deterioration in the risk profile of those that remain behind in the insured pool. This is similar to the

¹⁴⁹ Another limitation of the Fisher-type tests as proposed by Maddala and Wu (1999) and Choi (2001), is that these tests are designed under the assumption that the individual time series within the panel are cross-sectionally independently distributed.

mechanism employed by Dahlby (1983) to detect the presence of adverse selection within the Canadian automobile insurance market.

Therefore, if adverse selection is indeed present within the medical scheme industry, a decrease in the total number of medical scheme members one would expect a corresponding increase in the net claims incurred per average medical scheme beneficiary per month (effectively the benefits paid out or the coverage provided by the medical scheme) holding all other factors constant.¹⁵⁰ Hence, this empirical study proposes regressing net claims incurred (in real terms) per average medical scheme beneficiary per month (*NCLMSA*) on the total number of medical scheme members (*MEM*), net contribution income (in real terms) per average medical scheme beneficiary per month (*NCONTA*), gross administration expenses (in real terms) per average medical scheme beneficiary per month (*ADMINA*) and the solvency ratio (*SOLVR*). Both a fixed-effects model and random-effects model has been estimated. Even though, the Hausman test suggests that FE model is superior to the RE model, the RE model is included as it captures the two structural breaks that were necessary to highlight the impact of the three distinct regulatory regimes. Before the estimation models were estimated, Fisher-type tests as proposed by Choi (2001) were run to test for unit roots (test for the stationarity) of the series included in the empirical specification.

The empirical specifications for both the fixed-effect model and random-effects models are described by Equation 4-8 and Equation 4-9 respectively.

¹⁵⁰ Note the total number of medical scheme members is proposed as a dependent variable in the empirical specification, since medical scheme members are effectively the decision-makers when it comes to electing to exit the medical schemes market or not.

Equation 4-8: Fixed-Effects Panel Data Specification to test for a Positive Coverage-Risk Correlation

$$NCLMSA_{it} = \alpha + \beta_1 MEM_{it} + \beta_2 NCONTA_{it} + \beta_3 ADMINA_{it} + \beta_4 SOLVR_{it} + u_{it}$$

where $NCLMSA_{it}$ refers to the net claims incurred (in real terms), per average medical scheme beneficiary per month (pabpm) for medical scheme i and observed at time t (1993 – 2009). MEM_{it} denotes the total number of medical scheme members in medical scheme i observed at time t ; $NCONTA_{it}$ refers to the net contribution income (in real terms) per average medical scheme beneficiary per month (pabpm) for medical scheme i and observed at time t ; $ADMINA_{it}$ refers to the gross administration expenses (in real terms) per average medical scheme beneficiary per month (pabpm) for medical scheme i and observed at time t ; and $SOLVR_{it}$ denotes the solvency ratio of medical scheme i and observed at time t . u_{it} denotes the disturbance term.

Equation 4-9: Random-Effects Panel Data Specification to test for a Positive Coverage-Risk Correlation

$$NCLMSA_{it} = \alpha + \beta_1 MEM_{it} + \beta_2 NCONTA_{it} + \beta_3 ADMINA_{it} + \beta_4 SOLVR_{it} + \beta_5 D_1 + \beta_6 D_2 + u_{it}$$

where $NCLMSA_{it}$ refers to the net claims incurred (in real terms), per average medical scheme beneficiary per month (pabpm) for medical scheme i and observed at time t (1993 – 2009). MEM_{it} denotes the total number of medical scheme members in medical scheme i observed at time t ; $NCONTA_{it}$ refers to the net contribution income (in real terms) per average medical scheme beneficiary per month (pabpm) for medical scheme i and observed at time t ; $ADMINA_{it}$ refers to the gross administration expenses (in real terms) per average medical scheme beneficiary per month (pabpm) for medical scheme i and observed at time t ; $SOLVR_{it}$ denotes the solvency ratio of medical scheme i and observed at time t ; D_1 denotes the first structural break that demarcates the pre-deregulation and post-deregulation periods (at time $t = 1994$); and D_2 denotes the second structural break that demarcates the post-deregulation period and the post-implementation of the Medical Schemes Act No. 131 of 1998 (at time $t = 2000$). u_{it} denotes the disturbance term.

In both, Equation 4-8 and Equation 4-9, the primary coefficient of interest is β_1 as it is the indicator of the presence of adverse selection. If $\beta_1 < 0$ and it is statistically significantly different from zero at the 95 percent confidence level, the implication is

that an decrease in the total number of medical scheme members leads to an expected β_1 unit increase in the net claims incurred per average medical scheme beneficiary per month (effectively the benefits paid out or the coverage provided by the medical scheme) holding all other factors constant. Thus, confirming the presence of adverse selection.

4.2. Empirical Findings

Before undertaking the estimation of Equation 4-8 and Equation 4-9, it is necessary to test the panel for stationarity. Although, panel methods to test for unit roots and cointegration are based upon the time series derivatives of these tests, they nevertheless do require additional considerations as outlined previously (Cameron & Trivedi, 2009). In this study, Maddala and Wu (1999) and Choi (2001)'s Fisher-type test (based upon Augmented-Dickey-Fuller tests) is employed to test for unit roots in a panel data context.¹⁵¹ One is testing the null hypothesis that all panels within the dataset contain unit roots (that is they are non-stationary) against the alternative hypothesis that at least one panel is stationary. The results are presented in Table 4-1.

¹⁵¹ Refer to Equation 4-8 and Equation 4-9 for an outline of the specification, and refer to Maddala and Wu (1999) and Choi (2001) for a detailed exposition.

Table 4-1: Fisher-Type Unit-Root Test Results (based upon Augmented Dickey-Fuller Tests)¹⁵²

Number of Panels	234				
Asymptotics	$T_i \rightarrow \infty$				
ADF Regressions	1 Lag				
AR Parameter	Panel-Specific				
H_0 : All Panels contain Unit Roots					
H_1 : At least one Panel is Stationary					
	<i>NCLMSA</i>	<i>MEM</i>	<i>NCONTA</i>	<i>ADMINA</i>	<i>SOLVR</i>
Inverse Chi-Squared Test Statistic (P)	384.4074*	700.2675**	282.0400	506.1093**	857.8220**
p -value	0.0414	0.0000	0.9880	0.0000	0.0000
Degrees of Freedom	338	340	338	340	338

Source: Output generated within Stata™. * indicates statistical significance at the 95 percent confidence level (or 5 percent significance level), whilst ** indicates statistical significance at the 99 percent confidence level (or 1 percent significance level).

The results from Table 4-1 suggest that the null hypothesis that all the panels contain unit roots for each of the series (with the exception of (*NCONTA*)) is rejected at the 95 percent confidence level (*NCLMSA*) and in certain cases even at the 99 percent confidence level (*MEM*, *ADMINA*, *SOLVR*). This implies that these series are all indeed stationary (*NCLMSA*, *MEM*, *ADMINA*, *SOLVR*). Despite adjusting for consumer inflation, the results in Table 4-1 suggest that net contribution income per average medical scheme beneficiary per month (*NCONTA*) do appear to have a unit root for each of the 234 corresponding panels (the null hypothesis cannot be rejected), which is indicative of nonstationarity. This may be indicative of the fact that net contribution income per average medical scheme beneficiary per month has consistently outpaced consumer inflation and in addition, annual increases in contribution income have far-exceeded those increases that would be suggested by the corresponding increases in net claims incurred. Once a time trend term is included in the Fisher-type test for the *NCONTA* series, the Inverse Chi-Squared Test Statistic (P) is equal to 409.26 with a corresponding p -value of 0.0047. Thus, the hypothesis of all panels containing unit roots would be rejected, implying stationarity.

¹⁵² Panel means were included, but a time trend and drift term were excluded from the estimation.

Before progressing with the fixed-effects estimation procedure it is useful to examine the descriptive statistics of the relevant variables, which are presented in Table 4-2. The variables take on values within the expected ranges, except perhaps in the case of the solvency ratio. Due to a few outliers, for example, a solvency ratio of 665.44, which implies that the medical scheme has a solvency ratio in excess of 66500 percent was not erroneous. For certain small medical schemes, their accumulated funds expressed as a percentage of gross contribution income, may indeed “exaggerate” the solvency position of the scheme.

Table 4-2: Summary of the Descriptive Statistics for Panel Data Variables

Variable	Obs	Mean	Std. Dev.	Min	Max
id	2282	1355.585	420.7717	1001	11581
year	2282	2001.671	4.86638	1993	2009
name	0				
mem	2282	18217.13	53858.32	2	917580
benef	2282	45684.02	129944.2	7	2041908
solvrr	2243	1.466461	17.95076	-1.27	665.44
nconta	2278	707.1358	338.2774	.6635127	5967.875
nclmsa	2268	643.8344	325.7581	.7880425	3880.09
admina	2280	63.68664	120.3532	0	4328.769
d1	2282	.3540754	.478337	0	1
d2	2282	.5692375	.4952915	0	1

Source: Output generated within Stata™.

Now under a panel dataset, it is important to recognise that dependent and explanatory variables may potentially vary over both time and cross-sectional units (medical schemes). Variation over time or a particular medical scheme is referred to as *within* variation, whilst variation across medical schemes is referred to as *between* variation. A one-way FE model will only have explanatory power if each medical scheme’s net claims incurred (pabpm) above or below the medical scheme’s mean value for net claims incurred (pabpm) is significantly correlated with the medical scheme’s explanatory variables values above or below the medical scheme’s vector of mean values for the explanatory variables. It is precisely for this reason, that it is termed the *within estimator*, since it depends upon the variation *within the unit* (medical scheme). If a coefficient of a regressor with little *within* variation will be imprecisely estimated

and if at the extreme, there no within variation it will not be identified at all. Therefore, any characteristic that does not vary over time for each medical scheme should not be included in the model. Before progressing with the fixed-effects and random-effects estimation, it is necessary to examine the relative importance of *within* variation and *between* variation of a variable. These results are presented in Table 4-3 for the relevant variables utilised in the estimation models.

Table 4-3: Summary of *within* and *between* variation for a Selection of the Panel Data Variables utilised in both the FE and RE Estimation Models¹⁵³

Variable		Mean	Std. Dev.	Min	Max	Observations
nclmsa	overall	643.8344	325.7581	.7880425	3880.09	N = 2268
	between		246.907	42.01705	2157.237	n = 234
	within		218.3928	-799.0138	3898.539	T-bar = 9.69231
mem	overall	18217.13	53858.32	2	917580	N = 2282
	between		38728.12	16.5	442242.8	n = 234
	within		28872.78	-422802.7	493554.3	T-bar = 9.75214
nconta	overall	707.1358	338.2774	.6635127	5967.875	N = 2278
	between		241.7246	54.61786	1412.789	n = 234
	within		243.3142	-500.964	5650.381	T-bar = 9.73504
admina	overall	63.68664	120.3532	0	4328.769	N = 2280
	between		60.92193	0	684.3386	n = 234
	within		108.2964	-618.1652	3708.117	T-bar = 9.74359
solvr	overall	1.466461	17.95076	-1.27	665.44	N = 2243
	between		13.33038	-.266225	191.3691	n = 234
	within		12.92185	-163.1143	475.5374	T-bar = 9.58547
d1	overall	.3540754	.478337	0	1	N = 2282
	between		.2644329	0	1	n = 234
	within		.4396969	-.479258	1.263166	T-bar = 9.75214
d2	overall	.5692375	.4952915	0	1	N = 2282
	between		.3394262	0	1	n = 234
	within		.4306384	-.3398534	1.42638	T-bar = 9.75214

Source: Output generated within Stata™.

Time-invariant regressors would have a zero *within* variation and individual-invariant regressors would have zero *between* variation. Only the regressors *MEM* and *SOLVR*, have more variation across medical schemes (*between* variation) than over time (*within* variation), so *within* estimation may lead to some efficiency loss.

¹⁵³ Note Stata™ uses lowercase *n* to denote the number of medical schemes and uppercase *N* to denote the total number of individual-time observations.

The estimation results for the one-way fixed-effects panel specification (Equation 4-8) are presented in Table 4-4. Additional diagnostic tests were run on the FE specification, to test for heteroscedasticity and autocorrelation. Serial correlation is typically not a problem with panels that have short time horizons (in this case $T = 17$). From the residual plots there is no discernible indication that autocorrelation is present. The regression results are robust and do not appear to suffer from any misspecification bias. Weak form multicollinearity between the variables *MEM*, *NCONTA* and *ADMINA* was uncovered. This is likely the result of the transformation, whereby both net contribution income and gross administration expenses were divided by the total number of medical scheme beneficiaries to obtain the relevant explanatory variables *NCONTA* and *ADMINA*.

Table 4-4: Regression Output - Fixed-Effects Model

Fixed-effects (within) regression			Number of obs		=	2230
Group variable: id			Number of groups		=	234
R-sq: within = 0.5797			Obs per group: min		=	1
between = 0.7635			avg		=	9.5
overall = 0.6683			max		=	17
corr(u_i, Xb) = 0.2744			F(4,1992)		=	686.86
			Prob > F		=	0.0000

nclmsa		Coef.	Std. Err.	t	P> t	[95% Conf. Interval]

mem		-.0000105	.0001151	-0.09	0.927	-.0002362 .0002151
nconta		.6884398	.0137345	50.12	0.000	.6615044 .7153752
admina		-.0564018	.0465238	-1.21	0.226	-.1476421 .0348385
solvr		-.1268806	.2648344	-0.48	0.632	-.6462621 .3925008
_cons		162.5896	9.918471	16.39	0.000	143.138 182.0413

sigma_u		130.52776				
sigma_e		149.3946				
rho		.43290455	(fraction of variance due to u_i)			

F test that all u_i=0:			F(233, 1992) =	6.34	Prob > F = 0.0000	

Source: Output generated within Stata™.

The regression results in Table 4-4 indicate that the standard deviation of the individual effect (μ_i) (from Equation 4-5) is reflected in the Stata™ output by the value for *sigma_u*, which in this case equates to 130.52776. *sigma_e* gives value of the standard deviation of the remainder disturbance term (or idiosyncratic error),

which is equal to 149.3946. In Table 4-4, ρ represents the fraction of unexplained variance due to differences among medical scheme variables. The proportion of variance due to the individual-specific (medical scheme) effect is 43.29 percent. Thus, the balance is due to the idiosyncratic error, which implies that the majority of the variance in net claims incurred (pabpm) (*NCLMSA*) is caused by the regressors.

The $R^2(\text{within})$ value from Table 4-4 suggests that the variation within the medical schemes explains 57.97 percent of the variation in the data. Thus, the regression fit of the company over time is poor. The $R^2(\text{between})$ which measures the squared-correlation between cross-sectional unit means and the predicted values of the independent variable, shows that 76.35 percent of the variation in net claims incurred (pabpm) (*NCLMSA*) is explained by the explanatory variables. $R^2(\text{overall})$, which is a measure the *goodness-of-fit* for the entire regression specification, indicates that model explains only 66.83 percent of the variation in net claims incurred (pabpm) (*NCLMSA*). Finally, the F -test statistic ($F(1, 62) = 686.86$) indicates that the coefficients attaching to the regressors are statistically significantly different from zero simultaneously at the 99 percent confidence level ($p\text{-value} < 0.01$); thus, implying that the model is a good fit.

From Table 4-4, it is possible to substitute the estimation results into Equation 4-8 to obtain Equation 4-10.

Equation 4-10: Fixed-Effects Estimation Results

$$NCLMSA_{it} = 162.5896 - 0.0000105(MEM_{it}) + 0.6884398(NCONTA_{it}) \\ - 0.0564018(ADMINA_{it}) - 0.1268806(SOLVR_{it})$$

where $NCLMSA_{it}$ refers to the net claims incurred (in real terms), per average medical scheme beneficiary per month (pabpm) for medical scheme i and observed at time t (1993 – 2009). MEM_{it} denotes the total number of medical scheme members in medical scheme i observed at time t ; $NCONTA_{it}$ refers to the net contribution income (in real terms) per average medical scheme beneficiary per month (pabpm) for medical scheme i and observed at time t ; $ADMINA_{it}$ refers to the gross administration expenses (in real terms) per average medical scheme beneficiary per month (pabpm) for medical scheme i and observed at time t ; and $SOLVR_{it}$ denotes the solvency ratio of medical scheme i and observed at time t .

The only explanatory variable that is highly significant is net contribution income (pabpm) ($NCONTA$). The null hypothesis that the coefficient is statistically significantly different from zero is soundly rejected at even the 99 percent confidence level. The coefficient attaching to $NCONTA$ indicates that a one unit increase in net contribution income (pabpm) leads to an expected 0.6884 unit increase in net claims incurred (pabpm) holding all other factors constant. This is in-line with one's *a priori* expectations, since as policyholders pay more for medical scheme coverage; they may be induced to increase healthcare expenditure by increasing their utilisation of their current medical scheme benefits. Although, the direction of causality is commonly viewed in reverse, the relationship does still hold true, if one considers that net contribution income does influence net claims incurred.

The key coefficient of interest is the one attaching to the total number of medical scheme beneficiaries (MEM). Although, $\hat{\beta}_1 = -0.0000105$, is negative implying that a decrease in the total number of medical scheme beneficiaries leads to an expected increase in net claims incurred (pabpm) holding all other factors constant. Therefore, it would appear to be supporting the view that adverse selection is indeed present in South Africa's private health insurance market. Since it suggests that those members who would be the first to opt out of medical scheme coverage would be the

low risks – this would lead to a deterioration in the risk profile of those that remain behind in the medical scheme pool, which gets reflected in higher net claims incurred (pabpm). However, the null hypothesis that $\hat{\beta}_1$ is statistically significantly different from zero cannot be rejected at the 95 percent confidence level. Hence, there would appear to be no discernible adverse selection present in South Africa’s private health insurance sector.

The estimation results for the one-way random-effects panel specification (Equation 4-9) are presented in Table 4-5.

Table 4-5: Regression Output - Random-Effects Model

Random-effects GLS regression		Number of obs	=	2230		
Group variable: id		Number of groups	=	234		
R-sq:	within = 0.5910	Obs per group:	min =	1		
	between = 0.7532		avg =	9.5		
	overall = 0.6673		max =	17		
Random effects u_i ~ Gaussian		Wald chi2(6)	=	3589.31		
corr(u_i, X) = 0 (assumed)		Prob > chi2	=	0.0000		

	nclmsa	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]

	mem	-.0001135	.0000948	-1.20	0.231	-.0002994 .0000724
	nconta	.6819877	.0153989	44.29	0.000	.6518065 .7121689
	admina	-.0117186	.0461817	-0.25	0.800	-.1022331 .0787958
	solvr	.2789999	.2422866	1.15	0.250	-.1958731 .753873
	d1	2.810207	13.41257	0.21	0.834	-23.47795 29.09837
	d2	46.1219	14.7822	3.12	0.002	17.14931 75.09449
	_cons	137.6362	14.69897	9.36	0.000	108.8268 166.4457

	sigma_u	80.024676				
	sigma_e	146.46916				
	rho	.22988516	(fraction of variance due to u_i)			

Source: Output generated within Stata™.

The estimation of the RE model produces almost identical results to the FE model with one key exception; namely, the RE specification included two dummy variables to take into account the structural breaks in regulatory regime identified previously. Only the second dummy variable proves to be statistically significantly different from zero at the 95 percent confidence level. This indicates that the implementation of the Medical Schemes Act No. 131 of 1998 did cause a sustained structural change to the

private health insurance sector in South Africa. In particular, since 2000, net contribution income (pabpm) increased by R46.12 per average medical scheme beneficiary per month purely as a result of the change in regulatory regime (holding all other factors constant).

From Table 4-5, it is possible to substitute the estimation results into Equation 4-9 to obtain Equation 4-11.

Equation 4-11: Random-Effects Estimation Results

$$\begin{aligned}
 NCLMSA_{it} = & 137.6362 - 0.0001135(MEM_{it}) + 0.6819877(NCONTA_{it}) \\
 & - 0.0117186(ADMINA_{it}) + 0.2789999(SOLVR_{it}) \\
 & + 2.810207(D_1) + 46.1219(D_2) + u_{it}
 \end{aligned}$$

where $NCLMSA_{it}$ refers to the net claims incurred (in real terms), per average medical scheme beneficiary per month (pabpm) for medical scheme i and observed at time t (1993 – 2009). MEM_{it} denotes the total number of medical scheme members in medical scheme i observed at time t ; $NCONTA_{it}$ refers to the net contribution income (in real terms) per average medical scheme beneficiary per month (pabpm) for medical scheme i and observed at time t ; $ADMINA_{it}$ refers to the gross administration expenses (in real terms) per average medical scheme beneficiary per month (pabpm) for medical scheme i and observed at time t ; $SOLVR_{it}$ denotes the solvency ratio of medical scheme i and observed at time t ; D_1 denotes the first structural break that demarcates the pre-deregulation and post-deregulation periods (at time $t = 1994$); and D_2 denotes the second structural break that demarcates the post-deregulation period and the post-implementation of the Medical Schemes Act No. 131 of 1998 (at time $t = 2000$).

Importantly, once again the key coefficient $\hat{\beta}_1$ is negative implying that a decrease in the total number of medical scheme beneficiaries leads to an expected increase in net claims incurred (pabpm) holding all other factors constant. However, the null hypothesis that $\hat{\beta}_1$ is statistically significantly different from zero cannot be rejected at the 95 percent confidence level. Hence, it confirms the absence of adverse selection in South Africa's private health insurance sector.

A summary of the regression results for both the FE and RE models are presented in Table 4-6.

Table 4-6: Summary of FE and RE Estimation Results

Variable	FE	RE
mem	-.00001053	-.00011352
	.00011505	.00009485
nconta	.68843981	.68198769
	.01373446	.01539886
admina	-.0564018	-.01171862
	.04652375	.04618169
solvr	-.12688064	.27899993
	.26483441	.24228661
d1		2.8102074
		13.412572
d2		46.121899
		14.782203
_cons	162.58962	137.63622
	9.9184706	14.698972
N	2230	2230
r2	.57969821	
r2_o	.66826035	.66730164
r2_b	.763482	.75319802
r2_w	.57969821	.59100765
sigma_u	130.52776	80.024676
sigma_e	149.3946	146.46916
rho	.43290455	.22988516

Source: Output generated within Stata™.

As discussed previously, a useful test to determine which model, FE or RE, is more appropriate is the Hausman test. Under the null hypothesis that individual effects are random, the estimators under both the FE and RE should be similar because both are consistent. Essentially, it tests whether the individual-effects errors (μ_i) are correlated with the regressors and thus, the null hypothesis is that they are not.

In this study, an extension of the Hausman test is employed, whereby one specifies that both the covariance matrices are based upon the (same) estimated disturbance variance from the efficient estimator. This tends to be a preferred method, since in

certain cases, the conventional Hausman test may lead to variance estimates that are negative definite and therefore, estimated standard errors cannot be determined.

Table 4-7: Hausman Test - Fixed-Effects versus Random-Effects

---- Coefficients ----				
	(b)	(B)	(b-B)	sqrt(diag(V_b-V_B))
	FE	RE	Difference	S.E.
mem	-.0000105	-.0001135	.000103	.000071
nconta	.6884398	.6819877	.0064521	.
admina	-.0564018	-.0117186	-.0446832	.012788
solvr	-.1268806	.2789999	-.4058806	.1253236

b = consistent under Ho and Ha; obtained from xtreg				
B = inconsistent under Ha, efficient under Ho; obtained from xtreg				
Test: Ho: difference in coefficients not systematic				
chi2(4) = (b-B)'[(V_b-V_B)^(-1)](b-B)				
= 18.99				
Prob>chi2 = 0.0008				
(V_b-V_B is not positive definite)				

Source: Output generated within Stata™.

Table 4-7 suggests that the null hypothesis that the difference in the coefficients between the FE and RE estimates results is rejected as the p -value of the $\chi^2_{(4)}$ test statistic is less than 0.05. Therefore, the Hausman test results indicate that the FE model is preferred.

4.3. Reasons for the Absence of the Positive Coverage-Risk Correlation

Cohen and Siegelman (2010) examine several explanations for why a positive coverage-risk correlation may not be observed. These reasons include why members of medical schemes may not possess superior informational advantages over their corresponding medical schemes, or the inability of members to utilise whatever advantage they do possess. It is also necessary to discuss the factors that may have a negative coverage-risk correlation, which would have an offsetting effect upon the

positive coverage-risk relationship. Interactions between risk and risk aversion (*propitious selection*) are important as they have a direct influence upon the individual's decision to purchase insurance. Recall adverse selection theory emphasises the idea that high-risk individuals are more likely to purchase insurance than low-risk individuals. In addition, to the possible correlation between low-risk and high risk-aversion, low-risk may also be correlated with other variables associated with a greater tendency to purchase insurance (Fang, Keane & Silverman, 2008). Consider, for example, *cognitive ability*. The authors examine senior citizens in the United States and their choice to purchase supplementary *Medigap* insurance.¹⁵⁴ Fang, Keane and Silverman (2008)'s empirical results demonstrate that individuals with higher cognitive ability (as measured by survey analysis) are associated with increased demand for insurance, since it is suggested that more sophisticated individuals are better able to understand the need for insurance. Moreover, higher cognitive ability is associated with better health, because more sophisticated individuals take better care of themselves.

In a recent contribution, Koufopoulos (2007) identifies certain circumstances where the coverage-risk correlation does not arise. Koufopoulos (2007: 603) suggests that:

“If there is one loss level, in the presence of fixed administrative costs, asymmetric information plus competition among insurance companies implies a positive relationship between coverage and the accident probability only if, in equilibrium, all agents choose contracts offering strictly positive coverage”.

The implication is that if certain individuals opt for zero coverage, that is elect to remain uninsured, the positive coverage-risk relationship may not hold true. With multiple loss levels, he demonstrates that the positive relationship between coverage and the accident probability holds true if and only if the level of indemnity offered by the high-coverage policy is constant with respect to loss size. If the level of indemnity is a strictly monotonic function with respect to loss size, the coverage-risk correlation may be zero or negative (Koufopoulos, 2007). If the level of indemnity increases as loss size increases, he proves that even if the equilibrium contracts offer strictly

¹⁵⁴ Medigap insurance is a supplementary health insurance policy that serves as a ‘top-up’ to Medicare.

positive coverage and administration costs are negligible, the coverage-risk correlation may be zero or negative.

4.3.1. Absence of Private Information

Theoretical models of adverse selection, such as the analysis presented by Rothschild and Stiglitz (1976)'s analysis, do present the notion of information asymmetry in a highly-stylized manner according to Cohen and Siegelman (2010). The idea that somehow the insured has superior information about their risk profile than the insurance company is unrealistic. They assert that this is particularly the case in insurance lines such as automobile insurance. The extent to which the insured has private information about their risk profile is questionable. Indeed, Chiappori and Salanié (2000) argue that the absence of a positive coverage-risk correlation may be explained by the lack of superior information on the part of the insured.

Cohen and Siegelman (2010) present a useful example to illustrate. Consider an individual i 's expected loss is given by the following specification:

Equation 4-12: Expected Loss Cost for Individual i under insurance (Cohen & Siegelman, 2010)

$$E[L_i] = p(X_i) \cdot L(Y_i)$$

where $E[L_i]$ denotes the expected loss for individual i . X and Y are vectors of all the explanatory variables that describe the probability (p) and magnitude of loss (L). These include variables whose values are unknown to both the insurer and insured (Cohen & Siegelman, 2010).

The possible variables within vector X could include factors such as the insured's temperament, aggressiveness, average speed, number of kilometres driven within a predefined period, extent of traffic congestion en route to the place of employment

and so on (Cohen and Siegelman, 2010). Variables in vector Y could include the make and model of the insured's motor vehicle. Cohen and Siegelman (2010) suggest that there are several ways in which imperfect information may manifest itself within Equation 4-12. First, the insured may be completely unaware of certain factors in either X or Y , or the insured may only have incomplete information about certain factors; thus, generating possible random errors or biases. Second, the insured may not be aware of the functional forms that transform these factors, such as behaviour or environment, into estimates of the probability and magnitude of a loss. Lastly, the authors suggest that a factor completely unknown to anyone – *luck* – may influence the variance of the probability or size of the loss across all the insureds. Therefore, even if the insured has some private knowledge concerning certain factors within X and Y , it is doubtful that the insured will have a significant advantage over the insurer in determining one's individual expected loss (Cohen and Siegelman, 2010).

Moreover, it is further doubtful that the insured will be familiar with the value of these factors or their corresponding functional forms that influence the probability and expected magnitude of loss for other policyholders. In the case of automobile insurance, a driver may be aware of his or her own driving ability or style; however, in order to have superior private information, the insured would need to be aware about how his or her driving ability compares to those of others within the insured pool (Cohen and Siegelman, 2010). It is indeed debatable to suggest that the insured possesses such information. For example, Svenson (1981) finds that 80 percent of all drivers believe that their driving ability would place them within the top 20 percent of safest drivers (Cohen and Siegelman, 2010). The probability of having an accident is generally underestimated by individuals as people's perception of their driving ability is somewhat elevated. In the United Kingdom, Guppy (1993) finds that drivers underestimate the probability of having an accident.

It is argued by Cohen and Siegelman (2010), that even if a policyholder knew about their own objective values for the factors in vectors X and Y , it is certainly debatable that they would know how to translate these into a probability or to accurately

estimate the size of loss, and then compare these to the premium that they are being charged by the insurance company and subsequently, make an informed decision about what is their appropriate level of coverage. Even if the insured could determine the actuarially fair premium that they should be charged by the insurance company, this figure would most certainly deviate from the premium charged by the insurance company. The insurance company loads premiums to recover administration and other operational expenses, and reward shareholders for the cost of capital. Take for example, the total distance travelled by the insured during a month.¹⁵⁵ An individual may be fully aware of this distance, but they are naïve with respect to its significance in determining the probability of loss or magnitude of loss (Butler, 1996).

In addition, Cohen and Siegelman (2010) discuss the proficiency at which individuals can accurately predict uncertain outcomes. Outcomes with negligible probabilities are often poorly predicted by people. For example, if two possible outcomes *A* and *B* have probabilities of 0.0001 and 0.00001 respectively, people tend to group both outcomes together as being highly unlikely. However, outcome *B* is essentially ten times less likely than outcome *A*. In the case of predicting one's own life expectancy, people do appear to perform admirably, but their ability to predict what exactly their demise will be is generally unsuccessful (Cohen and Siegelman, 2010).

Therefore, the common assumption within adverse selection theory that individuals possess superior private information compared to the insurer may be violated in certain insurance segments or markets.

¹⁵⁵ Insurance companies have started offering policies whereby the premiums are directly determined by the distance that the insured travels within a given time period. This has become practical due to cost-effective global positioning system (GPS) devices that can be easily installed in a motor vehicle.

4.3.2. Private Information Distributed Unevenly Amongst Policyholders

Cohen (2005) finds that in Israel, a positive coverage-risk correlation can be detected amongst policies sold to drivers with 3 or more years driving experience. But when the author considers insurance policies sold to drivers with less than 3 years driving experience, the coverage-risk correlation is absent. Cohen (2005)'s empirical results are consistent with the hypothesis that only individuals with significant driving experience can glean enough private information to yield a coverage-risk correlation when they purchase an insurance policy from a new insurer. Therefore, Cohen and Siegelman (2010) suggests that whether policyholders have private information that can yield a positive coverage-risk correlation may depend upon the insurance product and particularly, the type of individuals purchasing the insurance policy. In other words, for a particular insurance product, there may be a subset of policyholders who possess private information, whilst the remaining pool of insureds does not possess any private information.

Therefore, Cohen and Siegelman (2010) assert that if one focuses upon the subset of insureds with private information, then it is likely that one would observe a positive coverage-risk correlation. However, this would not be the case for the members of the pool who do not possess private information. Hence, if one examined the entire pool and the members who do not have private information, were sufficiently large, then the authors contend that the coverage-risk correlation may not be uncovered.

4.3.3. Failure to Utilise Private Information

The next observation that may lead to the failure to detect a positive coverage-risk correlation concerns the situation whereby policyholders do not utilise private information that they do possess. Cohen and Siegelman (2010) maintain that for a coverage-risk correlation to occur:

“...it does not suffice for policyholders to have private information about their risk type; they must also adjust their purchasing decisions on the basis of this information” (Cohen & Siegelman, 2010: 65).

Policyholders that act upon their private information when it comes to insurance purchasing decisions would make the necessary adjustments to their coverage levels and associated premium levels through their insurance contract choices. Yet, behavioural economics suggests that individuals often do not actively pursue optimal choices when the stakes are not sufficiently high. The authors argue that this is precisely the case, for example, in automobile insurance, when individuals decide upon deductible levels. Further evidence of this lack of optimal decision-making is found in Hurd, Smith and Zissimopoulos (2002), where senior citizens do not utilise their private information concerning their mortality to make optimal decisions regarding their retirement and social welfare benefits.

Pauly, Withers, Viswanathan, Lemaire, Hershey, Armstrong and Asch (2003), as cited by Cohen and Siegelman (2010), note that individuals have more on their minds than attempting to take advantage of potential benefits that could extract within their insurance transactions. Furthermore, Pauly et al. (2003) prove that the demand for life insurance is insensitive to changes in either price or relative risk. Therefore, Cohen and Siegelman (2010) conclude that even when individuals can predict their own risk better than insurance companies, they choose not to act upon this information by adjusting their insurance coverage levels.

4.3.4. Insurers possess Superior Information

Another possibility that should be considered as to why one may not observe a positive correlation between coverage and risk is that insurers may actually possess superior information concerning the insured's risk profile – in other words, the traditional view of information asymmetry, but in reverse. Chiappori and Salanié (2000) do examine this possibility in their analysis of the automobile insurance

market in France. Historical claims experience, statistical and underwriting techniques that are based upon both subjective and objective criteria allow insurers to assess risk fairly accurately and thus, permit insurers to extricate superior information relating to the risk of the insured (Cohen & Siegelman, 2010).

Even though the insured may have relevant (hidden) information concerning their individual risk status, it is nevertheless possible for the insurer to utilise proxy measures that reveal this hidden information. Brockett and Golden (2007) illustrate this feature, when they examine the relationship between a person's credit score and risk.¹⁵⁶ The authors demonstrate how a person's credit score can be utilised as a proxy measure to predict the likelihood of a person experiencing a loss under an automobile insurance policy. The relationship between a person's credit score and their risk in relation to automobile insurance is remarkably robust (Brockett & Golden, 2007). Therefore, Cohen and Siegelman (2010) posit that an insurer may simply need an individual's credit score in order to predict the person's riskiness. In addition, the insurer's assessment may indeed prove to be better than the individual's assessment of their own risk even if the person possesses superior information about certain aspects of their own (risky) behaviour (Cohen & Siegelman, 2010).

Cohen and Siegelman (2010) also argue that it is plausible to suggest that the insurer may be in a better position to forecast risk (using the factors that it is aware of) than the insured.

¹⁵⁶ A credit score refers to a single measure that summarises the information contained in factors that affect a customer's default probability with respect to the obligations under a credit agreement that is often entered into with a financial institution.

4.3.5. Interaction between Risk and Risk Aversion (Propitious Selection)

Thus far, the previous reasons have focused upon the reasons why an individual may not possess, or may simply fail to utilise superior information *vis-à-vis* the insurance company. An alternative explanation for the absence of a positive coverage-risk correlation is that certain factors may in fact induce a negative coverage-risk correlation, which could offset any positive coverage-risk correlation due to adverse selection (Cohen & Siegelman, 2010). Hemenway (1990) coined the term of *propitious selection* to describe the situation where there is an observable negative relationship between coverage and risk. A low-risk individual may be induced to seek out more generous (comprehensive) private health insurance cover because they are highly risk averse. Conversely, high risk individuals may elect higher deductibles with respect to private health insurance as their degree of risk aversion is weak. The result of this behaviour favours insurers, since the insured population is less risky than overall population (Cohen & Siegelman, 2010).

As cited by Cohen and Siegelman (2010), DeMeza and Webb (2001) provide an elegant theoretical model to illustrate why individuals that are more cautious with respect to their driving habits may exhibit a tendency to purchase higher coverage levels with respect to automobile insurance. However, DeDonder and Hindricks (2006) argue that to sustain a negative coverage-risk correlation, it is not sufficient to show that strongly risk averse individuals are more likely to demand greater coverage levels – their risk may also appear lower as they take more precautions through their cautious behaviour. But it is also necessary to demonstrate that individuals, who are weakly risk averse, exhibit a decreasing willingness to pay for insurance (Cohen & Siegelman, 2010). Utilising simulation techniques, Karagyzova and Siegelman (2007) prove that even large negative values for the correlation between coverage and risk are unable to prevent a market from unravelling due to imperfect information.

Cohen and Siegelman (2010) discuss the findings of Cutler, Finkelstein, and McGarry (2008). Cutler et al. (2008) investigate the absence of a positive coverage-risk correlation by examining five different insurance markets; namely, life, health, annuities, long-term care, and Medigap. A key aspect underlying Cutler et al. (2008)'s analysis concerns an individual's tolerance for risk *vis-à-vis* their actual risk. The authors find that individuals may indeed vary their risk tolerance in addition to their riskiness. By examining, insurance coverage levels, several risk indicators and risk tolerance, they empirically test how these three factors are related to one another (Cohen & Siegelman, 2010). The results of their analysis suggest that individuals who elect not take precautions to reduce risk are less likely to purchase any of the five insurance types studied. Moreover, these same individuals display higher expected claims under life insurance and long-term care insurance but have lower expected claims for annuities (Cohen & Siegelman, 2010). Cutler et al. (2008) do not detect a systematic relationship with respect to expected claims for these individuals under private health insurance or Medigap.

Thus, Cohen and Siegelman (2010) conclude that models describing the ability of *propitious selection* to negate the positive coverage-risk correlation under adverse selection remain a contestable question.

4.3.6. Institutional and Regulatory Factors

An individual's decision to purchase insurance and indeed the selection of the appropriate coverage level may be influenced by "institutions" such as insurance intermediaries who attempt to match individuals with insurers (Cohen & Siegelman, 2010). Therefore, these influences may induce a negative correlation between coverage and risk. Cummins and Doherty (2006) argue that intermediaries may mitigate informational asymmetries and therefore reduce adverse selection. Although, the extent to which intermediaries (brokers) in South Africa's medical schemes sector, can mitigate informational asymmetries is somewhat limited by the regulatory

environment that prohibits risk classification. Nevertheless, medical scheme brokers can steer individuals towards certain benefit options and thus, they may have an undue influence upon the coverage levels selected by medical scheme members.

Cohen and Siegelman (2010) also contend that the marketing activities of insurers should be considered under institutional factors. In South Africa, it is precisely the marketing techniques of medical schemes that have allowed them to target their products (and benefit options) for a specific “audience” – this is in spite of the regulatory environment that enforces community-rating and open enrolment. As discussed previously, it these marketing activities that have permitted medical schemes to effectively cherry-pick lower risks.

Regulation prohibiting risk classification is also put forward by Cohen and Siegelman (2010) as another factor that may contaminate the relationship between coverage and risk. For example, in the case of South Africa, where medical schemes may not discriminate on the basis of current or past health status, this may induce high risk individuals to purchase more generous medical scheme cover (move to more comprehensive benefit options) not because of informational asymmetries, but rather as a result of opportunistic behaviour on the part of the high risks. Due to community-rating, high risks realise that their premiums are effectively subsidised by low risks and thus, it is merely rational economic behaviour to purchase additional medical scheme cover as they are essentially receiving the extra cover at a discount. It is the inability of medical schemes to utilise certain risk-related information that leads to the risk-related cross-subsidisation.

4.4. Conclusion

The empirical findings suggest that there is no discernible adverse selection present within the South African medical scheme market. Hence, the notion that the

promulgation of the Medical Schemes Act No. 131 of 1998 was necessary to address the negative consequences of adverse selection is proven to be false. Several reasons have been offered as to why one may detect no positive coverage-risk correlation in certain insurance markets, lines or segments. As an avenue for further empirical research, these reasons should be explored within the context of the medical scheme industry in South Africa.

5 REGULATORY CAPTURE

This chapter seeks to describe the theories of economic regulation that explain the observed behaviour of government intervention. This is achieved by reviewing the seminal literature within the area. Four distinct schools of thought regarding regulatory capture are discussed extensively.

Government has the ability to utilise public resources and exercise executive powers to bestow advantages (or disadvantages) upon industries. As Stigler (1971) argues the government can act as a potential resource or threat to any industry. Through the process of legislative policymaking and enforcement, the state has a further ability whereby it can create a framework to drive political or social objectives. Regulation is a mechanism that can incentivise individuals or institutions to behave in a particular manner that the state deems to be in the public interest (Stigler, 1971). When compared to the most powerful interest group or citizen, the state possesses the ability to coerce.

Levine and Forrence (1990) argue that the appropriateness of government intervention, whereby the activities of a firm or even the operation of an entire industry are regulated, should be judged solely upon the ability of the intervention to remedy market distortions where they exist. Laffont and Tirole (1991) focus upon the following primary distortions that potentially prevent markets to form or operate: destructive competition, monopoly pricing, information asymmetries and environmental externalities.

Economic regulation refers to the actions available to government whereby they can impose taxes, offer subsidies or directly institute legislative and administrative controls with respect to contribution rates (rate regulation), product innovation,

market activities, and entry and exit barriers. Stigler (1971: 3) describes the theory of economic regulation as an attempt to:

1. Explain the nature and form of regulation;
2. Describe the effect that regulation has upon the allocation of resources; and
3. Identify the participants that receive the benefits or suffer the burdens of regulation.

Two major theories have been offered to explain the motivation behind government intervention and regulation. The first of these theories, *public interest theory*, emphasises the role of central government to correct market distortions. According to Ippolito (1979: 55) public interest theory describes government regulation as an attempt to create:

“...a governmental corrective device that improves market performance in instances in which competition fails”.

The key premise behind public interest theory is that regulation is necessary to maximise social welfare. Although, market distortions may offer a convenient pretext to intervene, regulation should only be undertaken if it is in the *public interest*. The definition of what is in the *public interest* may indeed be an ambiguous construct.

The second, *capture theory*, and its natural extension, the *economic theory of regulation* provide an alternative view of why government may choose to regulate. Stigler (1971) brings attention to the role of special interest groups in the formulation of regulation. It is argued that special interest groups capture the regulator (or regulatory process) for their own benefit.

5.1 Theories of Economic Regulation

Theories of economic regulation that explain the observed behaviour of government intervention have been well-documented (Adams & Tower, 1994; Laffont, 1994; Posner, 1974; Priest, 1993; Shleifer, 2005; Stigler, 1971). Stigler (1971) broadly identifies two primary theories of economic regulation. Stigler (1971: 3) outlines the first as:

“...[being regulation that is] instituted primarily for the protection and benefit of the public at large or some large subclass of the public”.

Often referred to as *public interest theory*, the theory can be traced back to Pigou (1932), but it was only until the theory was adequately articulated by Stigler (1971) and Posner (1974) did it become possible to subject the theory to appropriate scrutiny and empirical investigation.

The second view is that the political process is somehow directed by “an imponderable shifting mixture of diverse forces” (Stigler, 1971: 3). The latter view may be misinterpreted as the belief that government intervention is tainted with uncertainty and directed by irrational behaviour on the part of its agents. It is in fact considerably more systematic. Whilst various versions of this view have been expounded, it is usually termed as *capture theory*. Capture theory’s origin can be traced back to Marx’s view that large corporations capture the regulatory institutions (and processes) to the detriment of labour. Stigler (1971) extended this theory by noting that even small business or industries may have the ability to usurp the regulatory process. Under capture theory:

“...regulation is supplied in response to the demands of interest groups, struggling among themselves [to maximise the benefits to their members]” (Posner: 1974: 336).

A third strand of regulation theory, namely, the *economic theory* of regulation, is derived as a version of capture theory from Stigler (1971), but includes elements from both public interest theory and capture theory. Adams and Tower (1994) declare that the economic theory of regulation proves to be theoretically and empirically superior in explaining government intervention. But this view is not shared by all contemporary researchers; Priest (1993) uses the early history of public utility regulation as a lens through which to view and to critique the competing public interest and capture (including economic) theories of regulation. He finds that neither the public interest theory nor capture theories of regulation are adequate.

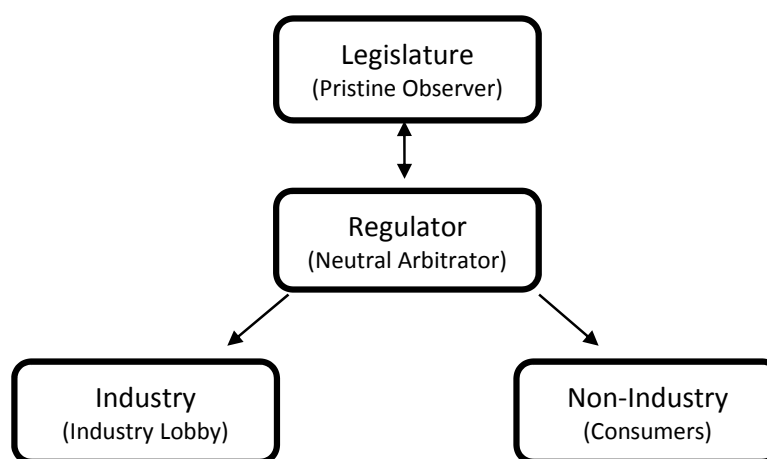
5.1.1 Public Interest Theory

Public interest theory developed from two premises. First, that economic markets are fragile and if left unhindered these markets tend to behave in an inefficient (or inequitable) manner (Posner, 1974). Second, government regulation is seen as relatively costless. Hence, government intervention is a response to “public demands for the rectification of palpable and remediable inefficiencies and inequities in the operation of the free market” (Posner, 1974: 336). Public interest theory thus endorses the view that the government intervenes to maximise social welfare through the process of instituting regulation to rectify market imperfections. Regulation of insurance companies “has traditionally been promoted as an exercise in safeguarding the public interest” (Adams & Tower: 1994: 165).

Government’s role is to correct market imperfections, such as monopoly pricing and environmental externalities (Laffont & Tirole, 1991). Corporate failures, such as the collapse of the automobile insurance company, Vehicle and General, in 1971 caused the United Kingdom government to improve financial disclosure and tighten solvency regulations (Finsinger, Hammond & Tapp, 1985). Hence, regulation is seen as reactionary rather than pre-emptive. At the extreme, regulation may be a response to a perceived market failure (Wilson, 1964). The underlying principle behind public

interest theory emphasises the role that regulation plays as a redistribution mechanism to enhance social welfare by correcting the misallocation of resources (Wilson, 1980: 138). While regulatory agencies suffer from informational asymmetries, they are nonetheless viewed as benevolent maximisers of social welfare (Laffont & Tirole, 1991: 1089). According to Adams and Tower (1994), under the public interest theory, regulators are viewed as independent and neutral arbitrators responding to the public's call to correct market imperfections and failures. A simple illustration of the conceptual model is represented in Figure 5-1.

Figure 5-1: Adams and Tower (1994: 165)'s Public Interest Theory – Conceptual Model



The concept of regulation to serve the public interest does generate ambiguity with respect to what exactly defines “in the public interest”. One definition proposed for the public interest is that it describes “...the best response to a situation in terms of all the interests and of the concepts of value which are generally accepted in our society” Redford (1954: 1108). Therefore, the author suggests that somehow public interest encapsulates a common set of values that society possesses. It is precisely these shared values that underpin a set of principles, which rather than mechanical rules, are flexible enough to allow government to drive society’s interactions to improve overall societal welfare. Bonbright (1961) promotes the view that the protection of consumers is paramount and any regulatory action that does so can, by definition, be regarded as

acting within the public interest. He is of the opinion that by linking the public interest to the protection of consumers renders the government "...as an instrument for the attainment of [societal] welfare" (Bonbright, 1961: 29).

In fact, the promulgation of the Medical Schemes Act No. 131 of 1998 is argued by some to be precisely an attempt to adhere to the principle of consumer protection. The introduction of *community rating* and *open enrolment* were argued by the South African government as necessary to overcome the situation whereby increasingly the elderly and unhealthy were being excluded from health insurance coverage (Reekie, 1999). Therefore, the new regulatory framework governing medical schemes in South Africa was seen as a response to an observable deterioration in the coverage levels of those who were precisely in need of health insurance cover the most. It was a response to protect consumers (albeit a subset of consumers) and thus, the new legislation was in the public interest.

A parallel can be drawn with the situation that Friendly (1962) discusses with respect to the regulation of the short-term insurance market in the United States. The author's review of a number of regulatory commissions uncovers a deep-seated concern for anti-competitive monopoly pricing practices and cartel-type pricing arrangements through so-called rating bureaus that can ultimately lead to destructive competition. Regulators at the time were concerned that new entrants to the market were driving prices (premiums) downwards and at the same time the new insurers were able to press service advantages over established insurers.¹⁵⁷ It rendered established insurers unable to compete. Regulators were afraid that insurance cover was being priced effectively "at a discount" (premiums were inadequate) and this generated greater insolvency risk. Thus, regulators decided that the public interest was best served by prescribing a pricing scheme that maintained a *fair rate of return* but at the same time

¹⁵⁷ This included having alternative distributional channels for the sale of insurance – direct writers versus independent agents.

discouraged unfair and destructive pricing practices (Friendly, 1962).¹⁵⁸ The author hints at *capture theory* here, as he observes that special interest groups can have a significant influence upon the formation of regulatory policy.

This raises a related issue with respect to the South African government's stated goal in the provision of health insurance cover to encourage the subsidisation of high risk members by low risk members through *community rating* and *open enrolment*. Is this goal consistent with the concept of *in the public interest*? Government's defence of this stated goal is that a form of market distortion was evident in South Africa's health insurance sector prior to the implementation of the Medical Schemes Act No. 131 of 1998; namely, adverse selection. Adverse selection was argued to be self-evident in the inability of the elderly and infirm to obtain affordable health insurance cover. As outlined in Chapter Two, the idea that low risks drive out higher risks from the health insurance market is a complete misinterpretation of the concept of adverse selection.

Furthermore, to argue that any activity that protects consumers is in the public interest relegates the theory of public interest to a mere mechanical rule that dictates that basis for in the public interest is housed within the primary objective of protecting consumers. Buchanan and Tullock (1962) concur with the view that the notion of in the public interest is ambiguous. This ambiguity itself has played an influential role in the evolution of insurance regulation (Kimball, 1961).

Harrington and Niehaus (1999) examine the evolution of insurance regulation in the United States and from their discussion it is evident that the regulation of insurance evolved as a response to two competing ideologies as to who ought to be responsible for insurance regulation; namely, state versus federal regulation. This debate is one example of the general debate in political philosophy and economics over the

¹⁵⁸ Friendly (1962) vehemently disagreed with view that the new pricing arrangement could be seen as being in the public interest. Friendly (1962) questions just how can the service offerings of an insurer that attract a lower price be deemed as counterproductive? The concept of a *fair rate of return* also raises further concerns as how it is determined.

advantages and disadvantages of centralisation of government power versus decentralisation (Harrington & Niehaus, 1999). Therefore, people that favour a decentralisation of power tend to view state regulation more kindly than do those who favour greater centralisation of power. A method to evaluate either approach, is to consider the comparative efficiency of state versus federal regulation, but uncertainty concerning the magnitude of the costs and benefits associated with each system leads to a diversity of opinion (Harrington & Niehaus, 1999).

*Paul v. Virginia*¹⁵⁹, the *Southeastern Underwriters Association* case and the *McCarran-Ferguson Act* were all key events that drove the direction of insurance regulation in the United States. The United States Supreme Court judgement in the case of *Paul v. Virginia* implied that states had the power to regulate insurance and that the federal government did not. This decision was upheld for approximately 75 years in many other cases that argued that insurance constituted interstate commerce. During the 1870s, numerous insurance companies became insolvent as a result of major fires that occurred in Boston and Chicago (Harrington & Niehaus, 1999). The increased number of insurer insolvencies helped spur the development of insurance rating bureaus, precursor organisations to the insurance advisory organisations that are currently found in the United States. Rating bureaus set property (short-term) insurance rates that would be charged by most companies, in principle to ensure adequate prices and therefore reduce insolvency risk. Here the public interest argument was evoked whereby ensuring that insurance prices were adequate protected consumers by reducing insolvency risk. Indeed, within the United States many states either permitted or encouraged the development of rating bureaus, and in some cases state regulators began to regulate their activities. In certain circumstances, state regulators even determined the rates that had to be charged by all insurance companies (Harrington & Niehaus, 1999).

¹⁵⁹ Mr. Paul was an agent for a group of New York fire insurers. Virginia law required out-of-state insurers and their agents to be licensed by the state. According to Harrington and Niehaus (1999), Mr. Paul refused to pay the security deposit that was required to obtain a license and he kept on selling fire insurance policies. He was subsequently arrested and fined \$50. Mr. Paul and the New York insurers challenged the conviction, arguing in part that the sale of insurance across borders was interstate commerce and thus that the Virginia law was unconstitutional because it interfered with interstate commerce. In 1868, the United States Supreme Court reached an interesting verdict – it held that insurance was not commerce and therefore was not subject to laws affecting interstate commerce. Since insurance was not commerce, Virginia's law was not unconstitutional (Harrington & Niehaus, 1999).

Harrington and Niehaus (1999) discuss that in response to a request from the Missouri Attorney General in 1942, the antitrust division of the United States Department of Justice began to investigate the activities of a large rating bureau known as the Southeastern Underwriters Association. This association was subsequently indicted by the United States Attorney General for alleged violations of the federal *Sherman Antitrust Act*. The Sherman Antitrust Act prohibits price fixing and related non-competitive activities. The charges against the Southeastern Underwriters Association included restraining and monopolising commerce, fixing prices and agents' commissions, attempting to force buyers to buy from member insurers, denying non-member insurers access to reinsurance from member insurers, and refusing to do business with agents who represented non-member insurers. The Southeastern Underwriters Association's defence against this legal action was that the *Sherman Antitrust Act* did not apply to insurance because according to *Paul v. Virginia* insurance was not commerce. Although, the Southeastern Underwriters Association argued that many of these practices were beneficial given the nature of insurance, they decided not enter these arguments into their defence against the legal action. A federal district court upheld this view and dismissed the case. However, the United States Attorney General appealed to the United States Supreme Court.

In 1943, the United States Supreme Court did not decide on the merit of the charges against the Southeastern Underwriters Association, but rather on the validity of the precedent emanating from *Paul v. Virginia*. It was adjudged that insurance was indeed commerce and that when insurance was sold across state borders then insurance would be regarded as interstate commerce. The implication of this decision gave the United States federal government jurisdiction in the regulation of insurance and further, the *Sherman Antitrust Act* applied to insurance. Harrington and Niehaus (1999) point out that the decision did not prohibit state regulation, but it held that state regulation that were in breach of federal regulation were invalid. The outcome of the Southeastern Underwriters Association case led to considerable uncertainty with respect to the scope of state regulation and the taxation of insurers. In addition,

questions were raised as to the legality of the operations of the insurance industry, particularly with respect to utilisation of rating bureaus (Harrington & Niehaus, 1999).

According to Harrington and Niehaus (1999), the uncertainty that was created by the Southeastern Underwriters Association decision, mobilised representatives from both the insurance industry and state insurance regulators to seek out legislation aimed at clarifying the role of state regulation with respect to insurance. In 1945 the United States Congress responded by enacting the McCarran-Ferguson Act. Two key provisions within the McCarran-Ferguson Act are that:

1. continued regulation and taxation of insurance by the states is in the public interest; and
2. insurance is to be exempt from federal antitrust law, provided that the relevant activities are subject to state regulation and do not involve “any agreement to boycott, coerce, or intimidate, or act of boycott, coercion, or intimidation” (Harrington & Niehaus, 1999).

The key issue underlying *Paul v. Virginia*, the *Southeastern Underwriters Association* case and the *McCarran-Ferguson Act* involved who was indeed responsible for the regulation of insurance – state versus federal regulation. Harrington and Niehaus (1999) provide a detailed account of the arguments in favour and against state and federal regulation. But as the McCarran-Ferguson Act argues continued regulation and taxation of insurance by state regulators is in the public interest. As stated by the authors, insurance regulation reflects the objective of promoting safety and soundness, given that unregulated market incentives and conditions might not produce an efficient level of solvency risk.

Kimball (1961) contends that only with tangible objectives can regulation truly be made to serve the public interest. Kimball (1961: 477) suggests that in terms of insurance, primary role of regulation should be to “...facilitate the successful

operation of the enterprise itself". Harrington and Niehaus (1999) concur with Kimball (1961) arguing that central to the successful operation of an insurer is the insurer's solvency. Specifically, Kimball (1961) maintains that insurer solvency can best be managed through the regulation of insurance premiums. In other words, the author advocates the implementation of rate regulation within the insurance context. Two types of rate regulation are common within the insurance market:

1. *Rate Compression*: this type of rate regulation refers to the situation where insurance companies are prevented from setting premium based upon an individual's characteristics such as age, gender, current health status and other risk-related characteristics.
2. *Rate Suppression*: can be defined broadly as government suppression of insurance rates below levels that would exist otherwise without rate regulation.

Harrington (1992) considers why certain insurance lines may be especially vulnerable to rate suppression. Higher insurance costs, particularly when higher-than anticipated claims costs are experienced, can encourage regulatory authorities to reduce insurance rates or limit insurance rate increases via regulation. Once again this may be viewed as being in the public interest. He contends that within a competitive insurance market, persistent rate suppression should lead to a reduction in product quality or exit by insurers. Therefore, both outcomes should discourage regulators from implementing rate suppression. This can be summarised in the statement:

"If quality responses are constrained, the threat of widespread exit alone should give politicians and regulators considerable pause before persistently suppressing rates" (Harrington, 1992: 186).

Harrington (1992) finds that there is a pattern of insurers exiting the market in certain states in the United States, but the pattern has not been widespread or rapid in any particular state. Thus, by implication the negative consequences of rate suppression that are anticipated may simply be an over-exaggeration, but the author argues that

this is difficult to reconcile with the poor underwriting results experienced and increased number of exits from the certain lines by insurance companies. A possible explanation may be that rate suppression only suppresses premiums to a level that provides a fair rate of return for insurance companies.

The idea behind rate regulation is to protect consumers by ensuring that premiums are adequate to mitigate insolvency risk. Indeed, the concern for solvency manifests itself in the stipulation that premiums must not be “inadequate” (Kimball, 1961: 482). Kimball (1961: 480) describe the first reason as to why rate regulation protects consumers as:

“Insurance is one important modern way that man seeks security in a world in which it does not exist naturally. It is thus not surprising that all systems of insurance regulation regard the financial solvency of the insurance enterprise as the central aim, for if nothing else, insurance must ‘insure’”.

Ippolito (1979) maintains that the costs associated with insolvency, but not limited to the time of the actual event of insolvency, can be so excessive that the threat of insolvency alone justifies government intervention. Indeed, the costs of financial distress just prior to insolvency may induce an insurer to go insolvent.

Frech and Samprone (1980) examine the welfare implications of rate regulation using the deregulation that had taken place in certain states in the United States between 1947 and 1974. They explain that rate regulation in the property-liability insurance market at that time took place according to the following procedure: insurers or rating bureaus would submit insurance rates to the state insurance department for approval prior to them being utilised. These rates would become effective once approved by the state insurance department or once a pre-determined amount of time had passed. As soon as the rates were approved, an insurance company was not permitted to change them unless a deviation request was filed and approved by the state insurance department (Frech & Samprone, 1980). The authors suggest that rate regulation protects consumers in two distinct ways:

1. If pricing is left to competitive forces it can lead to aggressive price competition that will result in inadequate insurance rates and therefore rate regulation is an instrument of the government to protect against insolvencies; and
2. Since many insurers were permitted and encouraged to set insurance rates in concert by subscribing to rating bureaus, rate regulation serves as a means to monitor and prevent abusive cooperative pricing practices.

Frech and Samprone (1980) point out that while rate regulation generally translates into higher insurance premiums for consumers (implies a negative effect upon societal welfare), the benefits derived from increased non-price competition (through perhaps the additional services that insurers may provide) can lead to welfare gains for consumers. An examination of two lines of automobile insurance (automobile liability and automobile physical damage insurance) in the United States provide mixed evidence that rate regulation leads to a societal welfare loss (Frech & Samprone, 1980).

Buchanan and Tullock (1962) maintain that due to the ambiguity of what can be ascribed as the motive to act in the public interest, regulation in the public interest should focus upon the efficiency (or lack thereof) of the market and not as a redistributive tool to correct injustices between producers and consumers. MacAvoy (1979) contests that the United States government enacted legislation that was solely concerned with market efficiency. Acting within the public interest could be viewed as introducing regulation that addresses the concerns of society that producers do not fully account for the social costs of their activities. Indeed, he maintains that legislation, such as health and safety regulations and pollution restrictions, are amongst several reasons why government may intervene in a market.

Redford (1954) also highlights that in the pursuit of the public interest, a compromise amongst competing values is required to address conflicts that arise between short and

long-term interests. The role of knowledgeable specialists in the regulatory process is well-documented by the author. Many issues involving the regulation of health insurance require a certain degree of technical competence, it is therefore necessary for the authorities to employ skilled staff that is able to understand the implications of certain regulations. Regulatory agencies need to monitor implementation and enforce compliance. Redford (1954) encourages the involvement by specialists (even those outside of the regulatory agency such as consultants) in the regulatory process as he argues they are more likely to provide practical solutions.

Schubert (1957) extends Redford (1954)'s assertions by implying that regulators act as benevolent entities that create the appropriate environment in order to pursue the public interest. Thus, the deployment of technical expertise within the regulatory agency, which may occur via contracting out to consultants with the requisite knowledge to assist with certain technical matters, could be viewed as one of Schubert (1957)'s so-called "creative forces" that help in the pursuit of the public interest. It is questionable to believe that regulators possess the aptitude and perhaps more importantly, willingness to pursue the public interest. Individuals who are appointed as regulators often have undue influence in setting the agenda to promote the public interest and therefore, Friendly (1962: 1294) stresses the importance of appointing regulators that have "...higher intellectual power and moral courage". The author emphasises that regulators should have an arms-length relationship with the regulated industry (or market). The role of the regulator should be seen as an adjudicator rather than as "...a business manager" (Friendly, 1962: 1285).

Important within the context of the regulatory environment in the health insurance sector, Adams and Tower (1994) argue that a lack of information (information asymmetries) and disclosure can warrant regulatory intervention. The assumption that individuals possess superior information with respect to their risk than the insurance company, and their unwillingness to share such information may not be due to some "moral defect" in character, but rather as the authors put it individuals' ignorance with respect to the value of such information to the insurer in order to assess an

individual's risk. They argue that the absence of adequate information with respect to either party to a transaction may cause a breakdown in market signals and therefore lead to malfunctions within the market. It is difficult to accept such a view in the health insurance context when one considers individuals that would elect not to share information with the insured simply because they are unaware of the relative importance of such information to the insurer. However, if one considers the situation where consumers lack information concerning the financial soundness of an insurer or technical provisions within a policy, it is quite acceptable to argue that the regulator should ensure that consumers have enough information to make optimal purchasing decisions.

In fact, Adams and Tower (1994) argue that the primary function of the regulator should be to ensure that consumers and producers have all the relevant information timeously so as to make optimal decisions. The regulator should not attempt to manage the activities of either producers or consumers (Adams & Tower, 1994; Booth, 1997). The regulator's objective of reducing information asymmetries is not designed to mitigate the risk of insolvency. In truth, if the regulator acts as a facilitator for the free-flow of information it is conceivable that consumers (or for that matter producers) may misinterpret this information. Consumers may simply ignore information made available to them. Moreover, it is not the purpose of regulation to eliminate insolvencies, insolvencies are a natural feature of any market whereby inefficient producers are 'forced' to exit. As Booth (1997: 680) points out "...insurance failure should, therefore, never be regarded as a *prima facie* case for further regulation".

Posner (1974) argues that if public interest theory were correct, regulation would mainly be observable in highly-concentrated industries or industries that generate significant externalities. Theoretical and empirical research suggests that this is not the case (Posner, 1974). The regulatory process is often marred by inefficiencies and unintended consequences, and thus the assumptions behind the public interest theory and its associated explanatory power is questionable. It is suggested that the

disappointing performance in the regulatory process can be explained by weaknesses in personnel and procedures rather than inherent flaws in the nature of the process or its objectives.

According to Booth (1997), there are three key shortcomings of the public interest view: First, it only considers one aspect of market failure. The second shortcoming is that the public interest view assumes that the correction of a market imperfection will automatically improve economic welfare. The author contends that the final shortcoming is that under the public interest view there is a temptation to consider any market problem as a market failure requiring regulatory intervention. Therefore, authorities may be inclined to introduce any number of regulations to direct any activity by insurers (or consumers) that is considered to be in the public interest. This is underscored by the following statement:

“There is more incentive for the law to grow, finding more and more ‘special cases’, which require regulation to correct failures in the market...it may be tempting for regulators to intervene to try to perfect what they regard as an imperfect market at every possible opportunity” (Booth, 1997: 683).

Booth (1997) is merely reiterating John Stuart Mill who articulated this idea more than a century ago:

“...the strongest of all the arguments against the interference of the public with purely personal conduct is that when it does interfere, the odds are that it interferes wrongly and in the wrong place” (Mill, 1859/2010: 54).

A reformulated version of the public interest theory proposes that regulatory agencies are created for *bona fide* public purposes. But these agencies are then mismanaged leading to potentially unfavourable (and unintended) consequences. This reformulation of the public interest theory remains inadequate since socially undesirable results of regulation are frequently sought by interest groups that have an influence over the development of the regulatory framework (Posner, 1974). Further, the evidence to suggest that mismanagement occurs within the regulatory agency is

surprisingly weak. Much of this evidence is consistent with the rival, capture theory of regulation:

“...where the typical regulatory agency operates with reasonable efficiency to attain deliberately inefficient or inequitable goals set by the legislature” (Posner, 1974: 337).

Doron (1979) departs from the public interest theories of regulation by proposing yet another reformulation. He suggests that in some cases while regulation may be initiated to advance the public interest, the practical consequences may simply be compatible with the self-interest of the industry. To demonstrate this view the author examines the effects of regulation imposed upon the tobacco industry.

Posner (1974) reports that the neutrality of the regulator often breaks down and this leads to a departure from the objectives of public interest theory. Meier (1991: 708) proposes three possible reasons for the regulator's failure to protect and promote the public interest:

1. bureaucratic incompetence;
2. the lack of skills and resources internally; and
3. the complexity of technical issues

First, the regulatory agency will often fall victim to disinterest by staff and this may translate into incompetence. Wilson (1980: 147) suggests that bureaucratic incompetence is rather the consequence of inflexibility and additional constraints that are a feature of the political system. Second, Mitnick (1980) attributes the inability of regulatory agents to operate effectively to a lack of required skills, capacity and financial resources. Meier (1991) cites the lack of regulatory capacity as the major contributing factor for the sharp rise in insolvencies in the US insurance sector during the 1980s. The third reason relates to the technical complexity that is often associated

with insurance products and issues, particularly, in the case of life and health insurance.

Smith, Louberge and Shapiro (1986) put forward a fourth possible reason for the failure by regulators to safeguard the public interest. Regulators are not benevolent participants, but rather they pursue their own self-interest. Mayers and Smith (1981) explains that regulators may deliberately introduce additional rules and encourage further regulation to preserve their function and satisfy their political masters. Posner (1974: 340) identifies that a serious flaw in any version of the public interest theory, is that it contains no linkage or mechanism by which a perception of the public interest is translated into legislative action.¹⁶⁰ It is not sufficient to presume that a voter will vote for a particular candidate, who promises to carry out policies that are perceived to be in the public interest, there may be other policies that benefit the individual voter more (Posner, 1974: 340). According to Posner (1974), the public interest theory fails as a conceptual framework to explain the observed behaviour of participants in the regulatory process. Rahman (1992) concludes that public interest theory has two shortcomings; it does not take into account economic realities and it does not acknowledge that governments are made up of self-motivated interest groups.

5.1.2 Capture Theory

Capture theory maintains that theory of economic regulation is not remotely about the public interest, but rather a process by which interest groups seek to promote their own self-interests (Posner, 1974: 341). Regulation is viewed as “a partisan political process conferring benefits upon politically effective groups, which capture and dominate the regulatory process” (Adams & Tower, 1994: 167 cite Noll & Owen, 1983 and Reagan, 1987). Posner (1974) distinguishes between two hypotheses of capture theory; namely, the Marxist/Muckraker view and the Political Scientist perspective.

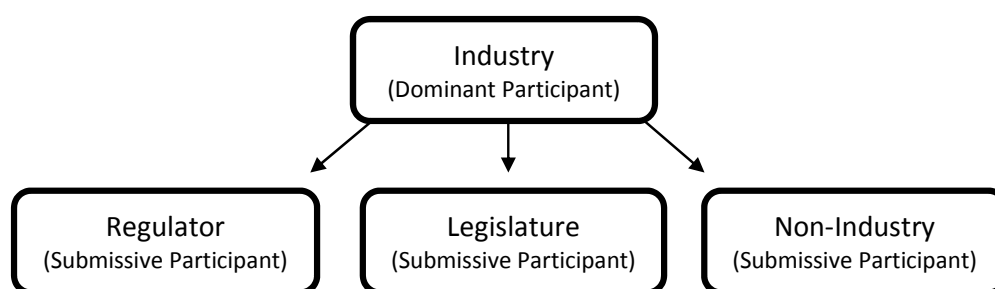
¹⁶⁰ “In the theory of markets, the efforts of individuals to promote their self-interest through transacting brings about the efficient allocation of resources” (Posner, 1974: 340).

The Marxist/Muckraker view is a more radical interpretation of capture theory, whereby capitalists are considered to be a dominant participant in the regulatory process. The regulatory process promotes the interests of the capitalist whilst simultaneously alienating labour. Adams and Tower (1994: 167) suggests that where markets have relatively unregulated industries and weak protection for consumers, it is indicative from the radical perspective that capitalist (corporate) interests are dominant to the detriment of the public. The idea that large corporations ultimately control regulation is defective, given that a great deal of economic regulation also serves the interests of smaller firms or non-industry groups (Posner, 1974 and Stigler, 1971).

Political scientists do not subscribe to the radical view advocated by the Marxists and Muckrakers. But rather considers the regulatory process as the outcome of opportunistic capture by politically effective groups. Figure 5-2 illustrates the conceptual capture theory model ascribed to by the political scientists. Industry groups are successful in capturing regulatory favour, because according to Feroz (1987) they possess:

1. superior economic resources;
2. interest in the beneficial outcome that can be derived by controlling the regulatory process;
3. sound organisational capabilities; and
4. a comparative advantage with respect to industry-specific technical knowledge.

Figure 5-2: Adams and Tower (1994: 168)'s Political Scientist View of Capture Theory – Conceptual Model



Stigler (1971: 4-6) discusses the four primary reasons why industry groups devote resources in an attempt to capture the regulatory process:

- *Direct subsidy* – An obvious avenue for industry groups is to seek a direct cash subsidy from government. For example, in the 1960s almost all universities in the United Kingdom were publically funded. Although more recently there has been an irrevocable change in the balance between public and private funding, direct public subventions still account for on average around two thirds of total funding for universities in the United Kingdom (Greenaway & Haynes, 2003).¹⁶¹ Stigler (1971) points out that industry groups with significant influence do not tend to seek out cash subsidies, because unless the number of beneficiaries can be restricted through an effective device, the subsidy funding will be distributed more thinly amongst a growing number of rivals.
- *Control over Entry* – Many devices are sought by industry groups to limit the rate of entry by new rivals into established oligopolistic industries. These include peculiar pricing (rate) regulation and vertical integration. It is likely that this form of pricing regulation introduced through the concept of community rating in South Africa's health insurance sector, has permitted medical schemes and their associated administrators to consolidate their dominant positions in the marketplace.
- *Substitutes and Complements* – Industry groups will often make an effort to influence regulation with respect to substitutes and complements. Stigler (1971)

¹⁶¹ Greenaway and Haynes (2003) also discuss how the mechanisms to disburse public funding to tertiary educational institutions have changed. There has been a shift away from block grants to fund teaching and research, to earmarked funding, which is partly formulaic and partly performance-based.

uses the example of butter producers, who wish to suppress margarine and encourage bread production.

- *Price Controls* – Price controls are essential to achieve competitive rates of return. Stigler (1971) notes that if there are an adequate number of firms within an industry, price discrimination becomes problematic in the absence of government intervention.

Authors such as Gormley (1986), suggest that the ability of industry groups to capture the regulatory process is directly related to the degree of complexity within an industry. Politicians that find issues too technical or complex within an industry are likely to discount their importance. Therefore, under these circumstances politicians would favour a system of self-regulation (Meier, 1991).

Posner (1974) questions the validity of capture theory viewing it as a hypothesis rather than a theory with theoretical foundations. Posner (1974: 342) takes exception to the characterisation of the interaction between the regulatory agency and the industry group as a “metaphor of conquest”. Rahman (1992) criticizes the political scientist’s view of capture theory since it has weak explanatory and predictive power. Moreover, Rahman (1992: 113) states that:

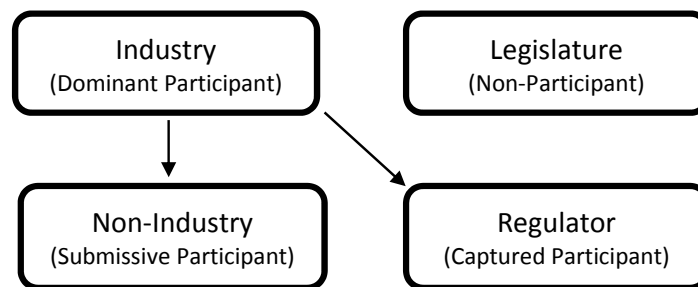
“...no reason is suggested as to why the regulated industry should be the only interest group able to influence a regulatory agency...Lastly, it ignores a good deal of evidence that the interests promoted by regulatory agencies can also be in accord with those of other groups rather than that of the regulated firms alone”.

5.1.3 Economic Theory of Regulation

The economic theory of regulation was first articulated by Stigler (1971). At first glance, it appears to be a refined version of capture theory. It discards the assumption of “pristine legislative purpose” and admits the possibility of other interest groups

“capturing” the regulatory process (Posner, 1974: 343). Stigler (1971) expands the tenets of capture theory to embrace the concept of public choice. Public choice is described by Mueller (1976: 395) as “... the economic study of non-market decision making or simply the application of economics to political science”. The theory perceives regulation to be an economic good governed by the traditional laws of supply and demand. Figure 5-3 describes the relationship between participant groups under the economic theory of regulation.

Figure 5-3: Adams and Tower (1994: 171)’s Economic Theory of Regulation – Conceptual Model



Stigler (1971)’s contention with the notion of a “pristine legislator” was its inconsistency with actual experience. He discusses an example in the United States where oil import quotas were introduced. The quotas were designed to protect the public by ensuring that there was a stable oil supply during times of conflict. The author contends that a tariff-based system would have been better as it would generate a significant stream of revenue for the government, and these funds could be allocated in such a way as to shore-up the oil supply chain to protect the United States economy during times of conflict. The oil quotas had the effect of merely benefitting the local oil producers at the expense of the consumers. Moreover, he answers the question as to why a well-organised and powerful special interest group like the oil industry, which is able to take advantage of the regulatory environment, would seek the coercive powers of the state rather than a cash subsidy.

This example serves to exemplify the underlying principles behind Stigler (1971)'s argument in favour of the economic theory of regulation. If the government provided a generous cash subsidy to local oil producers, the industry would see an influx of new entrants eager to receive these subsidies and thereby hinder incumbents' ability to extract economic rents (Stigler, 1971). The allure of cash subsidies may be a motivator for companies to actively lobby for cash subsidies from the state, but the dilution of these subsidies through the influx of new entrants suggests that there are even greater gains to be had from cultivating the coercive powers of the state. Stigler (1971: 4) describes the coercive powers of the state in the following bold statement:

“The state has one basic resource which in pure principle is not shared with even the mightiest of its citizens: the power to coerce. The state can seize money by the only method which is permitted by the laws of a civilised society, by taxation. The state can ordain the physical movements of resources and the economic decisions of households and firms without their consent. These powers provide the possibilities for the utilisation of the state by an industry to increase its profitability.”

An example of the regulation that favours one particular industry over another competing industry is discussed by Stigler (1971) in his examination of the railroad industry versus freight transportation by road. In the United States at the time of the *Great Depression*, public investment in infrastructure benefited the interstate and highway road network, which in turn saw the trucking industry gain market share in the long-haul freight business. This prompted the railroad industry to lobby for stricter regulation of the long-haul freight business. The railroad industry managed to gain assistance and protection from the government, but according to the author the costs to society were not offset by the benefits that the railroad industry enjoyed.

To understand how it is possible for certain special-interest groups to acquire the coercive powers of the state, it is necessary to examine the political process and its interaction with markets. Stigler (1971) points out that the political process is vastly different from the processes that govern markets. If a referendum is called on a

particular issue, say to prohibit the drinking of tea, and the outcome of the referendum is that the majority is in favour of this prohibition. The result of referendum would have to be followed by the entire population of the country, despite your individual vote (and objection) against such a proposal. In this example, it is clear that the political process underlying a democracy requires everyone to take a decision simultaneously. The political process incurs great costs as everyone's opinion has to be gauged to decide a particular issue at a single point in time. This cost may simply be transferred when voters elect representatives that are employed to take decisions on their behalf. Therefore, Stigler (1971: 10) argues that:

“...the political decision does not predict voter desires and make preparations to fulfil them in advance of their realisation”.

A further observation is that the political process requires the involvement of society as a whole, and not merely individuals who are directly involved or those that have an inherent interest in the outcome of the political decision. This is in contrast to the market where this is not required. For example, individuals who are afraid of flying, do not need to make choices between air travel and a train journey (Stigler, 1971). Companies that transport goods over long distances, however, would need to continuously evaluate the costs and benefits from both modes of transport before arriving at a decision. As Stigler (1971) suggests the company would essentially be “voting” on the issue on a continual basis. The political decision making process must include everyone, it cannot “...allow participation in proportion to interest and knowledge” (Stigler, 1971: 11).

The political process therefore suffers from a potential lack of participation by voters (apathy if you will). This non-participation may occur as a result of a number of reasons – it may simply be the case that individuals are unfamiliar or disinterested in certain issues. To overcome this problem Stigler (1971) suggests that government can effectively segment itself through the decentralisation of decision-making. The segmentation may take place with respect to elected representatives. For example, certain representatives may take up specific causes or merely devote themselves to particular policy issues involving education as opposed to healthcare. The existence of

political parties is simply another method by which politically-charged individuals can organise themselves according to a predefined set of policies. It allows persons that run for political office to seek out the support from specific groups of individuals and should they commit themselves to implement the wishes of their constituency, they also conceivably increase their likelihood of being elected.

The desire to be re-elected and remain in office, may encourage certain representatives to withdraw their support for a particular economic policy that would be seen to be harmful to their constituents (Stigler, 1971). One example is some politicians may seek favour by voting against austerity measures, which increase taxes and cut welfare benefits or social grants. He does suggest that a representative who could be assured of re-election if they voted against say austerity measures would undoubtedly do so. Special interest groups are not precluded from benefiting from the desires of representatives to be re-elected. For example, if a representative were to vote against legislation that favours a certain industry through subsidies or other financial support may encourage these industry members to support other candidates. Stigler (1971: 12) states it bluntly “[t]he industry which seeks political power must go to the appropriate seller, the political party”. In summary, Stigler (1971)’s view is that regulation is an economic good and the political process that supports this regulation has associated costs which must be met by industries who seek to capture the coercive powers of the state. The author’s argument appears to support the view that government intervention is driven by the desire of elected representatives to maintain political office.

Peltzman (1976: 212) contends that industry groups’ interests still prevail in the market because they possess informational advantages over consumers or politicians. Regulation is supplied by policymakers as long as the demand from politically effective groups exceeds the opposition to the regulation (Rahman, 1992: 115).

Unlike public interest theory, the economic theory of regulation asserts that government intervention does not attempt to remedy market imperfections, but rather like capture theory it accommodates the idea that regulation exists to promote the interests of politically effectual industry participants (Adams & Tower, 1994: 170 cites Stigler, 1971). It is further suggested that within an industry that has many competitors, the economic theory of regulation predicts that the demand for regulation will be greatest. Conversely, where there is a highly-concentrated industry with fewer companies, it is likely that an industry's interests will be best served through self-regulation (Adams & Tower, 1994).

Posner (1974) acknowledges that Stigler (1971)'s economic theory of regulation requires refinement and he goes further to highlight the absence of empirical support favouring this view of regulation. However, he does admit that the economic theory of regulation is far more compelling than the public interest view of regulation. Posner (1974) does not dismiss the possibility that regulation (and their corresponding regulatory agency) may be created with good intentions and it can be viewed as an honest attempt to pursue the public interest. It is difficult to believe though that even if a market failure is correctly identified, regulators will possess the appropriate interventions to correct this market failure and bring about a more efficient outcome. He argues that the lack of technical knowledge and the sheer complexity of markets is a significant obstacle in preventing regulators from achieving what may be their well-placed objectives.

In addition, the costs of legislative supervision of regulators are generally prohibitive. The process whereby legislators engage with the private sector via negotiation and consultation incurs further costs. These costs may not simply be of a financial nature and can include after all the need to devote an inordinate amount of time to understand the complexities of industry dynamics, engage in difficult negotiations and develop appropriate regulatory interventions. Therefore, Posner (1974) proposes that legislators will increasingly delegate duties to regulatory agencies and as a result they will relinquish control over them. The so-called life-cycle theory of regulation

promotes the idea that in the early stages of a regulator's life-cycle, legislators are prepared to devote time to tackle regulatory-related issues and oversee the functioning of the regulatory agency. However, over time this interest and willingness dissipates as attention is drawn away to other areas of concern. Considering all these factors, he quite cynically comments that it is no wonder that regulatory agencies fail in fulfilling their key objectives. In fact, there failure may distort the efficient functioning of regulated markets (Posner, 1974).

5.2 Chicago School of Regulatory Capture

Stigler (1971)'s economic theory of regulation does not preclude the possibility of regulatory capture. Unlike the authoritarian notion of capture by industry groups in capture theory, economic theory argues that through the superior information that regulated firms possess they are able to capture the regulatory process for their own benefit (Peltzman, 1976). Boehm (2007) distinguishes between *regulatory capture* and *regulatory opportunism*. *Regulatory capture* occurs when regulated firms capture reforms for their own narrow interest. *Regulatory opportunism* on the other hand describes the situation whereby regulations are captured by politicians, who abuse their regulatory powers for their own purposes. Corruption, regulatory capture and regulatory opportunism represent transaction costs of regulation, and they undermine the stated objectives of reforms (Estache & Martimort, 1999).

The capture or interest-group theories can be traced back to Montesquieu and Marx. Karl Marx promoted the view that large corporations controls institutions and consequently controls regulation, which adversely affects the working classes (Laffont & Tirole, 1991). Stigler (1971) extended the paradigm to include the possibility that smaller organisations and interest groups may also have a direct influence upon the regulatory process. Moreover, utilising, Buchanan (1965) and Olson (1965)'s theory of collective action, it was possible for Stigler (1971) to explain how "regulation is acquired by the industry, and designed and operated primarily for its benefit" (Stigler, 1971: 3). Stigler and Friedland (1962) was a precursor to Stigler

(1971)'s seminal paper on the theory of economic regulation, where the authors investigated the influence regulators had upon prices in the electricity sector in the United States. Stigler and Friedland (1962) undertook a regression analysis, where prices were a function of various variables and a dummy variable for regulation. It was concluded that there was no discernable difference between regulated and non-regulated public utilities. Therefore, they proposed the hypothesis that regulation's objective is to promote the interests of industry groups rather than maximise social welfare.

George Stigler (1971), Sam Peltzman (1976), Richard Posner (1971, 1974, and 1975) and Gary S. Becker (1983, 1986) constitute the main contributors from the Chicago School. These contributions are also referred to as the interest-group theories, since they are based upon Olson (1965)'s collective action argument. The emergence of powerful consumer groups and the regulatory experience during the 1970s led some researchers such as Peltzman (1976) to take a broader view of Stigler (1971)'s contribution (Laffont & Tirole, 1991: 1090). Government agents arbitrate among competing interests and this is not necessarily in favour of business.

Although, Peltzman (1976) formalises Stigler (1971)'s economic theory of regulation, he reaches a different conclusion. He asserts that frictions that are present within the political process hinder a dominant group's expansion and their ability to extract all the potential benefits. The notion is predicated on the view that well-placed special-interest groups that wield considerable influence upon regulators, cannot completely capture a regulatory agency because the regulator will have sufficient political pressure to "...exclusively [serve] a single economic interest" (Peltzman, 1976: 211). Peltzman (1976: 212) contends that industry groups' interests still "prevail" in the market because they possess informational advantages over consumers or politicians. The author put forward the following equation whereby the legislator attempts to maximise net votes or a majority M :

Equation 5-1: Expected Number of Votes for the Legislator (Peltzman, 1976)

$$M = nf - (N - n)h$$

where M represents the expected number of votes that the candidate would receive, that is the chance of re-election, n is the number of potential voters in the beneficiary group, N is total number of potential voters, f represents that net probability that a beneficiary will grant support and h represents the net probability that he who is taxed ($N - n$) opposes the legislator.

The model Peltzman (1976) sets out a regulatory process whereby a transfer of wealth is at stake. More particularly, he considers the political process as if control of the relevant taxing authority rests upon direct voting. He assumes that there are two groups, the beneficiary group who ultimately benefits from political favours by the legislator and a non-beneficiary group that receives no favours. Now, beneficiaries pay with both votes and dollars, where the dollars for the legislator serve to mitigate against opposition (Peltzman, 1976). In Peltzman (1976: 214)'s model, "...direct political support [votes] is the object sought directly by the regulator". The legislator seeks to maximise net votes or a majority in his favour. The author asserts that greater majorities are assumed to imply greater security of tenure. According to Peltzman (1976), the crucial decision that the regulator (or would-be regulator) has to make is with respect to the numerical size of the group he promises favours to and hence, the reciprocal group that he or she taxes.

Further assumptions that Peltzman (1976) makes include that:

- f and h are not equal to zero or unity, and depend upon the amount of the group member's gain or loss;
- Gains and losses are equal per capita within groups;
- Ignorance does not lead to perverse or biased voting – in other words if a beneficiary is not sufficiently informed to vote for his or her benefactor, then his or her voting decision are not biased in favour of or against the benefactor. Peltzman (1976) suggest that the person either abstains or votes by tossing a fair coin.

Peltzman (1976) defines the probability of support as follows:

Equation 5-2: Probability of Support (Peltzman, 1976)

$$f = f(g)$$

Where g is the per capita net benefit, and is given by:

$$g = \frac{T - K - C(n)}{n}$$

where f represents that net probability that a beneficiary will grant support, T is the total dollar amount transferred to the beneficiary group, K represents the dollars spent by beneficiaries in campaign funds, lobbying etc. to mitigate opposition and $C(n)$ represents the cost of organizing both direct support of beneficiaries and efforts to mitigate opposition. This organization cost increases with n , but no restrictions are stipulated with respect to the shape of the marginal cost curve (Peltzman, 1976).

From Equation 5-2, the number of votes in support depends upon n , in the following two offsetting ways: a larger n translates into a broader base of support, but at the same time dilutes the net gain per member and therefore the probability of a member's support would also decline (Peltzman, 1976). He proposes that the legislator chooses both K and T . Now, the total dollar amount transferred to the beneficiary group T is described in Equation 5-3.

Equation 5-3: Total amount transferred to the Beneficiary Group (Peltzman, 1976)

$$T = tB(N - n)$$

where T is the total dollar amount transferred to the beneficiary group, t represents the corresponding tax rate on B , which is the wealth of each member outside of the beneficiary group. n is the number of potential voters in the beneficiary group and N is total number of potential voters (Peltzman, 1976).

Peltzman (1976: 216) proposes that the “[o]pposition is assumed to be generated by the tax rate and mitigated by voter education expenditures per capita (z)”. Therefore,

h , which represents the net probability that he who is taxed (from the non-beneficiary group) opposes the legislator, can be determined as follows:

Equation 5-4: Probability of Opposition (Peltzman, 1976)

$$h = h(t, z)$$

Where z , the per capita expenditure on voter education, can be determined by the following:

$$z = \frac{K}{(N - n)}$$

where K represents the dollars spent by beneficiaries in campaign funds, lobbying etc. to mitigate opposition, n is the number of potential voters in the beneficiary group and N is total number of potential voters (Peltzman, 1976).

To summarise, Peltzman (1976)'s model demonstrates that the legislator will attempt to maximise Equation 5-1, that is maximise M , by considering the competing interest groups; namely the beneficiary group (n) and the non-beneficiary group ($N - n$). It will be necessary for the legislator to assess the relative probabilities of political support provided by the beneficiary group and opposition from the non-beneficiary group, denoted by f and h , respectively (Peltzman, 1976). From Equation 5-1, it is clear that he incorporates his view that a legislator does not serve a single economic interest – one has both the beneficiary and non-beneficiary groups competing for political support (or opposition as it may be). Furthermore, it is noted that as the non-beneficiary group (or public-at-large) increases their involvement in political activity (as the total number of potential voters increase, implicitly so does the total number of members of the non-beneficiary group) there will be a greater chance that the non-beneficiary group's interests being pursued. Yet, if the probability of non-beneficiary opposition (h) declines then it is more likely that the beneficiary group will benefit from the favours of the legislator and by inference have the ability to mould legislation to their benefit (Peltzman, 1976).

Crucial to Peltzman (1976)'s analysis is the assumption that the public is informed and politically active. Stigler (1971) articulated that without a vocal public, there is little incentive for a regulator to act in the public interest. Booth (1997) successfully argues that the regulation of insurance is likely to be an inconsequential issue for voters that are considering the re-election of a legislator.

Becker (1983) presents a model similar to Peltzman (1976), which once again considers the influence of various interest groups within the political or regulatory process. He argues that the influence and ultimately the policy outcomes depend upon the strength of the respective pressure groups. Becker (1983: 372) proposes that the competition:

“...between these pressure groups for political influence determines the equilibrium structure of taxes, subsidies and other political favours”.

Therefore, the crucial outcome from Becker (1983)'s model is that a regulator will accommodate broader economic interests, which contradicts the:

“... all-or-nothing outcomes implied by many other formal models of political behaviour, where the ‘majority’ clearly wins and the ‘minority’ clearly loses”
(Becker, 1983: 372 – 373).

Along the lines of Peltzman (1976), Becker (1983) for simplification purposes, proposes that there are two competing groups, those that pay taxes and those that receive subsidies. As in Peltzman (1976), the model introduces a political process that centres around a transfer of wealth mechanism. Both groups attempt to maximise their income. The author specifically concentrates on impact that the deadweight costs have on the competition between the two groups. For example, an increase in deadweight cost of a subsidy will translate into a lower revenue figure from taxes, which in turns lowers the contributions that can be made to the subsidised group. Hence, the subsidised group will have fewer incentives to lobby for subsidies going forward. Equally, high deadweight costs associated with tax will stimulate greater political pressure from the taxed interest group to reduce their tax liability.

The ability of any group to engender political influence is a function of each group's relative efficiency in putting forward their agenda compared to the other competing groups. What is more the relative efficiency of each group is itself a function of the capacity of the group to restrict *free-riders*. Free-riders create an externality upon the other members of a group, because they effectively increase the costs of lobbying, since they are able to avoid their obligations. Becker (1983) maintains that a group, which is able to restrict the activities of free-riders, will benefit greatly from improved efficiency. As a consequence, he argues that the more successful groups are likely to be small and relatively homogenous. Generally, homogenous groups would imply that the members of each group are identical with respect to all aspects. In the context of Becker (1983), a homogenous group refers to a group where the costs and benefits of pressure are equal. He does acknowledge that larger firms may be able to take advantage of economies of scale and therefore achieve greater success in the political process. This translates into the observation that groups who seek subsidies will generally be small relative to the groups that pay taxes. The author points to the political success of farmers in rich countries and urban dwellers in poor countries as examples. As the total number of taxpayers rise, he argues that the deadweight costs of tax would decline since there is a reduction in the tax paid per capita. Boehm (2007) states that as opposed to Stigler (1971), a single interest group cannot completely dictate their agenda to the legislator since other interest groups are able to exert political pressure as well. If the two groups outlined in Becker (1983)'s model were homogenous, it would imply that both group's political influence would be offsetting – thus, the aggregate influence would be zero.

Becker (1986) examines the issue of regulatory capture and finds support for Peltzman (1976)'s assertions that while the political clout of an industry acquiring regulation is important, Becker (1986: 230) concludes that "...it is not always decisive. The public's interest can be and is maintained in most states". Peltzman (1976) and Becker (1983)'s extensions to Stigler (1971)'s model do demonstrate that the public interest view of regulation is not unattainable. Empirical findings that examine the regulation of the insurance industry are ambiguous at best. Empirical

results published in Joskow (1973), and Frech and Samprone (1980) do favour Stigler (1971)'s hypothesis, whereas Ippolito (1979)'s results are more consistent with Peltzman (1976)'s assertions.

Boehm (2007) notes an apparent shortcoming of Stigler (1971)'s economic theory of regulation, namely, how can the theory explain the proclivity towards deregulation by industry groups. In the South African health insurance sector prior to the promulgation of the Medical Schemes Act No. 131 of 1998, it was evident that there was a period of deregulation. It permitted medical aid schemes to apply principles of sound risk management and managed care in the design of products, benefit coverage levels, contribution rates and healthy-lifestyle incentive schemes. Boehm (2007) argues that these periods of deregulation can be seen as a period of *re-regulation*. In South Africa, the 1989 and 1993 Medical Scheme Amendment Acts are a case in point, where the deregulation of medical schemes was achieved through the reformulation of the legislation.

5.3 Virginia School of Regulatory Capture

The Virginia School has its origin in Public Choice economics and its scholars have the flavour of Austrian economists such as von Hayek, Mises and Schumpeter. Whilst also considering competition among pressure groups for political influence, the Virginia School focuses upon the social welfare implications ensuing from these activities – the *theory of rent seeking*. Boehm (2007) suggests that in a narrow sense, public choice is concerned with the failures of government. The belief is that all political-empowered participants, including politicians and civil-service servants, are self-interested actors that follow their own interests. Moreover, government is thought to have insurmountable difficulties in correcting market failures or, at least, if they can correct these market failures, it is only possible at a cost that is in excess of the costs emanating from the market failures themselves. In a wider sense, he contends that public choice is the application of economic methods of analysis to political institutions and governmental decision-making.

The Austrian school of thought does provide one of the clearest counterarguments against the assertions underpinning the public interest view of regulation. According to the Austrian philosophy, the public interest view is unattainable. The argument that somehow regulators are in possession of all the required knowledge in terms of costs and consumer preferences is fallacious. The public interest will be pushed aside by the motives of self-interested regulators who are led by a narrow incentive set. Another key focus of the Austrian view is that market is the most effective disseminator of information. According to Booth (1997: 693) "...[variety] can exist in a market which can lead to the satisfaction of more preferences than in a regulated market".

The Virginia School acknowledges that some form of insurance regulation will always be present, but it is unlikely that the regulation will address market mischiefs. Hayek (1960: 222) remarks that a:

“...functioning market economy presupposes certain activities on the part of the state”.

The magnitude of regulation is perhaps less of concern when compared to the actual activities underlying government intervention. Hayek (1960) maintains that a centrally planned economy will lead to a market that does not function properly.

Hayek (1960) goes further by arguing that a framework of general laws is the most prudent course of action by government legislators since the impact will be more predictable. A general set of laws should not allow for bureaucratic discretion because any civil servant does not possess all the necessary information to take discretionary steps (Hayek, 1960). Moreover, by permitting bureaucrats to operate in a discretionary environment, it spawns the potential for market participants to receive contradictory signals as to how they should operate. Therefore, he asserts that a more appropriate regulatory framework would be based upon a set of predefined principles that favour generalities rather than specifics. The author contends that regulation for the most part will simply impose additional costs in terms of production and hinder

product innovations. Therefore, the objective of regulation must indeed be meaningful considering the considerable costs that it imposes.

Within an environment of general rules and principles, Hayek (1976) considers the ability of market participants to learn from their mistakes and this capacity to assimilate knowledge allows for the development of a sophisticated economy. He does not believe that the accumulation of knowledge is attainable by a central authority. The following statement emphasises this point:

“Even in the modern welfare societies the great majority and the most important of the daily needs of the great masses are met as a result of processes whose particulars government does not and cannot know” (Hayek, 1976: 2).

Hayek (1976) is adamant that a market should be allowed to develop within an appropriate set of regulations. The author views the development of an economy as a dynamic process, which evolves as market participants adapt to changing circumstances, evaluate promising avenues of commerce and abandon those that prove to be unsuccessful.

As a further extension to this discussion, Hayek (1988) argues that civilisation has only progressed due to ability of individuals to learn through their interactions with one another in terms of trade and exchange. According to the author, government intervention evidenced in Ancient Greece impeded the improvement and cultural evolution that was taking place at the time. He asserts that markets are propelled not by the provocation through government intervention, but rather by the desire for knowledge by individuals. According to Hayek (1988: 46), the development of any institution occurs because of the:

“...experimental process of adaptation to unforeseen change by the observation of abstract rules which, when successful could lead to the increase of numbers and the formation of regular patterns”.

Booth (1997) concurs with Hayek (1976, 1988)'s view concerning the importance of knowledge accumulation and suggests that the "litmus-test" for the introduction of any regulation should be whether or not it undermines the ability of market participants to acquire knowledge once mistakes are made. He maintains that the presence of deficiencies that hinder participants from acquiring knowledge from adverse events is the true test to determine if regulation is warranted. For example, the introduction of measures to avoid the repetition of past failures may prevent market participants from acquiring this knowledge.

Booth (1997) strongly objects to any attempt by regulators to protect investors from risk. The global financial crisis that began in 2008 is a point in question, whereby governments came to the rescue of financial institutions that essentially failed to adequately measure and manage the inherent risks they took on in their "blind" endeavours to earn superior returns. He argues that regulations to safeguard investors from their own decisions can easily be construed as protecting investors from market mechanisms. Attempts to shield investors from risk may simply be political and according to Booth (1997) the timeframe in which investors may acquire knowledge to correct past mistakes may be too prolonged and therefore untenable. Hence, the desire by regulators to respond promptly to financial scandals encourages regulatory action.

Tollison (1998) states that interest-group theory is concerned with lobbying and the theory of rent seeking is concerned with the costs associated with lobbying. Tullock (1967) demonstrated that in addition to the welfare loss attributable to the market power of a monopoly, there is additional wastage of resources to protect the monopoly's position against competition and regulation (Boehm, 1997). Other contributors to the Virginia School include Buchanan and Tullock (1962) and Downs (1957). Downs (1957) attempts specifically to integrate government and private decision-makers into a single general equilibrium theory. To summarise, regulation is

created not to serve the public interest, but rather to create and protect oligopolistic rents.¹⁶²

5.4 Tollbooth Theories

Politicians and bureaucrats are viewed as active participants in the capture of the regulatory process, whereby they attempt to extort rents from private firms and industries (Soto, 1989; Shleifer & Visny, 1994; and Porta, Lopez-de-Silanes, Shleifer & Visny, 1998). Under the Tollbooth theory, the political establishment and regulatory agents may erect obtrusive regulation to simply extract benefits from regulated industries.

Corruption becomes an avenue for regulated industries to circumvent these inefficient regulations. In the process of his research, Soto (1989) and a group of researchers undertook an experiment whereby they established a small business in Lima, Peru. The idea was to attempt to establish a small business by adhering to all regulations and without paying any form of bribe to hasten regulatory procedures (Soto, 1989). During this process, the researchers were approached no less than ten times by unscrupulous regulatory agents seeking to extort bribes. In two of the cases, it was necessary to pay the bribes otherwise it would have been impossible to establish the business. The process to establish the business took all of 289 days (Soto, 1989). Thus, according to the view espoused by the Tollbooth theories, corruption is often considered as necessary to overcome unfair or inefficient legislation and regulations (Boehm, 1997).

Shleifer and Vishny (1994) examine the privatisation of state-owned enterprises and create a bargaining model that investigates why regulators seek to maintain control over privatised firm through the regulatory process. In line with the Tollbooth theories

¹⁶² According to rent-seeking theories, these oligopolistic rents may be fully or partially wasted in the industry's attempt to compete for these rents.

of regulatory capture, regulators are reluctant to relinquish control since they are able to extract rents from the privatised firm through corrupt practices. Examples of which occur in many countries and Shleifer and Vishny (1994: 1007) observe that the heightened regulatory control:

“...might shed light on the large amount of corruption [present] in countries [such as] Italy or the Philippines, where firms are privately owned...and [these firms] pay enormous bribes to politicians who control them through regulation.”

Boehm (2007) suggests that Shleifer and Vishny (1994), however, remain vague concerning the type of regulation imposed upon newly privatised-firms in order to extract these rents through coercive means. Shleifer and Vishny (1994) also remain silent with respect to the structure of the industry in which the privatised firm finds itself and according to Boehm (1997), Shleifer and Vishny (1994) fail to address the significant trade-off between the opportunities for corruption through introduced regulation and the costs related to market imperfections if privatised firms are left unregulated.

Djankov, Porta, Lopez-de-Silanes and Shleifer (2002) empirically test for the existence of the tollbooth theory. Regulations considered by Djankov et al. (2002) include entry regulation, start-up requirements, number of procedures, official costs of following these procedures and minimum official time for a start-up to begin operations. As stated by Boehm (1997), the regulations considered should ensure for example, a minimum level of quality of goods that are supplied by new entrants, and further hinder entrants from operating according to *fly-by-night* strategies. Djankov et al. (2002) find that more regulation induces higher levels of corruption and larger underground economies. Furthermore, Djankov et al. (2002) find that efficient governments do not impose significant entry regulations and according to Djankov et al. (2002) provides evidence against the public interest theories of regulation in favour of the public choice view.

Rose-Ackerman (1999) warns against inferences that perceive certain regulations to be ‘unfair’ or ‘inefficient’. Rose-Ackerman (1999) argues that it provides for example, an excuse for multinational companies that operate in developing economies to engage in corrupt practices, since they view certain regulations as being unfair (or inefficient) and therefore opportunities to circumvent these become justifiable. This gives firms license to decide which laws and regulations they consider to be just and efficient and therefore by implication those that should be adhered to. Rose-Ackerman (1999) insists that such conduct would certainly not be tolerated in most developed economies. Many believe environmental or health and safety regulations are unfair as they confer competitive disadvantages upon certain companies (Boehm, 1997).

The consideration that somehow a well-placed payoff is justified as it leads to higher profits in the future is intolerable (Rose-Ackerman, 1999). These corrupt practices, according to Rose-Ackerman (1999), are extremely harmful to the development of credible democratic institutions, which assist in stimulating economic growth and the general upliftment of the population within developing economies. Boehm (1997: 8) conclude that:

“...[e]ven if it is certainly true and perhaps even inevitable that certain laws are inefficient or even unfair – [where does one draw the line?]”.

5.5 Toulouse School of Regulatory Capture

Laffont and Tirole (1991, 1993) assimilate both the Chicago and Virginia Schools to arrive at a theory of regulatory capture that considers two important elements; namely, asymmetric information and the principal-agent conflict. These contributions have been broadly grouped as part of the Toulouse School of Regulatory Capture. They state that in the absence of information asymmetries, firms would be unable to extract rents and thus, the theory of regulatory capture based upon rent-seeking would fall short. Spiller (1990) discusses the potential principal-agency conflict that can arise between regulator and legislature.

Levine and Forrence (1990) examine the merits of the public interest and capture theories of regulation within the context of agency and information theory. The notion that governing institutions operate as a “black box” is a drawback of both the public interest and capture theory (Levine & Forrence, 1990: 171). The authors suggest though that capture theory is far more successful in describing regulatory activities and serves as a suitable theoretical construct from which further research can be undertaken. It is the contention of the authors that most governmental institutions (including regulatory agencies) are able to operate without much oversight by the public or by legislators, who should themselves answer to the electorate. Thus, the difficulty with regulation is

“...the inability of voters or their intermediaries to effectively limit and control regulation within the complex political system” (Levine & Forrence, 1990: 171).

Levine and Forrence (1990) acknowledge the concerns raised by Posner (1974), whereby they also recognise the cost of acquiring accurate information with respect to a regulated industry and the uncertainty that pervades interactions between regulator, industry representatives and the public-at-large. It is these concerns that form the basis for the analysis undertaken by Levine and Forrence (1990), whereby they attempt to explain how the electorate’s concerns are translated into policy.

Kalt and Zupan (1984) attempt to judge:

“...the empirical importance of the altruistic, publicly interested goals of rational actors in determining legislative and regulatory outcomes” (Kalt & Zupan, 1984: 279).

Levine and Forrence (1990) utilise Kalt and Zupan (1984)’s framework to examine the situation where the principal-agent relationship between the voter and the politician ultimately leads to what can be referred to as *slack* (frictions resulting from

agency conflicts).¹⁶³ This feature permits a politician to side-step his or her responsibility to the electorate. Levine and Forrence (1990) assert that *slack* enables special interest-groups to persuade legislators to treat them favourably. Levine and Forrence (1990) do recognise that this particular principal-agent view does deviate somewhat from the traditional capture theory of regulation in the following manner:

“...unlike special-interest policies, they are not ‘sold’ to subsets of the polity in return for support. Rather, these acts or policies are other-regarding” (Levine & Forrence, 1990: 177).

This implies that under certain circumstances, where agency costs together with other informational costs are significant, the public would be unable to recognise that a particular issue is in the public interest and therefore, they might now demonstrate support for the regulator (Levine & Forrence, 1990). Levine and Forrence (1990) suggest that under these circumstances, regulatory policy would not be formed to simply cultivate support from a particular industry so as to gain political traction.

Stigler (1971) indicates that *slack* encourages politicians to maximise their own private utility, which is precisely the view shared by Levine and Forrence (1990). In addition, Levine and Forrence (1990) attest to the fact that regulators who seek out industry support do so not only to remain in office but also to enhance their future employment opportunities in the private sector. Regulators do have a choice to either cultivate *slack* or indeed consume it, and therefore Levine and Forrence (1990) maintain that regulators do not necessarily “default” to the capture route. According to Levine and Forrence (1990) this can be seen:

“...when a regulator has slack, she can invest it in office holding or wealth by pursuing special interest policies, or she can consume it by pursuing other-regarding policies not favoured by her relevant polity. This slack is valuable either way, and it should not be surprising that regulation is often conducted so as to create or increase it” (Levine and Forrence, 1990: 180).

¹⁶³ Kalt and Zupan (1984) assess the nature and significance of publicly-interested objectives in a particular instance of economic policymaking. They consider the voting patterns of the United States Senate with respect to coal strip-mining regulation.

Levine and Forrence (1990) conclude that the propensity for regulatory capture is in fact a function of the degree to which *slack* is present. In other words, if increased publicity about a certain issue does not meaningfully reduce the level of slack present then the possibility of capture remains. It is not necessarily true that the simple presence of slack will translate into regulatory capture. Legislators may also pursue their own private interests or agenda, which does not coincide with the view held by the public (Levine & Forrence, 1990).

The model described by Laffont and Tirole (1991, 1993: 465) assumes that the regulatory agency (agency) regulates firms' rate of return and prices. The regulated firm (agent) provides the agency with information concerning its costs. The agency has the capacity and time to discover the true nature of the firm – is the firm operating efficiently (at low cost) or inefficiently (at high cost)? The legislature (principal) on the other hand has to accept the information provided by the regulator and hence, an opportunity arises for the agency to collude with the agent to hide information from the principal. Laffont and Tirole (1991, 1993)'s primary contribution to the theory of regulatory capture concerns the incorporation of information asymmetries, which Stigler (1971), Peltzman (1976) and Becker (1983) do not adequately take this into account. Laffont and Tirole (1991) posit that these information asymmetries enable a regulator to favour either the industry or consumers. The superior information conferred upon the regulator enables regulators to entrench their positions. Regulators may utilise this private information to seek higher office or simply improve future employment opportunities within the private sector (Laffont & Tirole, 1991, 1993).

Laffont and Martimort (1997) explore the implications of collusion between agents and its impact on the regulatory process and market behaviour. They consider the problem when information is distributed amongst several agents whose objectives are not aligned with that of their principal. The mechanism that allows agents to align their behaviour to that of the principal is a key cornerstone of organisational design (Laffont & Martimort, 1997). Laffont and Martimort (1997) first outline the

Revelation Principle, which they suggest is a fundamental building-block in understanding this mechanism design problem. According to Laffont and Martimort (1997: 875), the success of this principle in explaining the mechanism is two-fold:

1. "...[I]t provides a simple way to characterize the set of implementable allocations when information is decentralized; [and]
2. [I]t gives the right framework for conducting normative analysis under asymmetric information, i.e., for comparing different allocation mechanisms" (Laffont & Martimort, 1997: 875).

However, Laffont and Martimort (1997) do acknowledge that the *Revelation Principle* suffers from several implausible assumptions:

1. "...[i]t presumes that communication is costless between the principal and the agents" (Laffont & Martimort, 1997: 875);
2. It postulates a *Bayesian-Nash* behaviour between agents, whereby it is assumed that agents behave non-cooperatively. This implies that binding agreements between them are unenforceable or at least "...the principal can prevent these agreements at no cost" (Laffont & Martimort, 1997: 875).

Laffont and Martimort (1997) deviate from the *Revelation Principle* mechanism by considering a mechanism design problem where agents can communicate between themselves and ultimately collude under asymmetric information. Consistent with Laffont and Tirole (1991), Laffont and Martimort (1997) stress the importance of the information structure between agents; private knowledge is a significant determinant of agents' ability to extract rents.

A further extension of the Toulouse School is the so-called life-cycle theory of regulatory agencies as suggested by Martimort (1999), and Estache and Martimort (1999). Boehm (2007: 10) describes the life-cycle theory as follows:

"...a new regulatory agency undergoes a life-cycle: when established during a regulatory reform, the agency is subject to close scrutiny by the government and even by the general public, but with time the attention focuses on other topics and the day-to-day activities of the regulator are less in the spotlight of public attention".

The theory emphasises that at the beginning the regulator faces intense pressure to operate effectively and protect consumers. But this pressure soon dissipates while the pressure by the industry to protect its interests remains constant. Therefore, the regulator becomes more likely to be dominated by the interests of the regulated firms.

Martimort (1999) attests to the fact that the dynamic life-cycle hypothesis is an adequate framework to explain regulatory capture. Over time greater opportunities will avail themselves to interest groups to collude with a regulator due to the accumulation of private information, which is inherent in the relationship between the interest group and the regulator (Martimort, 1999). Therefore, legislators will respond by implementing more stringent rules and limiting the regulator's discretion. As a consequence, a regulated industry would find themselves submerged in bureaucratic layers, which he describes as *bureaucratisation*. The author describes the growth in red tape as "...an optimal dynamic response to the threat of capture" (Martimort, 1999: 931).

More recent contributions concerning the economic theory of regulation and regulatory capture have included Besley and Coate (2003) who find that elected regulators may favour consumers rather than a regulated industry. They recognise that when regulators are appointed, there is a proclivity for regulation to become "bundled" with other issues (Besley & Coate, 2003: 1177). The model that Besley and Coate (2003) propose studies the situation where one has a majority of the population that are consumers and a minority of representatives from a regulated industry. They go further to assume that members of political parties organise themselves along lines other than via regulation. Therefore, Besley and Coate (2003: 1177) assert that when:

"...regulators are appointed, [political] parties may be tempted to field candidates who would appoint pro-stakeholder regulators to further their interests in the public spending dimension ... By contrast, if regulators are elected, their stance on regulation is the only salient issue so [from a political party's perspective] the electoral incentive is to run a pro-consumer candidate".

Grace and Phillips (2008) utilise Besley and Coate (2003)'s contribution when they examine the career prospects of legislators involved in the regulation of the United States automobile insurance market. Grace and Phillips (2008) find that the career prospects of legislators have an observable influence upon how an industry is regulated. In particular, they note that regulators who are not career-politicians, in other words, those who anticipate in the future that they will move into the private sector are most likely to be lenient towards the regulated industry. Naturally, this type of strategic behaviour will not harm their future employment prospects.

But Grace and Phillips (2008) observe that a regulator who simply adopts a soft-touch to regulation *per se* may experience little benefit in terms of future employment prospects in the private sector. It can be argued that a regulator may need to behave far more strategically, in the sense that they need to portray a considerable shift towards a regulated industry. Therefore, starting from a point that is extremely prejudicial towards the regulated industry and then moving gradually (or even rapidly) towards the regulated industry may indeed provide greater benefits as this demonstrates a considerable shift in favour towards the regulated industry that should be rewarded. A further point that Grace and Phillips (2008) make, is before regulators can hope to secure future employment opportunities within a regulated industry they will need to demonstrate their technical competence coupled with their favourable disposition.

For career politicians that will not rely upon future employment opportunities within the private sector, a contrary conclusion is observed whereby they are found to be more likely to favour consumers and thus, apply more stringent regulation (Besley & Coate, 2003; Grace & Phillips, 2008). Grace and Phillips (2008) observe that insurance premiums are found to be almost 5% lower for the case where a regulator has been identified as a consumer advocate. It is clear that the private interests of the regulator are best served currying favour with the electorate that the regulated industry if there is no intention to vacate political office. This is reaffirmed in the case of elected rather than appointed regulators (Grace & Phillips, 2008). However, Grace

and Phillips (2008) do not find that elected legislators will automatically favour consumers over regulated industries. This result is contrary to the widely-held belief that elected officials will always act in the interests of consumers. To illustrate the authors outline the institutional differences between insurance and public utility regulation within the United States. They point out that:

“State public utility commissions are composed of multiple commissioners and the size of the board typically ranges from three to seven commissioners. In contrast, the number of commissioners who exercise authority over insurance rate regulation is generally one” (Grace and Phillips, 2008: 129).

It is believed to be far easier for an industry to monitor and influence the behaviour of a single commissioner as opposed to multiple commissioners. Thus, Grace and Phillips (2008) suggest that as the complexity of the regulatory agency grows and the number of commissioners increase, the marginal cost of influencing the regulatory process will rise as well. Moreover, a regulator’s propensity to be either strict or lenient will also depend upon the extent of monopoly power they hold over private information.

5.6 Conclusion

As Stigler (1971) asserts, government intervention does not attempt to remedy market imperfections, but rather regulation is promulgated to promote the interests of politically-effectual industry participants. According to Peltzman (1976), economic theory argues that through the superior information that regulated firms possess they are able to capture the regulatory process for their own benefit.

Therefore, this thesis concurs with the proposition advocated by Stigler (1971), whereby the new regulatory framework instituted by the South African government through the implementation of the Medical Schemes Act No. 131 of 1998, may have been acquired by the industry or at least by the providers of healthcare.

6 DISCUSSION AND RECOMMENDATIONS

6.1. Contributions and Theoretical Implications

South Africa's private health insurance sector has undergone two major reformulations of the regulatory environment since the late 1980s. The first significant regulatory intervention took place during the late 1980s and early 1990s, when the government sought to deregulate the medical scheme industry in order to address the consecutive underwriting losses that were commonplace amongst medical schemes during the 1980s. As Doherty and McLeod (2003) argued, the Medical Schemes Amendment Act No. 19 of 1989 and the Medical Schemes Amendment Act No. 23 of 1993 was a direct response to calls by the medical schemes industry to free it from the controls that it claimed hampered its ability to address cost escalation.

Under the Medical Schemes Amendment Act No. 19 of 1989, medical schemes were able to apply the concept of risk-rating in their determination of contribution rates and benefit coverage options. The Medical Schemes Amendment Act No. 23 of 1993 removed the statutory minimum benefits and guaranteed payment of claims. Thus, Reekie (1999) believed that the deregulation created a regulatory environment that promoted competition, and removed the obstacles that hindered the application of insurance and risk management principles within the private healthcare financing sector. Evidence to support this view was implicitly observed when one considers the degree of concentration of the medical schemes industry, as measured by the Herfindahl-Hirschman Index (HHI), as it remained relatively sanguine during the 1990s and by 1997, the HHI touched a low of only 0.0265.

A key observation that emerged after the deregulation was that medical scheme coverage became increasingly unaffordable for those who required healthcare services the most; namely, the elderly and unhealthy. Despite Reekie (1999: 4)'s assurances

that “...the [healthcare] cost explosion had been reigned in” once the principle of managed care had become a legal possibility with the deregulation of the late 1980s and early 1990s; Doherty and McLeod (2003) observed that the healthcare cost escalation lingered well into the 1990s. Therefore, the first democratically-elected government in South Africa considered that the Medical Schemes Act No. 72 of 1967 and in particular, the subsequent amendments, were inconsistent with the objective to provide universal access to affordable healthcare for all South Africans. In particular, the government considered that the phenomenon, whereby medical schemes were actively *cherry-picking* the low risks and deliberately excluding those who represented the highest risk to the scheme – the elderly and unhealthy, as a manifestation of adverse selection.

The seminal literature on the theory of adverse selection (Akerlof, 1970; Spence, 1973, 1974; Rothschild & Stiglitz, 1976) is very clear about what constitutes adverse selection. Adverse selection theory predicts that policyholders who are high risk, and have better knowledge about their risk characteristics than the insurance company are likely to purchase higher insurance coverage (lower deductibles). Thus, central to the discussion of adverse selection is the notion that the premiums charged to an individual (or homogenous risk group that they belong to) should reflect their risk status and therefore, align appropriately to the expected benefits that they are likely to receive under a medical scheme policy. If this is not the case, individuals who end up subsidising other individuals that have a higher propensity to require healthcare services than themselves, will be more likely to opt out of insurance. Thus, government’s view of what constitutes adverse selection does not coincide with the seminal literature.

Nevertheless, government advocated the key reason behind the redrafting of the medical scheme legislation was to address the negative consequences of adverse selection. The promulgation of the Medical Schemes Act No. 131 of 1998 and accompanying regulations sought to bring about government’s objective of providing universal access to affordable private healthcare for all South Africans through the

application of three key elements – open enrolment, community-rating and the PMBs package.

Therefore, the primary purpose of this research was to provide an alternative explanation as to why the South African government promulgated the Medical Schemes Act No. 131 of 1998 and accompanying regulation. Therefore, it was necessary to ascertain if adverse selection was indeed the reason behind the implementation of the current regulatory framework governing the private healthcare financing sector in South Africa. To answer the research question, the central issue centres around proving whether the medical schemes industry is indeed plagued by the adverse selection phenomenon. To do this it was first necessary to demonstrate that South Africa's medical scheme industry is sufficiently competitive despite the legislation. After this was confirmed, it meant that the seminal literature discussing the theory of adverse selection would indeed be applicable.

An empirical methodology was thus devised to test for the presence of adverse selection within South Africa's medical scheme industry. An examination of prior empirical studies in numerous jurisdictions and different insurance markets suggested focussing upon using an empirical procedure that would detect a positive coverage-risk correlation. According to the predictions underlying adverse selection theory, policyholders who are high risk, and have better knowledge about their risk characteristics than the insurance company are likely to choose higher insurance coverage (lower deductibles). Therefore, one would expect that there would be discernible coverage and risk relationship, in particular, a positive correlation between coverage and risk. While not a sufficient condition to confirm the presence of adverse selection, it is nevertheless a necessary condition (Cohen & Siegelman, 2010). Hence, the absence of a positive coverage-risk correlation is adequate to discount the presence of adverse selection.

Utilising an empirical methodology that would directly test for the absence of a positive coverage-risk correlation in South Africa's medical scheme industry proved difficult to implement given the aggregated medical scheme data available through the Council for Medical Schemes.¹⁶⁴ Therefore, this thesis utilised a similar procedure to Dahlby (1983, 1992) that applied Akerlof (1970)'s *Lemons Principle* to aggregated data from the Canadian automobile insurance sector. The central premise of Dahlby (1983, 1992)'s empirical investigation was to suggest that if adverse selection is present, it implies that as people exit the insurance market, it is precisely the low risk who would opt of insurance, and this would be reflected in a deterioration of the risk profile of those who remained behind in the insured pool.

Therefore, the empirical procedure adopted in this thesis was to test whether a decrease in the total number of medical scheme members would translate into an increase in net claims incurred per average medical scheme beneficiary per month – thus, implying a deterioration in the risk profile of those who continued to have medical scheme coverage. The empirical findings of this thesis indicated that there was no discernible negative relationship between the total number of medical scheme members and net claims incurred (pabpm). Therefore, adverse selection does not appear to be a feature of South Africa's private healthcare financing sector.

Thus, the South African government's claim that adverse selection (as a consequence of the deregulation that took place during the late 1980s and early 1990s) is responsible for the deterioration in medical scheme coverage for the elderly, unhealthy or poor is fallacious.

¹⁶⁴ The empirical studies that directly tested for the positive coverage-risk correlation utilised individual insured data that could detect implicitly the link between an insured's contract (and coverage) choices relative to their observed risk. Refer to empirical studies such as Chiappori and Salanié (2000), Cardon and Hendel (2001), Finkelstein and Poterba (2004), Cohen (2005), Saito (2006), and Fang, Keane and Silverman (2008).

6.1.1 Regulatory Capture as Alternative Explanation

Since the absence of adverse selection in South Africa's medical scheme industry has been established, it can be argued that the current regulatory framework through the application of open enrolment and community-rating has failed to deliver upon government's objective of universal access to affordable healthcare for all South Africans. The anecdotal evidence provided in this thesis suggests that rather than becoming more affordable, medical scheme coverage has become far more expensive for the average medical scheme beneficiary per month and the medical scheme benefits (coverage) received by beneficiaries has deteriorated further under the current regulatory environment. This raises doubts concerning the real reasons behind why government instituted the Medical Schemes Act No. 131 of 1998.

It can be argued that medical schemes have managed to continue to operate successfully within an environment that prohibits risk-rating; this may be partly due to careful marketing practices and astute product offerings that allow medical schemes to continue to *cherry-pick* lower-risk individuals. At the same time, the costs of healthcare expenditure have continued to increase unabated, and this suggests that the healthcare providers remain incentivised to over-supply healthcare services due to the continued reliance upon the fee-for-service reimbursement arrangement. Therefore, one may argue that the regulatory framework governing the South African private health insurance sector has been "captured" by either the medical schemes (and their associated administrators) or by private healthcare providers (or both). Therefore, government's reasoning that the pervasiveness of adverse selection within the medical scheme sector, required government intervention (through the enactment of the Medical Schemes Act No.131 of 1998), as it was in the *public interest* to mitigate adverse selection, does not hold up to scrutiny.

This thesis concurs with the proposition advocated by Stigler (1971), whereby the new regulatory framework instituted by the South African government through the implementation of the Medical Schemes Act No. 131 of 1998, although not originally sought by the medical scheme industry, may have been acquired by the industry or at least by the providers of healthcare. Thus, even though the industry protested and made their objections heard at the time the regulation was formulated, the explanation that the industry has managed to operate the new regulatory framework primarily for its benefit is certainly plausible. This has been achieved through careful adjustment to their product offerings and coverage plans, coupled with astute marketing practices, and potentially certain other exogenous factors such as risk aversion (propitious selection) that influences the individual's demand for health insurance. Therefore, it is postulated that regulatory capture may be the alternative explanation for why the current legislation governing the medical scheme industry came about but perhaps, more importantly, on how it is currently being exploited by medical schemes and healthcare providers.

Whilst in the case of the South African medical scheme industry, it may be suggested that it was not the case that the industry sought out particular regulation to extract some form of benefit, prior to the promulgation of the Medical Schemes Act No. 131 of 1998. But rather the regulation was thrust upon the private health insurance sector with theoretically adverse consequences. It is this thesis' assertion that despite the theoretical predictions that the new regulatory framework would induce an adverse selection *death spiral* through opportunistic behaviour on the part of medical scheme beneficiaries, medical schemes and their associated administrators have managed to take advantage of the lack of adverse selection in the medical scheme sector and exploit the regulatory framework through various astute adjustments to their menu of policy options and benefit plans. In addition, medical schemes have intelligently formulated their marketing practices and activities to target in particular low-risk individuals.

It is this thesis' view that the risk of regulatory capture within South Africa's health insurance sector may manifest itself in the spirit of the Toulouse School. The merit of the Toulouse School and authors such as Spiller (1990) is the acknowledgement of the principal-agent relationships inherent to any type of regulation (Boehm, 2007: 11). Spiller (1990) identifies that regulated firms and regulators can collude to extract and divide rents from the regulator's principal (legislature). However, in South Africa's health insurance sector, it is argued that the risk of regulatory capture may take the form of passive capture of the regulator. The objectives sought by the South African government through the promulgation of the Medical Schemes Act No. 131 of 1998 was to create a private healthcare financing system that offered affordable and universal health insurance coverage to the public-at-large. In particular, the aim was to provide affordable cover for those who had previously fell outside of the private healthcare financing net and those who were increasingly excluded from coverage; namely, the elderly and unhealthy. In essence, the government sought to establish a national health insurance system to be funded via private medical scheme contributions where higher income (and healthier) individuals subsidised lower income (and less healthy) segments of the population. Far from achieving these objectives, the reverse has occurred partly as a result of the misinterpretation of the concept of adverse selection by the South African government, but more importantly as a result of the ability of the industry to 'capture' the regulatory framework.

The risk of regulatory capture is increased as the medical scheme industry can make use of their superior technical knowledge and informational advantages over the regulator to extract benefits and pursue their own self-interest. Further, the assumption that the regulator can uncover the real nature of a regulated medical scheme appears to be overly-optimistic (Laffont & Tirole, 1991, 1993). Grace and Phillips (2008) argue that in the case of insurance regulation it is unlikely that other government officials (those not directly involved in the regulatory process) will be able to acquire accurate information regarding the profitability of the industry. This would translate into a lower marginal cost to influence the regulator. It suggests that regulatory capture or even collusive-type behaviour is quite possible within the insurance context.

Government's stated objective in the provision of healthcare in South Africa, is for the healthy (and wealthy) to subsidise the unhealthy (and poor), but this is incompatible with the functioning of a competitive market and therefore, it raises the question whether government's goal is at all consistent with the concept of *in the public interest*. Indeed, Peltzman, Levine and Noll (1989) attempt to gauge the efficacy of the economic theory of regulation and compare its development with that of the actual regulatory environment in various sectors. Upon reflection, Peltzman et al. (1989) assert that the economic theory of regulation generally fits the regulatory experience better than the public interest theory. The graphical representation of the theory of economic regulation by Beard, Kaserman and Mayo (2003) also supports Peltzman et al. (1989)'s view. The failure of the new regulatory framework to meet the objectives of universal access and affordable private health insurance for all, does indeed raises questions concerning the inability of the regulatory reforms to translate into desired outcomes.

Even though the four distinct schools of thought (Chicago School, Virginia School, Tollbooth Theories and the Toulouse School) relating to the economic theory of regulation all have merit, it is this thesis's view that regulatory capture is likely to have taken place within the spirit of the Toulouse School – whereby in South Africa's private health insurance sector medical schemes (and/or healthcare providers) have passively captured the regulator or regulatory environment. But this assertion can only be confirmed through further empirical research that is beyond the scope of this thesis.

6.2. Limitations of the Research

The primary limitation of the research stems from the fact that individual medical scheme beneficiary data was unavailable from the Council for Medical Schemes or indeed forthcoming from medical schemes themselves. Individual medical scheme beneficiary data, such as contribution payments, benefit plan and coverage choices, and claims experience would have been most useful in confirming the thesis' conclusion that adverse selection is absent in South Africa's private health insurance sector. The lack of individual medical scheme beneficiary data prevented one from directly testing for the positive coverage-risk correlation as undertaken by studies such as Chiappori and Salanié (2000), Cardon and Hendel (2001), Finkelstein and Poterba (2004), Cohen (2005), Saito (2006), and Fang, Keane and Silverman (2008). While not a sufficient condition to confirm the presence of adverse selection, a positive coverage-risk correlation is nevertheless a necessary condition (Cohen & Siegelman, 2010). Hence, the absence of a positive coverage-risk correlation would have been adequate to confirm the absence of adverse selection.

Another limitation identified centres around examining more specifically the reasons why the absence of adverse selection may be observed in South Africa's private health insurance sector. Cohen and Siegelman (2010) provide an excellent discussion surrounding the possible reasons for the absence of a positive coverage-risk correlation and therefore, the absence of adverse selection. This thesis would have benefited from using individual medical scheme beneficiary data to ascertain, which precise reason(s) can explain the observed absence of adverse selection in South Africa's medical schemes industry.

From the seminal theoretical literature on adverse selection, it is often assumed that positive coverage-risk correlation is conditional upon the menu of coverage offered by the insurer, which through the self-selection mechanism as described by Salop and Salop (1976), induce policyholders to reveal their risk type. Cohen and Siegelman

(2010) posit an important aspect that was not considered within this thesis' exposition of the theoretical adverse selection literature; namely, that perhaps additional private information may be revealed by the insured's choice of insurance contract and level of coverage (Cohen and Siegelman, 2010).

6.3. Recommendations for Future Research

Many reasons are discussed by Cohen and Siegelman (2010) to explain why one may not observe a positive coverage-risk correlation within insurance markets; however, an interesting area of theoretical and empirical research could be undertaken to examine the situation whereby despite the presence of a positive coverage-risk correlation, adverse selection may still not be present. Recall a positive correlation between coverage and risk is not a sufficient condition to confirm the presence of adverse selection, but it is nevertheless a necessary condition. They suggest that such a correlation may arise even in the absence of adverse selection due to moral hazard. The authors provide an insightful statement into the distinction between adverse selection and moral hazard:

“Unlike adverse selection, which has to do with “hidden information,” moral hazard has to do with “hidden action.” The risk of an accident and the losses it produces may be a product not only of the policyholder's (unalterable) “type” but also of his or her [behaviour]. A policyholder may invest in precautions that reduce the probability or the size of the loss. The purchase of insurance diminishes the policyholder's incentives to invest in such precautions because the insured policyholder no longer captures their full benefits, at least part of which accrue to the insurer instead” (Cohen & Siegelman, 2010: 71).

Therefore, another avenue of research in South Africa's private health insurance sector could involve investigating the extent to which moral hazard is a feature of the medical scheme sector. In particular, the extent to which it affects the behaviour of medical scheme members and their dependents. Since this thesis establishes the absence of adverse selection, data permitting, further research could investigate whether indeed a positive coverage-risk correlation does exist. If so, it may explain

the presence of moral hazard within the medical schemes sector. Under health insurance, it is plausible to believe that medical scheme beneficiaries may behave in such way that does not reduce the motive to prevent loss despite having medical scheme coverage. A claim triggered under health insurance cover is usually accompanied by a physical ailment or medical incident that is most often deliberately avoided by individuals.

A key assumption behind much of the adverse selection theory revolves around the assumption that the insured has better information about their risk type than the insurer does – this is not necessarily valid for all insurance markets and products (Cohen & Siegelman, 2010). Indeed, part of the reason why one negates the presence of adverse selection in South Africa's medical scheme industry may be due to the violation of this assumption. The extent to which individuals know their health status and the likelihood of requiring healthcare treatment for certain medical conditions may be limited. Therefore, it could be useful to seek out those benefit plans (options) and products where members can be expected to have relevant private information.

The empirical section of this thesis proved the absence of adverse selection for the entire population of registered medical schemes. However, it may be possible that within a certain subset (or subclass) of medical schemes, adverse selection could be present. Indeed, this may even be the case for specific options offered by medical schemes. Therefore, separate tests should be undertaken to investigate whether the relationships estimated under the fixed-effects model are consistent when subsamples or subsets based upon certain criteria (size of medical scheme, geographical concentration of members, and demographic characteristics of members) are considered. Cohen (2005) find that policyholder's driving experience may be a key determinant as to whether a positive coverage-risk correlation is present or not.

Risk aversion's role in influencing an individual's decision to purchase medical scheme cover, their choice of medical scheme option or coverage levels is yet to be

adequately investigated in South Africa. Indeed, the extent to which propitious selection results in low-risk individuals opting for more generous medical scheme plans or comprehensive options, may explain the absence of adverse selection in the private health insurance market. Therefore, the ability of propitious selection to offset the so-called adverse selection *death spiral* would be a most interesting avenue for future research.

The empirical investigation undertaken by this thesis examined net claims incurred and net contribution income (net of claims incurred or contributions paid into medical savings accounts (MSAs)), so as to focus upon the health “insurance” aspect. But an interesting observation from the assessment of the medical schemes sector in Chapter 2, uncovers that increasingly members of medical schemes are bearing a significantly greater proportion of their healthcare-related expenditure. This is despite the restrictions imposed upon MSAs under the current legislative framework. Restrictions such as upper limits appear to have had little impact in reducing the utilisation of these risk-sharing arrangements. Another key determinant in the transfer of risk from the medical schemes back onto members is the deterioration in benefit levels through careful manipulation of the menu of available coverage options to members. Thus, it appears that medical scheme members are bearing a greater proportion of their risk, but at the same time paying more for their reduced cover. Further research into these observations would be beneficial in extracting the underlying causes for the high rates of increase in contributions relative to the modest increases in claims.

In the United States, many people reacted unenthusiastically to the cost-control methods that *managed care organisations* initiated. The reason for this was the perception that the cost-control methods would damage the quality of healthcare provision. This is perhaps part of the reason why *managed care* has received a less than favourable reception in many circles. While the deregulation experienced in South Africa during the late 1980s and early 1990s paved the way for a *managed care* system to develop, it failed to gain widespread implementation. South Africa faced numerous obstacles that hindered the adoption of a managed care framework, such as

the resistance from healthcare providers and capacity constraints. Reekie (1999) argues that only through the widespread implementation of managed care principles can South Africa address the excessive medical inflation that has been endemic within the private healthcare sector for many decades. However, Doherty and McLeod (2003) also mention that the incentives, specifically the fee-for-service reimbursement system, encourages healthcare providers to over-supply healthcare services and medical scheme beneficiaries to over-utilise healthcare benefits; and these are therefore, identified as key cost accelerants. This has been further exacerbated by the outflow of medical expertise and the scarcity of quality healthcare facilities. To overcome the problem of over-utilisation of healthcare in general, it is possible for healthcare providers to be offered financial incentives to restrict care. Eggleston (2000: 174) states that:

“Since providers have considerable influence over treatment decisions, supply-side cost sharing can be a powerful instrument for controlling healthcare costs”.

Indeed, Reekie (1999) and Doherty and McLeod (2003) would concur with Eggleston (2000)’s sentiments. The South African government has also responded by proposing the creation of a statutory price determination authority with the intention of bringing back multilateral tariff negotiations. The idea is to steer medical schemes and healthcare providers towards equitable and affordable outcomes through a mandatory arbitration mechanism. If implemented, it will be appealing to investigate the extent to which this mechanism will be successful in arresting private healthcare financing costs.

Moreover, further research into how the proposed National Health Insurance (NHI) system for South Africa will impact upon the private health insurance sector should be undertaken. Perhaps, as further details are published concerning the NHI, it will be possible to ascertain the exact cost of such a system for the country, the extent to which medical scheme operations will be affected and how the objective of universal coverage will be achieved.

6.4. Concluding Remarks

McIntyre et al. (2007) argued that the quintessential challenge for the South African healthcare system was to develop a mechanism to promote cross-subsidisation between the healthy, younger (and wealthier) population and the unhealthy, elderly (and poor) population. Thus, McIntyre et al. (2007) advocate that it was necessary to reformulate the healthcare regulatory environment and in particular the legislation governing medical scheme business, to create a legislative framework that would function as a redistribution device to address the inequitable distribution of resources (and quality of care delivered) between the public and private healthcare sectors in South Africa. Indeed, this idea was reflected in the stated objective of the government at the time when the Medical Schemes Act No. 131 was implemented. The principles of open enrolment and community-rating contained within the new legislation sought to create the mechanism whereby certain medical scheme members (young, healthy or wealthy) would subsidise others (elderly, unhealthy or poor). It was believed that these principles were in the public interest. However, the regulatory reforms underpinning the Medical Schemes Act No. 131 of 1998 move away from the version of equity that advocates that people should pay according to the expected benefits that they are likely to enjoy under medical scheme coverage. Many would argue that indeed risk-related and income-related cross subsidisation will lead to inefficient outcomes that hinder the functioning of a competitive market. Therefore, it raises the question of whether the regulatory reforms are in the public interest at all.

Government argued that the primary reason behind the redrafting of the medical scheme legislation towards the end of the 1990s was to address the negative consequences of adverse selection that it believed had manifested itself in the inability of the elderly, poor and unhealthy to obtain affordable private healthcare cover. Therefore, the promulgation of the Medical Schemes Act No. 131 of 1998 and accompanying regulations sought to bring about government's objective of providing universal access to affordable private healthcare for all South Africans. Considering this thesis' empirical findings it is evident that adverse selection does not appear to be

a feature of South Africa's private healthcare financing sector. Hence, the government's claim that adverse selection (as a result of the deregulation that took place during the late 1980s and early 1990s) is responsible for the deterioration in medical scheme coverage for the elderly, unhealthy or poor is simply false.

Having established that the ostensible reason for the current legislation does not stand up to scrutiny, the alternative explanation provided is that the current regulatory environment may have been acquired by the medical scheme industry itself or at least by the healthcare providers benefiting from its financing arrangements (Stigler, 1971).

Furthermore, despite the theoretical predictions that the current regulatory framework (applying the principles of open enrolment and community-rating) would induce an adverse selection *death spiral*, the medical scheme market continues to operate successfully albeit under significant cost escalation. The reason suggested is that medical schemes have managed to take advantage of the absence of adverse selection as they continue to exploit the regulatory framework by *cherry-picking* lower-risks through various judicious alterations to benefit options and astute marketing activities.

Government has once again responded to the failure of the current legislation to improve medical scheme coverage and arrest rampant cost escalation. Three key reforms that are about to be rolled-out in an attempt to re-assert government's objective of universal coverage include the Risk Equalisation Fund (REF), gradual implementation of the National Health Insurance (NHI) system and the creation of a statutory price determination authority.

The Risk Equalisation Fund (REF) will institute financial transfers across medical schemes, whereby the REF would receive funds from medical schemes with lower risk profiles and make corresponding payments to medical schemes with higher risk profiles (McLeod et al., 2004). Neuhaus (1995) argues that this is essential if the

principle of community-rating is applied within an insurance market. The NHI system would guarantee a full range of healthcare services to everyone, regardless of their ability to pay. In addition, any medical treatment received under the NHI standard benefit package would be free at the point of service. Finally, the statutory price determination authority would bring medical schemes and healthcare providers together to negotiate, with a mandatory arbitration mechanism, appropriate pricing for healthcare services. Medical schemes will have to justify what they can afford to pay for healthcare services, whilst healthcare providers will have to justify what they believe they are entitled to receive for their healthcare services.

South Africa is therefore about to institute another set of regulatory reforms that represent the most significant reform of the private health insurance sector in more than a decade. Will these reforms achieve government's stated objective of universal health coverage, on what timescale and at what cost? Only time will tell.

7 REFERENCES

- Aaron, H. J. (1991). *Serious and unstable condition : financing America's health care*. Washington, D.C.: Brookings Institution.
- Adams, M. B. & Tower, G. D. (1994). Theories of Regulation: Some Reflections on the Statutory Provision of Insurance Companies in Anglo-American Countries. *Geneva Papers on Risk and Insurance*, 19, 156-177.
- African National Congress. (2009a). *National Health Insurance Policy Proposal*. ANC National General Council. Released 22 June 2009.
- African National Congress. (2009b). National Health Insurance: A unified, equitable and integrated national health system that benefits all South Africans. *ANC Today*, 9(29) 24–30 July 2009.
- African National Congress. (2009c) *National Health Insurance (NHI)*. ANC National General Council. Released September 2010.
- African National Congress (2010) Additional Discussion Documents. ANC National General Council. Released September 2010. Available at: <http://www.anc.org.za/docs/discus/2010/aditionalo.pdf> (p.14)
- Akerlof, G. A. (1970). The Market for "Lemons": Quality Uncertainty and the Market Mechanism. *Quarterly Journal of Economics*, 84(3), 488-500.
- Altman, D., Cutler, D. M., & Zeckhauser, R. J. (1998). Adverse Selection and Adverse Retention. *American Economic Review*, 88(2), 122-126.
- Aristophanes, Barrett, D., & Dutta, S. (2007). *Frogs and other plays*. London: Penguin Books.

- Armitage, A. (1952). *The world of Copernicus*. New York: New American Library.
- Arrow, K. J. (1963). Uncertainty and the Welfare Economics of Medical Care. *American Economic Review*, 53(5), 941-973.
- Arrow, K. J. (1971). Insurance, Risk and Resource Allocation. *Essays in the Theory of Risk-Bearing*, 134-143.
- Ausubel, L. M. (1999). *Adverse Selection in the Credit Card Market*. Working Paper, Department of Economics, University of Maryland. Retrieved May 9, 2008, from <http://www.ausubel.com/creditcard-papers/adverse.pdf>.
- Baker, L. C. (2001). Managed care and technology adoption in health care: evidence from magnetic resonance imaging. [Article]. *Journal of Health Economics*, 20(3), 395-421.
- Baker, L. C., & Phibbs, C. S. (2002). Managed Care, Technology Adoption, and Health Care: The Adoption of Neonatal Intensive Care. *RAND Journal of Economics*, 33(3), 524-548.
- Baltagi, B. H. (2008). *Econometric analysis of panel data* (4th ed.). Chichester, UK ; Hoboken, NJ: John Wiley & Sons.
- Barrett, G. E, & Conlon, R. (2003). Adverse Selection and the Decline in Private Health Insurance Coverage in Austria: 1989-95. *Economic Record*, 79(246), 279-296.
- Beard, T. R., Kaserman, D. L., & Mayo, J. W. (2003). A Graphical Exposition of the Economic Theory of Regulation. *Economic Inquiry*, 41(4), 592-606.
- Becker, G. S. (1983). A Theory of Competition Among Pressure Groups for Political Influence. *Quarterly Journal of Economics*, 98(3), 371-400.

- Becker, G. (1986). The Public Interest Hypothesis Revisited: A New Test of Peltzman's Theory of Regulation. *Public Choice*, 49, 223-234.
- Becker, G. S. (1993). *Human capital : a theoretical and empirical analysis, with special reference to education* (3rd ed.). Chicago ; London: University of Chicago Press.
- Becker, G. S. (1995). *Human capital and poverty alleviation*. [Washington, D.C.]: World Bank, Human Resources Development and Operations Policy.
- Becker, G. S. (2006). Health and Human Capital: The Inaugural T. W. Schultz Lecture. *Review of Agricultural Economics*, 28(3), 323-325.
- Belli, P., & World Bank. Development Research Group. Public Economics. (2001). *How adverse selection affects the health insurance market*. Washington, D.C.: World Bank, Development Research Group, Public Economics.
- Besley, T., & Coate, S. (2003). Elected Versus Appointed Regulators: Theory and Evidence. *Journal of the European Economic Association*, 1(5), 1176-1206.
- Biermann, J. (2004). *Certificates of need are a recipe for chaos*. Free Market Foundation. Feature Article\2 March 2004 - Policy Bulletin / 30 June 2009. Retrieved July 3, 2010 from <http://www.freemarketfoundation.com/ShowArticle.asp?ArticleType=Bulletin&ArticleID=1915>.
- Boehm, F. (2007). *Regulatory Capture Revisited – Lessons from Economics of Corruption*. Working Paper, Internet Centre for Corruption Research, July.
- Bonbright, J. (1961). *Principles of Public Utility Rates*. New York: Columbia University Press.
- Booth, P. (1997). The Political Economy of Regulation. *British Actuarial Journal*, 3(3), 675-707.

- Bradshaw, D. & Steyn, K. (2001). *Poverty and chronic diseases in South Africa*. Technical Report. Cape Town: South African Medical Research Council.
- Breitung, J., & Das, S. (2008). Testing for Unit Roots in Panels with a Factor Structure. *Econometric Theory*, 24(1), 88-108.
- Brockett, P. L., & Golden, L. L. (2007) Biological and Psycho-Behavioral Correlates of Credit Scores and Automobile Insurance Losses: Towards an Explanation of Why Credit Scoring Works. *Journal of Risk and Insurance*, 74, 22-63.
- Brown, D. M., & Warren-Boulton, F. R. (1988). *Testing the Structure-Competition Relationship on Cross-Sectional Firm Data*. Discussion paper 88-6. Economic Analysis Group, U.S. Department of Justice.
- Browne, M. J., & Doeringhaus, H. I. (1993). Information Asymmetries and Adverse Selection in the Market for Individual Medical Expense Insurance. *Journal of Risk and Insurance*, 60(2), 300-312.
- Buchanan, J. M. (1965). An Economic Theory of Clubs. *Economica*, 32(125), 1-14.
- Buchanan, J. M., & Tullock, G. (1962). *The calculus of consent, logical foundations of constitutional democracy*. Ann Arbor,: University of Michigan Press.
- Buchmueller, T., & DiNardo, J. (2002). Did Community Rating Induce an Adverse Selection Death Spiral? Evidence from New York, Pennsylvania, and Connecticut. *American Economic Review*, 92(1), 280-294.
- Buchmueller, T. C., Couffinhal, A., Grignon, M., & Perronnin, M. (2004). Access to physician services: does supplemental insurance matter? Evidence from France. *Health Economics*, 13(7), 669-687.

- Burgon, J. W. (1965). *The life and times of Sir Thomas Gresham, knt., founder of the Royal Exchange; including notices of many of his contemporaries*. New York,: B. Franklin.
- Butler, P. (1996). Automobile Insurance Pricing: Operating Versus Ownership Costs, in: *Women's Travel Issues: Proceedings From the Second National Conference*, U.S. Department of Transportation, Federal Highway Administration Publ. FHWA-PL-97-024.
- Cameron, A. C., & Trivedi, P. K. (2009). *Microeconometrics using Stata*. College Station, Tex.: Stata Press.
- Cardon, J. H., & Hendel, I. (2001). Asymmetric information in health insurance: evidence from the National Medical Expenditure Survey. *RAND Journal of Economics (RAND Journal of Economics)*, 32(3), 408-427.
- Cawley, J., & Philipson, T. (1999). An Empirical Examination of Information Barriers to Trade in Insurance. *American Economic Review*, 89, 827-846.
- Chamberlain, G. (1984). Panel Data. in Z. Griliches, & Intriligator, M. D. (eds), *Handbook of Econometrics* Vol. II, 1247-1318. Amsterdam: Elsevier.
- Chassagnon, A., & Villeneuve, B. (2005). Optimal Risk-Sharing under Adverse Selection and Imperfect Risk Perception. *Canadian Journal of Economics / Revue canadienne d'Economie*, 38(3), 955-978.
- Chiappori, P.-A. (1999). Asymmetric Information in Automobile Insurance: An Overview, in: G. Dionne and C. Laberge-Nadeau, eds., *Automobile Insurance: Road Safety, New Drivers, Risks, Insurance Fraud and Regulation* (Norwell, MA: Kluwer), 1-12.
- Chiappori, P.-A. (2000). Econometric Models of Insurance under Asymmetric Information. in Dionne (ed.) *Handbook of Insurance*, Norwell, Mass: Kluwer. 365–394.

- Chiappori, P.-A., & Salanié, B. (1997). Empirical contract theory: The case of insurance data. *European Economic Review*, 41(3-5), 943-950.
- Chiappori, P.-A., & Salanié, B. (2000). Testing for asymmetric information in insurance markets. *Journal of Political Economy*, 108(1), 56.
- Chiappori, P.-A., & Salanié, B. (2003). Testing Contract Theory: A Survey of Some Recent Work, in: M. Dewatripont, L. P. Hansen, and S. Turnovsky, eds., *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress*. Cambridge: Cambridge University Press), Vol. 1, 115-149.
- Chiappori, P.-A., Jullien, B., Salanié, B. & Salanié, F. (2006). Asymmetric Information in Insurance: General Testable Implications. *RAND Journal of Economics*, 37, 783-798.
- Choi, I. (2001). Unit root tests for panel data. *Journal of International Money and Finance*, 20, 249–272.
- Claxton, G., DiJulio, B., Whitmore, H., Pickreign, J. D., McHugh, M. Osei-Anto, A. & Finder, B. (2010). Health Benefits In 2010: Premiums Rise Modestly, Workers Pay More Toward Coverage. *Health Affairs*, 29(10), 1942-1950.
- Cohen, A. (2005). Asymmetric Information and Learning: Evidence from the Automobile Insurance Market. *Review of Economics & Statistics*, 87(2), 197-207.
- Cohen, A. & Siegelman, P. Testing for Adverse Selection in Insurance Markets. *Journal of Risk and Insurance*, 77(1), 39-84.
- Cornell, J., Goudge, J., McIntyre, D. & Mbatsha, S. (2001) *South African National Health Accounts: The Private Sector*. Pretoria: National Department of Health.
- Council for Medical Schemes. (1994). *Annual Report 1993*. Pretoria: Council for Medical Schemes.

Council for Medical Schemes. (1995). *Annual Report 1994*. Pretoria: Council for Medical Schemes.

Council for Medical Schemes. (1996). *Annual Report 1995*. Pretoria: Council for Medical Schemes.

Council for Medical Schemes. (1997). *Annual Report 1996*. Pretoria: Council for Medical Schemes.

Council for Medical Schemes. (1998). *Annual Report 1997*. Pretoria: Council for Medical Schemes.

Council for Medical Schemes. (1999). *Annual Report 1998*. Pretoria: Council for Medical Schemes. pp. 1-35.

Council for Medical Schemes. (2000). *Annual Report 1999*. Pretoria: Council for Medical Schemes. pp. 1-49.

Council for Medical Schemes. (2001). *Annual Report 2000*. Pretoria: Council for Medical Schemes. pp. 1-52.

Council for Medical Schemes. (2002). *Annual Report 2001*. Pretoria: Council for Medical Schemes. pp. 1-76.

Council for Medical Schemes. (2003). *Annual Report 2002-3*. Pretoria: Council for Medical Schemes. pp. 1-100.

Council for Medical Schemes. (2004). *Annual Report 2003-4*. Pretoria: Council for Medical Schemes. pp. 1-121.

Council for Medical Schemes. (2005). *Annual Report 2004-5*. Pretoria: Council for Medical Schemes. pp. 1-136.

- Council for Medical Schemes. (2006). *Annual Report 2005-6*. Pretoria: Council for Medical Schemes. pp. 1-145.
- Council for Medical Schemes. (2007). *Annual Report 2006-7*. Pretoria: Council for Medical Schemes. pp. 1-143.
- Council for Medical Schemes. (2008). *Annual Report 2007-08*. Pretoria: Council for Medical Schemes. pp. 1-177.
- Council for Medical Schemes. (2009). *Annual Report 2008-2009*. Pretoria: Council for Medical Schemes. pp. 1-203.
- Council for Medical Schemes. (2010). *Annual Report 2009-2010*. Pretoria: Council for Medical Schemes. pp. 1-239.
- Crocker, K. J., & Snow, A. (1986). The Efficiency Effects of Categorical Discrimination in the Insurance Industry. *Journal of Political Economy*, 94(2), 321-344.
- Cummins, J. D., & Doherty, N. A. (2006). The Economics of Insurance Intermediaries. *Journal of Risk and Insurance*, 73, 359-396.
- Cutler, D. (1996). Public Policy for Health Care. *NBER Working Paper 5591*, May.
- Cutler, D. M., & Reber, S. J. (1998). Paying for Health Insurance: The Trade-Off between Competition and Adverse Selection. *Quarterly Journal of Economics*, 113(2), 433-466.
- Cutler, D. M., & Zeckhauser, R. J. (1998). Adverse Selection in Health Insurance, *Forum for Health Economics and Policy*, 1: Article 2.
- Cutler, D. M., & Zeckhauser, R. J. (2000). The Anatomy of Health Insurance. In A. J. Culyer & Newhouse, J. P. (Eds.) *Handbook of Health Economics*, Amsterdam: Elsevier Science, Vol. 1, Ch. 11, 563–643.

- Cutler, D. M., McClellan, M., & Newhouse, J. P. (2000). How Does Managed Care Do It? *RAND Journal of Economics*, 31(3), 526-548.
- Cutler, D. M., Finkelstein, A. & McGarry, K. (2008). Preference Heterogeneity in Insurance Markets: Explaining a Puzzle. *American Economic Review Papers and Proceedings*, 98(2), 157-162.
- Dahlby, B. G. (1983). Adverse selection and statistical discrimination : An analysis of Canadian automobile insurance. *Journal of Public Economics*, 20(1), 121-130.
- Dahlby, B. G. (1992). Testing for Asymmetric Information in Canadian Automobile Insurance. In G. Dionne (ed.) *Contributions to Insurance Economics*, Boston: Kluwer Academic Publishers, 423–443.
- Davie, E. G. & Urbach, J. (2006). *Determination of a maximum logistics fee*. Free Market Foundation of Southern Africa. Retrieved September 6, 2009, from http://www.healthpolicyunit.org/downloads/Logisitcs_Fee.pdf.
- DeDonder, P., & Hindriks, J. (2006). *Does Propitious Selection Explain Why Riskier People Buy Less Insurance?* CEPR Working paper DP5640.
- DeMeza, D., & Webb, D. C. (2001). Advantageous Selection in Insurance Markets. *RAND Journal of Economics*, 32(2), 249-262.
- Department of Health. (1997). *White Paper for the Transformation of the Health System in South Africa*. Pretoria: Government Printer.
- Department of Health. (2002). *Inquiry into the Various Social Security Aspects of the South African Health System*. Pretoria: National Department of Health.
- Department of Health. (2004). *South Africa Demographic and Health Survey 2003, Preliminary Report*. Pretoria: Measure DHS / ORC Macro and Department of Health.

- Department of Health. (2006). *A National Human Resources Plan for Health to Provide Skilled Human Resources for Healthcare Adequate to Take Care of all South Africans*. Pretoria: National Department of Health.
- Department of Health. (2009a). *National Health Insurance Plan for South Africa*. Pretoria: National Health Insurance Task Team – Department of Health.
- Department of Health (2009b) *Establishment of the National Health Insurance Advisory Committee*. Pretoria: Government Gazette, 11 September 2009.
- Department of Health. (2010). *Discussion Document – The Determination of Health Prices in the Private Sector*. Pretoria: National Department of Health.
- Department of Social Development. (2002). *Transforming the Present – Protecting the Future: Consolidated Report of the Committee of Inquiry into a Comprehensive System of Social Security for South Africa*. Pretoria: National Department of Social Development.
- Dionne, G., & Doherty N. A. (1992). Adverse Selection in Insurance Markets: A Selective Survey, in: G. Dionne, ed., *Contributions to Insurance Economics*, Boston: Kluwer Academic Publishers, 97-140.
- Dionne, G., Doherty N. A., & Fombaron, N. (2000). Adverse Selection in Insurance Markets. In G. Dionne (ed.) *Handbook of Insurance*, Boston: Kluwer Academic Publishers, 185–243.
- Dionne, G., Gouriéroux, C., & Vanasse, C. (2001). Testing for Evidence of Adverse Selection in the Automobile Insurance Market: A Comment. *Journal of Political Economy*, 109(2), 444-453.
- Djankov, S., Porta, R. L., Lopez-de-Silanes, F., & Shleifer, A. (2002). The Regulation of Entry. *Quarterly Journal of Economics*, 117(1), 1-37.

- Doherty, J. and McLeod, H. (2003). *Medical Schemes*. In: Ijumba P, Ntuli A, Barron P, editors. South African Health Review 2002. Durban: Health Systems Trust. Retrieved January 30, 2007 from <http://www.hst.org.za/uploads/files/chapter3.pdf>.
- Doron, G. (1979). Administrative Regulation of an Industry: The Cigarette Case. *Public Administration Review*, 39(2), 163-170.
- Dorrington, R., Bradshaw, D., Johnson, L. & Daniel, T. (2006) *The Demographic Impact of HIV/AIDS in South Africa: National and Provincial Indicators for 2006*. Cape Town: Centre for Actuarial Research, Burden of Disease Research Unit (Medical Research Council) and Actuarial Society of South Africa.
- Downs, A. (1957). An Economic Theory of Political Action in a Democracy. *Journal of Political Economy*, 65(2), 135-150.
- Dudley, R. A., Miller, R. H., Korenbrot, T. Y., & Luft, H. S. (1998). The Impact of Financial Incentives on Quality of Health Care. *Milbank Quarterly*, 76(4), 649-686.
- Econex (2009a) *Key Features of the Current NHI Proposal*. Econex NHI Note 14 September 2009. Retrieved June 8, 2010 from http://www.econex.co.za/index.php?option=com_docman&task=doc_download&gid=42&Itemid=60.
- Econex (2009b) *Supply Side Constraints*. Econex NHI Note 4 November 2009. Retrieved June 8, 2010, from http://www.econex.co.za/index.php?option=com_docman&task=doc_download&gid=46&Itemid=60.
- Eggleston, K. (2000). Risk Selection and Optimal Health Insurance-Provider Payment Systems. *Journal of Risk and Insurance*, 67(2), 173-196.

- Erasmus, M. and Fitchen, E. (2010) *Health Expenditure in South Africa*. Econex Occasional Note November 2010. Retrieved November 22, 2010, from http://www.econex.co.za/index.php?option=com_docman&task=doc_download&gid=76&Itemid=60.
- Estache, A. and Martimort, D. (1999). *Politics, Transaction Costs, and the Design of Regulatory Institutions*. Policy Research Working Paper 2073, World Bank.
- Ettner, S. L. (1997). Adverse selection and the purchase of Medigap insurance by the elderly. *Journal of Health Economics*, 16(5), 543-562.
- EuropeAid Development and Cooperation. (2010). Annual Report 2010. *Annual Report*, 1(1), 1-213.
- Fang, H., Keane, M. P., & Silverman, D. (2008). Sources of Advantageous Selection: Evidence from the Medigap Insurance Market. *Journal of Political Economy*, 116(2), 303-350.
- Feroz, E.H. (1987). Financial Accounting Standards Setting: A Social Science Perspective. *Advances in Accounting*, 5, 3-14.
- Fetter, F. W. (1932). Some Neglected Aspects of Gresham's Law. *Quarterly Journal of Economics*, 46(3), 480-495.
- Finkelstein, A., & Poterba, J. (2004). Adverse Selection in Insurance Markets: Policyholder Evidence from the U.K. Annuity Market. *Journal of Political Economy*, 112(1), 183-208.
- Finsinger, J., Hammond, E., Tapp, J., & Institute for Fiscal Studies (Great Britain). (1985). *Insurance : competition or regulation? : a comparative study of the insurance markets in the United Kingdom and the Federal Republic of Germany*. London: Institute for Fiscal Studies.

- Firebaugh, G. (2003). *The new geography of global income inequality*. Cambridge, Mass.: Harvard University Press.
- Folland, S., Goodman, A. C., & Stano, M. (2010). *The Economics of Health and Health Care* (6th ed.). Upper Saddle River, New Jersey: Prentice Education Inc.
- Frech, H. E. F., & Samprone, J. C., Jr. (1980). The Welfare Loss of Excess Nonprice Competition: The Case of Property-Liability Insurance Regulation. *Journal of Law and Economics*, 23(2), 429-440.
- Friedman, B., & Steiner, C. (1999). Does managed care affect the supply and use of ICU services? *Inquiry*, 36(1), 68-77.
- Friendly, H. (1962). The Federal Administrative Agencies: The Need for Better Definition of Standards. *Harvard Law Review*, 75(7), 1263-1318.
- Fuchs, V. R. (1982). *Economic aspects of health*. Chicago: University of Chicago Press.
- Fuchs, V. R. (1996). *Individual and social responsibility : child care, education, medical care, and long-term care in America*. Chicago: University of Chicago Press.
- Giffen, R. (1891). The Gresham Law. *Economic Journal*, 1(2), 304-306.
- Giffen, R. (1892). *The case against bimetallism*. London and New York,: G. Bell & sons.
- Gilson, L. & McIntyre, D. (2001) South Africa: Addressing the legacy of apartheid. in Evans, T., Whitehead, M., Diderichsen, F., Bhuiya, A. & Wirth, M. (Eds.) *Challenging inequities in health: From ethics to action*. New York: Oxford University Press.

- Gini, C. (1912). I fattori demografici dell'evoluzione delle nazioni *Biblioteca del Metron Serie B*, 8 Retrieved September 2, 2008 from <http://purl.oclc.org/DLF/benchrepro0212>
<http://catalog.hathitrust.org/api/volumes/oclc/5789472.html>.
- Gini, C. (1937). *Die probleme der internationalen bevölkerungs- und rohstoffverteilung*. Jena,: G. Fischer.
- Gormley, W. T., Jr. (1986). Regulatory Issue Networks in a Federal System. *Polity*, 18(4), 595-620.
- Gottfried, J., & Sloan, F. A. (2002). The Quality of Managed Care: Evidence from the Medical Literature. *Law and Contemporary Problems*, 65(4), 103-137.
- Grace, M., & Phillips, R. (2008). Regulator Performance, Regulatory Environment and Outcomes: An Examination of Insurance Regulator Career Incentives on State Insurance Markets. *Journal of Banking and Finance*, 32, 116-133.
- Gradison, B. (1995). Insurance market reform and the cost-containment imperative. *Academic Medicine*, 70(3), 214-215.
- Greenaway, D., & Haynes, M. (2003). Funding Higher Education in the UK: The Role of Fees and Loans. *Economic Journal*, 113(485), F150-F166.
- Greene, W. H. (2008). *Econometric analysis* (6th ed.). Upper Saddle River, N.J.: Pearson/Prentice Hall.
- Grobler, P., Theron, H. & Cooper, M. (2003). *Technical Report: Risk Equalisation in South African Medical Schemes*. Technical report prepared by Medscheme Integrated Care and submitted to the Risk Equalisation Fund Task Group of the Department of Health, June 2003.
- Grossman, M. (1972). On the Concept of Health Capital and the Demand for Health. *Journal of Political Economy*, 80(2), 223-255.

- Guppy, A. (1993). Subjective Probability of Accident and Apprehension in Relation to Self-Other Bias, Age, and Reported Behavior. *Accident Analysis and Prevention*, 25(4), 375-382.
- Hadri, K. (2000). Testing for stationarity in heterogeneous panel data. *Econometrics Journal*, 3, 148-161.
- Hall, D. J. (2006). *Adverse Selection in the Market for Health Insurance*. Mimeo: Brigham Young University Department of Economics.
- Harrington, S. E. (1992). Presidential Address: Rate Suppression. *Journal of Risk and Insurance*, 59(2), 185-202.
- Harrington, S. E., & Niehaus, G. (1999). *Risk Management and Insurance*. Boston; London: Irwin McGraw-Hill.
- Harrington, S. E., & Danzon, P. M. (2000). Rate Regulation, Safety Incentives, and Loss Growth in Workers' Compensation Insurance. *Journal of Business*, 73(4), 569-595.
- Harris, R. D. F. & Tzavalis, E. (1999). Inference for unit roots in dynamic panels where the time dimension is fixed. *Journal of Econometrics*, 91, 201-226.
- Hartnedy, J. (1994). You Can't Buy Insurance When the House Is on Fire: Community Rating Kills Health Insurance. *Postgraduate Medicine*, 95(7), 75-87.
- Hausman, J. A. (1978). Specification Tests in Econometrics. *Econometrica*, 46(6), 1251-1271.
- Hayek, F. A. v. (1960). *The Constitution of Liberty*. Chicago: University of Chicago Press.

- Hayek, F. A. v. (1968). *Der Wettbewerb als Entdeckungsverfahren*. Kiel.
- Hayek, F. A. v. (1976). *The mirage of social justice*. Chicago: University of Chicago Press.
- Hayek, F. A. v. (1988). *The Fatal Conceit: The Errors of Socialism*. (I. W. Bartley, Ed.). London: Routledge.
- Hemenway, D. (1990) Propitious Selection. *Quarterly Journal of Economics*, 105, 1073-1069.
- Hirschman, Albert O. (1964). "The Paternity of an Index". *American Economic Review*, 54(5), 761.
- Hoy, M., Orsi, F., Eisinger, F., & Moatti, J. P. (2003). The Impact of Genetic Testing on Healthcare Insurance. *Geneva Papers on Risk & Insurance - Issues & Practice*, 28(2), 203-221.
- Hsiao, C. (2003). *Analysis of panel data* (2nd ed.). Cambridge ; New York: Cambridge University Press.
- Hurd, M. D., Smith, J. P., & Zissimopoulos, J. M. (2002). *The Effects of Subjective Survival on Retirement and Social Security Claiming*. NBER Working Paper No. 9140.
- Im, K.S., Pesaran, M.H. & Shin, Y. (2003). Testing for unit roots in heterogeneous panels. *Journal of Econometrics*, 115, 53-74.
- Ippolito, R. (1979). The Effects of Price Regulation in the Automobile Insurance Industry. *Journal of Law and Economics*, 22(1), 55-89.
- Joskow, P. (1973). Cartels, Competition and Regulation in the Property-Liability Insurance Industry. *Bell Journal of Economics and Management Science*, 4 (2), 375-427.

- Kalt, J. P., & Zupan, M. A. (1984). Capture and Ideology in the Economic Theory of Politics. *American Economic Review*, 74(3), 279-300.
- Kaiser Commission on Medicaid and the Uninsured. (2010). *The Uninsured: A Primer*. Retrieved December 22nd, 2010, from <http://www.kff.org/uninsured/7451.cfm>.
- Kaiser Family Foundation/Health Research and Educational Trust. (2000). *Employer Health Benefits 2000 Annual Survey*. Retrieved September 10th, 2010, from <http://ehbs.kff.org/>.
- Kaiser Family Foundation/Health Research and Educational Trust. (2006). *Employer Health Benefits 2006 Annual Survey*. Retrieved September 10th, 2010, from <http://ehbs.kff.org/>.
- Kaiser Family Foundation/Health Research and Educational Trust. (2010). *Employer Health Benefits 2010 Annual Survey*. Retrieved September 10th, 2010, from <http://ehbs.kff.org/>.
- Karagyozyova, T., & Siegelman, P. (2007). *Propitious Selection in Insurance Markets: A Simulation Study Using Continuous Distributions*. Working paper, University of Connecticut Law School.
- Kimball, S. (1961). The Purpose of Insurance Regulation: A Preliminary Inquiry in the Theory of Insurance Law. *Minnesota Law Review*, 45, 471-524.
- KPMG (1988). *Survey of Employer-Sponsored Health Benefits*. Health Insurance Association of America (HIAA).

- KPMG (1993). *Survey of Employer-Sponsored Health Benefits*. Health Insurance Association of America (HIAA).
- KPMG (1996). *Survey of Employer-Sponsored Health Benefits*. Health Insurance Association of America (HIAA).
- Koufopoulos, K. (2007). On the positive correlation property in competitive insurance markets. *Journal of Mathematical Economics*, 43(5), 597-605.
- Laffont, J.-J. (1994). The New Economics of Regulation Ten Years After. *Econometrica*, 62(3), 507-537.
- Laffont, J.-J., & Martimort, D. (1997). Collusion Under Asymmetric Information. *Econometrica*, 65(4), 875-911.
- Laffont, J.-J., & Tirole, J. (1991). The Politics of Government Decision-Making: A Theory of Regulatory Capture. *Quarterly Journal of Economics*, 106(4), 1089-1127.
- Laffont, J.-J., & Tirole, J. (1993). *A theory of incentives in procurement and regulation*. Cambridge, Mass.: MIT Press.
- Levine, M., & Forrence, J. (1990). Regulatory Capture, Public Interest, and the Public Agenda: Toward a Synthesis. *Journal of Law, Economics, & Organization*, 6, 167-198.
- Levin, A., Lin, C. & Chu, C.J. (2002). Unit root tests in panel data: Asymptotic and finite-sample properties. *Journal of Econometrics*, 108, 1-24.
- Löfgren, K.-G., Persson, T., & Weibull, J. W. (2002). Markets with Asymmetric Information: The Contributions of George Akerlof, Michael Spence and Joseph Stiglitz. *Scandinavian Journal of Economics*, 104(2), 195-211.

- MacAvoy, P. (1979). *The Regulated Industries and the Economy*. New York: W. W. Norton & Company, Inc.
- MacIntyre, D. M. (1962). *Voluntary health insurance and rate making*. Ithaca, N.Y.: Cornell University Press.
- Macleod, H. D. (1858). *The elements of political economy*. London,: Longman, Brown, Green, Longmans, and Roberts.
- Maddala, G. S. & Wu, S. (1999). A comparative study of unit root tests with panel data and a new simple test. *Oxford Bulletin of Economics and Statistics*, 61, 631-652.
- Magennis, R. H. & van Zyl, J. (2009). *Making Health Insurance Work for the Low-Income Market in South Africa: Cost Drivers and Strategies*. Discussion Document – FinMark Trust and the Centre for Financial Regulation and Inclusion (Cenfri).
- Martimort, D. (1999). The Life Cycle of Regulatory Agencies: Dynamic Capture and Transaction Costs. *Review of Economic Studies*, 66(4), 929-947.
- Martin, S. (1993). *Advanced industrial economics*. Cambridge, Mass.: Blackwell.
- Mayers, D., & Smith, C. W., Jr. (1981). Contractual Provisions, Organizational Structure, and Conflict Control in Insurance Markets. *Journal of Business*, 54(3), 407-434.
- McIntyre, D. (1997) *Health Care Financing and Expenditure in South Africa: Towards Equity and Efficiency in Policy Making*. PhD Dissertation. Cape Town: University of Cape Town.
- McIntyre, D., Bloom, G., Doherty, J. & Brijlal, P. (1995) *Health Expenditure and Finance in South Africa*. Durban: Health Systems Trust and the World Bank.

- McIntyre, D. & Gilson, L. (2002) Putting equity in health back onto the social policy agenda: Experience from South Africa. *Social Science and Medicine*, 54, 1637-1656.
- McIntyre, D. & Doherty, J. (2004). Health care financing and expenditure: Progress since 1994 and remaining challenges. in Van Rensburg, H. (Ed.) *Health care in South Africa*, Pretoria: Van Schaik Publishers.
- McIntyre, D., Thiede, M., Nkosi, M., Mutyambizi, V., Castilo-Riquelme, M., Gilson, L., Erasmus, E., & Goudge, J. (2007). *A Critical Analysis of the Current South African Health System*. Health Economics Unit, University of Cape Town and Centre for Health Policy, University of the Witwatersrand.
- McKeown, T. (1976). *The modern rise of population*. New York: Academic Press.
- McKeown, T., Brown, R. G., & Record, R. G. (1972). An Interpretation of the Modern Rise of Population in Europe. *Population Studies*, 26(3), 345-382.
- McLeod, H. (2004). *Social Health Insurance: An introduction for practitioners*. Pretoria: School of Health Systems and Public Health, University of Pretoria.
- McLeod, H., Matisonn, S., Fourie, I., Grobler, P., Mynhardt, S. & Marx, G. (2004). *The Determination of the Formula for the Risk Equalisation Fund in South Africa*. Report prepared for the Risk Equalisation Fund Task Group under the auspices of the National Department of Health. Pretoria: National Department of Health.
- McLeod, H. and Ramjee, S. (2007). *Medical Schemes*. In: Ijumba P, Ntuli A, Barron P, editors. South African Health Review 2007. Durban: Health Systems Trust.
- Meier, K. J. (1991). The Politics of Insurance Regulation. *Journal of Risk and Insurance*, 58(4), 700-713.

- Mill, J. S. (1859/2010). *On Liberty and Other Essays*. New York: Digireads Publishing.
- Miller, R. H., & Luft, H. S. (1994). Managed care plan performance since 1980. A literature analysis. *Journal of the American Medical Association*, 271(19), 1512-1519.
- Miller, R. H., & Luft, H. S. (1997). Does managed care lead to better or worse quality of care? *Health Aff (Millwood)*, 16(5), 7-25.
- Mitnick, B. M. (1980). Review: Myths of Creation and Fables of Administration: Explanation and the Strategic Use of Regulation. *Public Administration Review*, 40(3), 275-286.
- Mueller, D. C. (1976). Public Choice: A Survey. *Journal of Economic Literature*, 14(2), 395-433.
- Mundell, R. (1998). Uses and Abuses of Gresham's Law in the History of Money. *Zagreb Journal of Economics*, 2(2), 3-38. Retrieved June 14, 2009 from <http://www.columbia.edu/~ram15/grash.html>.
- Nash, J. (1953). Two-Person Cooperative Games. *Econometrica*, 21(1), 128-140.
- Nash, J. F., Jr. (1950). The Bargaining Problem. *Econometrica*, 18(2), 155-162.
- Nash, J. F., Kuhn, H. W., & Nasar, S. (2002). *The essential John Nash*. Princeton, N.J.: Princeton University Press.
- National Treasury. (2010). *Medium-Term Expenditure Framework 2010*. Pretoria: National Treasury. Retrieved, October 30th, 2010 from <http://www.treasury.gov.za/documents/mtbps/2010/default.aspx>.
- Naylor, C. D. (1988) Private Medicine and the Privatisation of Health Care in South Africa. *Social Science and Medicine*, 27, 1153-1170.

- Neuhaus, W. (1995). Community Rating and Equalisation. *JASTIN Bulletin*, 25(2), 95-118.
- Neuhaus, W. (1996). Community Rating and Equalisation. *Journal of Risk & Insurance*, 63(1), 138-138.
- Newhouse, J. P. (1992). Medical Care Costs: How Much Welfare Loss? *Journal of Economic Perspectives*, 6(3), 3-21.
- Noll, R. G., & Owen, B. M. (1983). *The political economy of deregulation : interest groups in the regulatory process*. Washington, D.C.: American Enterprise Institute for Public Policy Research.
- Olson, M. (1965). *The logic of collective action; public goods and the theory of groups*. Cambridge, Mass.,: Harvard University Press.
- Pauly, M. V. (1974). Overinsurance and Public Provision of Insurance: The Roles of Moral Hazard and Adverse Selection. *Quarterly Journal of Economics*, 88(1), 44-62.
- Pauly, M.V., Withers, K.H., Viswanathan, K. S., Lemaire, J., Hershey, J.C., Armstrong, K. & Asch, D. A. (2003). *Price Elasticity of Demand for Term Life Insurance and Adverse Selection*. NBER Working Paper No. 9925.
- Pearmain, D. (2000). Impact of Changes to the Medical Schemes Act. *South African Health Review*, 8, 183-200.
- Peltzman, S. (1976). Toward a more General Theory of Regulation. *Journal of Law & Economics*, 19(2), 211-240.
- Peltzman, S., Levine, M., & Noll, R. (1989). The Economic Theory of Regulation after a Decade of Deregulation. *Brookings Papers on Economic Activity. Microeconomics*, 1989, 1-59.

- Pigou, A. C. (1932). *The economics of welfare* (4th ed.). London,: Macmillan and co., limited.
- Porta, R. L., Lopez-de-Silanes, F., Shleifer, A., & Vishny, R. W. (1998). Law and Finance. *Journal of Political Economy*, 106(6), 1113-1155.
- Posner, R. A. (1971). Taxation by Regulation. *Bell Journal of Economics and Management Science*, 2(1), 22-50.
- Posner, R. A. (1974). Theories of economic regulation. [Article]. *Bell Journal of Economics & Management Science*, 5(2), 335.
- Posner, R. A. (1975). The Social Costs of Monopoly and Regulation. *Journal of Political Economy*, 83(4), 807-827.
- Pillay, A. (2006). *International Benchmarking of Medicine Prices in South Africa*. Paper presented at the Department of Health, Pretoria, South Africa. Retrieved September 6, 2009, from <http://www.doh.gov.za/docs/presentation/inter.html>.
- Priest, G. L. (1993). The Origins of Utility Regulation and the "Theories of Regulation" Debate. *Journal of Law and Economics*, 36(1), 289-323.
- Puelz, R., & Snow, A. (1994). Evidence on Adverse Selection: Equilibrium Signaling and Cross-Subsidization in the Insurance Market. *Journal of Political Economy*, 102(2), 236-257.
- Rahman, R. (1992). *The Australian Accounting Standards Review Board. The establishment of its participative review process* (ed.). New York: Garland Publishing.
- Reagan, M. D. (1987). *Regulation: The politics of policy* (ed.). Boston: Little Brown.

- Redford, E. (1954). The Protection of the Public Interest with Special Reference to Administrative Regulation. *American Political Science Review*, 48(4), 1103-1113.
- Reekie, W.D. (1999). *Towards Better Health Care: A Review of South African Health Care Policy*. South African Foundation. February.
- Rejda, G. E. (1995). *Principles of risk management and insurance* (5th ed.). Reading, MA: Addison Wesley.
- Rejda, G. E. (2005). *Principles of risk management and insurance* (9th ed.). Boston ; London: Addison Wesley.
- Republic of South Africa. (1956). Friendly Societies Act No. 25 of 1956.
- Republic of South Africa. (1965). Medicines and Related Substances Control Act No. 101 of 1965.
- Republic of South Africa. (1967). Medical Schemes Act No. 72 of 1967.
- Republic of South Africa. (1974). Pharmacy Act No. 53 of 1974.
- Republic of South Africa. (1974). Health Professions Act No. 56 of 1974.
- Republic of South Africa. (1975). Medical Schemes Amendment Act No. 43 of 1975.
- Republic of South Africa. (1982). Allied Health Professions Act No. 63 of 1982.
- Republic of South Africa. (1989). Medical Schemes Amendment Act No. 19 of 1989.
- Republic of South Africa. (1993). Medical Schemes Amendment Act No. 23 of 1993.
- Republic of South Africa. (1995). Pharmacy Amendment Act No. 6 of 1995.

Republic of South Africa. (1997). Pharmacy Amendment Act No. 88 of 1997.

Republic of South Africa. (1997). Medicines and Related Substances Control Amendment Act No. 90 of 1997.

Republic of South Africa. (1998). Long-Term Insurance Act No. 52 of 1998.

Republic of South Africa. (1998). Short-Term Insurance Act No. 53 of 1998.

Republic of South Africa. (1998). Medical Schemes Act No. 131 of 1998.

Republic of South Africa. (1998). Medicines and Medical Devices Regulatory Authority Act No. 132 of 1998.

Republic of South Africa. (2000). Pharmacy Amendment Act No. 1 of 2000.

Republic of South Africa. (2000). Allied Health Professions Amendment Act No. 6 of 2000.

Republic of South Africa. (2001). Medical Schemes Amendment Act No. 55 of 2001.

Republic of South Africa. (2002). Medicines and Related Substances Control Amendment Act No. 59 of 2002.

Republic of South Africa. (2003). National Health Act No. 61 of 2003.

Republic of South Africa. (2005). Nursing Act No. 33 of 2005.

Republic of South Africa. (2007). Health Professions Amendment Act No. 29 of 2007.

Republic of South Africa. (2008). Medical Schemes Amendment Bill No. 58 of 2008.

- Republic of South Africa. (2008). Prevention of and Treatment for Substance Abuse Act, No. 70 of 2008.
- Republic of South Africa. (2008). Medicines and Related Substances Control Amendment Act No. 72 of 2008.
- Richaudeau, D. (1999). Automobile Insurance Contracts and Risk of Accident: An Empirical Test Using French Individual Data. *Geneva Papers on Risk and Insurance - Theory*, 24(1), 97-114.
- Rispel, L. & Behr, G. (1992). *Health Indicators: Policy implications*. Johannesburg: Centre for Health Policy, University of the Witwatersrand.
- Rolnick, A. J., & Weber, W. E. (1986). Gresham's Law or Gresham's Fallacy? *Journal of Political Economy*, 94(1), 185-199.
- Rose-Ackerman, S. (1999). *Corruption and government : causes, consequences, and reform*. New York: Cambridge University Press.
- Rothschild, M., & Stiglitz, J. (1976). Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information. *Quarterly Journal of Economics*, 90(4), 629-649.
- Saito, K. (2006). Testing for Asymmetric Information in the Automobile Insurance Market under Rate Regulation. *Journal of Risk and Insurance*, 73(2), 335-356.
- Salop, J., & Salop, S. (1976). Self-Selection and Turnover in the Labor Market. *Quarterly Journal of Economics*, 90(4), 619-627.
- Samuelson, W., & Marks, S. G. (2006). *Managerial economics* (5th ed.). Hoboken, N.J.: Wiley.

- Sanders L. and Lloyd A. (2005). *Human Resources: International Context*. In Barron P, Ijumba P. editors. South African Health Review 2005. Durban: Health Systems Trust. Retrieved April 22nd, 2009 from http://www.hst.org.za/uploads/files/sahr05_chapter6.pdf.
- Schelling, T. C. (1960). *The strategy of conflict*. Cambridge,: Harvard University Press.
- Scherer, F. M., & Ross, D. (1990). *Industrial market structure and economic performance* (3rd ed.). Boston: Houghton Mifflin.
- Schubert, G. J. (1957). "The Public Interest" In Administrative Decision-Making: Theorem, Theosophy, or Theory? *American Political Science Review*, 51(2), 346-368.
- Selgin, G. (1996). Salvaging Gresham's Law: The Good, the Bad, and the Illegal. *Journal of Money, Credit and Banking*, 28(4), 637-649.
- Seog, S. H. (2009). Insurance Markets With Differential Information. *Journal of Risk & Insurance*, 76(2), 279-294.
- Shavell, S. (1979). On Moral Hazard and Insurance. *Quarterly Journal of Economics*, 93(4), 541-562.
- Shleifer, A. (2005). Understanding Regulation. *European Financial Management*, 11(4), 439-451.
- Shleifer, A., & Vishny, R. W. (1994). Politicians and Firms. *Quarterly Journal of Economics*, 109(4), 995-1025.
- Shortell, S. M., Gillies, R. R. & Anderson, D. A. (1994). The New World of Managed Care: Creating Organized Delivery Systems. *Health Affairs*, 13, 46-64.

- Siegelman, P. (2004). Adverse Selection in Insurance Markets: An Exaggerated Threat. *Yale Law Journal*, 113(6), 1223-1281.
- Smith, M. L., Louberge, H., & Shapiro, A. F. (1986). Insurer Management and Regulation. *Journal of Risk & Insurance*, 53(1), 178-178.
- Soto, H. D., & Instituto Libertad y Democracia (Lima Peru). (1989). *The other path : the invisible revolution in the Third World* (1st ed.). New York: Harper & Row.
- Spence, M. (1973). Job Market Signaling. *Quarterly Journal of Economics*, 87(3), 355-374.
- Spence, M. (1974). *Market signalling: information transfer in hiring and related screening processes*. Cambridge: Harvard University Press.
- Spence, M. (1976). Informational Aspects of Market Structure: An Introduction. *Quarterly Journal of Economics*, 90(4), 591-597.
- Spence, M. (1978). Product differentiation and performance in insurance markets. *Journal of Public Economics*, 10(3), 427-447.
- Spiller, P. T. (1990). Politicians, Interest Groups, and Regulators: A Multiple-Principals Agency Theory of Regulation, or "Let Them Be Bribed". *Journal of Law and Economics*, 33(1), 65-101.
- Statistics South Africa. (2000). *Mid-year Population Estimates 2000*. Pretoria: Statistics South Africa. Retrieved February 13, 2010 from <http://196.25.65.5/publications/P0302/P03022000.pdf>.
- Statistics South Africa. (2009). *Mid-year Population Estimates 2009*. Pretoria: Statistics South Africa. Retrieved February 13, 2010 from <http://www.statssa.gov.za/publications/P0302/P03022009.pdf>.

- Stigler, G. J. (1971). The theory of economic regulation. *Bell Journal of Economics & Management Science*, 2(1), 3.
- Stigler, G. J., & Friedland, C. (1962). What Can Regulators Regulate? The Case of Electricity. *Journal of Law and Economics*, 5, 1-16.
- Svenson, O. (1981). Are We All Less Risky and More Skillful Than Our Fellow Drivers? *Acta Psychologica*, 47(2), 143-148.
- Theron, N. & van Eeden, J. (2009). *The Extent of Current Cross-Subsidisation in the SA Health System*. Econex Research Note 13 July 2009. Retrieved August 17, 2009, from http://www.econex.co.za/index.php?option=com_docman&task=doc_download&gid=40&Itemid=60.
- Tollison, R. D. (1998). The Interest-Group Theory of Government. In A. Owens and C. Rowley (Eds.). *The Locke Luminary*, Vol. I, No. 1 (Summer 1998), Part 4.
- Tullock, G. (1967). The Welfare Costs of Tariffs, Monopolies, and Theft. *Western Economic Journal*, 5, 224-232.
- United Nations Development Programme. (1990). *Human Development Report 1990*. New York: Oxford University Press.
- United Nations Development Programme. (1996). *Human Development Report 1996*. New York: Oxford University Press.
- United Nations Development Programme. (2010). *Human Development Report 2010*. New York: Oxford University Press.
- Valsamakis, A. C., Vivian, R. W., & Du Toit, G. S. (2000). *Theory and Principles of Risk Management* (2nd ed.). Durban: Heinemann.

- van den Heever, A. M. (2007). *Evaluation of the merger between Network Healthcare Holdings and Community Healthcare*. Pretoria: Council for Medical Schemes.
- Venter, C., Erasmus, M. and Theron, N. (2010). *Updated GP and Specialist Numbers for SA*. Econex Health Reform Note 7 October 2010. Retrieved November 22, 2010, from http://www.econex.co.za/index.php?option=com_docman&task=doc_download&gid=72&Itemid=60.
- Villeneuve, B. (2003). Mandatory Pensions and the Intensity of Adverse Selection in Life Insurance Markets. *Journal of Risk and Insurance*, 70(3), 527-548.
- Volckart, O. (1997). Early Beginnings of the Quantity Theory of Money and Their Context in Polish and Prussian Monetary Policies, c. 1520-1550. *Economic History Review*, 50(3), 430-449.
- Von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton,: Princeton university press.
- Von Neumann, J., & Morgenstern, O. (2007). *Theory of games and economic behavior* (60th anniversary ed.). Princeton, N.J. ; Woodstock: Princeton University Press.
- Wadee, H. & Kahn, F. (2007). *Human resources for Health*. In: Ijumba P, Ntuli A, Barron P, editors. South African Health Review 2007. Durban: Health Systems Trust.
- Wambach, A. (2000). Introducing Heterogeneity in the Rothschild-Stiglitz Model. *Journal of Risk and Insurance*, 67(4), 579-591.
- Weisbrod, B. A. (1991). The Health Care Quadrilemma: An Essay on Technological Change, Insurance, Quality of Care, and Cost Containment. *Journal of Economic Literature*, 29(2), 523-552.

- Wilson, C. A. (1976). *Equilibrium in a Class of Self-Selection Models*. PhD. Thesis: University of Rochester.
- Wilson, C. A. (1977). A Model of Insurance Markets with Incomplete Information. *Journal of Economic Theory*, 16(2), 167-207.
- Wilson, G. (1964). The Effect of Rate Regulation on Resource Allocation in Transportation. *American Economic Review*, 54(3), 160-171.
- Wilson, J. Q. (1980). *The Politics of regulation*. New York: Basic Books.
- Wilson, W. (1913). *The new freedom : a call for the emancipation of the generous energies of a people*. London: Chapman and Hall.
- World Bank. (2009). *World Development Indicators*. Data retrieved July 17, 2010, from World Development Indicators Online (WDI) database.
- World Health Organisation. (2000) .*World Health Report*. Geneva: World Health Organisation.