

https://dspace.org/

The use and impacts of a corporate data warehouse: the case of the national health laboratory service

Cohen, Lyrice

2018
https://hdl.handle.net/10539/27827

Downloaded from a DSpace repository. https://dspace.org/

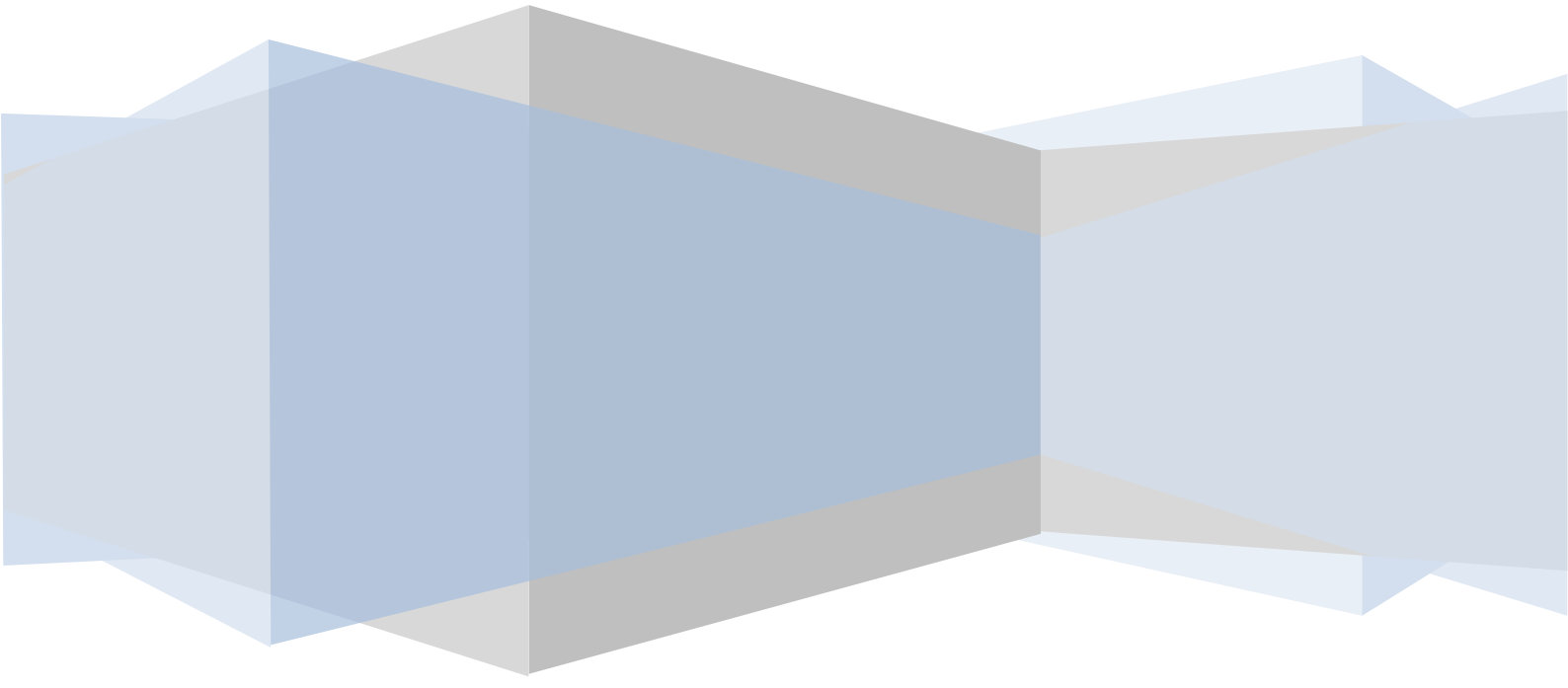
University of the Witwatersrand

The Use and Impacts of a Corporate Data Warehouse: The Case of the National Health Laboratory Service

By Lyrice Cohen (Student No: 7726556)

Submitted in fulfilment of the requirements for the degree of Masters of Commerce (MCom) at the University of the Witwatersrand

March, 2018



ABSTRACT

The promotion of population health is a key concern of government authorities and various health institutions. With the prevalence in South Africa of communicable diseases like AIDS and Tuberculosis, the increase in the incidence of non-communicable diseases like Cancer, Cardiovascular disease and Diabetes, as well as health care access, management and insurance issues, it is clear that the area of population health in South Africa is in need of urgent attention. Decisions about interventions that may impact on the health of the population tend to be far reaching and must be made with care. Basing these decisions on facts drawn from relevant data collected on various aspects of population health would be highly advantageous.

This study explores the potential and application of Business Intelligence and Data Warehouse technology for the population health space. Specifically, the National Health Laboratory Services (NHLS) was identified as the context for the study. The National Health Laboratory Services (NHLS) is responsible for laboratory tests done at public health institutions throughout South Africa, and as such keep records for 80% of the population, all of which are accessible from the NHLS Corporate Data Warehouse (CDW) as a BI tool. The full impact of the NHLS CDW cannot be realised by simply providing the technology and hoping that it is well used. The CDW can only make a difference if the information it contains is utilised by people who base their decisions on it. An understanding of the issues and challenges experienced by these people in obtaining information from the CDW would facilitate positive outcomes which would in turn promote a positive impact on population health.

Drawing on Activity Theory, this study explored and described the ways in which the NHLS CDW is being used, and the impact that this use is perceived to be having on population health in South Africa. To this end, the following research questions were addressed:

RQ1: What are the current applications of the NHLS CDW?
RQ2: What are the issues and challenges in the use of the NHLS CDW?
RQ3: What are the individual outcomes for the NHLS CDW user and data requestor?
RQ4: What are the perceived benefits (impacts) of the NHLS CDW use for the NHLS as an institution and for population health in general?

After a preliminary session with the staff of the CDW department, ten interviews were conducted with members of staff from four separate parts of the NHLS – the National Institute for Communicable Diseases, the National Cancer Registry, the National Priority Program and operational management. In relation to RQ1, the NHLS was found to be an exceptionally large organisation involved in many areas of population health. Consequently, a broad range of applications of the NHLS CDW were identified in areas as diverse as population health tracking, medical research and administration of laboratory performance. In relation to RQ2, issues and challenges were identified in terms of difficulties obtaining a patient-oriented view of the data, data quality, staffing issues, time delays, restrictions due to privacy and security and the need to access external data sources. In relation to RQ3, it was found that the NHLS CDW was considered by all participants interviewed as an essential and relevant tool in the accomplishment of their jobs – most are highly appreciative of it. In relation to RQ4, the general perception is that the CDW is invaluable to the NHLS as an organisation and for the improvement in population health in general. Evidence stored within the CDW clearly shows the impact on population health of NHLS population health programs that focus HIV/AIDS and Tuberculosis. These programs depend heavily on the CDW for their effective operation. There was, however, a definite sentiment amongst those interviewed that however much is already being done with the CDW, there is still potential for so much more.

This research has contributed to the body of knowledge and the body of evidence on data warehouses, particularly on data warehouses implemented in South Africa and/or in the population health context. The extension of the use of Activity Theory into the study of data warehouses as information systems, and the demonstration of Activity Theory use in elucidating issues and challenges in CDW use, also constitute a theoretical contribution. Contributions to practice include possible NHLS use of study findings to resolve identified issues and challenges, or utilisation by other organisations of research results as input to decisions about existing or potential datawarehouse applications.

Declaration

I declare that this is my own, unaided work except as acknowledged in the text. It has not been submitted before for any other degree at this or any other university.

Lyrice Cohen 29 March 2018

Acknowledgements

It has been an absolute privilege to have Prof Jason Cohen as a supervisor, and I thank him for his enthusiasm, his subtle and patient teaching, and his remarkable responsiveness - despite a heavy work load. Jason's lucidity of thought, and depth of knowledge and understanding are remarkable and I am deeply grateful to have him as a mentor.

I would also like to thank all the members of NHLS staff who allowed me to interview them and provided me with so much information and so many insights.

The support and encouragement of my colleagues in the Information Systems department was invaluable, and this, too, is deeply appreciated.

Dedication

I dedicate this report to my very special family.

To my precious husband, who has endured evening after evening and weekend after weekend without my presence, thank you for being the sensitive and kind person that you are.

To my wonderful children who bring light and energy into my life.

And last but not least, to my very special parents, without whose help and support (especially doing lift schemes) I would never have been able to complete this report.

Table of Contents

List of Figures xii

List of Tables xiv

1 Introduction 1

1.1 Introduction to Business Intelligence and Population Health 1

1.2 Problem Statement and Research Context 2

1.3 Research Questions 4

1.4 Research Objectives 5

1.5 Contributions of this Study 6

1.6 Report Structure 6

2 Literature Review 8

2.1 Questions Guiding the Systematic Review 8

2.2 Methods for the Literature Review 10

2.2.1 Literature Search Strategy 10

2.2.2 Study Selection and Characteristics 12

2.3 Results 15

2.3.1 Question 1: What applications of BI for better decision making in the area of population health have been implemented? 15

2.3.1.1 Tracking and Surveillance of Population Health 16

2.3.1.2 Sharing Data and Resources 21

2.3.1.3 Enabling Fact-based Decision Making 22

2.3.2 Question 2: What evidence links the use of BI to better population health outcomes? 22

2.3.3 Question 3: What challenges and issues have been faced in the implementation and use of BI for the betterment of population health? 24

2.3.3.1 Privacy and Security 24

2.3.3.2 Data Quality 25

2.3.3.3 Development and Maintenance of BI infrastructure 27

2.4 Discussion 28

2.4.1 Applications of BI for Population Health 28

2.4.2 Evidence Linking the Use of BI for Population Health to Better Outcomes 29

2.4.3 Challenges and Issues in Implementing BI for Population Health 29

2.5 Limitations 30
2.6 Recommendations for Practice 30
2.7 Conclusion and Implications for the Study 31

3 Theoretical Underpinning : Activity Theory 34
3.1 Theoretical Framework Chosen for the Study 34
3.2 Activity Theory 34
3.3 Activity Theory and Information Systems Usage 37
3.4 An application of Activity Theory to a Study on Sustaining Technology Innovations in Teacher Education 38
3.5 An application of Activity Theory to a Study on the Tensions in a Nepali Rural Telecentre 40
3.6 The Context of the NHLS CDW 41
3.7 Developing an Activity System for a Request for Data from the NHLS CDW 43
3.8 Chapter Conclusion 44

4 Methodology 46
4.1 Research Philosophy 47
4.2 Research Approach 47
4.3 Research Strategy 48
4.4 Data Collection 50
4.5 Interviews 53
4.6 Data Analysis 57
4.7 Method Evaluation 59
4.8 The Role of the Researcher 61
4.9 Reliability and Validity 62
4.10 Ethical Clearance 63
4.11 General limitations 64
4.12 Chapter Conclusion 64

5 Research Findings 65
5.1 Profile of Participants and Information Requests 65
5.2 Profile of Documents Collected 66
5.3 Themes Discovered and the Activity System for an Information Request . 66

Literature, eg. 69

Rizi and Roudsari (2013) 69

5.4 RQ1: What are the current applications of the NHLS CDW? 70

5.4.1 General Applications of the CDW 71

5.4.2 Applications of the CDW for Information Requests 74

5.4.2.1 Applications of the CDW discovered at the NICD 74

5.4.2.1.1 NICD Applications of the CDW for Public Health Tracking 75

5.4.2.1.2 NICD Applications of the CDW for Research 76

5.4.2.2 Applications of the CDW discovered at the NCR 76

5.4.2.2.1 NCR Applications of the CDW for Public Health Tracking..... 76

5.4.2.2.2 NCR Applications of the CDW for Research..... 77

5.4.2.3 Applications of the CDW discovered from Operational Management..... 77

5.4.2.3.1 Operational Management Applications of the CDW for Administration 77

5.4.2.4 Applications of the CDW discovered at the NPP 78

5.4.2.4.1 NPP Applications of the CDW for Public Health Tracking..... 79

5.4.2.4.2 NPP Applications of the CDW for Administration..... 80

5.4.2.5 NPP Applications of the CDW for Research..... 80

5.4.2.6 RQ1 Summary 80

5.5 RQ2: What are the issues and challenges in the use of the NHLS CDW? ... 81

5.5.1 Relationship with the CDW Team 82

5.5.2 Staffing 86

5.5.3 Time Delays 89

5.5.4 Outsourcing 93

5.5.5 Data Quality..... 94

5.5.6 Patient-oriented View 100

5.5.7 Privacy and Security 104

5.5.8 Direct Access..... 108

5.5.9 Metadata..... 110

5.5.10 Ease and Efficiency..... 112

5.5.11 External Data 114

5.6 RQ3: What are the individual outcomes for the NHLS CDW data requestor? 117

5.7 RQ4: What are the perceived benefits (impacts) of NHLS CDW use for the NHLS as an institution and for population health in general? 120

5.8 Chapter conclusion..... 127

6 Discussion of Results 128
6.1 RQ1: What are the current applications of the NHLS CDW? 128
6.1.1 NHLS CDW Applications Classified by Purpose 130
6.1.1.1 NHLS CDW Applications for Public Health Tracking 130
6.1.1.2 NHLS CDW Applications for Administration and Management 131
6.1.1.3 NHLS CDW Applications for Medical Research 132
6.1.2 NHLS CDW Applications Classified by Request Type 132
6.1.2.1 Information Requests made through a Customised Application 133
6.1.2.2 Information Requests made through a CDW Service Representative 133
6.1.2.3 Information Requests made through a Data Analyst 133
6.1.2.4 Information Requests made through a High level CDW access tool 134
6.1.2.5 Information Requests made through a Low level CDW Access Tool 134
6.1.3 NHLS CDW Applications Classified by Originating Division 135
6.2 RQ2: What are the issues and challenges in the use of the NHLS CDW? .. 136
6.2.1 'Privacy and Security' 137
6.2.2 'Data Quality', 'Data Governance' and 'Patient-oriented View' 137
6.2.3 'Staffing', 'Relationship with CDW Team', 'Time Delays', 'Outsourcing' and 'Direct Access' 141
6.2.4 'Metadata' and 'Ease and Efficiency' 143
6.2.5 'External Data' 144
6.2.6 Most Serious Challenges 145
6.3 RQ3: What are the individual outcomes for the NHLS CDW user and data requestor? and
RQ4: What are the perceived benefits (impacts) of the NHLS CDW use for the NHLS as an institution and for population health in general? 145
6.4 Chapter Conclusion 148
7 Conclusion 149
7.1 Summary of the Study 149
7.2 Contributions of the Study 150
7.3 Recommendations for the Future 151
7.4 Limitations of the Study 152
7.5 Conclusion 153
References 154
APPENDIX A: Interview Protocol 159

APPENDIX B: First Page of Template for Interview Transcription..... 164

APPENDIX C: Illustration of Unitising Transcribed Interviews and Colour Coding for Themes..... 165

APPENDIX D: Participant Information Sheet 167

APPENDIX E: Interview Consent Form..... 169

APPENDIX F: Recording Consent Form 171

APPENDIX G: Letter of Permission from the NHLS..... 173

APPENDIX H: Final Ethics Clearance Certificate 174

APPENDIX I: NHLS CDW Information Request Form 175

List of Figures

Figure 1: SLR Information Flow Diagram. 11
Figure 2: Common reformulation of Vygotsky's mediated act (Engestrom, 2001) 35
Figure 3: Hierarchical Levels of an Activity (Kuutti, 1996) 35
Figure 4: The Structure of Human Activity (Adapted from(Engestrom, 1987)) 36
Figure 5: Activity System for Study on Sustaining Technology Innovations in Teacher Education (Johnson, 2004) 39
Figure 6: Activity System featuring four sub-triangles (Johnson, 2004) 39
Figure 7: Production Triad Highlighted (Adapted from (Johnson, 2004)) 40
Figure 8: Smaller Activity Theory Triangle. (Lee, 2010) 41
Figure 9: Smaller Activity Theory triangle in context of larger Activity Theory triangle (Lee, 2010) 41
Figure 10: Summary Structure of an Information Request Activity 43
Figure 11: Structure of an Information Request Activity 45
Figure 12: The Research 'Onion' (Saunders et al., 2009) 46
Figure 13: Original Quota Sampling Matrix 52
Figure 14: Information Requests Sampled 53
Figure 15: Summarised Structure of an Information Request Activity 55
Figure 16: Information Requests Sampled 66
Figure 17: Allocation of Themes to Activity System for an Information Request 70
Figure 18: Primary Functions of the NHLS CDW (Source: National Health Laboratory Service Presentation 'Data Warehouse Overview') 70
Figure 19: Information Requests Categorised by Sampling Categories 72
Figure 20: Mobile Clinic 73
Figure 21: Allocation of Themes to Activity System for an Information Request 82
Figure 22: Sub triangle of the activity system assigned to 'Relationship with CDW Team' (Through Customised Application) 82
Figure 23: Sub triangle of the activity system assigned to 'Relationship with CDW Team' (Through CDW Service Representative or Data Analyst) 83
Figure 24: Sub triangle of the activity system assigned to 'Staffing' 86

Figure 25: Sub triangle of the activity system assigned to 'Time Delays' (due to Staffing and the Information Request Iterative Cycle) 89
Figure 26: Sub triangle of the activity system assigned to 'Time Delays' (due to Technical Problems) 89
Figure 27: Sub triangle of the activity system assigned to 'Time Delays' (due to Prioritisation) 90
Figure 28: Sub triangle of the activity system assigned to 'Outsourcing' 93
Figure 29: Sub triangle of the activity system for 'Data Quality' (within CDW responsibility) 95
Figure 30: Sub triangle of the activity system for 'Data Quality' (Outside of CDW responsibility) 95
Figure 31: Sub-triangle of the Activity System for 'Patient-oriented View' 100
Figure 32: Presentation Slide showing CDW Capabilities 102
Figure 33: Sub-triangle of the Activity System for 'Privacy and Security' 104
Figure 34: Conditions Specified on NHLS Request for Data Form 105
Figure 35: Sub-triangle of the Activity System for 'Direct Access' 108
Figure 36: Sub-triangle of the Activity System for 'Direct Access' 110
Figure 37: Sub-triangle of the Activity System for 'Ease and Efficiency' 112
Figure 38: Sub-triangle of the Activity System for 'External Data' 114
Figure 39: Part of the Activity System for 'Individual Outcome' 117
Figure 40: Part of the Activity System for 'NHLS and Population Health Outcome' 120
Figure 41: First Four Google Scholar Results for Search on ("National Health Laboratory Service" "Corporate Data Warehouse") 122
Figure 42: Information Requests Categorised by Application Purpose, Information Request Type, and Division of Origin 129
Figure 43: Allocation of Themes to Activity System for an Information Request 136
Figure 44: Sub-triangle of the Activity System for 'Patient-oriented View' – Altered Perspective 140

List of Tables

Table 1: Terms in title considered to refer to a BI Intervention 12

Table 2: Terms in title considered to refer to a Health Outcome 14

Table 3: Applications of BI for Population Health and Aspects of BI Use 17

Table 4: Ways of Requesting Information Found in the Literature 51

Table 5: Possible Information Request Purposes 52

Table 6: Mapping of Interview Protocol questions to elements of the Information Request Activity 56

Table 7: Interviewees by Length of Time Making Information Requests 65

Table 8: Interviewees by Job Function 65

Table 9: Documents Collected 67

Table 10: Colour Coded Themes 69

1 Introduction

This study explores the potential and application of Business Intelligence and Data Warehouse technology for the population health space through a case study done at the National Health Laboratory Services (NHLS). The NHLS is responsible for laboratory tests done at all public health institutions throughout South Africa, and keep records of test results in the NHLS Corporate Data Warehouse (CDW) as a BI tool.

1.1 Introduction to Business Intelligence and Population Health

"Business Intelligence" (BI) is an area of Information Technology (IT) that involves the collection, analysis and presentation of large amounts of data. The concept of "Business Intelligence" was first promoted in the early 1990s by Howard Dresner of the Gartner group, when he used the term to encourage Businesses to exploit the data that had been collected by their IT systems (Shollo and Kautz, 2010). Subsequently, the term has come to describe a range of applications and technologies that collect data to make it available for access and analysis that will aid in decision making(Watson, 2009). BI systems thus combine data gathering, storage, and management with analytical tools to enable the presentation of complex internal and external information to organisational decision makers (Negash, 2004). Analysis requires data – the richer the better. It is clear from the two definitions given that the gathering and storage of the data is a fundamental part of BI. Information needed for analysis is generally spread across multiple disparate data sources both within and external to an organisation. Inmon (2005) describes how the productivity and data credibility problems that arise from attempting to retrieve required data from these sources led him to suggest the concept of a data warehouse, "an integrated, subject-oriented, time-variant, non-volatile collection of data that provides support for decision making" (Inmon, 1995: 1). The data warehouse, also known as the Enterprise Data Warehouse (EDW) or the Corporate Data Warehouse (CDW), provides a central pool of comprehensive data for analysis and is thus a core part of an organisation's BI infrastructure.

These systems have been successfully applied to promote good decision making in a variety of environments, eg. Retail (Agarwal, 2014, Phan and Vogel, 2010), Manufacturing (Gangadharan and Swami, 2004, Barakat et al., 2013), Banking (Posthumus and Bank, 2008, Watson, 2010), and even Air and Rail Travel (Watson et al., 2006, Wixom et al., 2011). Despite some work, eg. Kulkarni et al. (2017), their broader application in fields such as health care has received less attention. The promotion of population health is a key concern of government authorities and various health institutions.[1] Decisions about interventions that may impact on the health of the population tend to be far reaching and must be made with care. Basing these decisions on facts drawn from relevant data collected on various aspects of population health would be highly advantageous (Krumholz, 2014, Mooney et al., 2015, Pfeiffer and Stevens, 2015, Shani, 2000, Musa et al., 2013, Olszak and Batko, 2012, Asokan and Asokan, 2015).

[1] The field of population health is concerned with the various social, economic, biological and environmental determinants that may contribute to shaping population health on the one hand, and on the other hand as the definition and measurement of "the health outcomes of a group of individuals, including the distribution of such outcomes within the group" (Kindig, 2003: 380).

With the prevalence in South Africa of communicable diseases like HIV/AIDS and Tuberculosis (TB), the increase in the incidence of non-communicable diseases like Cardiovascular disease and Diabetes[i], as well as health care access, management and insurance issues, it is clear that the area of population health in South Africa is in need of urgent attention. Initiatives in this regard are likely to benefit from the use of trustworthy and relevant information about population health that decisions can be based on. Business Intelligence and the underlying data warehouse infrastructure can enable the collection, extraction and analysis of information to inform such decision making. Without the actual application of these technologies in the field of population health, however, these benefits cannot be realised.

1.2 Problem Statement and Research Context

Applications of data warehouses in commerce, are concerned primarily with assisting the organisations that own them to earn more money, eg. Anderson-Lehman et al. (2004). In contrast, applications of data warehousing for population health are concerned with the public good, eg. reduction of diseases; better responses to disease outbreaks; better preventative measures, and could consequently result in a healthier population (see also Kindig (2007)). The potential applications of data warehouses for the betterment of population health are numerous – the literature review has shown that similar collections of data are often used for the tracking and surveillance of population health, to promote early detection, diagnosis and treatment, managing and monitoring of various diseases (Berndt et al., 2003, Ben Said et al., 2005, Wisniewski et al., 2003, Rizi and Roudsari, 2013). Though, as shown later in the literature review, publicly available information gives some idea as to what some of the applications of data warehousing are within public health, such as sharing data across different medical entities, enabling fact based decision making, and tracking and surveillance of public health, no overall study of the use of a data warehouse in the South African public health care context has been found, and none that explore the mechanisms, challenges and issues involved in obtaining required information from it, and certainly none in a national laboratory context. The lack of studies of this nature constitutes the literature gap which has been addressed by this research.

Given this problem, the context selected for this study is the Corporate Data Warehouse (CDW) of the National Health Laboratory Services (NHLS) in South Africa.

The National Health Laboratory Services (NHLS) was established by the South African government in an Act of Parliament in 2001 which essentially brought together all laboratories in the public health system. This involved the amalgamation of what was then the South African Institute for Medical Research (SAIMR), the National Institute for Virology, the National Centre for Occupational Health, and the university and provincial pathology laboratories.[2] The NHLS has approximately 7000 members of staff and over 260 laboratories[3] in all 9 provinces of South Africa, and is responsible for all laboratory tests done at public health institutions throughout the country - covering an estimated 80% of the national population.[4]

[2] http://www.nhls.ac.za/?page=who_are_we&id=17
[3] http://www.nhls.ac.za/
[4] http://www.nhls.ac.za/?page=about_us&id=16

The NHLS developed a CDW to support its aims to contribute to the building of "a healthy nation through early detection, diagnosis and monitoring to prevent, manage and treat diseases to reduce the disease burden, and promote health and wellness, thereby ensuring a 'long and healthy life for all South Africans'". The NHLS CDW houses data on laboratory tests and patient demographics data for 80% of the South African population. This information is vitally important for the National and Provincial Departments of Health, enabling them to monitor their health programs and measure outcomes of health interventions (NHLS, 2015) .

The use cases for the CDW support the various programs within the NHLS for disease monitoring and prevention and health promotion. For example, the extremely high incidence and burden of disease of both HIV/AIDS and Tuberculosis (TB) in South Africa led the NHLS to create a National Priority Programs (NPP) division in 2010. The original mission was "to provide affordable, accessible HIV and TB diagnostic services that yield accurate, reliable, relevant and timely results aligned with National Department of Health (NDoH) strategy and are based on current scientific knowledge and international norms."[5] This has since been extended to include active facilitation of and assistance with national programs like early infant HIV diagnosis, HIV drug resistance, TB GeneXpert testing, and support for Correctional Services and for mining and peri-mining communities. The NPP employs program managers to oversee several of these programs, and the team is also assisted by experts including scientists, pathologists, biomedical engineers and a health economist. Some of the programs receive funding from several sources, including the Global Fund, and there is collaboration with many organisations like the World Health Organisation (WHO), the US Centre for Disease Control (CDC), the Gates Foundation, the Clinton Foundation and others.[6]

Aside from the pathology infrastructure, the NHLS has four important additional divisions[7]. These are the National Institute for Communicable Diseases (NICD), the National Institute for Occupational Health (NIOH), the National Cancer Registry (NCR) and the South African Vaccine Producers (SAVP) .[8]

The NICD is a centre for disease and surveillance control and provides information and expertise on outbreak and communicable disease burdens to southern African regions. The NICD also undertakes research on communicable diseases in Southern Africa.[9]

The NIOH analyses occupational environments and investigates occupational diseases. The services it offers include health hazard evaluations, information provision and the provision of statutory autopsies.[10]

---

[5] https://www.wits.ac.za/pathology/divisions/molecular-medicine--haematology/research-and-analytical-services/national-priority-programme-npp/

[6] *ibid*

[7] In one of the interviews conducted for this study it was brought to light that there are currently plans to remove the NICD, the NIOH and the NCR from the NHLS and put them into a new public institution called the National Public Health Institute of South Africa (NaPHISA). There is a bill to this effect that should be passed in Parliament in 2018, and NaPHISA should become operational in 2019. For the time period covered by this study, however, the NICD, NIOH and NCR remain part of the NHLS.

[8] http://www.nhls.ac.za/?page=divisions&id=20

[9] http://www.nhls.ac.za/?page=national_institute_for_communicable_diseases&id=36

[10] http://www.nioh.ac.za/?page=services&id=19

The NCR was first established in 1986. It is a surveillance unit that collates, analyses and records all newly reported cancer cases, thus contributing significantly to the awareness of the growing problem of cancer in South Africa.[11]

The SAVP is responsible for the manufacture of the anti-venom used to treat snake bites, spider bites and scorpion stings. Actually a subsidiary of the NHLS, they are the only manufacturers of this type of anti-venom in sub-Saharan Africa (NHLS, 2017).[12]

All NHLS Laboratories use a standardised Laboratory Information system (LIS) called TrakCare to record laboratory tests done and their results. Test information is also transmitted from all the different laboratories in South Africa to the NHLS Corporate Data Warehouse (CDW) in Johannesburg. Since NHLS laboratories are responsible for all laboratory tests done at public health institutions throughout the country, laboratory test records for approximately 80% of the population are accessible from the NHLS CDW. It is not surprising, therefore, that the NHLS regards the CDW as "the organisation's gem and treasure" and employ it to help the national and provincial departments of health to monitor health programs and interventions (NHLS, 2015). Data from the NHLS CDW has also been used for various research projects, including studies on mother to child HIV/AIDS transmission (Sherman et al., 2014b), Tuberculosis (Erasmus et al., 2008, Nanoo et al., 2015), a Cholera outbreak (Archer et al., 2009) and the status of Hepatitis C in South Africa (Prabdial-Sing et al., 2013).

An understanding of the issues and challenges involved in the use of the CDW would facilitate positive outcomes which would in turn promote a positive impact on population health. This research, therefore, has a broader focus than the information and how it could lead to better decision making, but rather includes also the issues and challenges involved in the actual activity of requesting and obtaining information from the CDW. This activity of obtaining information is referred to as 'use' in the study. The full impact of the NHLS CDW cannot be realised by simply providing the technology and hoping that it is well used. The CDW can only make a difference if the information it contains is utilised by people who base their decisions on it.

1.3 Research Questions

This study explores and describes the ways in which the NHLS CDW is being used, the issues and challenges encountered in this use, and the impact that this use is perceived to be having on population health in South Africa. This study is carried out from an activity theory perspective, and from this perspective the users of the CDW can be referred to as 'information requestors'. They function within the social context of the organisation, its people and its rules. Information requestors may be academics doing medical research, officials making decisions about population health or administrative staff seeking to improve management of laboratories. Any understanding of the issues and challenges concerning information requests would therefore need to be considered within the context that the information requestors find themselves in. Consequently, the unit of analysis for this study will be the activity of an information request, which, in addition to the user

[11] http://www.nhls.ac.za/?page=national_cancer_registry&id=38
[12] http://www.savp.co.za/

(information requestor) and the information being requested, includes social (other employees, rules) and technical (CDW and related) aspects of the environment.

The following research questions are addressed in the study:

RQ1: What are the current applications of the NHLS CDW?

RQ2: What are the issues and challenges in the use of the NHLS CDW?

RQ3: What are the individual outcomes for the NHLS CDW data requestor?

RQ4: What are the perceived benefits (impacts) of NHLS CDW use for the NHLS as an institution and for population health in general?

1.4 Research Objectives

To address the abovementioned research questions the following research objectives were met.

Firstly, a systematic literature review of the application of BI in the field of population health was conducted. This enabled the identification of possible uses of BI for the benefit of the health of the population, and also exposed the researcher to issues and challenges that occur in these types of BI implementations.

Secondly, a theoretical framework was chosen for the study. Activity Theory provides for the study of an activity within the context of the organisation. Since the activity of requesting information from the NHLS CDW is done within the context of an organisation, and should be understood within that context, Activity Theory was considered to be the more appropriate framework to use.

Thirdly, Interpretivism was chosen as an appropriate paradigm for this study as the use of Activity Theory required a subjective engagement with people in the organisation requesting information from the NHLS CDW. This subjective engagement took the form of interviews where the interview protocol was constructed taking cognisance of both the information discovered in the literature review and the different elements of Activity Theory. The textual nature of the data led to the decision that qualitative, rather than quantitative, analysis would be used.

Next, data was collected. Since this research was exploratory in nature, it was necessary to hold a preliminary discussion session with staff of the CDW team to establish some basic information and determine how to proceed.

Semi-structured interviews were then held with information requestors in four different divisions of the NHLS. There were ten interviews which took place and during this process documents, screens, forms and relevant research publications were collected when possible so that they could be used to verify the interview results.

Once collected, it was necessary to analyse and present the data. Interviews were recorded and then subsequently transcribed so that they could be analysed qualitatively. Common themes began to emerge during the interview process, but were clarified after transcription when the data was unitised and the units assigned to different themes. All units corresponding to the same theme were then gathered together so that the results could be structured and presented. Aspects of method evaluation were considered during the analysis of results, to improve the rigour with which it was done.

Finally, once results had been presented and interpreted the research was concluded with recommendations for practice and theory and a discussion of the limitations of the research.

1.5 Contributions of this Study

Contributions to theory have been made by this study in various ways. Firstly, the research gap has been addressed by reporting on a data warehouse implemented in the health sector and used for the benefit of population health and thus adding to the body of existing knowledge on data warehouses and business intelligence technologies, the sectors they are implemented in and their uses. The empirical analysis of issues and challenges encountered while making an information request also adds to the body of evidence on data warehouse used for population health – the systematic literature review revealed only one empirical study, and that was not focused on the data warehouse itself, but rather on OLAP, which is a tool for accessing the data warehouse (Tremblay et al., 2007). Secondly, although studies using activity theory as a lens through which to study information systems were found, none of them focused specifically on data warehouses, so a contribution has been made by extending activity theory into the study of data warehouses as information systems. Finally, this study contributes to the knowledge on data warehouses implemented in the South African context.

In terms of contributions to practice, this research could enlighten officials concerned with population health about the benefits of data warehousing, which may result in the construction of other similar facilities for public benefit. Organisations that are currently developing or maintaining a data warehouse may also benefit from the experiences of the NHLS CDW implementation and the issues and challenges identified. Organisations with data warehouses may be encouraged to use additional business intelligence tools and approaches that participants have reported as effective when used with the NHLS CDW, and organisations considering building a data warehouse may utilise this study's result as an input in making a decision.

1.6 Report Structure

The remainder of this report is divided into the following chapters:

Chapter 2: Systematic Literature Review

This chapter is a systematic literature review of literature that has been published on the use and impacts of BI and data warehouses on population health. The broader topic of BI as opposed to just

data warehousing was chosen for the review since data warehousing (and thus the CDW) constitutes only one aspect of BI that is used in conjunction with many other aspects of BI. The review examines applications of BI that have been implemented for the benefit of population health and which aspects of BI they use, looks for evidence that links the use of BI for population health to better outcomes, and explores what challenges and issues have been faced in the implementation and use of BI for the betterment of population health.

Chapter 3: Theoretical Underpinnings

This chapter discusses Activity Theory and looks at two of the applications of Activity Theory in Information Systems related research. The choice of Activity Theory as a framework for this study is then justified. Using information discovered during the course of the study, the Activity System for an information request from the NHLS CDW is formulated.

Chapter 4: Methodology

This chapter describes the methodology that was used to do the research. After exploring and justifying the choices of research philosophy, approach and strategy, it reviews the steps and approach involved in data collection. Chosen methods of sampling are explained, and a description is given of the type of interviews conducted. The development of the Interview protocol is then outlined. Steps involved in the qualitative analysis done are detailed and the means of ensuring rigour with method evaluation are considered. Finally, the process of ethics clearance is reported on, followed by a discussion of the general limitations applicable to this type of study.

Chapter 5: Presentation of Results

After creating a profile of participants interviewed and of other documents and research publications collected as data, this chapter presents the results as derived from that data. Each research question is tackled in turn. Evidence relating to each common theme is discussed in turn and where possible the identified theme is mapped on to the activity system for an information request that was developed in chapter 3.

Chapter 6: Interpretation of Results

Results presented in chapter 5 are now interpreted, taking cognisance of any related literature. Attention is paid to highlighting any contradictions that may have been found within the activity system for an information request.

Chapter 7: Conclusion

This chapter begins with a synopsis of the study, describing the overall objective, the process executed and the results obtained. Finally, conclusions are drawn and recommendations given for both theory and practice and limitations specific to this study are pointed out.

2 Literature Review[13]

The purpose of this study was to examine the uses, issues and challenges, and impacts on population health of a particular application of BI, namely the Corporate Data Warehouse (CDW) implemented by the National Health Laboratory Service (NHLS) in South Africa. The CDW, or data warehousing in general, can be seen as one aspect of BI that is used in conjunction with several of the other aspects of BI, eg. reporting, analytics, etc. Proper treatment of this research thus necessitated a prior investigation into other existing applications of BI in the population health field. Consequently a systematic literature review was carried out that explored what uses have been made of BI in the population health field, what impacts this use has had on population health, and what issues and challenges were encountered in this use. The results of that systematic literature review are discussed in this chapter.

First, the set of questions that guided the review are presented. Then the literature search strategy and the process of selecting relevant articles from those returned by the database searches are described. The chapter then reports on information found in the selected articles. Following this, the limitations of the review are discussed along with the implications for this study.

2.1 Questions Guiding the Systematic Review

The following questions guided the systematic review and were intended to provide a broad overview of the effort that is being made in the application of BI in the field of population health, as well as a more specific summary of evidence connecting the BI intervention with success. The resulting picture of the current status of the use of BI in population health will facilitate the identification of gaps in the research and will highlight ways in which BI can be applied to assist government authorities and health institutions in making fact based decisions that will lead towards a healthier population. Results will also inform potential implementers about the areas in which there is previous experience, as well as the issues and challenges they may face when initiating a new application.

The questions for the review were:

SLRQ1 : What applications of BI for better decision making in the area of population health have been implemented?
SLRQ2: What evidence links the use of BI for population health to better outcomes?
SLRQ3: What challenges and issues have been faced in the implementation and use of BI for the betterment of population health?

[13] This chapter is an updated version of the author's paper published as Cohen, L. (2017, September). Impacts of business intelligence on population health: a systematic literature review. In Proceedings of the South African Institute of Computer Scientists and Information Technologists (p. 9). ACM. It has been reformatted and extended for the purposes of this dissertation.

Answering SLRQ1 revealed existing applications of BI in the population health field that were then compared to applications of the CDW found at the NHLS. This was important in answering the study's first research question. BI aspects used in these applications were also detailed and are shown in table 3 on page 17. Data warehouses had been implemented for seven out of the eleven applications found.

Answering SLRQ2 identified empirical evidence linking the use of BI to better population health outcomes and thus gave an idea of what empirical research had been done in this area.

Answering SLRQ3 highlighted issues and challenges that had been faced in the implementation and use of BI. This was extremely helpful in identifying possible issues and challenges for NHLS CDW use and in the initial development of the interview protocol for this research. Issues and challenges identified during the research were also compared to those found in the literature review.

2.2 Methods for the Literature Review

2.2.1 Literature Search Strategy

The development of the protocol for the literature review was informed and adapted from the preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 Statement (Moher et al., 2015).

In order to find articles published in both the medical and information technology areas, both the Pubmed and Science Direct (Elsevier) electronic databases were searched, although the Science Direct search was restricted to three relevant categories, being: “Decision Sciences”, “Medicine and Dentistry” and “Nursing and Health Professions”. In addition, searching the ACM and IEEE databases, with their focus on computing oriented research fields, helped ensure a broader inclusion of computing literatures and also ensured that conference papers, and thus research in progress, would be included. This was considered important since the application of BI in the population health area is relatively new. In all cases, searches were restricted to English journal articles whose full text was available to the authors.

The flow diagram shown in figure 1 on page 11 documents the search strategy and process of document selection that was carried out.

The search string used on all four databases included the terms (“Population Health” OR “Public Health”) AND (“Business Intelligence” OR “Data Warehouse” OR “Data Warehousing” OR “Big Data” OR “Data Science”) AND (Mortality OR Morbidity OR “Well being” OR Disability OR “Health State” OR “Health Status” OR “Quality of Life” OR Policy). These terms ensured that reviewed literature covered the context of population health within which the study is done, the intervention being investigated, namely Business Intelligence (BI) technologies, and the population health outcomes of BI. Outcome measures of population health suggested by Kindig (2007) are Mortality rate (number of deaths within a prescribed time), Morbidity (departure from a state of well being), Disability (reduction of capacity to function in society), Health state/health status (health of an individual at a point in time), Quality of Life (concerning all aspects of an individual’s existence), Health-related quality of life (impact of health aspects on quality of life) as well as two measures that are summaries of the afore mentioned. Kindig (2007) also points out that although health policy is not itself an outcome measure, it exerts an important influence on the actual population health outcomes. These outcomes were thus included in the search terms.

The four databases searched returned the following number of results: Science Direct 409, Pubmed 175, ACM 158 and IEEE 400, making a total of 1142. These were all loaded into the Endnote reference management tool, where duplicates and incomplete citation records were removed, leaving a total of 1124 entries. Relevant articles could now be selected from these entries, first by looking at the abstracts of the articles, then at the full text.

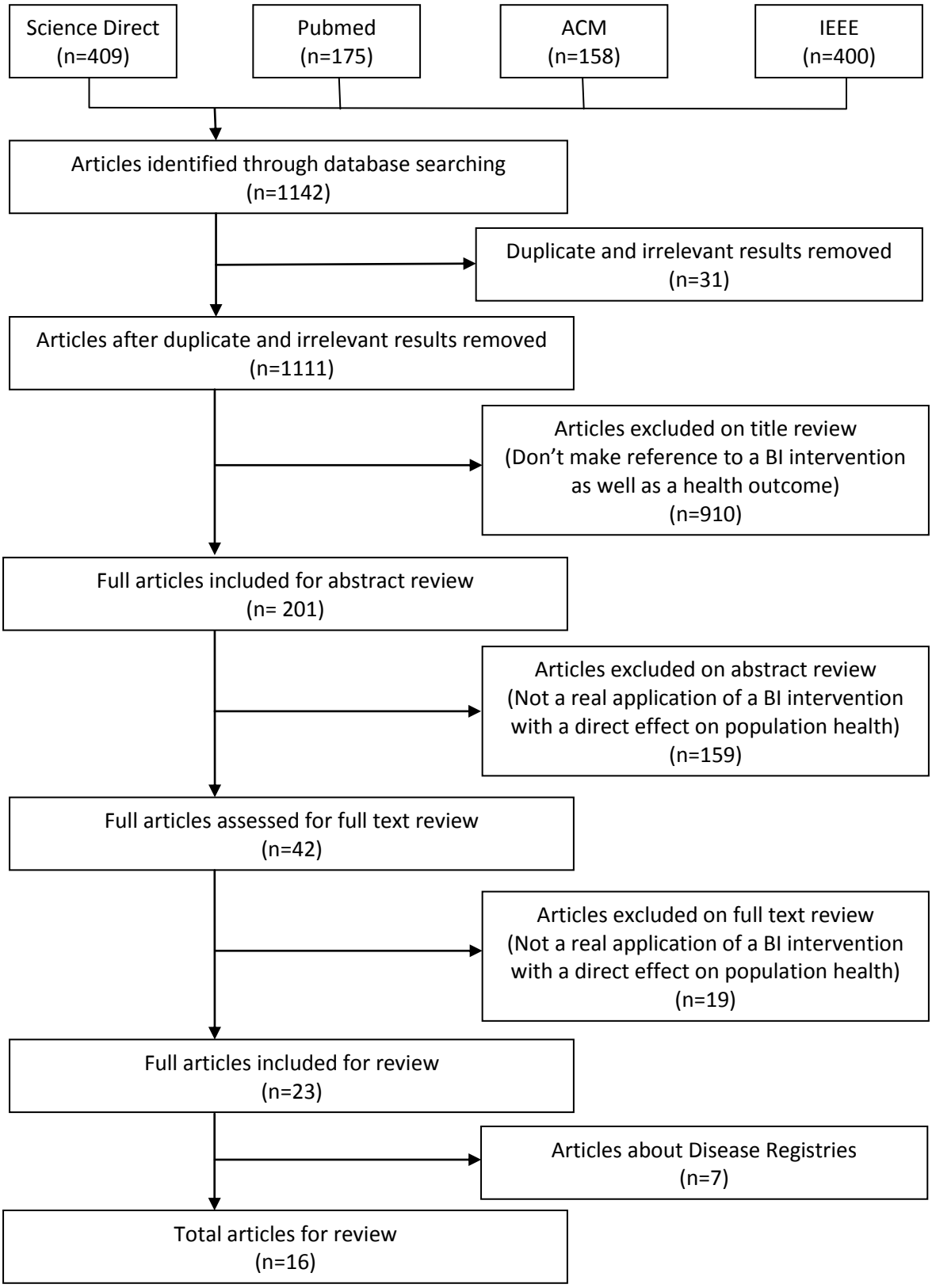


Figure 1: SLR Information Flow Diagram.

2.2.2 Study Selection and Characteristics

All titles in the results list were reviewed, and were only retained if they made reference to a BI intervention as well as a health outcome.

Articles were considered to reference a BI intervention if their titles included words or phrases like 'business intelligence', 'data warehouse', 'big data', 'data science', 'registry', 'repository' , 'learning health system', 'surveillance' (referring to the monitoring of data), 'data' in conjunction with another word like 'collecting', 'sharing', 're-using', 'quality', 'security', 'privacy' , or if their titles made mention of any of the plethora of BI tools and techniques, like Data Mining, OLAP, GIS etc. Table 1 contains a full list of all terms that were considered to refer to a BI intervention. Based on the above rationale, 910 papers were excluded, leaving 201 articles selected on the basis of their titles.

General	Business Intelligence Big Data
Data Collection and Integration	Data Warehouse / Warehousing Registry Repository Data collection Data sharing Data integrating Data quality Data privacy Data security Federated data Federated querying Harmonising databases Community information Distributed heterogeneous information Surveillance
Data Analysis and Presentation	Data Science Online Analytical Processing (OLAP) Cube (often used to refer to OLAP) Data Mining Predict / Prediction Analysis / Analytics Geographical Information Systems (GIS) Model Decision making / support Text search Text mining Visualisation Data exploring / exploration Indicator / Measure / Index Bio-intelligence Collective intelligence

Table 1: Terms in title considered to refer to a BI Intervention

References to a health outcome included all the items in the last bracketed term of the search string, as well as any generic words like ‘health’, ‘medicine’, etc. Diseases, health-related behaviour, medical conditions and medical interventions as possible health outcomes were also included; a full list of those encountered is given in table 2, which lists all terms considered to be a health outcome. With the intention of omitting articles whose focus is not on population health, titles including terms like ‘medical insurance’, ‘healthcare costs’, and ‘healthcare admin’ were rejected. Titles speaking of ‘research’ were only accepted if there was some other indication of population health in the title.

The next phase involved the examination of the abstracts of these 201 articles. Articles were retained if the abstract indicated that they were looking at an actual application of BI for population health. Where it was difficult to ascertain from the abstract whether the article should be kept or not, the article was retained so that a final decision could be made during the impending review by full text. Inspection of the 201 abstracts against these criteria yielded 42 selected articles. The full text of each of these articles was then looked at, applying the same selection criteria, but this time basing the choice on fuller information. This process resulted in the selection of 23 articles. 7 articles were further excluded as they discussed registries but did not explicitly discuss BI techniques used.

As a result, a total of 16 articles satisfied all the criteria for final selection.

Search String	Mortality Morbidity Well being Disability Health State / Status Quality of Life Policy
General	Health Medical / Medicine Dental / Dentist Nurse / Nursing Hospital Public organisation Blue Cross Patient Wellness Epidemiology
Disease / Health-related Behaviour / Medical Condition / Medical Intervention	Addiction Adverse drug reaction Aging Ambulatory care Antibiotic resistance Blood bank Brain Tumor Cancer Cardiovascular Cataract surgery Child development / welfare Childhood lead poisoning Depressive disorder Diabetes / Diabetic Dietary Ebola Human mobility Infectious Disease Influenza Live donor risk Lupus Mental Health Obesity Radio therapy Radiobiology Renal disease Salmonellosis Smallpox Substance use screening Suicide Tuberculosis Urologist

Table 2: Terms in title considered to refer to a Health Outcome

2.3 Results

2.3.1 Question 1: What applications of BI for better decision making in the area of population health have been implemented?

Each of the chosen articles discusses at least one type of implemented application of BI for the betterment of population health. Some of the applications are discussed in more than one article. In total, there are 11 applications, which have been classified according to their overall purpose or intended contribution to population health. This classification, together with an indication of the types of BI facilities used in each application, is illustrated in table 3.

Types of BI facilities mentioned in selected articles were:

- **Data Warehouse:** A store of data that has been sourced from various places and integrated into a consistent form so that it can be used for decision making
- **Analytics:** Various forms of software that enable the analysis of data as an aid for decision making. Often includes advanced algorithms, numerical techniques, and statistics.
- **Reports:** Textual representation of specifically requested data, formatted to be easily readable and often including summaries
- **Charts:** Visual representation of specifically selected data in a chosen chart type. Possible charts include line graphs, bar charts, pie charts, and others.
- **Data Warehouse Browser:** Software tool that provides the user with a description of what information is available in the database and allows them to view and summarise bywhichever parts of the data they choose.
- **On-line Analytical Processing (OLAP):** Software tool that provides spreadsheet-like multi-dimensional views of the data, allowing the user to select the data they wish to see and which dimensions they wish to summarise the data on. 'Drilling down' to deeper levels of detail in data displays is facilitated.
- **Graphical Information Systems (GIS):** Software that uses longitude and latitude co-ordinates to enable analysis and display of information by location
- **Dashboards:** Several visual representations of key performance indicators on a single screen. Visuals can be in many forms including charts, maps, symbols, and text.
- **Alerts:** Messages sent to pertinent people when certain data thresholds are met, eg. the number of people infected with a certain disease reaches a level associated with an epidemic
- **Free-form SQL:** 'Structured Query Language', a standard computer language used for accessing data in a database or data warehouse. Requires a level of technical expertise.

The majority of the applications (8 applications - discussed in 13 of the articles) constitute some type of tracking or surveillance of population health. A further article discussing a single application mentions tracking or surveillance as a side benefit, despite a different overall focus. Sharing data and resources was the intention of 2 applications and enabling fact-based decision making is discussed in 1 article.

Describing the types of BI facilities reported to be implemented by each of the different applications is another form of categorisation. Data warehousing formed part of 7 of the applications. Pre-

defined custom-made reports are used in 5 applications, dashboards in1 application, alerts in 3 applications, data warehouse browsing tools (ORACLE Discoverer in this case) in 1 application, on-line analytical processing (OLAP) in 2 applications, graphing and charting facilities in 3 applications, graphical information systems (GIS) in 3 applications, more sophisticated analytics in 8 applications and free-form SQL to access data directly from the data warehouse is used in 1 application. These are, of course, only the reported uses of certain BI facilities in the applications – it is possible that further BI facilities are being utilised but have not been reported on.

The applications will now be described, grouped according to their overall purpose or intended contribution to population health.

2.3.1.1 Tracking and Surveillance of Population Health

The most mature and comprehensive application of BI discussed in the articles was implemented to facilitate the Comprehensive Assessment for Tracking Community Health (CATCH) methodology. Development of the CATCH methodology began at the College of Public Health at the University of South Florida in 1991, in response to two influential reports from the Institute of Medicine (IOM) that emphasised the need to systematically build up a health profile of American communities. The methodology involves the collection, analysis and presentation of hundreds of health and social indicators for each local community. It has been successfully applied in more than a dozen counties in Florida, as well as selected places in other states. Health issues within each county are identified and ranked by comparing each county indicator value with two averages for the same indicator: one based on three peer counties and one for the whole state. Collection of the necessary data and creation of the indicators and averages was first implemented manually – it was to expedite this process that the CATCH data warehouse was initially developed (Berndt et al., 2000). The value of the data warehouse and associated BI facilities has far exceeded its initial purpose, and Berndt et al. (2003) refer to it as “a unique application of technology in the field of public health” (pg. 371). Core reports, inexpensively distributed to communities and public stakeholders in a variety of formats, can be used to identify problems that need attention. Because the data warehouse is built from detailed transactional data from a variety of data sources, facilities like data warehouse browsers and OLAP tools are used to drill down into the data to discover the reasons for the problems (Berndt et al., 2003). Alternatively, initial exploration using OLAP and data warehouse browsers sometimes highlights the need for the development of a new standard report (Berndt, 2001). Use of OLAP with the CATCH data warehouse by knowledge workers in regional health planning agencies was also found to be highly successful in satisfying information requests from outside parties (Tremblay et al., 2007). SQL has been used to retrieve data directly from the data warehouse, but end-users are cautioned not to use it unless they have appropriate SQL skills and an understanding of the data tables. There is discussion about investigation into using data mining as well, but nothing is implemented as yet (Berndt et al., 2003). All of the data warehouse and other associated BI facilities are aimed at providing high quality information that can be used in health care planning and policy formulation (Berndt, 2001). Berndt et al. (2003) believe that the CATCH data warehouse will result in the widespread availability of information not previously available in most communities, thus facilitating good decisions and having an important impact on the population health status. The

Pur-pose	Application	Application Description	Data Ware-house	Ana-lytics	Re-ports	Charts	DW Brow-ser	OLAP	GIS	Dash-boads	Alerts	Free-form SQL
Tracking and Surveil-lance	CATCH	Collection, analysis & presentation of hundreds of health and social indicators for each local community, used for identifying health issues. (Berndt et al., 2000, Berndt, 2001, Berndt et al., 2003, Tremblay et al., 2007)	✔	✔	✔		✔	✔				✔
	ESRD	French registry for tracking the health and treatment of patients with end stage renal disease. (Ben Said et al., 2005, Jean-Baptiste et al., 2005, Richard et al., 2005)	✔			✔		OLAP-type selection on GIS	✔			
	CARP	Combines data from several departments of three different hospitals to report trends in hospital infections due to antimicrobial use. (Wisniewski et al., 2003)	✔									
	FluDM	Data marts for use in influenza surveillance derived from Canada's British Columbia Health Services Authority data warehouse. (Rizi and Roudsari, 2013)	✔	✔	✔			✔ + Excel Pivots				
	GUARDIAN	Receives patient data from hospitals and uses analytics to determine the extent of each disease and send alerts when necessary. (Silva et al., 2013)		✔	✔						✔	
	RODS	Uses hospital emergency department admissions data and analytics to pick out anomalies and send alerts when necessary. (Tsui et al., 2003)		✔		✔			✔		✔	
	NRDM	National retail data on over the counter (OTC) health care product sales received daily and monitored for early disease outbreak detection. (Wagner et al., 2003)		✔		✔						
	NHPSS	New Hampshire pharmaceutical data OTC health care product sales received daily and monitored for early disease outbreak detection. (Xiaohui et al., 2004)	✔	✔	✔				✔		✔	
Sharing Data Across Entities	OCHIN	Non-profit collaborative of community health centers pool resources for shared IT facilities which include data warehouse and analytics. (Devoe and Sears, 2013)	✔	✔								
	BigMouth	A multi-institutional dental data repository formed by four dental schools to share data had already collected by their EHR systems. (Walji et al., 2014)	✔									
Fact-based Decisions	CCO	Use advanced analytics on over 170 internal and external data sources to support improvement in cancer prevention, screening and care. (Garay et al., 2015)		✔	✔					✔		

Table 3: Applications of BI for Population Health and Aspects of BI Use

CATCH data warehouse is seen as a “work in progress” with effort being made to improve both the technical facilities available and the available community health care applications (Berndt et al., 2003: 382).

Another application of BI constructed in response to the need for a health profile was developed in France as a response to an increase, over a decade, of the incidence and prevalence of end stage renal disease (ESRD). The Renal Epidemiology and Information Network (REIN), organised around a network of individuals involved in ESRD care, was formed to assess the precise magnitude of this problem. The resulting disease registry was implemented through a multi-source information system (MSIS) which includes a data warehouse as well as a graphical information system (GIS) – due to these BI features this particular registry has been included in this review. The MSIS collects ESRD information and places it in a production database. To facilitate reporting and use with the GIS, the information is then moved to the data warehouse and integrated with outside information sources on demography and geographical boundaries. A web-based application called “Systeme d’Information Geographique pour la Nephrologie” (SIGNe) provides geographical visualisations (maps) of the data, thereby facilitating the understanding of two aspects of the geography of the disease – the spacial and temporal trends of ESRD epidemiology (incidence and prevalence of the disease) and the distribution of the demand and supply of ESRD care – both of which have substantial bearing on population health decision making. The SIGNe user interface allows an OLAP-type selection of the dimensions of the data to be displayed on the map (eg. view disease incidence by region). Various graphs and charts are also available. In 2005 this application was successfully running in six regions of France (Ben Said et al., 2005, Jean-Baptiste et al., 2005, Richard et al., 2005).

On a more local level, a clinical data warehouse has been developed by the Chicago Antimicrobial Resistance Project (CARP) in an attempt to combat hospital infections caused by antibiotic resistance. These types of infections have been found to increase health care costs and patient mortality and morbidity, and the first step towards controlling them is the ability to report trends in antimicrobial use and resistance and hospital acquired infections. The CARP data warehouse combines data from the pharmacy, laboratory, radiology, medical records and emergency departments of three different hospitals, and also stores data from CARP studies and surveillance programs. Investigators now automatically receive the needed trend reports on their desktops, a feat that previously took many person hours to complete. New infections are automatically detected on a daily basis and an analysis detects patterns (called ‘antibiograms’) of the bacterial susceptibility of infections in each patient care unit. CARP staff can use these as a guide to the selection of antibiotics to prescribe. Combining the information from different sources in the data warehouse means that infection information can be analysed by a number of dimensions, including hospital unit, length of stay, bacterial species and even the part of the body where the infection occurred. In addition, analytics have been developed to detect the prescription, for the same patient, of two or more antibiotics that target the same or very similar bacteria. In a pilot run, physicians agreed with this information in 98% of cases, and amended prescriptions, which had the unexpected side benefit of substantial cost savings. The CARP data warehouse is considered to be a great success, with intentions for future development including the further automation of disease surveillance and an assessment of the value of data mining for the identification of infection control risk factors (Wisniewski et al., 2003).

As mentioned by da Silva et al. (2010), estimates of the high number of deaths per year due to influenza by both The World Health Organisation (WHO) and the American Centres for Disease Control and Prevention (CDC) highlight the importance of influenza surveillance. In response to the H1N1 influenza pandemic of 2009, the British Columbia Centre for Disease Control (BCCDC) in Canada realised that there was a need for a substantial improvement in their analytics processes. As an agency of the British Columbia Health Services Authority (PHSA) they were able to use the data in the existing PHSA data warehouse to create two data marts to be used for this analysis. (A data mart typically contains an extracted section of the data in a data warehouse that is dedicated to meeting the analytical needs of a specific group of people (Inmon, 2005). The data marts use both relational and multi-dimensional structures and OLAP cubes are provided. The project was completed in 2011 and the two data marts are currently being accessed regularly by surveillance analysts through Excel Pivot Tables (a form of OLAP), SSRS reports or SAS (Rizi and Roudsari, 2013).

In addition to the need for surveillance of naturally occurring infectious diseases like influenza, heed should be taken of the threat of large scale biological attacks with agents like anthrax, plague, tularemia and smallpox. Timely detection using surveillance and resulting in the prompt institution of quarantine, vaccination and/or antibiotic treatments could minimise the impact of these attacks. Ideally outbreaks should be recognised from the symptoms and signs of the infection, and from the infection's effect on human behavior, before the patient's first contact with the health care system. This type of surveillance is called syndromic surveillance (Tsui et al., 2003). Many sophisticated electronic syndrome surveillance systems have been created, eg. BioSense, ESSENCE, RODS, NC DETECT, New York City Department of Health and Mental Hygiene Syndromic surveillance system and GUARDIAN (Silva et al., 2013). Mention was found in the selected literature of the existence of two of these syndromic surveillance tools, GUARDIAN and RODS.

The Geographic Utilisation of Artificial Intelligence in Real Time for Disease Notification (GUARDIAN) is a relatively new surveillance system funded by the United States Department of Defense and deployed at various hospitals in the Chicago metropolitan area. GUARDIAN receives patient data through the existing (standardised) hospital ICT communication infrastructure, parses it, links it to other data for the same patient and stores it in a relational database. Analytics, in the form of national language processing, disease models, Bayesian networks and geospatial clustering of cases are used to determine the extent of the incidence of each disease, and, when appropriate, to create alerts and reports that are sent to hospital and public health personnel for review (Silva et al., 2013). Though there is no explicit mention of the use of business intelligence techniques in this discussion of GUARDIAN, it involves both the collection of data and its analysis as an aid to decision making and is thus seen to be a business intelligence application.

Similarly, the Real-time Outbreak and Disease Surveillance (RODS) system is seen as a business intelligence application due to the use it makes of data collection and analysis for decision making – although it does include graphing and GIS components aswell. RODS is freely available for non-commercial use, and detects disease outbreaks from data that is routinely collected for other purposes, for example the chief complaints of patients at emergency departments and special clinics; and sales of over-the-counter (OTC) health care products (Tsui et al., 2003).

On registration of a patient at an emergency department or special clinic, the recorded information is sent to RODS through the existing (standardised) hospital ICT communication infrastructure. An

analytic in the form of a naïve Bayesian classifier is used to classify the patient's chief complaint as constitutional, respiratory, gastrointestinal, neurological, etc. A Recursive-Least-Square adaptive filter and heuristic search are further analytical techniques used for the detection of disease anomalies and consequent creation of alerts to relevant authorities. The RODS team has also established a technique of linking the complaint data to culture results, radiology reports and dictated emergency room notes which make case detection more specific. In addition, they have created a drill-down function that they call "look-back", that allows an investigator to view the electronic medical records for a particular case that needs to be examined. RODS has been in operation in Pennsylvania since 1999 and in Utah since January 2002 (Tsui et al., 2003).

The RODS laboratory has also been responsible for the development of the National Retail Data Monitor, a surveillance application that uses sales of over the counter (OTC) health care products for early outbreak detection. OTC products tend to be purchased for self-treatment early on in the course of the illness, and studies have suggested that purchase patterns can be used for early outbreak detection. Monitoring of OTC retail sales is made possible by the existence of internationally utilised Universal Product Bar Codes (UPC) for identifying each product. Through agreements with four chosen major retailers, the National Retail Data Monitor receives nationwide data from the retailers' national data warehouses on a daily basis. Products are then aggregated into categories through a specially designed algorithm, and the data is monitored for signs of a large number of contaminated people. The National Retail Data Monitor has been in continuous operation since December 2003. Its data is available to health departments in one of two ways – through secure raw data feeds and through the RODS web interface (Wagner et al., 2003).

The RODS user interface uses a password-protected, encrypted web site that allows the viewing of trends derived from both the emergency department complaints and OTC sales data – the main screen swaps between the two every two minutes. Line graphs show time series of recorded data and use of a GIS facility has enabled the provision of colour coded maps showing how the outbreak data is distributed geographically (Tsui et al., 2003).

Xiaohui et al. (2004) discuss the New Hampshire Pharmaceutical Sales Surveillance (NHPSS) system, another surveillance application that uses OTC sales data for disease outbreak detection. Pharmaceutical sales data received daily does not have as wide a scope as the data received by the National Retail Data Monitor, but it does serve a useful function for the New Hampshire area. When received, it is placed in a data warehouse that organises the data along logical dimensions that enable easy analysis. Sales are automatically categorised – on the basis of their active ingredients - according to the syndromes they medicate (eg. respiratory, gastro-intestinal, allergy-related, etc.) . Disease detection is achieved by comparing values derived analytically from incoming data with analytically calculated reference values. A GIS plays a key role in the data warehouse, performing spatial abnormality analyses with scanning and ranking. It also supports outputs mapping for risk assessment and provides comprehensive reports with possible alerts. Since its pilot application in 2002 which used data from pharmacies in 23 cities, NHPSS supported the detection of a large scale gastrointestinal disease outbreak – triggering an alert four days prior to recognition of the outbreak at state level. In addition, it triggered influenza alerts up to twelve days before schools were closed to combat the spread of an influenza outbreak (Xiaohui et al., 2004).

2.3.1.2 Sharing Data and Resources

The selected articles contained two applications that enable the sharing of data across different centers, and whose main focus is not the tracking and surveillance of population health.

The first is the Oregon Community Health Information Network (OCHIN). OCHIN, Inc. is a non-profit collaborative that was formed in 2001 by community health centers and other community organisations to facilitate the adoption of IT to help improve the quality of care for vulnerable populations. Through the collaborative, resources of various centers could be pooled in order to create an IT infrastructure that could serve all of them. Patients, community advocacy organisations, public health leaders, community health centre clinicians, informatics experts and policy makers all worked together, using the pooled resources to create this shared IT infrastructure which included the implementation of Electronic Health Records (EHR), and the use of a centralised data warehouse and analytics to aid in decision making. The collaborative also facilitated interfacing with sources of external data such as laboratory test results and medical images (Devoe and Sears, 2013).
Shared data in the data warehouse allows for the measurement and comparison of various quality metrics across all the member clinics, thus identifying those that need assistance and those that are able to offer assistance. The approach and processes of successful clinics can be reviewed and provided as recommendations for those that are not doing as well. Analysis of the effect on clinic performance of the demographics of the various clinic patient populations can be done. All of this information is highly relevant to population health decision makers, who would wish to shape policies that incentivise clinics to care for vulnerable populations. In addition, the impact of practice and policy changes can be gauged through analysing the community data in the data warehouse. Although the main objective of OCHIN Inc was to improve the quality of care for vulnerable populations, the pooled data does also enable public health surveillance, eg. tracking of influenza outbreaks (Devoe and Sears, 2013).

The OCHIN collaborative is reported to include nearly 70 members with more than 300 clinic sites who care for more than 1 million patients annually. Vulnerable populations served by these clinics include underinsured and uninsured patients, racial and ethnic minorities, migrant and seasonal farm workers, and patients living in poverty. With time, it is hoped that community members will learn how to fully harness the power of collective data to provide fair and equitable medical care (Devoe and Sears, 2013).

The second application that enables the sharing of information across different centers is BigMouth, a multi-institutional dental data repository. This repository was formed by four dental schools – at UT Health, UCSF, Harvard and Tufts – who wanted to share data that had already been collected by their EHR systems. They all use the same dental EHR system, axiUm, and belong to the Consortium of Oral Health Research and Informatics (COHRI) formed by axiUm users in 2007. The repository was implemented on the Informatics for Integrating Biology and the Bedside (i2b2) data warehousing platform and was launched with data on 1.1 million patients in August 2012. Access to the repository is only allowed to those institutions who have contributed data. A data governance framework that encourages institutions to contribute has been created in anticipation of the expansion of the repository to include other institutions that are members of COHRI (Walji et al., 2014).

2.3.1.3 Enabling Fact-based Decision Making

Only one of the selected articles was found to fit into this category.

Cancer Care Ontario (CCO) is an organisation that supports improvement in the quality of cancer prevention, screening and patient care. Through various implemented IT solutions, they have access to over 170 internal and external data sources for analysis. Historically this data has been analysed by three developed applications:

- The Cancer System Quality Index (CSQI) tool calculates quality measures to track progress in terms of service and outcomes, and informs policy makers and others about where improvements are needed.
- iPort and iPort Access are secure web-based tools provide information about both cancer surveillance and wait time for a surgical procedure. iPort and iPort Access have quite a few business intelligence functions, being able to generate detailed analytical reports based on geography, time, age and diagnosis time. Comprehensive dashboards show up-to-date information on health system performance, and a drill-down facility enables a more detailed analysis of the underlying data supporting each metric (Garay et al., 2015).

With the objective of making more use of the valuable data available, the external Strategic Analytics Advisory Group (SAAP), composed of nine thought leaders from across the health industry, was set up. This has led to a number of newly implemented projects:

- A model predicting the number of hip and knee replacements needed to meet and sustain wait time targets
- An interactive decision support tool to assess the impacts of follow up care, as well as repatriation of certain types of care to help determine number of medical professionals needed
- A project to predict the impact on the health system of changing to a new type of cancer screening test
- A planning tool that projects kidney disease patient population growth over the next 12 years and translates it into capacity requirements at each dialysis facility.

CCO see the establishment of SAAP as a “journey that is just beginning”, and they strive to continue to encourage development that will support clinical, management and policy decisions that positively impact health care performance (Garay et al., 2015).

Findings discussed in relation to SLRQ2 will be discussed later in the chapter, in section 2.4.1.

2.3.2 Question 2: What evidence links the use of BI to better population health outcomes?

Little empirical work has been done to assess the impact of BI on population health. Of the 16 selected articles, only one reported on a formal study.

The goal of this study by Tremblay et al. (2007) was to understand the impact of using on-line analytical processing (OLAP) at a regional health planning agency in Florida in the USA. Regional health planning agencies are responsible for developing district health plans based on data and

analyses, giving recommendations designed to improve access to healthcare, reduce disparities in health status, assist governments in the development of healthcare policies, and advocate on behalf of the underserved. Health planners working at these agencies exhibit typical characteristics of knowledge workers, in that they are highly skilled individuals with advanced degrees who willingly apply themselves to complex tasks that require creativity. Static reports from the CATCH data warehouse discussed above had already been used by these health planners for some time when OLAP was introduced as a tool to access the same data. OLAP interfaces provide spreadsheet-like multi-dimensional views of the data, and allow for extremely simple, flexible and powerful selection, aggregation and presentation of data to be viewed. The objective of the research was to understand firstly how knowledge workers like health planners could be aided by OLAP, and secondly the impact of this use of OLAP on the communities served (Tremblay et al., 2007).

A year-long longitudinal qualitative field study was done at a single research site – a Florida health planning agency. During this year the health planners at that agency were given access to the data warehouse in an iterative fashion, starting off with hard copy reports and finally having full OLAP access. Multiple data collection methods – interviews, observations, documents – were used for corroboration. Semi-structured interviews with health planners were held prior to initial implementation as well as after each implementation stage. Interview protocols were based on the Task-Technology Fit model (TTF). TTF reflects the Goodhue and Thompson (1995) belief that technologies “have a positive impact on performance only when there is a correspondence between their functionality and the task requirements of users” (pg. 1306). The study therefore attempted to understand the impact of OLAP use through considering the fit between the characteristics of the task, the characteristics of the health planners and the characteristics of the available technology, i.e. OLAP (Tremblay et al., 2007).

The fit between these aspects was discovered to be extremely positive, with health planners expressing a strong desire to learn as much as possible about the tool and declaring that they could no longer do their work without OLAP. Empowered by the ability to choose what data to view, their role in the agency changed from data collector to consultant. Instead of just exporting requested data, they could re-interpret user needs or cross-examine them on ambiguous requests and then find the appropriate data for them. They could dig deeper into the information than what the user asked for, and thus understand the situation and advise on what data would be most appropriate. The OLAP drill-down capability enabled them to get more specific diagnoses of problems, eg. when asked to investigate cardiovascular disease they could isolate which community or socio-economic class the disease was occurring in. They are also able to add new groupings to the data – eg. they could define their own ‘synthetic’ communities that highlight issues across racial composition, income levels or anything else they feel is important. Though it is difficult to estimate the impact of the use of OLAP on the population, feedback from clients requesting the information from the agency has been very positive. Indications are that requestors really happy with the data and satisfied with the quick turnaround time. They are especially pleased that the health council actually had the data they were looking for, and the number of information requests is on the rise, which is a positive sign. Health planners themselves feel they are functioning better, with one of them saying “Its quicker, its more effective, more efficient, gives a better picture”. Clearly this study has shown that OLAP has made a very strong contribution to the performance of the health planners at a Florida regional health planning agency (Tremblay et al., 2007).

2.3.3 Question 3: What challenges and issues have been faced in the implementation and use of BI for the betterment of population health?

A review the challenges and issues discussed in the 16 articles revealed that they could be classified into three areas: Privacy and Security, Data Quality and Development and Maintenance of BI infrastructure.

2.3.3.1 Privacy and Security

It is hardly surprising, given the sensitive nature of personal health information, that the privacy and security of health data used for BI purposes is of primary concern. Since most BI applications look at data aggregated over a whole set of personal data, the data is usable without knowledge of which particular individual it pertains to. For this reason many applications remove identifying information from the data before making it available for use. There are no personal identifiers of any kind kept in the CATCH data warehouse (Berndt et al., 2000), and both the RODS Surveillance system and the BigMouth dental repository de-identify data before use (Tsui et al., 2003, Walji et al., 2014). Other applications may also de-identify the data, but mention was not made of it in the chosen articles.

Additional precautions are also taken to avoid identification of individuals. Berndt et al. (2000) mention that care is taken when reporting from the CATCH data warehouse on rare events in certain geographical areas as they might allow a person to be identified from different sources. A similar example is given by Rizi and Roudsari (2013) who mention that, to guard privacy, users are strongly discouraged from linking data in different data marts in any way, and linkage on the patient level is actually prevented by technical barriers.

The sensitivity of medical data also necessitated compliance with government legislation in many of the discussed applications. Four cases – the RODS surveillance system (Tsui et al., 2003), the CARP project to combat hospital infections caused by antibiotic resistance (Wisniewski et al., 2003), the OCHIN community network (Devoe and Sears, 2013) and the BigMouth dental repository (Walji et al., 2014) – all took cognisance of the requirements of the United States Health Insurance Portability and Accountability Act (HIPAA). Interestingly, the National Retail Data Monitor surveillance system is not governed by HIPAA, as the aggregate of OTC sales of medical products does not constitute personal data (Wagner et al., 2003). The patient identification server for the REIN ESRD system complies with French privacy law, with necessary agreements of the Commission Nationale Informatique et Libertes (CNIL) having been obtained (Ben Said et al., 2005). Acknowledgement is also given by Cancer Care Ontario of the need to meet the obligations pursuant to the Canadian Personal Health Information Protection Act (PHIPA) and those of its regulator the Information and Privacy Commissioner of Ontario (IPC) (Garay et al., 2015).

By nature, BI systems collect data from various sources, and when dealing with population health, these data sources often belong to other organisations. The “owners” of the data have a fiduciary responsibility to keep the data confidential and secure and to ensure it is interpreted correctly (Wisniewski et al., 2003), and therefore it is necessary to have agreements with them to use the data. Different applications have had different ways of dealing with this.

In creating the data marts from the PHSA data warehouse, the British Columbia CDC avoided this problem by only allowing new access if it is covered by previously existing permissions or agreements (Wisniewski et al., 2003).

The OCHIN community network requires participating members to access information belonging to other members. Fortunately this is provided for by an Organised Health Care arrangement that is recognised by the HIPAA privacy rules – entities participating in joint activities can share patient information if it benefits the joint venture. Comment was made that the diversity of participating entities made the achievement of the required HIPAA-compliant data sharing agreements a challenge (Devoe and Sears, 2013).

For the BigMouth dental repository, where data from at least four dental schools is shared, extra controls are implemented. In addition to the HIPAA privacy rule requirement of Data Use Agreement (DUA) for each participating institution, only organisations that have contributed data may access the repository. They may query their own data as well as the combined data, but may not associate the data with the institution it came from. Control of their contributed portion of the data is maintained by them and they may remove it if it has not yet been used for research (Walji et al., 2014).

Obtaining institutional approval was cited as one of the biggest challenges in the planning of the CARP data warehouse built for hospital infection control. Data is brought in from several clinical departments in three different hospitals, and the CIO had to get approval from the CEOs of each hospital, as well as from each of the clinical department administrators (Wisniewski et al., 2003).

Data sharing agreements have been made between the RODS surveillance system and all participating health systems that contribute emergency admissions data. As an additional stringency, there is an annual meeting of RODS members at the University of Pittsburgh where they review their obligations and sign an agreement indicating that they understand and will abide by the terms of the data sharing agreements. Interestingly, the terrorist attacks (including anthrax) of 2001, prompted a number of new hospitals to contribute data (Tsui et al., 2003).

The University of Pittsburgh also has data sharing agreements with all retailers that send data for the National Retail Data Monitor (NRDM). These agreements permit the University to redistribute the data to participating Departments of Health (DoHs) and the CDC provided that data is first aggregated with similar other data by zip code. Each DoH can only access data relevant to its jurisdiction, although they can access overall aggregated data using RODS. To address retailer concerns about competitor access to their information, a clause in the agreement prohibits aggregated data from being shared back to any retailer (Wagner et al., 2003).

2.3.3.2 Data Quality

Good data quality has been found to be crucial to the successful implementation of any BI system. Very often, data quality problems in the systems that provide the source data for a BI implementation are surfaced by the BI initiative (Yeoh and Koronios, 2010). Although none of the

applications discussed in this review seem to have a comprehensive and formal approach to data quality management, several of the articles discussed ways in which the data quality issue is being addressed.

The team that created the British Columbia Influenza data marts has the view that any data quality problems should be remediated at source, which is where they relay any discovered issues to (Rizi and Roudsari, 2013). This could be viable, since the source of their data is a developed data warehouse itself. The BigMouth dental repository team also sees itself functioning as part of a learning lifecycle, where discovered errors can be corrected at the originating hospital (Walji et al., 2014).

Creators of the CATCH data warehouse have engaged in a comprehensive and ongoing effort to create preprocessing routines that check the quality of the data before it goes into the data warehouse. These have become known as 'filters' and they do cross-checking within and between data rows, checking for duplicates, comparing aggregates, etc. (Berndt et al., 2000, Berndt et al., 2003). Health planners using the CATCH data indicate that there is a high level of trust in the data, but they also make sure that they understand the data before disseminating it to requesting clients (Tremblay et al., 2007).

Wisniewski et al. (2003) describe the substantial and ongoing investment in time required for validating the data for completeness, continuity and accuracy. Missing data was found and data extraction programs for the CARP infectious disease data warehouse were improved by repeatedly comparing their results to the source data. Cross checks were done between the data from each of the three hospitals and the data warehouse, and reports from the data warehouse are compared with reports of individual hospital data on an ongoing basis.

The approach of the BigMouth dental repository is to execute a data quality checklist each time data are loaded into the repository. This identifies missing data and counts patients in each of a number of categories. Discrepancies with previous data loads are then investigated. The BigMouth team have also identified a need for the development of a standard dental diagnosis terminology, which they have begun working on (Walji et al., 2014).

Data checking for the REIN ESRD system is performed regionally by clinical research assistants, who compare the MSIS source data with the patient files, looking for any sign of missing or incorrect data (Jean-Baptiste et al., 2005, Richard et al., 2005).

The only mention of data quality checking in the RODS surveillance system is that incoming messages from emergency departments are checked for duplicates to be removed, and scheduling information, which must be filtered out (Tsui et al., 2003). The other RODS system, the National Retail Data Monitor, has an interesting aspect of the data to fix up, or "normalise". This is due to the difference in population or market share in different zip code areas. For example, raw sales data in midtown Manhattan in the absence of an epidemic could be far greater than those in rural New Jersey even if every person in the New Jersey town were sick. Sales may also be influenced by factors like store hours or bad weather. Completeness of the data is also an issue, but as pointed out, this is largely dependent on the extent to which contributing pharmacies cover the population (Wagner et al., 2003).

2.3.3.3 Development and Maintenance of BI infrastructure

Issues and challenges with regard to creating a data warehouse and the ongoing management of a BI environment, as well as factors contributing to a successful infrastructure were discussed.

According to Wisniewski et al. (2003), the hardware and software required for the building of a data warehouse is available and affordable but technical expertise, personnel time, administrative support and substantial work is also required. Perhaps the most difficult part of initial data warehouse development is the integration of data from multiple disparate data sources. Berndt et al. (2003) mention the difficulty, in the creation of the CATCH data warehouse, of integrating data across diverse sources that, due to the lack of standardisation, express health care data in a variety of different formats with different semantics. Wisniewski et al. (2003) also speak about difficulties with non-standardised data formats in the development of the CARP data warehouse, for example the drug name sometimes includes the preparation strength. Sometimes they found accurate data to be unusable for computation, for example the value of "dose" could be expressed as a weight, or as a volume, or as a number of pills – this had to be adjusted manually by the clinical staff. Understanding all the departmental databases that the data had to be integrated from also presented a problem. Walji et al. (2014) found a challenge in the building of the BigMouth dental repository in that even though all contributing dental schools had the same EHR, they had all been implemented and coded differently. They also bemoaned the fact that oral health is perceived to be different from general health in the USA and feel that the best way to remedy this would be the integration of medical and dental care and data.

On a different note, creators of the British Columbia influenza data marts reported data modeling problems due to too many unrelated attributes and multi-valued fields. They also reported difficulties extracting data from patient care systems without negatively disrupting those systems (Rizi and Roudsari, 2013).

The importance of the user interface is another factor to be considered in the creation of a data warehouse, as this is the primary determinant of success from a user perspective (Berndt et al., 2000, Berndt et al., 2003).

BI systems usually hold very large volumes of data that demanding queries are made against, which means that performance can become a problem. Some of the ways of dealing with this are including tables of pre-calculated aggregates so they don't have to be calculated at query time, using special indexing strategies and table partitions, and keeping smaller tables in cache (Berndt et al., 2003). The RODS surveillance system actually caches pre-aggregated counts that are refreshed every 30 minutes for fast retrieval (Tsui et al., 2003).

The National Retail Data Monitor has been set up with the assumption that in the event of a bioterrorism outbreak thousands of users will query the site continuously. To facilitate this they have built in fault tolerant features and the ability to create a second mirrored site (Wagner et al., 2003).

The most prominent success factor mentioned was having good relationships with people: with top management who could help with resource provision, change management and vision sharing (Rizi and Roudsari, 2013); with clinical department directors and staff managing local servers so that links

could be created with those servers (Wisniewski et al., 2003); between projects and the intervention programs they pertain to (Garay et al., 2015); and within an interdisciplinary development team with expertise in medical informatics, computer science, law and engineering (Wagner et al., 2003).

Early development of a data governance model was also recommended. This could be based on existing examples, and is seen to be an ongoing project that should be done in parallel with design and development for mutual benefit (Rizi and Roudsari, 2013).

2.4 Discussion

2.4.1 Applications of BI for Population Health

The eleven applications identified by this review attest to the fact that BI is being applied in the area of population health, particularly in the seven cases where the significant effort and cost of data warehousing has been undertaken. Additional applications are mentioned in the trade and electronic literature (news@orchard, 2015, Catalyst, 2016), and there are probably applications that exist but have not been documented. The fact that the majority of the eleven applications identified are applied to the tracking and surveillance of population health implies, perhaps, that this is the area of population health activity that can most benefit from BI intervention. Of the tracking applications, there are a large number (six) that are concerned with the (hopefully early) detection of communicable disease outbreaks with a view to better managing possible outbreaks and bioterrorism attacks, possibly reflecting the urgent concern of public health officials in this area. Only one application, the CATCH data warehouse, is used for the tracking of community health in general, and only one application tracks the treatment of patients with a non-communicable disease. No doubt these categories of population health benefit from being tracked and maybe with time there will be more development in this regard. Only three of the applications reported using BI for administrative or management purposes: REIN's ESRD registry to understand and address the dynamics of the demand and supply of renal care, OCHIN to compare the performance of different clinics and identify best practices, and CCO to improve health system performance by minimising wait times for surgical procedures

The applications reported on were all in developed countries, with no evidence emerging from this review of applications in developing countries, despite the higher burdens of communicable diseases alongside non-communicable diseases.

Also of interest is the extent to which the different BI aspects are being used within the applications. Although both data warehousing and analytics are employed by the majority, tools that empower end users to access the available data with flexibility and with advantageous visual impact, like data warehouse browsers, OLAP, GIS and Dashboards are not very prevalent. Data mining is only explicitly mentioned as being investigated for potential advantage; however some of the more complex analytics discussed could be construed as data mining even though not overtly classified as such.

2.4.2 Evidence Linking the Use of BI for Population Health to Better Outcomes

The evidence from the single empirical study found (Tremblay et al., 2007) implies there is a definite advantage in the use of OLAP tools. The fact that healthcare workers found their role to be changed from data collector to data consultant is indeed significant. OLAP gives them the flexibility to choose what data they want to see, without waiting for the IT department to develop a new report every time. The ability to "dig deeper" into the data empowers them to help clarify, for information requestors, what data is actually needed as opposed to what they thought they needed. It seems logical, then, to have received feedback indicating that users are especially pleased that the health council actually had the data they were looking for. By changing the role of the health planner, the OLAP tool has improved the service of the agency and this should ultimately impact on population health.

2.4.3 Challenges and Issues in Implementing BI for Population Health

Most applications acknowledged the challenge of ensuring privacy and security in a population health environment. Three areas of importance in this regard were identified: ensuring privacy through protecting the identity of the patients that the data pertains to, compliance with government regulation, and establishing agreements to cover the sharing of data between different parties. Attention has been paid to each of these issues by many of the applications, as discussed above.

Basing critical decisions for population health on bad quality data could potentially be disastrous. Therefore, many of the applications have recognised this and have instituted controls or procedures to ensure good data quality, obviously taking cognisance of the aspects of quality that are most relevant to their particular purpose. No acknowledgement has been made, however, of the substantial research into data quality that can be found in the literature. The research in the literature covers two aspects of ensuring data quality. Firstly, attention has been paid to which dimensions data quality should be measured by, and how they should be measured. .. Various dimensions of data quality have been suggested in healthcare (e.g. Weiskopf and Weng, 2013). Wand and Wang's set of dimensions usually underpin and are the basis for these works. . Examples of suggested dimensions include accuracy, completeness, consistency, timeliness, reliability, relevance and currency (Wand and Wang, 1996). Secondly, there are various methodologies for ensuring data quality that have been suggested and tried. These all involve some kind of procedural quality assessment and improvement cycle (Lee et al., 2002, Batini et al., 2009).

Cited issues and challenges in the development and maintenance of BI infrastructure for population health are not different to those experienced with BI infrastructures in other sectors. These include the importance of relationships with people, the requirement for technical expertise and personnel time, the difficulty in integrating data from multiple disparate sources, the difficulty in extracting data from source systems without impacting their current functioning, the need to optimise query performance and the importance of a good user interface (Wixom and Watson, 2001, Inmon, 2005, Kimball et al., 1998).

One unusual challenge was mentioned in connection with the National Retail Data Monitor, which uses OTC sales of medications for early detection of disease outbreaks. Here technological provision has been made to handle excessive surveillance query traffic in the possible event of a disease outbreak due to bioterrorism.

2.5 Limitations

This review has several limitations.

Firstly, though the use of the PRISMA protocol helped to facilitate a rigorous approach, the results may be limited by the use of key terms chosen, databases selected and inclusion criteria applied. Given these criteria and the focus on published work, there may be recent emerging trends that have not been covered.

Secondly, it could only reflect on those applications that have been written about and published. It is known, as mentioned above, that there are applications that have not been documented. Even within documented applications, mention may not have been made of all the different BI facilities that have been used. Further development may also have taken place and new aspects of BI added since the date of publication.

Furthermore, relevant articles may not have been found simply because they were not linked to any of the four databases that the search string was applied to.

The selection of articles may also have resulted from researcher bias, since there was only one researcher selecting the articles on the basis of title, abstract and full text.

Finally, though this review focused on applications of BI for population health, most of the issues and challenges encountered were the same as those encountered by BI applications in other areas. A more generic search for issues and challenges with BI implementations that might have revealed more empirical research was not done as it was not within the scope of this systematic literature review. Development of a conceptual model arising from the findings is also left for future work.

2.6 Recommendations for Practice

The applications discussed in this review have illustrated that positive benefit to population health can be derived from the use of BI. Public health officials should certainly consider use of BI in all aspects of population health decision making. In particular, the development of a data warehouse, which provides a large pool of clean, consistent and integrated data for decision making, equips an organisation with a solid base from which other BI facilities can be used. Organisations may be reticent about implementing a data warehouse due to the large amount of complexity, time and money involved. They should consider developing one, however, as the potential benefit from future analyses based on the data would be well worth the investment. In particular, the approximately 172 disease and medical device registries that were mentioned in articles rejected from this review should consider implementing data warehousing technology.

Perhaps due to the urgency of being able to address possible disease outbreaks, most population health tracking and surveillance systems in this review have focused on the spread of communicable disease. Non-communicable diseases like diabetes, cancer and cardiac disease, however, also constitute an accelerating problem in population health. Though there are registries tracking these kinds of diseases, the introduction of more BI into this area would be beneficial.

Users employing pre-built applications and reports to get to the data being analysed are restricted to what is provided for them. A requirement for different data necessitates a change to the software by the IT department, and this may take time. OLAP tools allow the user to choose what data is to be viewed and how it is to be summarised and presented, thus giving them the flexibility to really explore the data as they need to, and effectively empowering them to do their job much better. As mentioned previously, health planners using OLAP against the CATCH database actually found that it changed their role from data collector to data consultant. Organisations not currently using OLAP should consider doing so.

It is surprising that only three of the applications found are using GIS. A geographical representation of population health data clearly illustrates the distribution of disease, access to healthcare, health indicators, etc. The benefit of the SIGNe GIS system for monitoring end stage renal disease is highlighted in the articles that discuss it (Jean-Baptiste et al., 2005, Richard et al., 2005). GIS should be considered as an option for population health tracking systems.

Performance dashboards also do not feature prominently in the articles reviewed. Effectively a visual report of key performance indicators, the dashboard has become more prominently used in recent years and may have been developed subsequent to the writing of the published articles. They are nonetheless a useful addition to the BI facilities of an organisation, particularly if they allow for drilling down deeper into the data to explore problem areas.

Potential for data mining and other advanced analytics is endless in any field and for population health as well. Due to the high level of expertise needed for implementation and the challenge of identifying appropriate applications, development is bound to take time. Population health officials should bear in mind their potential utility.
Though many of the applications discussed have made a substantial attempt to ensure the quality of the data they are using for decision making, BI implementers in general should consult the literature on the different dimensions that can be used to measure data quality and the different methodologies that can be used to track and improve data quality.

2.7 Conclusion and Implications for the Study

This systematic literature review explored the literature available on the use of BI for population health to provide a synthesis of information available on this use, the impact of this use on population health, and the issues and challenges encountered in this use.

An interesting array of applications of BI for population health were found, the majority of which were concerned with tracking and surveillance. Other applications involved sharing information between different medical entities, and applying BI to enable fact-based decision making in a specific area. To discover which of these functions the NHLS CDW serves, the interview protocol for this study includes a question to ascertain the purpose for which the information is being requested. In addition, information gleaned from the NHLS Annual Report (NHLS, 2015) indicated that information from the NHLS CDW is used for public health tracking, administration, and medical research purposes. A quota sampling technique, where at least one or two information requests satisfying each of these purposes, was used in an attempt to ensure that data from information requests concerning all three of these uses was collected.

It is also worth noting that aside from the main focus of the reviewed applications being on public health tracking, only three of them had the administrative goal of improving health system performance. Moreover, no reviewed papers focused on the issues and challenges associated with the use of the BI applications for medical research. The empirical work done for this research project therefore makes an additional contribution to the literature in that it covers all three of those aspects, looking at uses of the NHLS CDW for public health tracking, administration and medical research.

Use of various different aspects of BI was described in the selected articles, namely data warehouses, analytics, reports, data warehouse browsers, OLAP, GIS, dashboards, alerts and even direct access to the data warehouse using SQL. To determine which of these aspects of BI are made use of by the NHLS CDW, they were categorised into customised applications (reports and dashboards), high level data warehouse access tools (data warehouse browsers, OLAP, GIS) and low level data warehouse access tools (SQL, analytics). For completeness, a further category for information requests that are made via a service representative of the CDW was included. A quota sampling technique was then used to try to ensure that data was collected for at least one or two information requests in each of these categories, thus encouraging a complete picture of the aspects of BI used in conjunction with the NHLS CDW.[14]

All applications indicated that they had had a positive impact on the management of population health, and there seems to still be substantial potential for further development in this regard. Exploring the impact of the NHLS CDW on the NHLS as an institution involved with population health management, and the impact on population health in general, is part of the problem that this research project aims to solve, and is addressed by the fourth research question, which is: “What are the perceived benefits (impacts) of NHLS CDW use for the NHLS as an institution and for population health in general?” Questions in the interview protocol enquire about perceived impacts, and any documents or any other evidence that may illustrate these impacts are requested. An additional question asking whether the information requestor felt the CDW was being used to its full potential revealed some enlightening ideas about possible further uses and impacts of the NHLS CDW.

No doubt not all applications of BI for population health have been written about and published. Researchers should document more of these applications so that a more complete picture can emerge. This is particularly true for applications of BI in developing countries, as there is currently

[14] Obviously, since the NHLS CDW is a data warehouse, that aspect of BI is central to this project.

very little documentation about BI use in this context. The examination of the NHLS CDW done in this research project tells the story of a BI application in South Africa, thus augmenting the available documentation in this area.

Only one empirical study was identified in this review and that focused on the very specific area of OLAP use. There is clearly a need for more empirical studies on implementation, use and benefit of all the different aspects of BI, as well as on the issues and challenges that arise when developing and using BI for population health. This research project constitutes one such empirical study, and thus contributes to the building up the body of knowledge on the use of BI for population health.

Issues and challenges arising from the development and use of BI facilities for population health were found to fit into three categories: Privacy and Security, Data Quality, and Development and Maintenance of BI infrastructure. Data quality and infrastructure challenges were very similar to those encountered in applications of BI outside of the field of population health, but privacy and security of medical information was found to be exceptionally sensitive. Questions in the interview protocol for this study were designed to explore whether these identified issues and challenges are pertinent in the use case of the NHLS CDW.

This study makes a further contribution by focusing on decision makers and decision making, and by making use of the Activity Theory framework to describe how decision makers interact with BI tools to achieve decision outcomes within the population health context.

The next chapter will examine the theoretical underpinning for this research project. Activity Theory, the framework that has been chosen as a lens through which to view the BI application being studied, will be explained. Thereafter, the use of Actvity Theory to create an activity system applicable to the study of the use and impacts of the NHLS CDW will be described.

3 Theoretical Underpinning : Activity Theory

This chapter presents Activity Theory as the chosen theoretical underpinning for this study. Starting with an exposition of Activity Theory and its history, it then moves on to illustrate ways in which Activity Theory has been applied in Information Systems Research. This is followed by a discussion of the application of Activity Theory as a theoretical framework for this study, justifying this choice of framework and explaining the development of the activity system for an information request.

3.1 Theoretical Framework Chosen for the Study

Two possible theoretical frameworks were initially considered for this study, the Task Technology Fit Theory (TTF) (Goodhue and Thompson, 1995) and Activity Theory (Engestrom, 1987), since both of them explore the interaction between an individual (in this case an information requestor) and a technology or tool being used (in this case the NHLS CDW and facilities) in the accomplishment of a task (in this case obtaining information)

TTF posits that the performance of an individual using technology depends on the extent to which that technology assists the individual with the task to be accomplished, in other words it depends on the degree of the "fit" between the individual, the task and the technology. This fit would be dependent on the correspondence between task requirements, individual abilities, and the functionality of the technology (Goodhue and Thompson, 1995). This framework was used by the only empirical study found in the literature review, which looked at the impact of the introduction of OLAP into a health planning agency (Tremblay et al., 2007). In the case of the NHLS CDW, the individual would be the data requestor, the task would be the obtaining of data from the CDW and the technology would be the CDW and associated facilities. Using the TTF framework could perhaps have uncovered how use of the CDW could impact on the performance of the data requestor. It would not, however, have included an examination of the impact on population health in general. In addition, no cognisance is taken of the context in which an information requestor is operating. TTF was therefore decided not to be an appropriate framework for this study.

The NHLS CDW is an information system used by humans who are working within the context of the organisation, performing the activity of requesting information. The application of Activity Theory to understanding NHLS CDW use situates this request for information within an activity system, and provides an ideal framework for understanding CDW use within its context. The context of the NHLS CDW is a relatively complex one, and as was the case with the study done by Johnson (2004) described above, Activity Theory will also help to understand the context itself. Activity also addresses the outcome of the activity, which in this case could include the impact on population health, one of the research questions for this study. Activity Theory was thus chosen as the theoretical underpinning for this study.

3.2 Activity Theory

Activity Theory has its roots in the work of the Russian psychologist Lev Semyonovich Vygotsky. Though he died at a young age of tuberculosis in 1934, he left behind a string of essays on various

aspects of human cognition and language, some of which were finally translated into English and published in 1978 (Parker, 1979). Vygotsky posited that human activity is different from non-human activity in that it is mediated by the 'tools' (in particular, language) used in its accomplishment. As such, activities can only be properly understood while considering the tools that mediate them (Crawford and Hasan, 2006). This idea was expressed using Vygotsky's famous triangular model which is now commonly represented as shown in Fig 2.

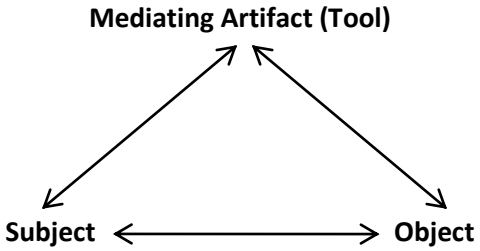


Figure 2: Common reformulation of Vygotsky's mediated act (Engestrom, 2001)

This model shows how the effect of an activity performed by a 'Subject' with a particular purpose or 'Object' is mediated by the tool used. The tool is essentially a cultural artefact as it would be produced by individuals in the society within which the subject functions. Thus the activity performed by an individual using the tool is placed within the cultural context of his/her society since the unit of analysis is no longer the subject on its own, but includes all three members of the triad (Engestrom, 2001).

According to Engestrom (2001) the limitation of this unit of analysis is that it is still individually focused – still looking at a single activity taking place in isolation from other activities. Vygotsky's student, Alexei Nikolaevich Leont'ev, expanded on this idea by pointing out that activities often consist of a series of actions, which in turn are composed of operations, as shown in Figure 3.

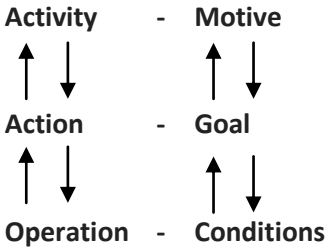


Figure 3: Hierarchical Levels of an Activity (Kuutti, 1996)

An activity consists of a network of co-operative actions, each of which has a goal that works towards the motive (object) of the activity. (Kuutti, 1996) explains how Leont'ev used his famous example of a primeval hunt to highlight the essential difference between an isolated action and one that is done co-operatively as part of an activity. Primitive hunters separate into two groups to perform the activity of catching game: catchers and beaters. In isolation, the actions of the beaters frightening the game do not seem rationally connected with the activity's motive of catching the game – only when it is realised that they are working co-operatively and that the beaters are moving the game towards the catchers does it make sense. Each action, therefore, is done within the

context of a community where individuals have specific roles to play and where a set of rules are followed.

Leont'ev never expanded Vygostky's triangular model to encapsulate this idea. Instead this was done much later by Engestrom (1987), a Finnish professor, who proposed the following structure of a human activity, illustrated in Figure 4.

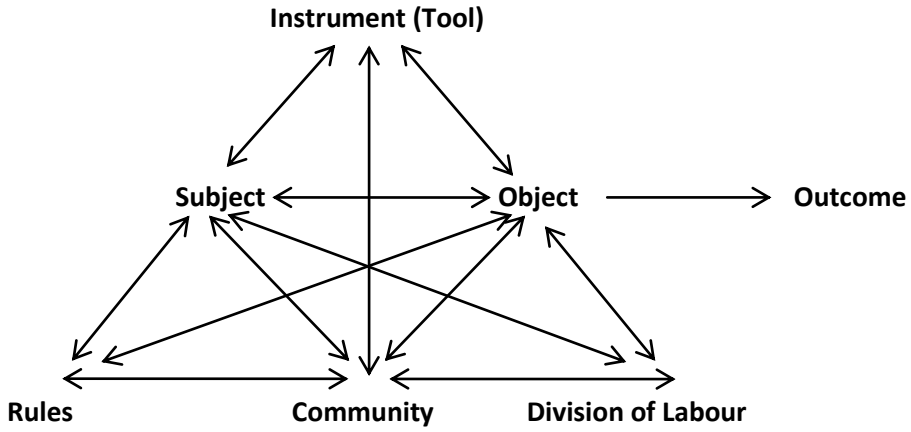


Figure 4: The Structure of Human Activity (Adapted from(Engestrom, 1987))

Here the subject and the object are placed within the context of the community that the activity is being performed in. The way that a subject functions within the community is governed by (and thus mediated by) a set of 'rules' which include the norms, conventions and social relations within the community. In addition, the achievement of an object within a community is mediated by the 'division of labour', the manner in which the community is organised in order to achieve its purpose (eg. catchers and beaters) (Kuutti, 1996). The activity system is part of a network of activity systems within the community (Engestrom, 2001).

The tangibility of an object may vary from something completely material to a shared idea or motive. It is the transformation of the object into the desired outcome that motivates the existence of the object. Both the object and its motive may undergo changes during the activity, being themselves revealed by the actual process of the activity itself. Tools also may evolve in the development of the activity, and thus reflect the historical development of the relationship between the subject and object (Kuutti, 1996). 'Rules' and 'Division of Labour' may also evolve. The entire activity system evolves with everyday practice over time and thus is appropriately described by the maxim "You are what you do" (Nardi, 1996).

Engestrom (2001) summarises what he calls "Cultural-Historical Activity Theory" with the help of five principles:

1. The prime unit of analysis is a "collective, artifact-mediated and object-oriented activity system, seen in its network relations to other activity systems" (pg. 136)

2. Multi-voicedness - the ‘division of labour’ creates different positions for different participants, each of which have their own points of view, traditions and interests. This is inevitably a source of conflict that can lead to positive innovation.

3. Historicity – activity systems can only be understood against the backdrop of their transformation over time

4. Contradictions, or “historically accumulating structural tensions within and between activity systems” (pg. 137), play a central role as a source of change and development in the activity system. Eg. Introducing a new element to the activity system from the outside, like a new tool, may lead to a contradiction with another element, like the division of labour, which may result in change.

5. Expansive transformation to the activity system may be made possible through the contradictions.

The fifth principle is the culmination of those that precede it and the basis for Engestrom’s (2001) proposed concept of expansive learning at work. He points out that the most intriguing types of learning are about solutions that are not defined or understood ahead of time. New forms of activity must be learned while they are being created. Contradictions in existing activity systems may lead to some of the individual participants questioning and deviating from the rules. Expansive transformation happens when groups of these individuals collaborate and reconceptualise a radically different object and activity system that addresses the contradictions (Engestrom, 2001).

3.3 Activity Theory and Information Systems Usage

As Kuutti (1996) mentions, computer technology is a tool that exists and was created for human usage – initially to ‘automate’ human operations and subsequently to ‘informate’ or assist humans to make sense of information. The humans who use the computer to achieve an objective are not purely mechanical; they are humans that function within the context of a community of other humans. Information Systems are advanced and complex technological tools that are used by humans in the setting of the organisations that they and other humans work in. Information systems and their human-computer interfaces are thus “socio-technical” in nature (Crawford and Hasan, 2006: 53), and their use and impact can only be properly understood if the dynamics of the environment are taken into account. Activity Theory is a well developed framework for studying the complex dynamics of settings, and thus a highly appropriate vehicle for the study of the use and impact of information systems and their interfaces (Crawford and Hasan, 2006, Hasan, 1999, Kuutti, 1996, Nardi, 1996, Kaptelinin, 1996).

All information systems usage involves a human ‘subject’ using the system to perform a task with an ‘object’ in mind. The lens of Activity theory allows us to reflect on this use within its organisational context, while acknowledging that the tool – in this case the particular information system being used (Kuutti, 1996, Kaptelinin, 1996) – has an impact on the way in which the ‘object’ will be achieved. The elements of an activity system are expected to evolve over time, which takes cognisance of the fact that while changes in the work environment may be enabled by an

information system, the information system itself may have to change to better suit the environment (Kuutti, 1996, Nardi, 1996).

Despite its name, Activity Theory is not constituted by a "fixed body of accurately defined statements" and is therefore not a *theory* per se, but rather a framework (Kuutti, 1996). As a framework, Activity Theory is holistic and insightful, as its focus is on the whole activity rather than just on the human performing the activity. The suitability of Activity Theory for research in Information Systems is illustrated by its use as a theoretical basis for studies on data warehousing applications (Hasan, 1999), the innovation uptake of Health Information Systems in African countries (Igira, 2008), global computer-based Knowledge Sharing systems (Joshi et al., 2007), the use of Geographical Information Systems in mathematics education (DeBay, 2013), sustaining technology innovations in teacher education (Johnson, 2004), the tensions in rural Telecenters (Lee, 2010). The last two of these were used to inform the approach to this study, and are therefore discussed in the next two sections.

3.4 An application of Activity Theory to a Study on Sustaining Technology Innovations in Teacher Education

Johnson (2004) used Activity Theory as the primary lens through which to explore the results of her study on sustaining technology innovations in Iowa State University's teacher education program. Innovations to integrate technology into the teacher training curriculum had been facilitated by funding from the government's "Preparing Tomorrow's Teachers to Use Technology (PT3)" program, which was being phased out. Johnson's (2004) study examined what strategies teacher education programs had put in place to maintain technology innovations accomplished through PT3 funds, what challenges had been encountered during this process, and what roles leaders had played in sustaining these innovations.

Two theoretical frameworks were used for the study: Activity Theory as the primary framework and Rogers Diffusion of Innovation as the secondary framework.

Rogers Diffusion of Innovation, which attributes the spread of an innovation to the communication of that innovation to different members of a social system over time (Rogers, 2010), was found to be relevant since this study involved the adoption of innovations.

Activity Theory was chosen because it enabled the understanding of the complex context of all the partners supporting teacher education that the sustaining of teacher innovations needed to take place in. The activity system developed for Johnson's research is shown in Figure 5 on page 39.

As can be seen in Figure 5, the activity of integrating technology into teacher education is set in the context of the community of partners supporting teacher education. As such, the activity is performed by a subject that comes from that community and is mediated by both the rules of the community and the division of labour. The activity system thus does help to clarify the relationships between supporting partners, enabling the activity to be better understood in that context.

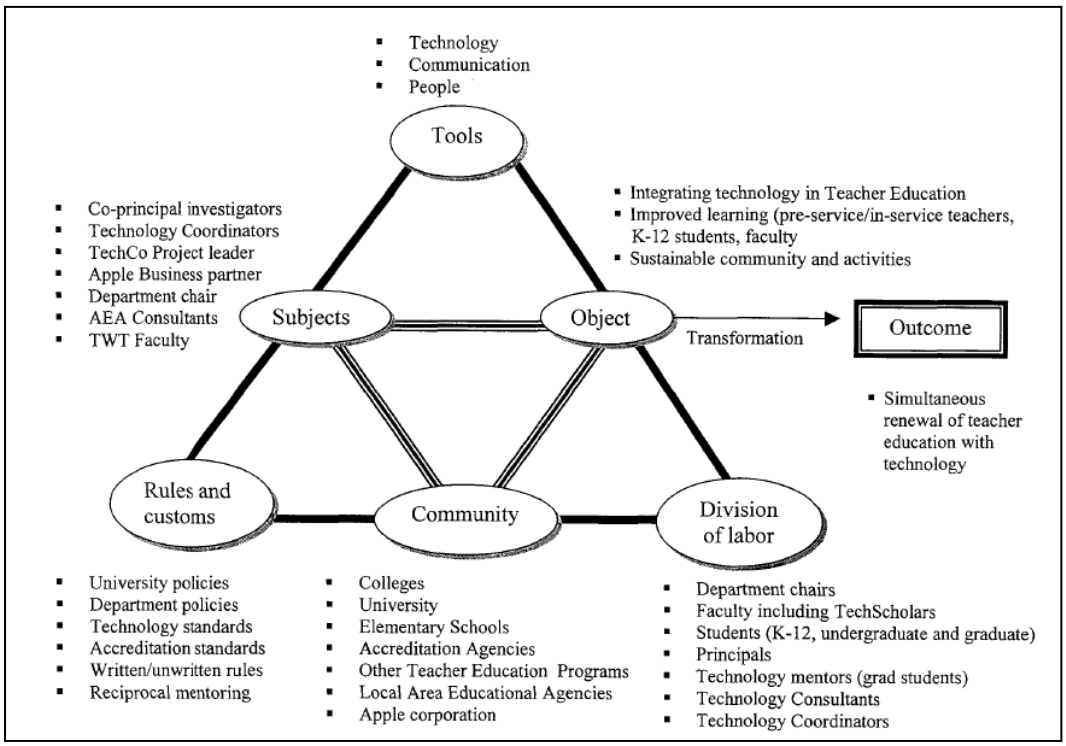


Figure 5: Activity System for Study on Sustaining Technology Innovations in Teacher Education (Johnson, 2004)

Johnson (2004) sees the activity system as being divided into four sub-triangles, as illustrated in Figure 6. Issues identified that concern the mediating effect of tools used on the activity performed by a subject to achieve a object would belong in the "Production" sub-triangle. Similarly, activities can be assigned to the "Exchange", "Consumption" or "Distribution" triangles when appropriate. After presenting the qualitative results of her study theme by theme, she interprets the results using the lens of Activity Theory and placing the issue being discussed in one of the four sub-triangles.

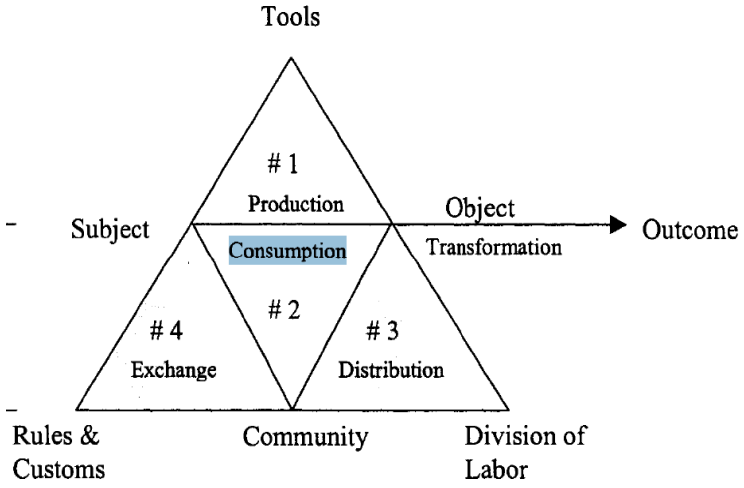


Figure 6: Activity System featuring four sub-triangles (Johnson, 2004)

After presenting her results, Johnson (2004) finds three major contradictions in the activity system: budget cuts, a new federal mandate and a change in leadership. She then discusses each of these in turn, placing them in the relevant triad of the activity system so that they can be viewed in context. For example, budget cuts involve money, which is seen as a mediating tool, therefore budget cuts belong in the "Production" triad, as illustrated in Figure 7.

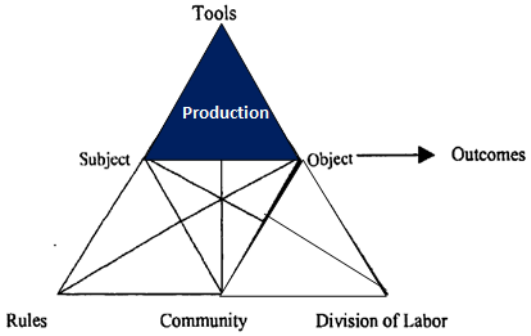


Figure 7: Production Triad Highlighted (Adapted from (Johnson, 2004))

Continuing in this way, Johnson (2004) used Activity Theory to get a full understanding of her research problem within its complex context.

3.5 An application of Activity Theory to a Study on the Tensions in a Nepali Rural Telecentre

Activity Theory was used by Lee (2010) as a framework for his ethnographic study on a Telecentre in Sankhu, a small village outside of Kathmandu in Nepal. The Telecentre, which was built by the government in an attempt to make technology available to the local population who would otherwise not be able to afford access to technology, was not achieving its purpose. The study explores possible tensions faced by the youth of Sankhu that may have prevented them from making use of the Telecentre. Activity Theory was chosen as a framework because it was considered to be a descriptive theory that fits properly with an ethnographic approach.

Like Johnson, Lee (2010) first coded his data to identify common themes. He then sorted the discovered themes according to the elements of the Activity Theory triangle that they corresponded to, i.e. Subject, Object, Outcome, Community, Rules and Division of Labour. Once sorted it was found that tensions (represented by the arrows on the Activity Theory triangle) emerged, and these could be charted on the Activity Theory triangle. For example, when it was found that gender expectations – essentially a societal rule - prevented female youth from using the Telecentre, this could be mapped on to the sub-triangle shown in Figure 8. Note that the production, exchange, consumption and distribution sub-triangles used by Johnson (2004) were not applied to this study. The approach of Lee (2010) also differed from that of Johnson (2004) in that results were interpreted as they were being presented, rather than being presented in one chapter and interpreted in another.

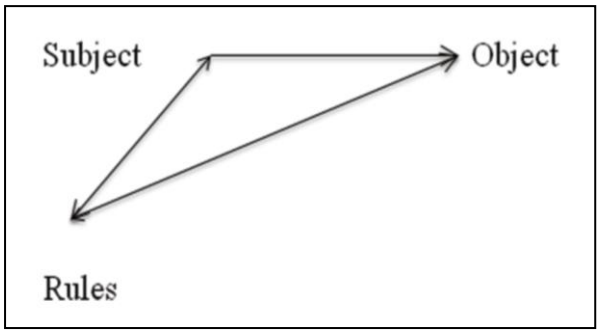


Figure 8: Smaller Activity Theory Triangle. (Lee, 2010)

This smaller triangle is shown as part of the larger Activity Theory triangle in Figure 9.

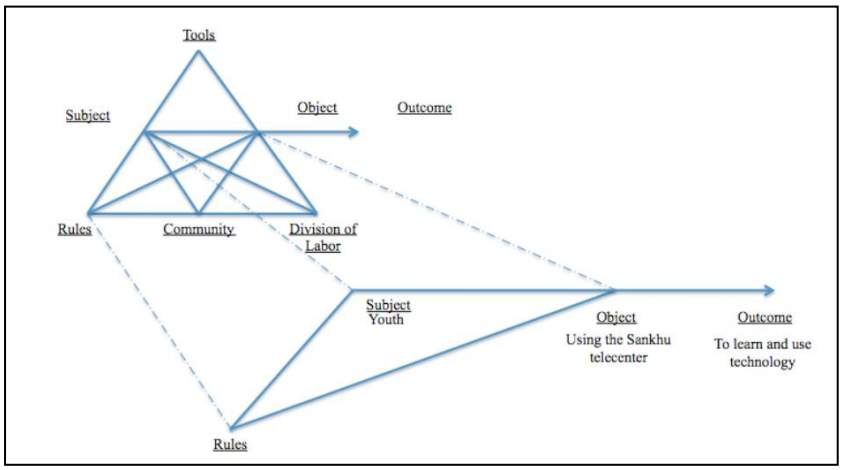


Figure 9: Smaller Activity Theory triangle in context of larger Activity Theory triangle (Lee, 2010)

Lee (2010) did not present an initial activity system for the project – instead Activity Theory was applied directly to the results so that the activity system emerged as the results progressed, helping to illuminate the discovery of the tensions in the Telecentre with in the context of the society.

Since the context of this study is relatively complex, it was considered advantageous to develop an activity system for this research, as was done by Johnson (2004). For this reason the context of the NHLS CDW, as discovered during the research process, is described below.

3.6 The Context of the NHLS CDW

The context of the NHLS CDW was discovered as data collection proceeded and the mechanisms for data flow through the NHLS and related organisations were better understood. Most of the information about these mechanisms was obtained in a preliminary session with CDW staff, but

some emerged more clearly during the interview process. The context is described here as it is pertinent to the research process.

The main 'client' of the NHLS is the Department of Health (DoH), who are responsible for the network of public health facilities that require laboratory tests to be done. As such, much of the data in the NHLS CDW originates from the doctors and other clinicians within the DoH facilities that request the tests on request forms. Test requests are then passed on to the "Receiving Office" of one of the NHLS's 260+ laboratories in the country, where the data on the lab request form is captured onto TrakCare, the standard Laboratory Information System (LIS) used in all NHLS laboratories. After the tests have been done, TrakCare is updated with the results and perhaps also with a pathologists report. Information from each TrakCare LIS at each lab is transmitted to the CDW on a regular basis.

Right time access to laboratory test results is required for two purposes. Firstly, DoH must receive results testing positive for anything designated as a category 1 notifiable medical condition (eg. Anthrax, Botulism, Cholera, Listeriosis, Malaria, Yellow Fever) within 24 hours.[15] This information may be used to detect possible outbreaks or to determine which patients in hospitals should be isolated. Secondly, mobile clinics need to receive lab test results quickly. In response CDW (in conjunction with an outside company) have developed a process whereby new laboratory results are transmitted to the CDW's operational data store (ODS)[16] within a short timeframe. From the ODS, pertinent information is transmitted to the DoH and the mobile clinics. This is done on a continual cyclical basis, and CDW staff claim that the DoH will receive information about category 1 notifiable diseases within two hours of test results being entered onto the TrakCare LIS. This requirement being satisfied, the ETL[17] process can be applied to the data on the ODS and with time the information is integrated properly with the data on the CDW.

To allow for a full picture of cancer incidence, data from private laboratories is also loaded onto the CDW. Information from the CDW is regularly downloaded to the National Cancer Registry (NCR) where it is "coded" by medical specialists using classification systems like ICD10. The CDW is then subsequently updated with these codes, but the data also remains in the NCR for use in profiling the disease.

An additional source of data for the CDW is the Finance Module of the Oracle ERP that is used to keep track of the NHLS billing. Billing information is transmitted to the CDW where it can be enquired on using BI tools, or where actual invoices can be produced.

Data in the NHLS CDW is valued and used by staff in the various divisions of the NHLS, as well as the DoH, non-governmental organisations (NGOs) like "Right to Care" and "Aurum" that support

---

[15] http://www.nicd.ac.za/index.php/nmc/notifiable-medical-conditions-list/

[16] Organisations often create a specialised data warehouse, known as an operational data store (ODS), which stores a subset of data that quick access is needed to - ECKERSON, W. 2006. *Performance Dashboards: Measuring, Monitoring and Managing Your Business,* New Jersey, Wiley. Aside from fact that speed of access is retarded due to the large size of a data warehouse, use of the data may be required before there is time to clean and integrate it sufficiently for inclusion in the data warehouse. Data is usually subsequently cleaned, integrated, and placed in the data warehouse.

[17] ETL – Extract, Transform and Load. This is the process whereby new data is prepared for inclusion in the data warehouse.

government programs, private companies that win tenders to do work for the government and organisations like the Global Fund and the World Health Organisation (WHO). Data is used for population health purposes, for administration and management and for medical research.

Data can be acquired from the CDW in a number of ways. There are a number of predefined reports and dashboards that can be accessed easily once permission has been granted. Ad-hoc or unusual requests for information can be made through a service representative at CDW. An additional department, the Surveillance Information Management Unit (SIMU) produces dashboards and reports that can be used with the CDW. Some of the departments, like the National Priority Program (NPP) employ their own data analysts. Members of those departments access the data through these data analysts, who function as intermediaries and deal with the CDW to get the data. These data analysts can use software called "Aginity" to get direct access to authorised parts of the data in the CDW via SQL. It is interesting to note that On-line Analytical Processing (OLAP), typically an end-user tool access to a data warehouse, is not provided to data requestors as a means of getting directly to the data in the warehouse. Instead, it is used by technical specialists as part of the ETL process.

3.7 Developing an Activity System for a Request for Data from the NHLS CDW

For ease of explanation, a summarised version of the activity system is shown in Figure 10. As shown, the 'subject' is the person requesting the information, and their 'object' is to get the required information. The intended 'outcome' of the subject receiving this information is that it should aid in decision making and ultimately have a positive impact on population health. The achievement of the 'object' is mediated by the 'tool' being employed; in this case the NHLS CDW and associated facilities. The 'community' that the 'subject' is situated in includes staff from the NHLS and the DoH, and the 'division of labour' would involve the way in which tasks are spread amongst these staff. The 'rules' would include the procedures users need to go through in order to make an information request and the types of requests allowed.

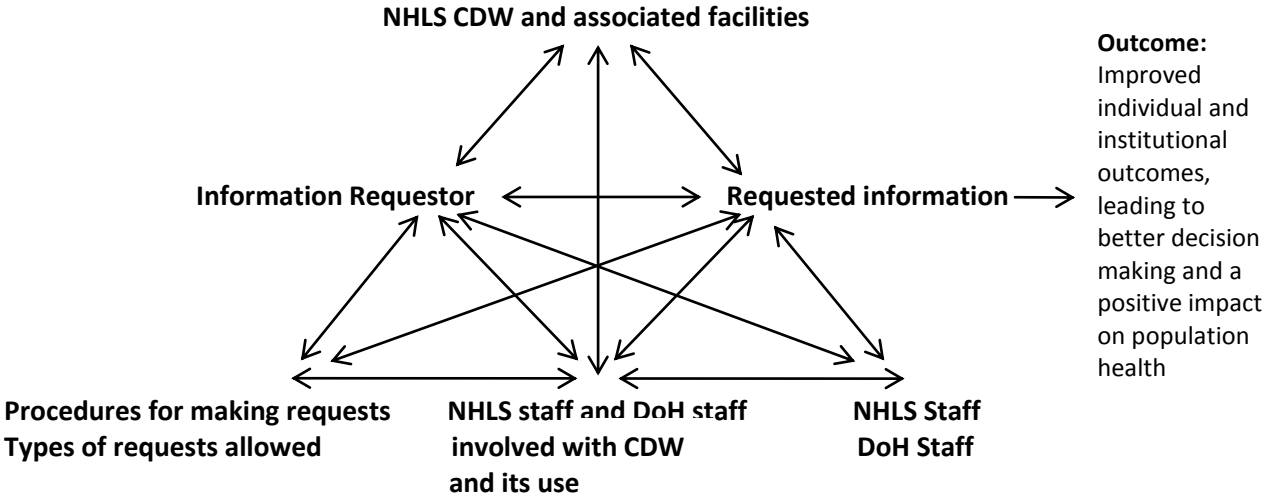


Figure 10: Summary Structure of an Information Request Activity

The extended version of this activity system that shows all the elements of the context described in the previous section is shown in Figure 11 on page 45. Following the approach of Johnson (2004), the activity system is used to capture and understand the complex context of the NHLS CDW by enumerating the participating entities of each activity system element. In conjunction with this, Lee (2010)'s approach of assigning each theme to the appropriate activity system sub-triangle (as exemplified in figure 9), was used to gain a deeper understanding of the issues and challenges of making information requests. The four sub-triangles used by Johnson (2004) were not found to add value to this study, and therefore they were not applied.

This proposed activity system formed the springboard for the exploration of the way in which the NHLS CDW is used. As it evolved during the study not only did a clearer picture of its elements emerge, but contradictions between the elements were revealed, thus allowing for constructive comment in the discussion and concluding sections of this report.

After identifying the uses of the NHLS CDW through examination of information requests (RQ1), the study will explore the issues and challenges in the use of the CDW (RQ2) by examining procedures for making information requests. Here an information request will be regarded as an activity performed by a subject (information requestor) to achieve an object (get the information), and activity theory will be used to understand it within its context - the NHLS CDW and surrounding facilities; the rules that govern the environment, eg. procedures for making requests; staff responsible for the NHLS CDW; and division of labour amongst the staff (RQ2). RQs 3 and 4 address the outcomes of the activity system for individuals, the institution and population health by exploring the benefits for individual decision makers and perceived population health impacts.

3.8 Chapter Conclusion

This chapter has explained the basic concepts behind Activity Theory as the chosen theoretical underpinning for this study, and explored different applications of Activity Theory in Information Systems Research. Thereafter the context of the NHLS CDW was described and the activity system for an information request developed.

The next chapter will examine the methodology that was used to conduct the study. Choices of research philosophy, approach and strategy are justified and the data collection process described. Methods of sampling and the development of the interview protocol are outlined. Attention is also given to the qualitative analysis done, method evaluation, the role of the researcher, ethical clearance and limitations of the study.

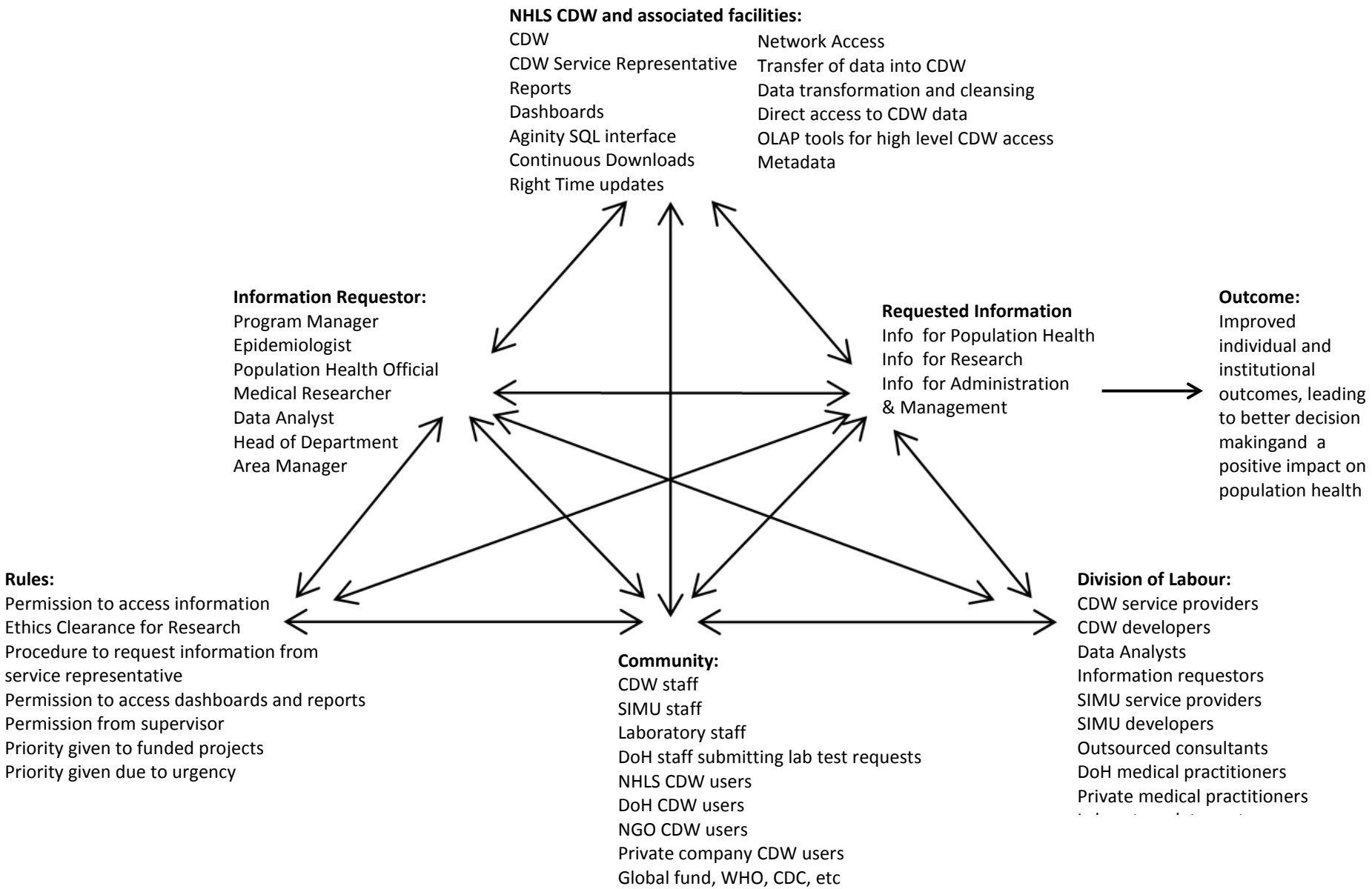


Figure 11: Structure of an Information Request Activity

4 Methodology

The methodology used to carry out the research is described in this chapter. First, the chosen philosophy, approach, and strategy of the research are explored, and then the approach towards data collection, the procedure of data collection, and choice of sampling method are explained. Thereafter the interview process is discussed, with attention paid to the type of interviews done and the development of the interview protocol. The chapter then outlines the stages of qualitative analysis performed, and considers method evaluation as a means of ensuring rigour. It then contemplates the role of the researcher, reports on the process of obtaining ethics clearance, and ends with a discussion of the general limitations applicable to this type of study.

The collection of empirical evidence has meaning only if it is done within the context of the ideological setting of the way in which the world is seen and the way in which that world is best studied. For this reason it is best to consider the philosophy, approach and strategy to be used in the research process before considering choices of data collection and analysis techniques. Saunders et al. (2009) use the onion depicted in Figure 12 to represent the different 'layers' that should be 'peeled away' or contemplated . before performing data analysis. This chapter will examine each of these aspects in turn.

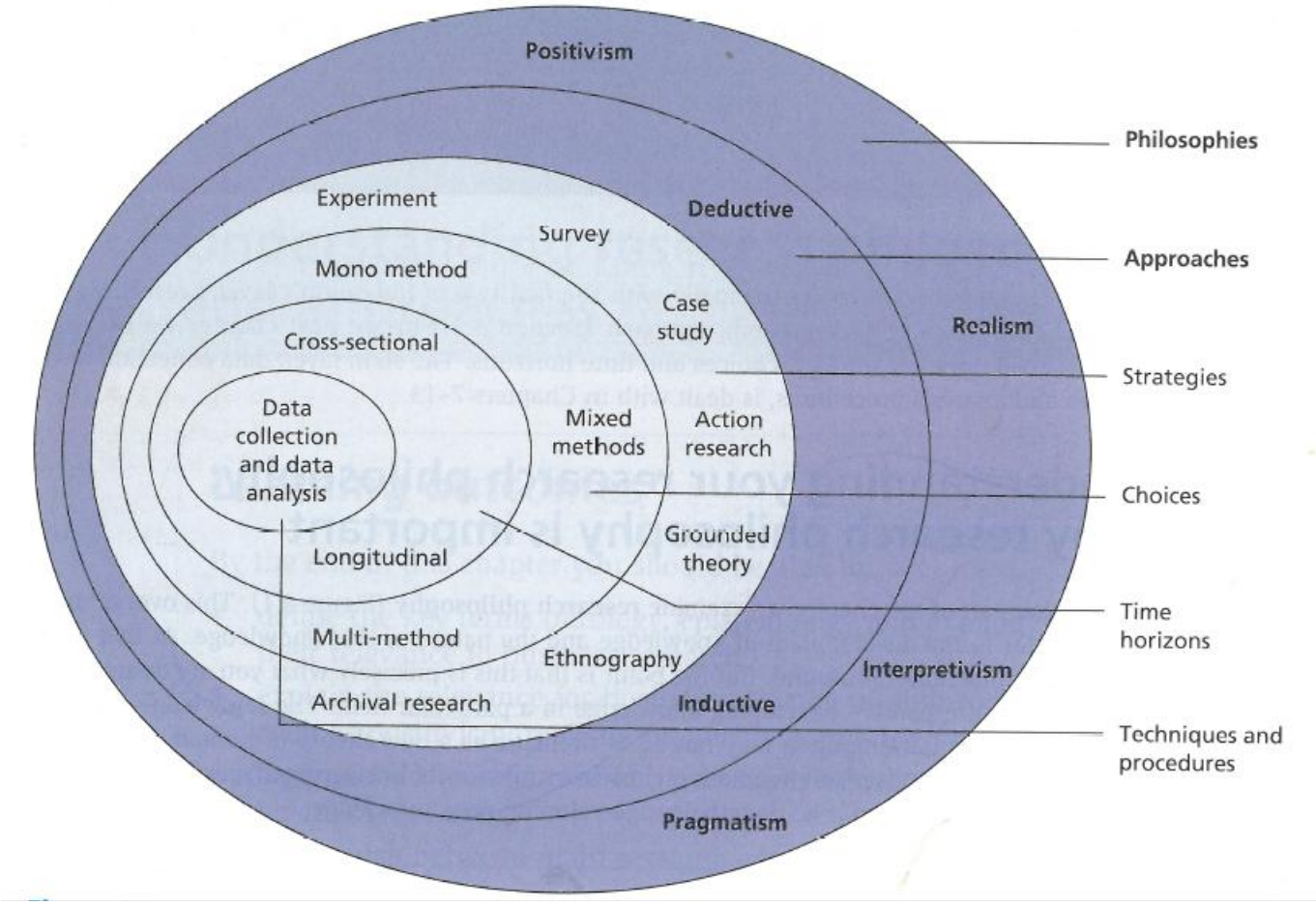


Figure 12: The Research 'Onion' (Saunders et al., 2009)

4.1 Research Philosophy

The underlying ontology – assumptions about the way we see the world – and epistemology – assumptions about the best way to study the world – for this research are reflected in the choice of philosophical paradigm, or mental model that has been used to organise reasoning and observations (Bhattacherjee, 2012). The two philosophical paradigms most commonly informing information systems research are post-positivism and interpretivism.

Positivism, a paradigm that has informed research for hundreds of years, views the phenomenon under study as being separate from the researcher, enabling the researcher to gain knowledge and understanding through objective observations and measurements. Post-positivism, a development of positivism, allows for the augmentation of objective observation and measurement with logical reasoning (Bhattacherjee, 2012). The requirement for objective measurement means that these paradigms are not always suited to studying the social world, or how humans think and act in the context of their communities (Oates, 2006). As pointed out earlier, the phenomenon under study is concerned with the social world of humans making information requests within the context of their organisation. The use of the Activity Theory framework to understand this activity requires the researcher to engage with these human 'subjects' in order to assess different aspects of the activity system. These engagements create a level of involvement with the human 'subjects' so that the researcher is not separate from the phenomenon of study and therefore not able to make objective observations. The positivism and post-positivism paradigms were therefore not considered appropriate for this study.

The subjective engagements with humans required for this study fits in very closely with the interpretivism paradigm, which postulates that the social world should be studied through the subjective interpretation of the different humans involved (Bhattacherjee, 2012). According to Oates (2006), applying interpretivism to information systems research concerns understanding the social context of an information system – to explore the relationships and interdependence between all its different factors. This clearly corresponds to the manner in which Activity Theory was used as a framework for this study. This research was thus informed by the interpretivist paradigm.

4.2 Research Approach

Quantitative research is based on numerical and standardised data analysed through the use of diagrams and statistics. Quantitative research results depend on meaning that is derived from numbers. Meaning, however, is not always dependent on numbers (Saunders et al., 2009). Saunders et al. (2009), citing Dey (1993), point out that working with ambiguous and inflexible concepts makes it more difficult to quantify data in a meaningful way. This study was concerned with individuals functioning in a social context and as such needed a richness of analysis that is not possible with numbers, so quantitative research was not considered to be an appropriate approach.

Qualitative research, on the other hand, is based on meaning that is derived from words. It provides an opportunity to explore a subject in 'as real a manner as is possible', resulting in a more rich and full understanding of the situation being explored (Saunders et al., 2009). Since a rich and full

understanding of the process of making a request for information from the NHLS CDW was the objective of this study, the qualitative approach was deemed to be the most appropriate. Individual perspectives were collected in largely textual format and then analysed qualitatively. Any documents collected were also analysed qualitatively.

Bhattacherjee (2012) explains that "all research is an iterative process of observation, rationalisation and validation" (pg. 22). *Observation* refers to the exploration of the phenomenon being researched, which is then made sense of or *rationalised* by pulling together and analysing what has been discovered. *Rationalisation* may lead to the formulation of a theory, which would be tested in the *validation* part of the process. Inductive research refers to that part of the process that attempts to rationalise the observations, whereas deductive research refers to the validation or testing of a theory that has been proposed during rationalisation.

This study was concerned with collecting the observations about information requests made to the NHLS CDW and then attempting to rationalise them. Since activity theory was used as an underpinning for this rationalisation, there was an element of deduction. However, the research was mainly inductive in nature as no proposed theory was tested.

4.3 Research Strategy

Yin (2003) suggests that there are three conditions that should be considered when selecting a research strategy:

- "the type of research question posed"
- "the extent of control the investigator has over actual behavioural events"
- "the degree of focus on contemporary as opposed to historical events"

The type of research question posed gives an indication of whether the research is to be exploratory, descriptive or explanatory. The purpose of an exploratory study is to find out *what* is happening – to understand a phenomenon and its associated problems (Oates, 2006). This research was concerned with finding out what happens at the NHLS with respect to information requests to the CDW, and with uncovering any associated issues and challenges. Consequently it was exploratory in nature. This concurred with the idea expressed by Saunders et al. (2009) and Yin (2003) that "what" research questions may imply explorative research - the four research questions for this study all started with the word "what".

Answering a "what" question may also entail assessing the dimensions of a problem or phenomenon, answering questions like "how much?" and "how many?" (Yin, 2003). Evaluating the data in this way, or alternatively synthesising ideas to more deeply understand the phenomenon and its associated problems, would result in the building up of a profile of the persons, events and situations involved. This would make the research descriptive in nature (Saunders et al., 2009). Activity Theory, the lens through which this study will attempt to synthesise the different perspectives of NHLS CDW use, is claimed by Nardi (1996) to be a "powerful and clarifying descriptive tool" (pg. 4). Through using activity theory to synthesise the ideas found in the data

collected for this study, a clearer profile of NHLS CDW use was built up, enabling constructive comment for improvement.

This study can thus be described as both exploratory and descriptive. An explanatory element has also been included in so far as the analysis has helped to explain the way in which elements of the activity system are influenced by one another.

Moving on to the second condition, which concerns the extent of control the researcher has had over behavioural events, it is evident that there was none. An independent researcher coming in to an organisation like the NHLS has no power to change the way that the information requestors behave.

The third condition requires an assessment of the degree of focus on contemporary as opposed to historical events. The main focus here was on contemporary work, although there was some examination of historical events.

Various research strategies were considered with the above in mind. Evidently an experiment was not a valid strategy here, as the researcher had no control over the behaviour of the human 'subjects'. Experiments and surveys were also more suited to a deductive rather than an inductive approach. An archival strategy was not fitting either, as it entails looking at documents and focusing on historical rather than contemporary events (Oates, 2006).

Exploratory and descriptive research of this nature could have been approached using any strategy, but a case study strategy was deemed to be the ideal approach to achieve the objectives of the study. Firstly, the intention of this research was to examine CDW use from within the natural setting of the NHLS. Benbasat et al. (1987) propose case methodology for situations where mainly contemporary events are to be examined in their natural setting. Secondly, the activity of requesting information from the NHLS CDW was to be studied as performed by a human subject in the social context of the organisation. Several authors indicate that a case study would be suitable for this, saying that the case study has the ability to "develop a nuanced understanding of the interlocking set of social, cultural and political factors" (Bhattacherjee, 2012: 41) and gives researchers a holistic view of what is really happening" (Yin, 2003). The case study would be classed as idiographic –i.e. as research that attempts to understand a phenomenon in its context (Benbasat et al., 1987), which is precisely what this study attempted to do. Lastly, both Benbasat et al. (1987) and Oates (2006) explain that case studies are used to study a single situation in depth so that a detailed understanding is acquired, as opposed to other strategies like surveys that acquire more shallow information over a greater breadth of situations. Again, this meant that the case study approach was fitting, as this research aimed to do a detailed investigation into a single situation – the use of the NHLS CDW. Note also that this was a single-site case study as no site other than the NHLS was looked into.

4.4 Data Collection

Interviews with people who request information from the NHLS CDW constituted the predominant method of data collection, although various documents were collected as well. Before interviews could begin it was necessary to set the scene by clarifying how the CDW department at the NHLS works, what modes of access information requestors have, and how many and what type of information requestors there are. A list of possible interview participants needed to be identified, and early information about CDW use and impact was expected be useful. Data collection therefore began with a preliminary discussion with CDW staff. This session was held at the CDW offices on 20th October 2017, and the above were discussed. Much of the description of the context of the NHLS CDW in section 3.6 of this report was derived from what was said in this preliminary session. Modes of access for information requestors mentioned were the use of customised dashboards and reports and the filling in of a form to request data from a CDW service representative. No exact number of information requestors was given, but different types of information requestors dealing with operations management, medical research and public health tracking were identified. A list with the contact details of 14 possible interviewees was later received from them. Documents and other possible sources of useful information were requested, but only one – a powerpoint presentation giving an overview the CDW and examples of available dashboards – was emailed.

In addition to interviews, publicly published medical research, reports, articles, and policy documents, as well as a form and slide presentation obtained from the NHLS during data collection, were examined. Information contained in these additional data sources were used to supplement, as well as verify the reliability of, interview data.

Before embarking on the interview process it was requisite to consider an approach to sampling the data to be collected. Activity theory, which is being used as the framework here, dictates that the unit of analysis is the activity (Engestrom, 2001). The activity in this case is a human subject making an information request from the NHLS CDW. The population from which the sample must be drawn includes all requests for information from the NHLS CDW. Access to knowledge about these requests can only be attained through interviewing the requestors. The choice of sampling frame, or accessible section of the target population (Bhattacherjee, 2012), depended on which requestors were made available for interviewing and what types of requests they make.

Requestors were selected from the sampling frame with the objective of optimising the research value of the sample. Hence the probability of each individual unit being selected was not known, and consequently required non-probability sampling (Bhattacherjee, 2012, Saunders et al., 2009). Two non-probability techniques were used to select the sample, namely quota sampling and purposive (judgement) sampling.

Selection of the requestors interviewed was done using purposive sampling, by selecting for inclusion those people who were most likely to provide valuable data to answer the research questions (Oates, 2006, Saunders et al., 2009). Consideration was given to the types and variety of requests that they were likely to be making of the CDW so as to ensure that as wide a range of requests as possible were uncovered.

Quota sampling involves segmenting the population into mutually-exclusive subgroups, and then deciding on a predefined quota that will be selected from each group (Bhattacherjee, 2012). In this case two dimensions were combined to create these groups. These are firstly, the way in which the information request is made, and secondly, the purpose for which the information is being requested.

The literature revealed several ways of making information requests that have been used in other applications of BI for population health. These are shown in table 4. Although only the first two of these had been identified as being used with the NHLS CDW in the preliminary discussion, it was still possible that use of the latter two would be discovered during the course of the interviews.

Ways of Making an Information Request	Source
A customised application, report or dashboard that has been specifically developed for the requestor's use	(Berndt, 2001, Berndt et al., 2000, Nambiar et al., 2013) A conversation with a NHLS CDW staff member revealed that information requestors make use of specially developed dashboards.
A service representative that processes an information request for the requestor	(Tremblay et al., 2009)
High level[18] CDW access tools like data warehouse browsers and OLAP	(Xiaohui et al., 2004, Berndt, 2001, Berndt et al., 2000)
Low level CDW access tools like SQL[19] and analytics	(Liyanage et al., 2014, Berndt et al., 2000)

Table 4: Ways of Requesting Information Found in the Literature

Bearing the functions of the NHLS in mind, as well as the type of information stored in the CDW, the purpose of an information request might possibly be classed as shown in Table 5.

[18] In this report, 'high level CDW access tool', denotes a way of accessing CDW that is closer to the way the user thinks and therefore easier to use. 'Low level CDW access tool' refers to tools that can only be used if a level of technical expertise is required first.

[19] SQL stands for 'Structured Query Language', and is the standard computer language used for accessing data in a relational database

Purpose of Information Request	Source
Medical research	A literature search on the NHLS CDW reveals at least 42 medical research articles acknowledging their use of the NHLS CDW. This suggests the NHLS CDW is being used to generate data for medical research purposes.
Administrative	(NHLS, 2015)
Public health tracking	(NHLS, 2015)

Table 5: Possible Information Request Purposes

Figure 13 depicts a sampling matrix showing the quota initially considered for each sub-category.

	Medical Research	Administrative	Public Health Tracking
Customised Application	**1**	**1**	**2**
Service Representative	**1**	**1**	**2**
High level CDW Access Tool	**1**	**1**	**2**
Low level CDW Access Tool	**1**	**1**	**2**

Figure 13: Original Quota Sampling Matrix

It depicts the initial aim to get a total of 16 in the sample. A higher quota was assigned to all public health tracking requests, as these were likely to be more closely aligned to a positive outcome for population health, which was of primary interest in this study (RQ4). A sample size of 16 is acceptable for a qualitative study of this nature, where collection and analysis skills are more important than the size of the sample (Oates, 2006).

During interviews, however, it was discovered that there is an additional way in which end users can make an information request, and that is through an intermediary called a 'data analyst'. These people get information from the CDW on behalf of other people, so the other people are accessing the CDW through them, whereas they are accessing the CDW directly. This was regarded as two requests – one made by the original requestor through the analyst, and one made directly to the CDW by the data analyst. An additional row was therefore added to the matrix to represent information requests done through data analysts. Unfortunately, since no researchers requesting information from data analysts were interviewed, there were no requests sampled at the intersection between data analysts and medical research. From interviewing the data analysts, however, it is known that they do handle requests for medical research purposes. It is also interesting to note that the only requestors using a low level CDW tool like SQL are the data analysts.

Furthermore, preliminary evidence suggested that no information requests fitted into the first cell of the matrix, representing information requests done for purposes of medical research through a customised application. On reflection this makes sense. This cell would never be filled, since each

new research project is different, so the information request would not be repeated and there would be no point developing a customised application.

In addition, no direct use of the CDW through high level CDW access tools like data warehouse browsers and OLAP was revealed during the interviews. Instead, information extracted from the CDW is often emailed to the requestors in Microsoft Excel format, so that they can use Excel to further analyse the data. In one case extracted data is actually stored on the department's own local SQL Server database so that Microsoft's Power Pivot tool can be used to do OLAP type work and other analyses of the data.

Thus the distribution of information requests actually sampled is shown in figure 14. Note that to simplify this categorisation, an information request was matched to the purpose the information was being requested for – each purpose representing a single request. This imposed standard helped to clarify the number of requests, since it was found that in some cases many reports or dashboards contributed towards satisfying a single purpose, and in other cases it was not clear whether information required for different purposes may have been combined into a single report. Some of the purposes given were more generic than others – in this case they were still considered to be a single sampled request.

Information requests recorded in this matrix represent only those mentioned by the participants in the interviews. It is of course quite possible that they make additional requests that were not spoken about.

	Medical Research	Administrative	Public Health Tracking
Customised Application	**0**	**6**	**8**
Service Representative	**4**	**3**	**2**
Data Analyst	**0**	**1**	**3**
High level CDW Access Tool	**0**	**0**	**0**
Low level CDW Access Tool	**2**	**2**	**2**

Figure 14: Information Requests Sampled

A total of 33 requests were therefore sampled with data on each request collected through ten interviews as discussed next.

4.5 Interviews

As mentioned, the primary mode of data collection was interviews, in line with general case study practice (Bhattacherjee, 2012, Oates, 2006). An interview is described by Saunders et al. (2009), citing Kahn and Cannell (1957), as a purposeful discussion between two or more people. For this research the discussion was held between the researcher and one other person, and would therefore be classed as a 'face-to face' interview (Bhattacherjee, 2012). Oates (2006) points out that

there is a tacit agreement that the researcher has control – there is an open objective of producing material for research. The researcher thus planned the meeting with a purpose and agenda in mind and steered the conversation to topics of interest. The content of the interview was accepted to be 'on the record' for later use and analysis.

Interviews could be structured, where predetermined, standardised and identical questions with often pre-coded answers would be asked of every respondent, semi-structured, where the list of themes or questions to be covered would be more flexibly used, or unstructured, where a topic would be introduced and the interviewee is given the opportunity to talk freely (Oates, 2006, Saunders et al., 2009). Semi-structured interviews were considered most appropriate for this study and were found to provide direction for the interview while still allowing the interviewer to probe into interesting or ambiguous responses to gain a deeper understanding. This probing sometimes was even found to lead the discussion into areas that had not previously been considered. Because of the semi-structured nature of the interviews, the time each interview took was variable – the shortest taking 24 minutes and the longest taking 1 hour and 10 minutes.

Yin (2003) points out that case study interviews are different to survey interviews in that the interview protocol is not targeted directly at the interviewee, but rather serves as a reminder to the interviewer of what questions need to be asked. Interviews were recorded as advised by Benbasat et al. (1987) and Oates (2006) and this did enable the interviewer to focus on the interview while being assured that none of the information was lost. Interviewees gave permission for this recording by signing a consent form before the interview. Bhattacherjee (2012) and Oates (2006) both recommend that provision be made for capturing field notes to record observations and comments on the context of the interview, as well as interviewer thoughts on the atmosphere, perceived intent of the interviewee, non-verbal communication, etc. Making notes during the interviews was found to detract from what was being said, but a short summary of researcher perceptions was noted down shortly after each interview.

As mentioned earlier, the activity theory framework has been used as a theoretical basis for various information systems research projects, including studies on data warehousing applications (Hasan, 1999), the innovation uptake of Health Information Systems in African countries (Igira, 2008), global computer-based Knowledge Sharing systems (Joshi et al., 2007), sustaining technology innovations in teacher education (Johnson, 2004), the tensions in rural Telecenters(Lee, 2010) and the use of Geographical Information Systems in mathematics education (DeBay, 2013). The general approach in these studies has been to map the data collected onto the different elements of the activity system triangle, and thus to synthesise the information and derive meaningful comments. The interview protocol for this study was based on the one used by Johnson (2004) in her study (described in chapter 3) on sustaining technology innovation in teacher education. In her protocol, each question is listed together with the literature source it was derived from, and the element(s) of the activity system that the question addresses.

The initial interview protocol evolved during the first few interviews, partially due to new interesting themes that were unveiled in those interviews. Experience doing the interviews also informed a slightly better approach to the ordering of the questions. After the first three interviews, the protocol stabilised and no further requirements for change were encountered. The final interview protocol is shown in appendix A.

The protocol was divided into three sections. The first section starts with some demographic questions about the subject being interviewed and ends with two general questions about the requests they make from the NHLS CDW. A list of one or more information requests made by the subject was obtained at this stage.

The second section lists questions about a specific information request or 'activity', which is the unit of analysis. Where possible and relevant, this section was repeated for each of the information requests in the list obtained in the first section. It may be noticed that this second section starts with questions about the outcome of the request. This was one of the changes made to the protocol – it was found that an initial explanation of the outcome and perceived impact of the request put the participant at ease and set the scene for the discussion of issues and challenges.

The third section of the interview protocol addresses possible concerns about use of the NHLS CDW in general – i.e. not specific to a particular information request. These questions were asked once only, after questions on all the requests had been completed.

In accordance with Johnson (2004), many of the questions are followed by other, probing questions (indented, and preceded with a bullet point) that can be used to further prompt the interviewee for more detailed information.

Bearing in mind the summarised structure of an information request activity system – shown here again in Figure 15 – Table 6 indicates which interview questions deal with each of its elements. Note that some of the questions deal with more than one of the activity system elements.

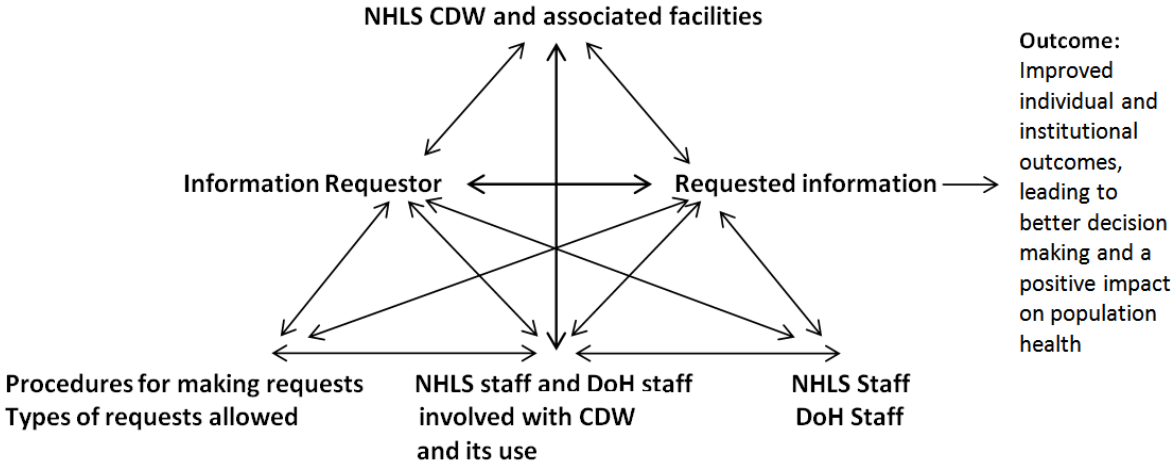


Figure 15: Summarised Structure of an Information Request Activity

Activity System Element	Activity System Element for this study	Questions
Subject	Information Requestor	10,11,12,13,14,15,16,17,18,24,26,27
Tools	NHLS CDW and associated facilities	10,13,16,17,18,19,20,21,22,24,26,27
Object	Requested information	16,19,20,21,22,23,24,26,27
Outcome	Better decision making, leading to a positive impact on population health	7,8,9,22,28
Community	NHLS staff and DoH staff involved with CDW and its use	10,11,12,14,15,26,27
Rules	Procedures for making requests Types of requests allowed	10,23,25,26,27
Division of Labour	NHLS Staff DoH Staff	11,12,14,15,26,27

Table 6: Mapping of Interview Protocol questions to elements of the Information Request Activity

The probing questions associated with question 7, "What information do you request from the NHLS CDW?" are specially designed to reveal the link between the activity and the eventual outcome.

The first two probing questions address the answer to RQ3, "What are the individual outcomes for the NHLS CDW user and data requestor?", and the others address the answer to RQ4, "What are the perceived benefits (impacts) of NHLS CDW use for the NHLS as an institution and for population health in general?"

They are:

- Why are you requesting this information?
- In what ways is this information helping you?
- In what ways is this information helping the NHLS?
- In what ways do you feel that the information you requested may be improving population health?

Asking these questions in this way encouraged the interviewee to disclose firstly the impact that this information has on his/her own objectives, then to move on to the possible impact on the NHLS as an organisation, and finally to discuss any possible impacts on population health in South Africa, and perhaps beyond.

The data for this study was collected once, by interviews conducted over a period of almost two months. This study was therefore not longitudinal in nature, but rather represented a cross-section of all possible information requests.

Prior to conducting interviews a pretest was done, which involved a review of the interview schedule by experienced and qualified researchers who could comment on the structure and content of the interview schedule. Feedback from the pretest was used to refine the final interview protocol.

Of the fourteen contacts for possible interview participants, eleven were purposively selected with the aim of covering as wide a range of information requests as possible. Each of the selected contacts was invited to participate via an email with a participant information sheet for the study attached. Only one of these contacts replied to this email and the interview was duly arranged. After following up with phone calls and also physically knocking on doors at the NHLS, a further nine interviews were arranged. In total ten interviews took place between 20th November 2017 and 12th December 2017, each interview only being done after consent to do the interview and consent to record the interview had been signed by the interviewee. The interviewees were from four different parts of the NHLS – the NICD, the NCR, the NPP and operational management. As mentioned above, a total of 34 information requests were encountered and explored during the interviews. These requests were made in different ways: through customised applications, through service representatives, through data analysts and through SQL, which has been classed here as a low level CDW access tool; and for different purposes: for medical research, administration and public health tracking. Thus most quota sampling requirements were met with ease. With the exception that no researcher requesting data from a data analyst was interviewed, all categories were either covered or a reason was established for why they would never actualise in the context of the NHLS CDW. The exact number of requests found for each category is shown in figure 14 above.

Once each interview had taken place and been recorded, the process of data analysis could begin.

4.6 Data Analysis

Data analysis took place both during and after data collection. Interview data for analysis included recordings of the interview, notes about the interview taken shortly afterwards, and any related documents collected during and/or after the interview.

After each interview the recording was transcribed. For this purpose the “Transcription Template” shown in appendix B was created. It is made up of a Microsoft Word table with four columns – one for a left margin, one to indicate whether the interviewer or participant is speaking, one for the transcribed text, and one for the right margin. It was found that this template sped up the transcription process and ensured a uniform format for all interviews. Because it is a Word table, its structure is flexible, so the margins could be stretched later on for the insertion of comments about the data, any amount of text could be typed in for each entry and gridlines could be removed or inserted at will. The “Express Scribe Transcription Software”, which – among other things – allows you to control the speed of audio play back, was used to assist in the transcription. Following advice from Saunders et al. (2009), each transcribed interview was saved in a separate file on a password protected computer, with a file name containing codified information so that it can be recognised yet still preserves anonymity.

Common themes began to be identified as data collection proceeded, and then, as recommended by Oates (2006), once all the data was transcribed it was be read through once in its entirety to get a general impression, and to identify more common themes. A list of themes was kept and updated as the research progressed and new themes were identified. Themes were developed with reference to - but not restricted by - the elements of the Activity Theory theoretical framework, with each one

being given a name to be identified with. Each theme was given a unique colour. Names of themes were derived from terms that emerged from the data, were used by the participants or found in existing theory and literature in line with the guidance of Saunders et al. (2009), citing Strauss and Corbin (2008).

The next step in the process involved 'unitising' the data, or dividing it into units that consisted of anything from a number of words or a sentence to a paragraph or more. Each one of these units was attached to one or more of the identified themes, and the text was given the same colour as the theme. Where there was more than one theme per unit, both colours were used intermittently, with an attempt to match sub-sections to the colour of the theme matching the sub-section, though this was not always possible. Theme names (in the appropriate colour) were placed in the margins of the transcript next to the matching unit. Additional notes for the researcher, eg. type and purpose of a information request being spoken about, were also placed in the margin, with a curly brace indicating which part of the text they apply to. By way of illustration, an excerpt from one of the interviews is exhibited in appendix C.

After all the interviews had been transcribed a new document was created for each theme. The units that matched that theme in any of the interviews were copied and pasted into that theme's document, with care being taken to note which interview each unit was taken from, and which information request the unit applied to. If a unit referred to more than one theme it was copied into the documents of each of the different themes that it corresponded to.

Data thus prepared was then ready for structuring and in accordance with the Activity Theory Framework and subsequent interpretation. For this, the approach of Lee (2010) described in chapter 3 was used. First the discovered themes were sorted according to the elements of the activity system for an information request that they corresponded to, i.e. Subject, Tools, Object, Outcome, Community, Rules and Division of Labour. Once sorted it was found that tensions emerged, and these could be charted on the activity system and thus seen in context. Some of the themes corresponded to more than one part of the activity system – this gave greater insight into the issues involved. There were some themes that did not fit into the activity system, and these fell into two categories. The first comprised only the 'Use' theme, corresponding to all units discussing what the current uses of the NHLS CDW are. This theme was used in answering research question 1. The second category, with themes 'Full potential' and 'Data governance model', did not map onto any issues, challenges or outcomes of an information request, but rather gave a possible indication of further improvements to the CDW and surrounding services.

Each research question was addressed as follows:

Research question 1, "What are the current applications of the NHLS CDW?" was addressed through identifying applications mentioned either in the interview units falling under the 'Use' theme or in documents obtained during the preliminary exploration of the context.

Research question 2, "What are the issues and challenges in the use of the NHLS CDW", was explored by examining units relating to the activity system elements of: Subject, Tools, Object, Community, Rules, Division of labour. Contradictions between the requirements of these elements of the activity system began to reveal issues and challenges, as predicted by Engestrom (2001).

Research question 3, “What are the individual outcomes for the NHLS CDW user and data requestor?” were addressed through themes that are concerned with the ‘object’ and ‘outcomes’ elements of the activity system. In particular, responses to the first two probes for question 7 in the interview protocol helped to shed some light on this.

Finally, research question 4, “What are the perceived benefits (impacts) of NHLS CDW use for the NHLS as an institution and for population health in general?” was attended to through the ‘outcome’ element of the activity system. The last two probes for question 7 in the interview protocol were designed specifically to address this question, and resulted in the disclosure of interesting information.

Presentation of the results led to their interpretation and a comparison with the literature, after which conclusions were drawn and recommendations made.

4.7 Method Evaluation

In an interpretive case study of this nature it is important that attention is paid to the rigour with which the qualitative data collected from interviews and other sources is analysed and interpreted. Cognisance must be taken of various factors and social interactions that may influence the ultimate conclusions of the researcher. Klein and Myers (1999) recommend seven principles that should be complied with to ensure rigour in interpretive field studies like this one. These principles, together with the way in which they were applied to this research study, are given below.

1 – The Fundamental Principle of the Hermeneutic Circle
The notion of a hermeneutic circle argues that the way in which we understand a complex whole is through our preconceived understanding of its constituent parts and the relationships between them. Similarly, we only properly understand the meaning of each part once we have understood the context of the whole that it belongs to. Furthermore, the now understood whole may itself be part of a larger complex whole that the same process would apply to. It is thus contended that full understanding can only be attained through iterating between the contemplation of the interrelated parts and the contemplation of the whole. This principle is fundamental and is a meta-principle of the other six principles (Klein and Myers, 1999).

This principle was applied during the course of each interview where consideration of each new piece of information that came to light improved the understanding of the use that the participant was making of the NHLS CDW. Each interview in turn formed part of a group of interviews belonging to a particular section of the NHLS, and understanding each participant’s role and use of the CDW enabled the understanding of the purpose and status of the section. Finally, the understanding of the purpose and status of each section within the organisation led to an understanding of why issues and challenges of CDW use were being differently experienced by participants working in different sections.

2 – The Principle of Contextualisation
The subject of the research – in this case the information requestor making an information request – must be viewed within its social and historic context. Critical reflection of the social context and the

history that caused certain situations to come about will enable better understanding and interpretation of research results (Klein and Myers, 1999).

From the outset this research has been concerned with understanding the activity of a subject making an information request from the NHLS CDW within the social context of the organisation, its people and its rules. Activity Theory was specifically chosen as a research framework because it is ideally suited to studying a phenomenon within its context. Therefore substantial attention was paid to the situational context and this enabled a more thorough understanding of the research.

Knowledge of the historic context of the formation of the NHLS in 2001 and the NPP in 2010[20] did facilitate a better understanding of the current situation and the roles and purposes of the information requestors interviewed. Histories of some of the interviewees during their employment by the NHLS also explained why certain employees had less of a need for assistance than others. Most importantly, the historic context of population health, where it was made known, was extremely valuable in that comparison with the present illustrated ways in which use of the NHLS CDW has helped to impact on population health.

3 – The Principle of Interaction between Researchers and Subjects

Researchers' perceptions of participants influence the way in which qualitative data is interpreted. Concepts discussed in the interview process may also be appropriated by participants, thus influencing their perceptions and the things they may say. It is therefore important for the researcher to critically reflect on the influence their interaction with the subjects may have had (Klein and Myers, 1999).

In the analysis, the possibility that participants' replies were influenced by the way in which questions were asked or by something that had been said previously in the interview by the researcher was taken into account. Wherever documents or other secondary data could be obtained, triangulation with the interview data was done to verify conclusions drawn.

4 – The Principle of Abstraction and Generalisation

Once details discovered during the research process have been contextualised (applying principles 1 and 2) it may become possible to associate them with abstract concepts that are applicable in other situations – i.e. it becomes possible to generalise. Principle 4 requires that an attempt be made to relate discovered idiographic details to general concepts concerning human understanding and social action (Klein and Myers, 1999).

This research project was exploratory in nature and the creation of theoretical abstractions was not considered part of its scope, so this was not done. An example of a possible generalisation, however, is given here. Consider a participant's comment that they don't have all the data because the doctors don't fill in their forms properly. This comment should be viewed in the context of the way the information flows through the organisation, where forms filled in by DoH doctors are sent to the laboratories where they are captured into the lab information system and later sent to the CDW. From this we could make a generalisation that the quality of the data in data warehouse is

[20] Discussed in chapter 1

dependent on the extent to which the correct data was captured into the operational systems in the first place.

5 – The Principle of Dialogical Reasoning
There is a chance that data that comes to light during the course of the research will contradict preconceptions held by the researcher that the original research design was based on. The researcher should be sensitive to this possibility and be prepare for cycles of revision if needs be (Klein and Myers, 1999).

The researcher was sensitive to this and in fact a few revisions to the interview protocol were necessary after participants mentioned issues that were not part of the researcher's preconceived list of possible issues (derived from the literature). Appropriate questions were added to the protocol to discover whether other participants were experiencing the same or similar issues. Similarly, some preconceived issues did not turn out to be issues at all in the research context. These were left in the interview protocol just to ensure that they would covered if they did apply to later participants.

6 – The Principle of Multiple Interpretations
Different accounts of the same phenomenon from different participants are not necessarily contradictory if they are understood in terms of the context that each individual finds themselves in. It is therefore important for the researcher to be sensitive to ways in which contextual influences may be colouring the interpretation of a particular participant (Klein and Myers, 1999).

Disparate participant perceptions of aspects of obtaining data from the CDW were carefully considered and ultimately attributed to the different departmental contexts of each participant. Sensitivity to this possibility actually helped to determine the cause of apparently contradictory experiences with the CDW.

7 – The Principle of Suspicion
Interview participants may have preconceptions that could cause them to misrepresent a phenomenon. The researcher is required to be sensitive to any possible biases or distortions that could arise from preconceptions of this nature (Klein and Myers, 1999).

The researcher attempted to be sensitive to this. Once found, biases and distortions could be resolved by comparing responses of different participants and by comparing participant responses with any documents collected.

4.8 The Role of the Researcher

Interpretive research involves a complex human process whereby the researcher accesses other people's interpretations of events and then, inevitably influenced by their own preconceptions, filters and amalgamates these into a new version of events that is then reported on. To carry out this work successfully, it is important for the researcher to have an understanding of their own role

in the process (Walsham, 1995). Two of the principles discussed in the previous section, the "principle of interaction between researchers and subjects" and "the principle of dialogical reasoning" are highly relevant in this regard.

The researcher performs an important role throughout the different steps of the data collection and analysis process. In the interview, for example, bearing the "principle of interaction between researchers and subjects" in mind, the interviewer should portray an interested but neutral reaction to participant responses to avoid a lead that may encourage bias. Interviewers should also encourage the flow of conversation, and should be aware that their own level of enthusiasm will affect this (Saunders et al., 2009). The researcher of this study made an effort to project an enthusiastic but neutral demeanor during the interviews.

Interviewers need to have good listening skills, and must be prepared to spend the time to understand the message of the respondent. Where necessary, the interviewer should probe participant explanations with further questions to clarify meaning. Biased and incomplete interpretations can also be avoided by presenting the researcher's summarised version of what the participant is saying so that it can be acknowledged or corrected (Saunders et al., 2009). The latter approach was found to be exceptionally useful. Researcher interpretations were often confirmed with respondents, and sometimes they were adjusted in the process. Where there was any uncertainty in comprehending participant explanations, further probing questions were asked to elucidate. Since the interviews were semi-structured in nature it was not necessary to stick to the protocol strictly, and sometimes it was necessary to veer off track for a short while in order to explore a different angle.

The influence of researcher perspectives extends to data analysis and reporting. Personal knowledge and experience gained through five years of working with database and decision support in the industry, as well as through ten years of teaching about Business Intelligence and Data Warehousing, was used to make sense of the material when unitising the data and assigning the units to themes.

In reporting the results, it is the researcher who decides which of the participants are given voice and to what extent, and what issues are given prominence. Researcher interpretations reported on will also be influenced by prior knowledge and perceptions. Most importantly, final conclusions of the study are dependent on researcher interpretation of the data (Fink, 2000). All of these considerations had to be taken into account in the analysis and reporting of the results. Although it is never entirely possible to avoid bias, an attempt was made to ensure that the research output was as unprejudiced as possible.

4.9 Reliability and Validity

The reliability and validity of qualitative research is assessed through the attributes of trustworthiness and authenticity. Trustworthiness is achieved through methods that establish the credibility of the research results (Johnson, 2004). Assuring credibility refers to the conscious effort to establish confidence in an accurate interpretation of the meaning of the data (Whittemore et al.,

2001). Research findings will be supported if evidence from multiple data collection methods is found to converge. (Benbasat et al., 1987) An effort was therefore made to triangulate findings from the interviews with those from documents collected during the preliminary exploration of the case study context (Johnson, 2004), and the documents were found to substantiate interviewee data.

According to Johnson (2004), authenticity concerns the different realities of the individuals represented in the case study. Citing Denzin and Lincoln (1998), Johnson (2004) lists the criteria for establishing authenticity as:

- *Fairness*: the degree to which the research represents the different viewpoints of the participants
- *Ontological authenticity:* The degree to which the research helps members of a social setting to better understand their own environment
- *Educative authenticity:* The degree to which the research helps members to understand the perspectives of other members
- *Catalytic authenticity:* The degree to which the research acts as an impetus for social action
- *Tactical authenticity:* The degree to which this research actually empowers action.

Authenticity has also been described by Walsham (2006) as "the ability of the text to show that the authors have 'been there', by conveying the vitality of life in the field" (pg. 326).

The above can be supplemented with the view of that in interpretive studies, researchers should be sensitive to any biases and distortions in participant data, multiple narratives and contradictions, and how the researcher's interaction with the participants influences the results.

Care was taken to represent the viewpoints of all the participants to ensure fairness, and the report was written with the intention of enabling people involved with the NHLS CDW to better understand their environment and thus establish ontological authenticity. Educative authenticity will result from the views of multiple individuals being expressed. With reference to catalytic and tactical authenticity, although the primary intention of this study is to explore the uses, issues, challenges and impacts of the NHLS CDW, and the underlying research philosophy is not critical, none-the-less there may be an opportunity for the results to create impetus for change and empower action.

4.10 Ethical Clearance

Ethical clearance was required for this study and was applied for from the Human Research Ethics Committee (HREC) (medical) at the University of the Witwatersrand, Johannesburg on 7th June 2017. A participation information sheet as well as two consent forms needed to be attached to this application.

Appendix D outlines the participant information sheet which highlights the ethical considerations of informed consent, anonymity and confidentiality, voluntary participation and the right of withdrawal with risk or loss of benefit. The participant information sheet describes the purpose of the interviews and provides for users to give their informed consent prior to participation using the form shown in appendix E. Separate consent was required for audio-taping and the form for this is in appendix F.

Notification of conditional ethics approval came on 31st July 2017, requiring that permission for the research be granted by the NHLS before full ethics clearance could be granted. This permission, shown in appendix G, was applied for in August and granted on 3rd October 2017. Full ethics clearance was then granted by the HREC (Medical) on 12th October 2017. The clearance certificate, with protocol number M170642, is shown in appendix H.

4.11 General limitations

Because this is a case study, generalisability is not a major objective. Instead, this study is done on a specific implementation of data warehousing at a specific organisation. Results are thus not intended to be generalisable to other situations. However there may be useful lessons for others, as suggested above.

A possible limitation of this research is due to the fact that data collected for this study was obtained by interviewing ten people who request information from the NHLS CDW. The NHLS is a very large organisation and there are many more people that request information than were interviewed. If a different subset of ten information requestors were interviewed the results may have been different. It may also be possible that each interviewee did not mention every request for information that they make.

A further and most important limitation is that bias could be introduced since the interpretation of the data depends on the researcher's perceptions, as well as on their ability and experience in interpreting qualitative data. This could be the case particularly since the researcher is the only person coding the data and there will be no opportunity to formally determine inter-coder reliability.

4.12 Chapter Conclusion

In this chapter the layers of Saunders et al. (2009)'s research 'onion' have been 'peeled' away as one by one the research philosophy, approach and strategy have been discussed, the approach towards and process of data collection have been reported on, the development of the interview protocol and the interviewing of the participants have been recounted, and the means of data analysis described (paying attention to ensuring rigour by applying method evaluation). In addition, this chapter then continued with the elucidation of the role of the researcher, a description of the procedure of obtaining ethics clearance, and a list of study limitations.

The next chapter will present the results of the study. To begin with, all data collected – in the form of interviews and documents – will be profiled. Themes discovered in data analysis will then be listed and mapped onto the activity system for an information request. Finally, results pertaining to each of the four research questions will be represented in turn, with attention focused on applicable themes one by one.

5 Research Findings

The purpose of this study was to explore the use and impact of the NHLS CDW, as well as any issues and challenges experienced in its use. This chapter presents the findings of the research. After profiling all interviews and documents s collected, the themes that were found during data analysis will be listed and described. Themes that fit in will then be mapped onto the activity system for an information request. The bulk of the chapter will then be spent discussing the results pertaining to each of the four research questions in turn.

5.1 Profile of Participants and Information Requests

The majority of data collection was done through semi-structured interviews with NHLS staff who request information from the CDW. One preliminary session was done with the CDW staff and ten interviews were done with information requestors in four different parts of the NHLS. A breakdown showing how long each interviewee has been doing this job is shown in table 7. As can be seen, all respondents have been making information requests for at least a year, with the majority having done so for between two and ten years. Respondents were therefore well qualified to discuss the information requests being made.

Years Performing Role	Number of Interviewees
Approximately 1 year	1
Approximately 2 years	3
Approximately 3 years	4
Approximately 6 years	1
Approximately 10 years	1

Table 7: Interviewees by Length of Time Making Information Requests

Table 8 indicates the breakdown of interviewees according to job function.

Job Role	Number of Interviewees
Epidemiologist	2
Data Analyst	3
Head of Division	1
Area Manager	1
Program Manager	3

Table 8: Interviewees by Job Function

As a result of the purposive sampling, there were a variety of job functions, which implied that there would be a good spread of different types of information requests being made, from both a request type and a request purpose point of view. It can be seen in the final sampling matrix shown as figure 4 in chapter 4 (shown again here) that this was found to be the case.

	Medical Research	Administrative	Public Health Tracking
Customised Application	**0**	**6**	**8**
Service Representative	**4**	**3**	**2**
Data Analyst	**0**	**1**	**3**
High level CDW Access Tool	**0**	**0**	**0**
Low level CDW Access Tool	**2**	**2**	**2**

Figure 16: Information Requests Sampled

This sampling matrix provided for fifteen possible categories of information requests. As discussed in chapter 4, four of those categories would probably never be filled – one because it does not make sense to create a customised application for medical research as the data requested will be different for each research project, and the other three just because that pattern of use, i.e accessing the NHLS CDW directly using a high level CDW access tool, does not seem to exist in the CDW context. Of the eleven remaining categories, the thirty three information requests were spread across ten of them. The data collected from the interviews therefore covered a good variety of different types of information requests and was expected to give a reasonably broad picture of the use and impact of the CDW, and of the issues and challenges that are being encountered.

5.2 Profile of Documents Collected

Various documents collected are summarised in table 9 on the 67. They form an alternative source of data that was used, where relevant, to triangulate what was said by the interviewees. Since interviewees were sometimes reluctant to provide internal documents without being granted permission first, many of the documents were sought out in the public domain. Interviewees themselves did direct the researcher to some of these public sources.

5.3 Themes Discovered and the Activity System for an Information Request

A total of fifteen themes were uncovered during the course of the data collection. Of these themes ten were expected at the outset, and five emerged as research progressed.

Type	Document	Source
Slide Presentation	Data Warehouse Overview Beyond Implementation: Health systems strengthening and focus on high risk populations: 2011-2014 GeneXpert Implementation in South Africa Public Sector	CDW Staff http://www.stoptb.org http://www.stoptb.org
Dashboard Report	Cryptococcal Antigen Trends	Interviewee
NHLS Form	Information Request Form	Interviewee
Web Site	National Health Laboratory Services (NHLS) National Institute for Communicable Diseases (NICD) University of the Witwatersrand (WITS) – section on Research and Analytical Services in Molecular Medicine and Haematology National Department of Health (NDoH)	http://www.nhls.ac.za http://www.nicd.ac.za https://www.wits.ac.za/pathology/divisions/molecular-medicine--haematology/research-and-analytical-services/ http://www.health.gov.za
Article	The South African National Cancer Registry: an update Establishment of a cancer surveillance programme: the South African experience	(Singh et al., 2013) (Singh et al., 2015)
Report	NHLS Annual Reports NICD Review Early infant diagnosis of HIV infection in South Africa: 2008 to 2010	http://www.nhls.ac.za/?page=annual_report&id=45 http://www.nhls.ac.za/?page=nicd_review&id=81 (Sherman and Lilian, 2011)
Population Health Policy Document	National Health Act, 2003 (Act no. 61 of 2003): Regulations relating to the Surveillance and the Control of Notifiable Medical Conditions Breast Cancer Prevention and Control Policy Cervical Cancer Prevention and Control Policy	http://www.nicd.ac.za/index.php/nmc/nmc-regulations/ http://www.health.gov.za/index.php/2014-03-17-09-09-38/policies-and-guidelines/category/327-2017po
Medical Research Article	The epidemiology of meningitis among adults in a South African province with a high HIV prevalence, 2009-2012 Effects of introducing Xpert MTB/RIF test on multi-drug resistant tuberculosis diagnosis in KwaZulu-Natal South Africa Laboratory information system data demonstrate successful implementation of the prevention of mother-to-child transmission programme in South Africa: prevention of mother-to-child transmission-Progress towards the Millennium Development Goals Toward elimination of mother–to–child transmission of HIV in South Africa: how best to monitor early infant infections within the Prevention of Mother–to–Child Transmission Program Missed diagnostic opportunities within South Africa's early infant diagnosis program, 2010–2015	(Britz et al., 2016) (Dlamini-Mvelase et al., 2014) (Sherman et al., 2014a) (Sherman et al., 2017) (Mazanderani et al., 2017)

Table 9: Documents Collected

Expected themes were ‘Use’, ‘Outcome’, ‘Relationship with CDW Team’, ‘Ease and Efficiency’, ‘Data Quality’, ‘Privacy and Security’, ‘Direct Use’, ‘Outsourcing’, ‘Data Governance Model’, and ‘Full Potential’.

The themes named ‘Use’ and ‘Outcome’ map directly onto the problem being solved by the study – the title of the study contains the terms ‘use‘ and ‘impact’ (which would be an outcome), so their emergence as recurrent themes comes as no surprise. Applications mentioned under the ‘Use’ theme will be used to answer research question 1, “What are the current applications of the NHLS CDW?”. Information falling under the ‘Outcomes’ theme will be used to answer research question 3, “What are the individual outcomes for the NHLS CDW user and data requestor?”, and research question 4, “What are the perceived benefits (impacts) of the NHLS CDW use for the NHLS as an institution and for population health in general?”. The ‘Outcome’ theme also matches directly with the element of the same name of the Activity Theory triangle that was initially intended to address research questions 3 and 4.

The remaining eight expected themes were revealed in the literature as having been encountered in other applications of BI for population health, and were thus probed for as part of the interview protocol.

Three additional themes that emerged from the first three interviews were ‘Patient-oriented View’, ‘External Data’ and ‘Metadata’. These themes were not *a priori* identified in the literature but given that they were suspected to be important issues, they were added to the interview protocol so that the perspectives of other interviewees with respect to these themes could be heard.

A further two themes emerged from the interviews. Although these were not explicitly probed, they arose frequently as part of answers to other questions. They are ‘Staffing’ and ‘Time Delays’. Thus there were a total of fifteen themes.

As mentioned previously, themes were colour coded in order to expedite the process of unitisation and thematic analysis. The labels for the fifteen identified themes, in their chosen colours, with their corresponding descriptions and sources are shown in table 10 on page 69.

Based on data collected, twelve of the themes were mapped on to the activity system for an information request. One of the themes, ‘Data Quality’, was mapped twice, and another theme, ‘Time Delays’, was mapped three times. Justification for each one of these mappings will be explained during presentation of the results for each theme in section 5.5. The mapping of themes to the different parts of the activity system is illustrated in figure 17 on page 70.

Of the remaining three themes, the ‘Use’ theme provided an answer to research question 1, and the ‘Full Potential’ and ‘Data Governance Model’ themes provided some input for recommendations at the end of this report.

The next section addresses Research Question 1.

Theme Name	Theme Description	Source of Theme
Use	Use or application of the NHLS CDW as indicated by the reason for making an information request.	RQ1
Relationship with CDW Team	Information requestors' experience of co-operation with the CDW team in obtaining the required data, and adequacy of available assistance and training.	Literature, eg. Rizi and Roudsari (2013)
Staffing	Sufficient staff with sufficient expertise to maintain the CDW and assist in obtaining information from the CDW.	Interview
Time Delays	Time lag between requesting the data from the CDW and receiving the data as it is required.	Interview
Outsourcing	Requirement of services of an outside company to assist in getting the correct information from the CDW.	Literature, eg. Wisniewski et al. (2003)
Data Quality	Extent to which the information requestor considers the data received from the CDW to be correct, complete, timely and relevant.	Literature, eg. Wisniewski et al. (2003)
Patient-oriented View	Importance of the ability to obtain a full picture per patient with a history of all laboratory tests done for that patient; and extent to which this is possible.	Interview
Privacy and Security	Level to which data is protected due to privacy and security concerns, and effect of this on the process of making information requests.	Literature, eg. Berndt et al. (2000)
Direct Access	Direct use of CDW data without having to go through a service representative or customised application, perhaps through high level CDW access tools like data warehouse browsers or OLAP tools.	Literature.eg. Tremblay et al. (2007)
Metadata	Facility often provided with data warehouses, where information about the data kept in the warehouse is maintained and can be queried.	Interview
Ease and Efficiency	Extent to which information requestor feels that it is easy and efficient to obtain the requested data	Literature, eg. Berndt et al. (2003)
External Data	Need for using data outside of the NHLS CDW to satisfy the information requestor's needs.	Interview
Outcome	Ways in which the information requests affect decision making and the resulting impact on the information requestor, the NHLS and population health in general.	RQ3 and RQ4
Full Potential	Extent to which interviewees feel that the full potential of the CDW is being realised, and suggestions for exploiting unused potential.	Literature, eg. Wisniewski et al. (2003)
Data Governance Model	Investigation of interviewee awareness of a data governance model for the CDW.	Literature, eg. Rizi and Roudsari (2013)

Table 10: Colour Coded Themes

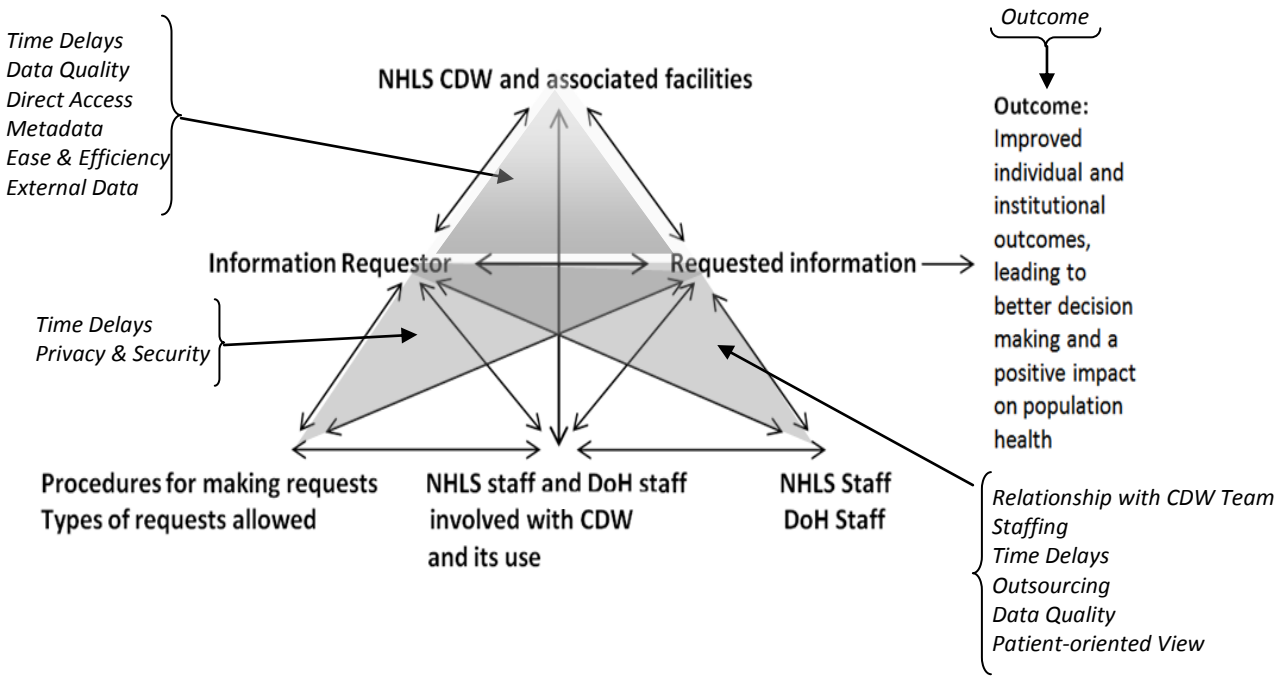


Figure 17: Allocation of Themes to Activity System for an Information Request

5.4 RQ1: What are the current applications of the NHLS CDW?

The current applications of the NHLS CDW were found to be many and diverse. An idea of the scope of the functions of the CDW can be seen in figure 18.

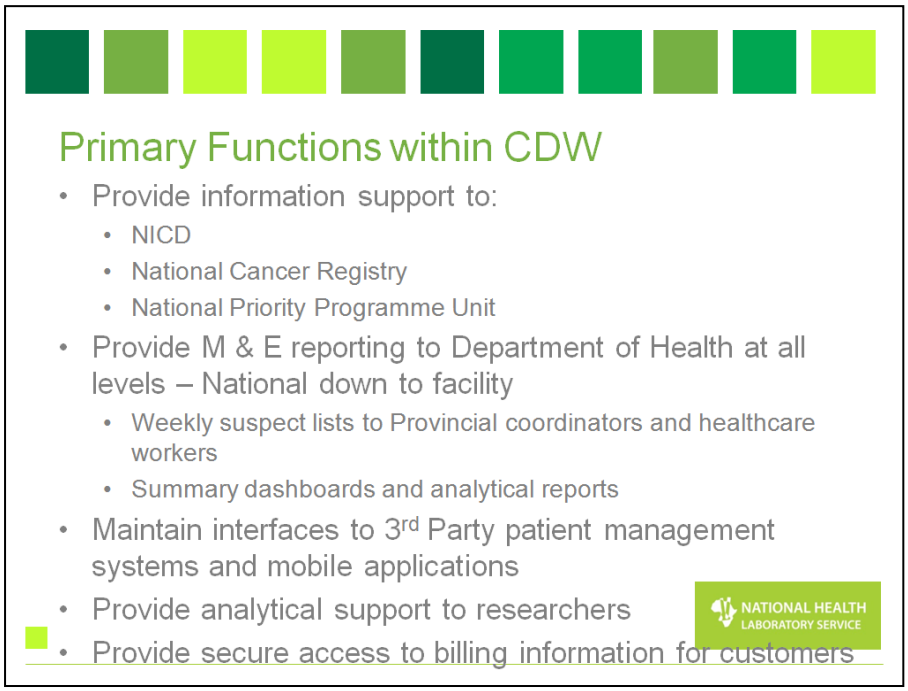


Figure 18: Primary Functions of the NHLS CDW (Source: National Health Laboratory Service Presentation 'Data Warehouse Overview')

According to this list of functions, the CDW provides information for public health tracking to NHLS divisions as well as to the DoH; it provides secure access to administrative, or billing, information; and it provides analytical support for researchers. In addition, although beyond the scope of this project, the CDW is said to be helping to support health care with its interfaces to third party patient management systems and mobile applications.

Most of these functions were encountered during the course of the research, either as sampled information requests, or as anecdotes or stories told in the interviews and preliminary session.

Applications of the CDW were indicated by information requests discussed in interviews with staff from the National Institute of Communicable Diseases (NICD), the National Cancer Registry (NCR), operational management and the National Priority Program (NPP). These applications were found to deal with research, administration and public health tracking and therefore matched the intended sampling categories for information requests, and the corresponding information requests are summarised in accordance with these categories in figure 19 on page 72. The NHLS division to which the request applies is indicated in square brackets after each request.[21] These applications are discussed below, after first describing general applications of the CDW unrelated to specific information requests.

5.4.1 General Applications of the CDW

In addition to the CDW applications indicated as information requests, other applications discovered during the course of the research were identified. These were found to be concerned with the DoH, the NGOs working with the DoH and NHLS billing.

Members of DoH staff were not interviewed, but the fact that NHLS CDW data was used and relied on by the DoH was mentioned several times. The following comments gave an example of the type of applications DoH is using the data for.

"even the Department also do request the information. .. . . . some of the reports we send them to the department as well, but they can request as well, when they want for purposes of operations, . . . or they want to see, for example when they look at the babies, how many of the babies' results are less than 200 . . . for them to be able to check if their programs are working or what, or patients are complying . . ."

" . . . this HIV testing and promotion of people to go and test for to know their status. Ya, so as they test they are using our labs so that information end up with us, so the Government will, at some stage wants to know how many people are being involved in these campaigns. Is it increasing, decreasing, those type of things."

[21] NICD: National Institute of Communicable Diseases
NCR: National Cancer Registry
O: Operational Management
NPP: National Priority Program

	Medical Research	Administrative	Public Health Tracking
Customised Application	*(unlikely that customised application would be developed for research as each research project requires different data)*	Monitor laboratory & regional performance through turnaround[0] Prevent wastage by monitoring rational use of laboratory tests[0] Monitor volumes of laboratory tests done[0] Monitor rejections due to incorrect or insufficient samples (by area manager and program manager, so 2 requests)[0][NPP] Monitor performance of laboratories at correctional services[NPP]	Track people with CD4 less than 100 for Crypto-Meningitis[NICD] Track anti-microbial resistance (by 2 epidemiologists, so 2 requests)[NICD] 'Continuous flow' of data from CDW to NCR for cancer tracking[NCR] Track HIV status of babies of HIV positive mothers[NPP] Monitor missed diagnostic opportunities[NPP] Track response to ARVs in correctional facilities[NPP] Track TB positivity and drug resistance in correctional facilities[NPP]
Service Representative	Request information required for Mycology research[NICD] Request Meningitis tests data for a research project[NICD] Request data for SA HIV Cancer Match Study[NCR] Request data for other research on cancer[NCR]	Get more detail on monthly report data to find problem[0] Request projections for planning next year's budget[0] Pinpoint slow episodes at correctional services laboratories[NPP]	Get baseline data on normal disease incidence to verify outbreak[NICD] Investigate disease outbreaks[NICD]
Data Analyst	*(This does occur, but was not included in the sample)*	Monitor performance of laboratories serving correctional services[NPP]	Request EID information for specific time period for trend analysis[NPP] Request viral load test results for women of childbearing age[NPP] Monitor TB positivity and drug resistance rates for GeneXpert tests[NPP]
High level CDW Access Tool	*(High level access tools like data warehouse browsers and OLAP were not found to be used directly against the CDW to satisfy information requests)*	*(High level access tools like data warehouse browsers and OLAP were not found to be used directly against the CDW to satisfy information requests)*	*(High level access tools like data warehouse browsers and OLAP were not found to be used directly against the CDW to satisfy information requests)*
Low level CDW Access Tool	Request data for medical research done by NPP team (by 2 Data Analysts, so seen as 2 requests sampled)[NPP]	Request data for admistration as required by NPP team (by 2 Data Analysts, so seen as 2 requests sampled)[NPP]	Request data for public health tracking as required by NPP team (by 2 Data Analysts, so seen as 2 requests sampled)[NPP]

Figure 19: Information Requests Categorised by Sampling Categories

" . . . the decentralisation of TB. TB drug resistance. Initially in government you would have one place that deal with drug resistant patients. Now they have decentralised the whole system into clinics and the likes, so people could get . . . access to treatment. So now government wants to see the impact" (This particular project was undertaken by a private company working on behalf of the DoH)

NHLS CDW is also responsible for fulfilling an exceptionally important requirement of government, concerning notifiable diseases (mentioned previously in chapter 3). This requirement is stipulated in one of the documents collected for the research, the "National Health Act, 2003 (Act no. 61 of 2003): Regulations relating to the Surveillance and the Control of Notifiable Medical Conditions". DoH must receive results testing positive for anything designated as a category 1 notifiable medical condition within 24 hours. This information may be used to detect possible outbreaks or to determine which patients in hospitals should be isolated. As discussed in chapter 3, a right-time infrastructure has been built by CDW to facilitate this, and DoH receives the results well within the prescribed 24 hour window.

To alleviate the shortage of health care services in rural areas, and sometimes to provide specialised services, various NGOs operate mobile clinics in certain regions. Figure 20 displays one such clinic. Results for tests requested by these clinics can be transmitted back to the mobile clinics using the CDW right time infrastructure.



Figure 20: Mobile Clinic

CDW staff mentioned that "NGOs that are supporting government programs", like Right to Care, Aurum and others, request data from them on a regular basis, but they did not give specific examples.

Staff in the finance department of the NHLS could not be interviewed as their contact details were not provided, but mention was made in the preliminary session of secure access to billing information that is available through the CDW:

"The billing – we also do financial, but the billing part for all the tests that happen . . . We have a BI tool where the DOH can go and check their accounts and their invoices and their CSVs – a

breakdown of that reports by transaction. . . .you can only see your own report, your own account, so you can't see other provinces. . . like, for Western Cape we send them billing for those who request billing for the whole of Western Cape, . . . but then certain individuals, even if they are in Western Cape they can only see specific accounts / facilities, and they can't see all of Western Cape."

Mention was also made in the preliminary session of using the CDW to do exploratory studies:

" . . . some people when they do exploratory studies, they say 'Ok. We want patients at this place who have been treated for these conditions, from this period onwards.' . . . we will give them the results . . . and on the other side we are busy profiling whether these people could be part of the subject of their studies or not. For example the other one that we recently did was kids who are exposed to people who have TB, so it's like now we've got results of parents who were diagnosed with TB . . . they are looking to see data to see whether these people did they infect their children"

In summary, the general applications of the CDW involved helping government track the impact of population health programs, alerting government to the incidence of notifiable diseases, sending timely test results to mobile clinics, NHLS billing enquiries and exploratory studies that determine whether data can be used for research.

5.4.2 Applications of the CDW for Information Requests

Having concluded discussion of general applications of the CDW, applications that were identified as information requests are explored in this section. Results pertaining to each of the four divisions explored will be presented in turn.

It is appropriate at this point to reiterate that there was not always a one-to-one relationship between an 'information request' identified as a finding here and an actual physical instantiation of the request for information like a report, a dashboard, or an Excel spreadsheet. Rather an 'information request', and consequently what is regarded as an application, is associated with a purpose or a function being achieved. The decision to do this was made by the researcher for two reasons. Firstly, sometimes it was not clear whether information to satisfy several functions was received with the results of a single physical request or not, and secondly, sometimes more than one physical request result was used to achieve a single purpose.

Where the same application is used by more than one person it is regarded as one information request per person, as issues and challenges experienced may vary from person to person. Different people may also have different uses for similar applications.

5.4.2.1 Applications of the CDW discovered at the NICD

The NICD is a centre for disease and surveillance control and provides information and expertise on outbreak and communicable disease burdens to southern African regions. The NICD also undertakes

research on communicable diseases in Southern Africa.[22] Applications corresponding to information requests made by interviewees from the NICD were found for both public health tracking and research.

5.4.2.1.1 NICD Applications of the CDW for Public Health Tracking

Use of the CDW to track people with CD4 less than 100 for Cryptococcus Meningitis
This application has its roots in the NPP, which deals with HIV/AIDS and TB. Cryptococcus Meningitis results from a fungal infection, however, which is why it is being attended to by the Mycology department within the NICD.[23] HIV/AIDS patients with a CD4 count of less than 100 were found to be susceptible to Cryptococcus Meningitis, which can be deadly if it is not treated in time. Patients reporting to local clinics presented with common symptoms like a headache and were often sent home with inappropriate medication while the disease progressed and was only finally correctly diagnosed when it was too late, and with the patient having suffered unnecessarily. This therefore impacted on morbidity and mortality, and placed an unwelcome load on the health services. So, as said by the interviewee,

> "*an agreement has been reached with NDoH and again with WHO that crypto should fall within the casket of care of HIV and TB. So . . . you need data that will enables you to understand how patients are managed within that category, which is patients who could be falling within stages of AIDS defined diseases, because they are the ones that are at risk of getting crypto meningitis. So . . . we normally give a threshold of CD4 less than 100 in our data output*"

A policy decision has been made for the laboratories to automatically test for Cryptococcus antigens for any patients with a CD4 count of less than 100. Patients whose results are positive can then be called back for further testing and care. This whole process is managed and monitored using information from the CDW, and in particular several dashboard reports have been made available for this purpose.

Use of the CDW to track anti-microbial resistance (done by 2 epidemiologists for different purposes)
This is a form of surveillance where the extent of anti-microbial resistance is tracked.
One epidemiologist interviewed uses dashboards that contain snapshots of anti-microbial information that are updated every four months to produce required national reports, and also assists in the implementation of the Global Antimicrobial Surveillance System (GLASS) by submitting data to the WHO.

A more senior epidemiologist gets a data extract on antimicrobial resistance every week and uses it to create a report to evaluate anti-biotic resistance.

Use of the CDW to get baseline data on normal disease incidence to verify outbreak
Outbreaks can be reported to the NICD through any health facility or doctor. It is then incumbent on NICD staff to determine whether the reported episode is really an outbreak. This is done by focusing

[22] http://www.nhls.ac.za/?page=national_institute_for_communicable_diseases&id=36
[23] Mycology is a branch of biology that deals with the study of fungi

on the particular facility and pathogen concerned, and requesting baseline data from the CDW on what the normal incidence of that disease is in that place.

Use of the CDW to investigate disease outbreaks
Once an outbreak has been confirmed, CDW data is requested as needed and is used to investigate the outbreak further.

5.4.2.1.2 NICD Applications of the CDW for Research

Use of the CDW to request information for Mycology research
The Mycology department of the NICD employs a data analyst who requests and manages CDW data required for various research projects in Mycology.

Use of the CDW to request Meningitis tests data for a research project
An epidemiologist working in a different department of the NICD required CDW data for her research project into Meningitis.

5.4.2.2 Applications of the CDW discovered at the NCR

The NCR was first established in 1986. It is a surveillance unit that collates, analyses and records all newly reported cancer cases, thus contributing significantly to the awareness of the growing problem of cancer in South Africa.[24] The NCR runs both a pathology based registry and a population based registry, and most of the data in the pathology based registry is derived from the CDW. Applications corresponding to information requests made by interviewees from the NCR were found for both public health tracking and research.

5.4.2.2.1 NCR Applications of the CDW for Public Health Tracking

Use of the CDW to provide a 'Continuous flow' of data from CDW to NCR for cancer tracking

An application has been created which facilitates data being extracted from the CDW and loaded into the NCR application on a continuous basis. The CDW also receives data from private laboratories, which it then passes on to the NCR, so that a more holistic picture of the cancer situation in South Africa can emerge. As said by an interviewee,

> *". . . since 2011 cancer is a reportable condition, so the NHLS has an obligation to report all cancers diagnosed by histology[25] and cytology[26] and haematology[27] to the NCR. It's a regulation.*

[24] http://www.nhls.ac.za/?page=national_cancer_registry&id=38
[25] Histology is a department of anatomy dealing with the microscopic structure of body tissues
[26] Cytology is a branch of medicine concerned with the structure and function of body cells
[27] Haematology is a branch of medicine involving the study and treatment of blood

The NCR relies on laboratory test data from the CDW in order to track all new incidences of cancer."

5.4.2.2.2 NCR Applications of the CDW for Research

Use of the CDW to request data for SA HIV Cancer Match Study

This current NCR research project is best described by the interviewee,

"With our HIV Cancer Match study we are creating a virtual cohort of HIV positive patients for research purposes – not just cancer related research, for any research purposes. It will be the largest virtual HIV cohort in the world, because we have the biggest burden of HIV. It is the first time anybody's doing this with CDW data"

Use of the CDW to request data for other research on cancer

The NCR runs other cancer related research projects that rely on data requested from the CDW.

5.4.2.3 Applications of the CDW discovered from Operational Management

With approximately 7000 members of staff and over 260 laboratories[28] in all 9 provinces of South Africa, managing the operations of the NHLS is a complex task. Data from the CDW is used to improve both the efficiency and effectiveness of this undertaking. Several administrative applications corresponding to information requests made by an operational area manager were found.

5.4.2.3.1 Operational Management Applications of the CDW for Administration

Use of the CDW to monitor laboratory & regional performance through turnaround
The NHLS has specified targets turnaround times for the providing of results of the different tests done. Data from the CDW, in the form of periodically emailed extracts, is used to monitor whether laboratories or regions are meeting the specific targets, so that action can be taken if they are not.

Use of the CDW to prevent wastage by monitoring rational use of laboratory tests
Previous experience where the results of requested tests were never used led to an agreement between the DoH and the NHLS on the rational use of laboratory tests. As a result there are rules governing whether certain tests are restricted, or whether repeats of the same tests will be allowed. Management uses periodically emailed data from the CDW to monitor tests rejected as a result.

Use of the CDW to monitor rejections due to incorrect or insufficient samples
The laboratories can only perform the required tests on samples if those samples are sufficient in certain respects. For example, for blood tests, the correct tube must be used and the volume of

[28] http://www.nhls.ac.za/

blood must be correct. Tests also cannot be done if there is something wrong with the blood. Using periodically emailed CDW data to monitor the rate of rejection of tests for these reasons enables management to detect problems and take mitigating action. After identifying the public health care facility responsible for submitting the sample, they can be engaged with and educated about how to submit sufficient samples of good quality.

Use of the CDW to monitor volumes of laboratory tests done
Monitoring the number of laboratory tests done enables management to check whether they are functioning at the expected capacity. This information is emailed periodically from CDW, and reasons for an increase or decrease in test volumes can subsequently be investigated.

Use of the CDW to get more detail on monthly report data to find problem
In the event of management being dissatisfied with the performance indicated on one of the periodical reports mentioned above, more detailed data about the issue in question is requested from CDW. This more detailed data can then be interrogated to isolate what and where the actual problem is, so that solutions can be ascertained and actioned.

Use of the CDW to request projections for planning next year's budget
When planning the budget for the following year, management are empowered to make better, fact-based decisions by using pertinent information requested from the CDW.

5.4.2.4 Applications of the CDW discovered at the NPP

The NPP is a division of the NHLS that was created in 2010 to address the extremely high incidence and burden of disease of both HIV/AIDS and Tuberculosis (TB) in South Africa. In addition to their initial mandate to provide affordable, accessible, accurate and efficient HIV and TB diagnostic services , they offer active facilitation of and assistance with national programs like early infant HIV diagnosis, HIV drug resistance, TB GeneXpert testing, and support for Correctional Services and for mining and peri-mining communities. Some of the programs receive funding from several sources, including the Global Fund, and there is collaboration with many organisations like the World Health Organisation (WHO), the US Centre for Disease Control (CDC), the Gates Foundation, the Clinton Foundation and others.[29]

Interviews were held with program managers of three of the NPP programs: early infant diagnosis, GeneXpert testing and support for correctional services.

The aim of the early infant diagnosis (EID) program is to attempt to eliminate the transmission of HIV from mother to child. To this end, early infant testing for HIV infection has been instituted so that HIV-infected infants can be identified and cared for, and factors contributing to transmission can be discovered and attended to.

Older TB tests were replaced with the newer GeneXpert platform, which is a faster and more accurate means of testing for TB, and which can test for TB and TB drug-resistance at the same time.

[29] https://www.wits.ac.za/pathology/divisions/molecular-medicine--haematology/research-and-analytical-services/national-priority-programme-npp/

The GeneXpert program involves monitoring the performance of and results from GeneXpert platforms at 203 sites around the country.

The correctional services program was created to address the substantially high level of TB in prisons as compared to that of the civilian population. Inmates are tested regularly and given the proper care, and positivity rates and response to treatment are monitored.

Applications corresponding to information requests made by interviewees from these NPP programs were found for public health tracking, administration and research.

5.4.2.4.1 NPP Applications of the CDW for Public Health Tracking

Use of the CDW to track HIV status of babies of HIV positive mothers
The Polymerase Chain Reaction (PCR) test is used to detect whether babies have been infected with HIV. Regular emails of these test results from CDW are used to track the status of babies with HIV positive mothers, so that the correct training, intervention and care can be provided.

Use of the CDW to monitor missed diagnostic opportunities
This regular email from CDW provides lists of PCR tests that were rejected by NHLS laboratories due to problems with the specimen submitted. This information is used to view statistics through a more realistic lens. For example, if 100 tests were submitted and there were 5 positive results, that doesn't amount to 5% if 20 of the tests were rejected. Rather 5 out of 80 tested positive, which is 6.25%.

Use of the CDW to track response to ARVs in correctional facilities
Part of the NPP work in the correctional facilities includes HIV programs, where CD4 counts and viral loads of inmates are monitored. Regularly emailed data extracts from CDW showing these test results give the program manager an idea of the response of inmates to anti-retroviral (ARV) therapy.

Use of the CDW to track TB positivity and drug resistance in correctional facilities
Regularly emailed data extracts from CDW showing GeneXpert test results facilitate the tracking of TB and TB drug resistance in correctional facilities.

Use of the CDW to request EID information for specific time period for trend analysis
When EID information needed for training, or for provincial, district or clinic meetings, does not match the information received by email on a regular basis, this information is specifically requested from CDW. The same type of information will be requested, but for different time periods, as per the requirements of the training or meeting.

Use of the CDW to request viral load test results for women of childbearing age
The EID program deals predominantly with the baby, but the viral load of the mother is also important as if it is not kept low enough HIV may be transmitted to the baby. Checking viral loads of women of childbearing age gives an indication of whether the mothers are being managed correctly, which consequently would affect the incidence of HIV infection.

Use of the CDW to monitor TB positivity and drug resistance rates for GeneXpert
The GeneXpert program manager requests GeneXpert test results from CDW through a data analyst. This data enables the monitoring of TB positivity and drug resistance rates for those tested on this platform.

5.4.2.4.2 NPP Applications of the CDW for Administration

Use of the CDW to monitor rejections due to incorrect or insufficient samples
This application is also used by operational management and is explained above. In this case it is used in connection with laboratories that serve correctional facilities, and once again can result in positive engagement and training to submit the correct specimens.

Use of the CDW to monitor performance of laboratories ***at*** correctional services
Seven of the correctional facilities have been kitted out with GeneXpert testing equipment. Regular information received from CDW by email allows for the monitoring of the performance of these specific laboratories in terms of the number of tests done and test turnaround time.

Use of the CDW to pinpoint slow episodes at correctional services laboratories
When a laboratory is showing a slow turnaround time, additional information is requested to pinpoint which testing episodes are causing the problem, so that they can be addressed.

Use of the CDW to monitor performance of laboratories ***serving*** correctional services
Not all correctional services have their own GeneXpert testing equipment. Those that do not are served by NHLS laboratories nearby. The program manager gets information for all laboratories serving correctional facilities from the data analysts. Again, this information allows for the monitoring of the performance of these laboratories in terms of the number of tests done and test turnaround time.

5.4.2.5 NPP Applications of the CDW for Research

Use of the CDW to request data for medical research done by NPP team (done by data analysts)
Data analysts in the NPP division spoke about providing CDW data to researchers working in national priority areas. Unfortunately it was not found to be possible to interview any of these researchers.

5.4.2.6 RQ1 Summary

Taken together, the various uses of the CDW can be understood in terms of the identified framework (Figure 19).

Given the aims of the NHLS to contribute to a healthy nation by detecting, managing and preventing disease, it was not surprising to find a large number of uses relate to Public Health Tracking. The prevalence of HIV/Aids and TB in the country has led to the identification of these diseases as

national priorities, and so it stands to reason that the majority of NHLS CDW public health tracking application focus on them. Public health tracking applications also include AMR monitoring, the monitoring and management of disease outbreaks and cancer surveillance. Aside from operational management who perform an administrative function only, all three of the other divisions covered in the interviews use public health tracking applications.

The NHLS CDW is also being used comprehensively for administration within the organisation. Both operational management and the NPP use CDW data to monitor laboratory performance, laboratory usage of capacity, quality control, and tests rejected to avoid wastage. Operational management also consult CDW data for budget planning. No administrative use of the data was found for the NICD.

Research uses of NHLS CDW data identified in the interviews were not comprehensive, and none of the research identified was done by the NPP. This may be because many of the people interviewed were not focused on research. Reference was made quite frequently in the interviews to research being done using CDW data by the NPP and others, and an on-line research revealed tens of research papers crediting the CDW as the source of the data. Supporting documents for the research shown in table 9 include articles on NICD and NPP research that used CDW data. These look at the epidemiology of meningitis when HIV rates are high, the effects of using new GeneXpert tests for TB, and the prevention of mother to child AIDS transmission.

In terms of the way in which the data was accessed, the majority of requests were made either through a customised application or a CDW service representative. Requests made through data analysts applied only to NPP staff. This may be because they have two data analysts employed for that purpose. No requests were done via a high level CDW access tool because this facility has not been made available, and only the NPP data analysts reported on requests made using SQL, which has been defined as a low level CDW data access tool.

Presentation of results in relation to research question 1 is now concluded. Research Question 2 is addressed next.

5.5 RQ2: What are the issues and challenges in the use of the NHLS CDW?

This study views the issues and challenges encountered in the use of the NHLS CDW through the lens of Activity Theory. An activity system for a request for information from the CDW within the context of the NHLS was developed in chapter 3.

During analysis of interview data fifteen common themes came up, and, earlier in this chapter, twelve of these themes were mapped on to the activity system for an information request. As a reminder, these mapping are shown again in figure 21. Eleven of the themes mapped onto the activity system, ‘Relationship with CDW’, ‘Staffing’, ‘Time Delays’, ‘Outsourcing’, ‘Data Quality’,

'Patient-oriented View', 'Privacy and Security', 'Direct Access', 'Metadata', 'Ease and Efficiency', and 'External Data', will be examined in terms of the issues and challenges that they do or do not infer.[30]

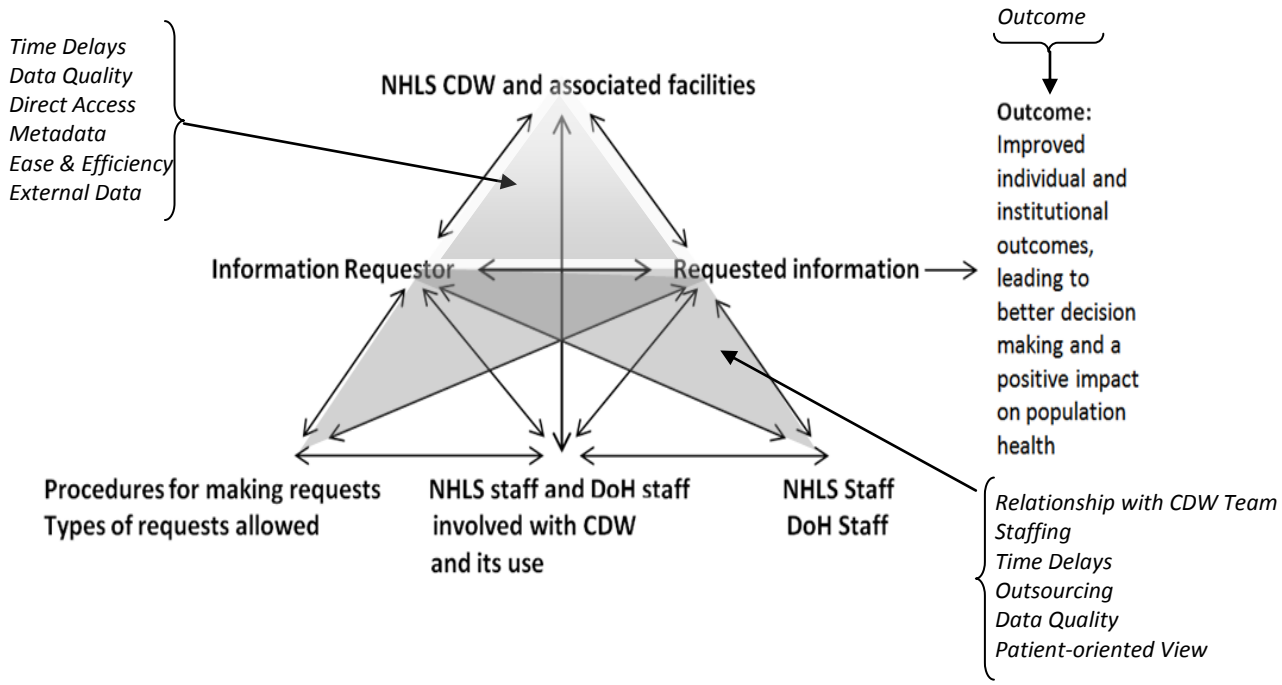


Figure 21: Allocation of Themes to Activity System for an Information Request

5.5.1 Relationship with the CDW Team

The first challenge associated with the use of the CDW is characterised by the theme 'Relationship with CDW Team'. This theme was placed in two different sub-triangles of the activity system for an information request shown in figure 22 and figure 23 respectively. This choice will be explained once the data for this theme has been presented.

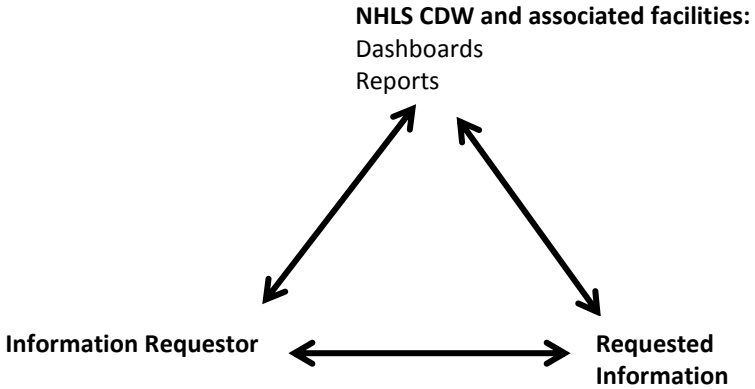


Figure 22: Sub triangle of the activity system assigned to 'Relationship with CDW Team' (Through Customised Application)

[30] Of the other three themes mapped onto the activity system, 'Outcome' will be used to address research questions 3 and 4, and although the use of a 'Data Governance Model' and 'Metadata' were questioned by the researcher, they were not found to definitely be present in the NHLS CDW environment. The latter two themes will therefore be used when suggesting recommendations for the future.

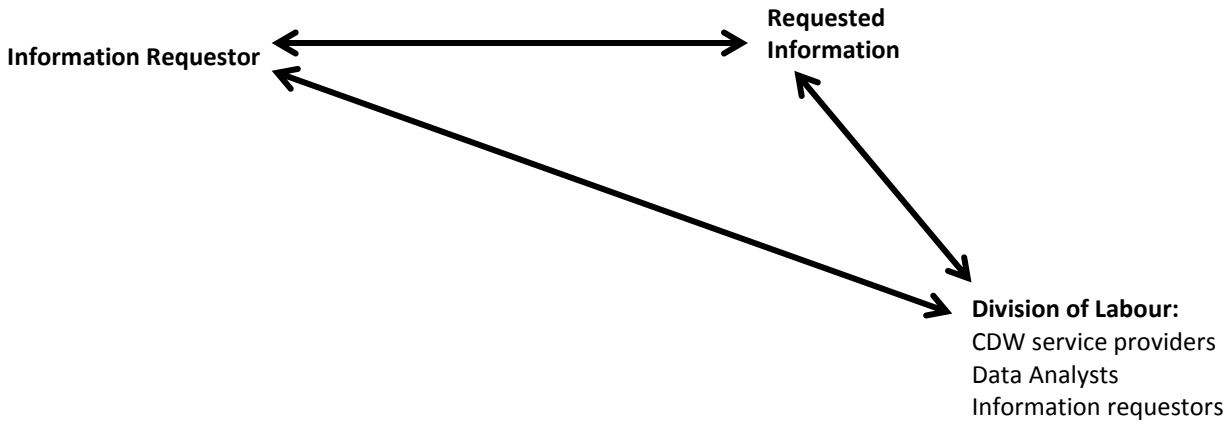


Figure 23: Sub triangle of the activity system assigned to 'Relationship with CDW Team' (Through CDW Service Representative or Data Analyst)

Since only data analysts access the CDW with low level CDW access tools like SQL, information requestors that are not data analysts have three ways in which they can request data through the CDW: through a customised application, through a CDW service representative and through a data analyst. The relationship with the CDW team differs for each of these means of access, so they will be discussed separately.

Information requests made through a customised application need very little interaction with the CDW team. The application has already been pre-prepared and can be used if registered for. A form, signed by a supervisor, is handed in to request access. Once access is granted, the application can be used as and when required by the information requestor.

In order to make an information request through a CDW service representative, the form showed in appendix I must be filled in and given to the IT help desk, who then pass it on to the CDW team. When a CDW service representative becomes available the data extract is produced and emailed to the requestor. Unfortunately this data extract very rarely, if ever, satisfies the requirements of the requestor. Instead, an iterative cycle of clarifying requirements and re-performing the extract takes place until the correct data is received. The clarifying of requirements can take place with emails or in person. Many of the interviewees discussed 'sitting' with the CDW representative to work things out, for example:

> *"It was difficult in the sense that I had to understand the data I was looking for myself first, and then I had to sit with the person extracting the data and explain to him what exactly it means, and what data he's supposed to extract for me. . . . So it was quite challenging and I felt that I had to – it was sort of an iterative process – I had to go many times and sit with them to get the right data."*

The number of iterations, and therefore the amount of time taken, depends very much on the type of request and the complexity of the data required. It seems that requests for research data generally have a longer cycle of iterations than those for administration and public health tracking data.

This iterative process requires close co-operation between information requestors and the CDW team, and a willingness on the part of the CDW service representative to spend the time to get it

right. Happily, most of the interviewees expressed their appreciation of the CDW team, and praised their responsiveness and willingness to help, with comments like

" [They are] Very very responsive"

"So I'm very pleased with people that are working with CDW so far."

"They're actually – for me – I think they're doing their best and I'm happy with their service which I'm getting from them because whatever requests I do get, I haven't really had the disappointment where I've failed to do my job because they couldn't provide me with the information timeously"

Interviewees from NICD were more guarded in their praise for CDW, saying, for example:

"I've had good service, ya. I think one has to realise the limitations, and . . . but I have had good service so far."

The 'limitations' refer to the fact that CDW staff are not medical people, and therefore do not easily understand the data and its implications in a medical context. It is thus a more difficult and time consuming process to express information request requirements to CDW staff member in a way that they can understand and action:

"So you need to sit them and look at the variables that they have, within that line of disease or whatever results you want and be able to shift their mind into that, so that they can be able to write coding that belongs to that output that you need. You cannot just make a list and say "bring me this" and they will do it, because they are not scientists, they are not medically to be able to give you all that. You are the ones who understand what they need."

Nonetheless it is still understood that the CDW service representative has their own area of expertise, which is needed here:

"I think the person I dealt with was helpful and he was trying, but I think he didn't also quite understand the data itself, and - but he did understand the process of extracting the data, which I didn't understand."

It was thought by the researcher that the quality of assistance and training offered by CDW would be a big factor in their relationship with information requestors. A question targeted at this revealed that this was not the case. Requestors tend to get assistance and training from colleagues who have previously dealt with CDW data. One data analyst actually mentioned that he himself is the trainer, "*going around the country training people on how it works".* Many of the interviewees cited familiarity with the data on the CDW because of previous experience working in other parts of the NHLS, particularly in the laboratories themselves. Newer employees, however, did have problems with finding out the meaning of the coded names given to each of the laboratory tests, saying that it took time to work out which test results they needed.

Information requestors dealing with the CDW team through data analysts seemed considerably more satisfied with having their requirements understood, as well as with turnaround times for

information requests and with the quality of data received. Though they have to go through a similar cycle of iteration to get the requested data right, data analysts are more accessible and readily available. Employed full-time to serve the needs of the information requestors in their department, they develop a good understanding of the function of the department and the information needs of the staff working there, thus expediting the information requesting process. Their data access skills are far more developed than the information requestors' and they are able to access data in the CDW directly using SQL, thus working with more flexibility and circumventing the bureaucracy and procedure of the formal data request route. They also get to *"understand exactly how the data is structured at CDW"*, which facilitates easier data extraction. Most importantly, dealing frequently with CDW team members promotes the development of a good working relationship - data analysts interviewed understood the importance of cultivating a good relationship with the CDW team:

"Well I'll say we have sort of fostered the relationship between the guys there, so, on average it take one day or two day. But previously it used to take some time . . . We have a good one, very good relationship"

"it's a relationship that you need to keep on building"

"communicating with CDW is not difficult because we have created that kind of a channel between us and CDW."

"I think we have been really improving year on year, which is a cool thing, and we have managed to have our bi-monthly meetings to explain those challenges with CDW if we experience any, so every 2 months we have got a meeting with CDW and we sit around the table and talk about things"

Considering the aforementioned points in the light of the activity system developed helps to put the relationship between the information requestor and the CDW team into context.

Where the request is through a customised application, the application is effectively a tool that has become part of the CDW facilities. In this case this theme would fit into the top sub-triangle, where the data received by the requestor is mediated by the application (eg. dashboard, Excel extract) used. This is illustrated in figure 22. Note that only those tools that are relevant to this theme are shown.

As mentioned earlier, the interaction between information requestors and the CDW team is minimal in this case, so for simplicity's sake this allocation has not been added to the diagram in figure 21 above.

Clearly in the cases where the information request has been made through either a CDW service representative or a data analyst there is a division of labour between the two or three parties involved. Though it seems that the parties are all very clear about what their roles in the process are, the efficiency and effectiveness of the activity depends on the skills and abilities of the individuals involved. The existence of a data analyst as a "middle man" will also promote success. Therefore this theme has been placed in the sub-triangle illustrated in figure 23. Only those parties who labour is being divided amongst for this theme are shown.

Perhaps the best way to summarise this theme – 'Relationship with the CDW Team' - is with the following comment made by one of the interviewees, which shows that the CDW team is really open to improvement:

> *"They are open to suggestions . . . I'll give you an example. . . . the NHLS form didn't have a field whereby we need to indicate when was this person screened . . . they took it, factored it into the LIS – even today it's still standing there, and then also produce an SOP[31] to be used nationally on how to do and capture that information . . . And so, for me, I really feel that you need to work with such kind of people who aren't saying we know it all . . . They are open to suggestions that will help them improve on their services, ya, for that I take the hat off for them."*

Despite the good relationship with the CDW team, frustrations were expressed with two important aspects of dealing with them. These were in the areas of 'Staffing' and 'Time Delays' which have themselves been identified as themes. The 'Staffing' theme is discussed next.

5.5.2 Staffing

The first challenge associated with the use of the CDW is characterised by the theme 'Staffing'. This theme was placed in the sub-triangles of the activity system for an information request shown in figure 24. This choice will be explained once the data for this theme has been presented.

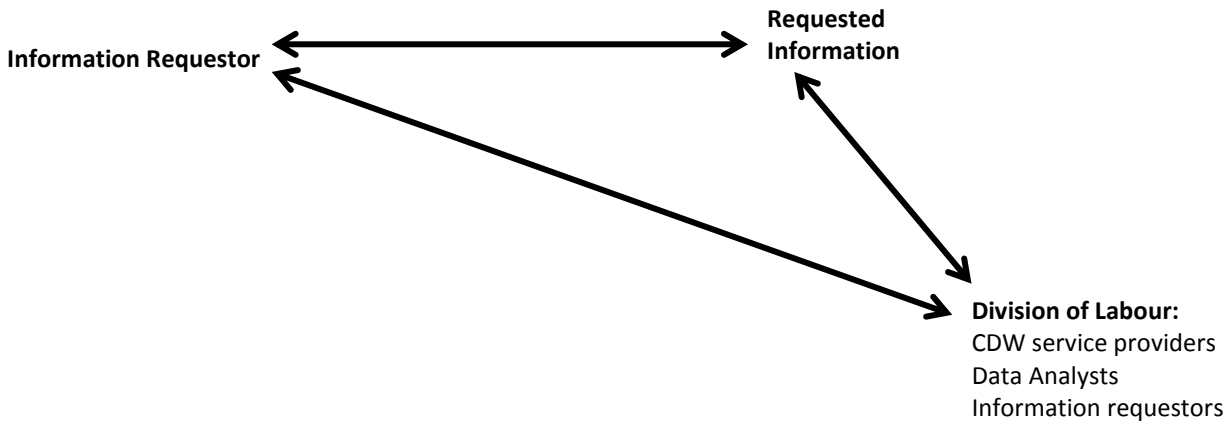


Figure 24: Sub triangle of the activity system assigned to 'Staffing'

The CDW team is tasked with the immense responsibility of maintaining the complex technical data warehouse infrastructure as well as answering to the information needs of a myriad of different information requestors both inside the NHLS and beyond. This load is felt by CDW staff, one of whom said:

> *"And you see we are quite a small team. We are making data for the whole country . . . . every time there are new requests, new tests coming through, we have to enhance the system to accommodate them."*

[31] Standard Operating Procedure guideline

It is clear that many of the interviewees perceive CDW to be subject to "*constraints with staff" or being "stretched to the limit*". This often results in time delays for requestors, particularly when CDW receives multiple requests concurrently. Sentiments in this regard expressed by the interviewees are:

"There's no staff to assist, or you're in a queue, there's other people requesting data"

"I think it's mainly the time it takes for requests to come. And I mean it's understandable because we know there aren't many staff working there and they've got numerous projects"

"Months! I waited a year – not even for a research, for them to dump my core surveillance data. Yup. Research. Months! There is no standard turnaround time, so it depends when they have somebody available to do it, when they have somebody with the skills to do it."

"Well, sometimes issues will come to being not given that data on time, and that you should be understanding, because their requests could be coming from multiple partners from different areas and they had to try and manage their time . . . because you cannot pressurise them. You pressurise them more and get something that is incorrect, that you'll send back, and it will be a back and forth kind of thing, so rather gave them and then wait for them to do the right thing. But – yes – with the volume of requests they might end up with delays."

One of the problems the CDW encounters with staff is the high turnaround, apparently not uncharacteristic of the Information Technology (IT) industry. This was referred to by one of the interviewees:

"I can say there are challenges which are . . . to do with our own internal issues of skills and all that . . . because . . . the IT - they are very mobile. They are very mobile so you build your skill base and then the next thing they've left to another . . . it could be one of our limiting factors sometimes . . ."

This turnaround can affect information requestors adversely in several ways. Firstly, any relationship cultivated with the member of CDW staff having left will have to be reformed with someone else, which takes time. When asked how he would rate the service of CDW staff one of the interviewees said:

"Look, this depends on the turnaround of the staff, because you might be having a good relationship with this person and he doesn't stay long, so normally you will have to start teaching the others so that they get to the level of where you were with the other person."

Another interviewee expressed frustration about a piece of programming code vital to providing important information being 'lost' when a member of staff left:

"They say that the person who previously wrote the code to bring the data out of the CDW . . . they can't find the code – the person has left, nobody knows what, where that code is sitting, so they are unable to apply it to draw the reports out."

This kind of situation is serious, and it causes the CDW and all information requestors that depend on it to be vulnerable, particularly if the data is needed to perform core functions. One interviewee expresses concern over this predicament and voices the conviction that there should be a backup in the event of someone leaving:

"Recently, one person that has been working with CDW left . . . And it somehow affected some of the information that we were requesting from CDW because of their absence. Well, the question that we had was, in their stay at CDW, did they not reproduce themselves? To have more people with their knowledge, you understand? To be person-centric it's very dangerous. . . . What if they leave tomorrow? What if they don't come back next year? Then what? What are we going to be? It's something that I still am very afraid, I do not know if they were able to let go of what they know so that they teach the rest of the crew so that even in their absence, work can and must continue. That's my fear – that's my biggest fear."

As mentioned in interviews, limited resources also prevent the CDW from being used to its full potential:

"I think they can do so much more with the data. I think you need the resources – at the moment there's very little resources that can help us do what we need to do in the time that we need to do it. In terms of staffing issues."

"Firstly they don't have the people to do the core work, and then when you're looking at nice to haves – which actually is not nice to haves, it is an extension of core work – but perhaps it's not existing at the moment."

The impact of these staffing issues on a department can be alleviated if there is funding to enable the department to hire their own staff, like data analysts, for example. As discussed, data analysts can improve both the efficiency and the effectiveness of the information request process. The NPP has funding from the Global Fund and other sources, and benefits greatly from the data analysts hired. Even within the NICD, funded projects are given priority. As a member in one of the NICD departments noted:

"right now we're trying to get our AMR dashboards optimised to – you know - add more elements to it. So since there isn't a dedicated person that, you know, like for our centre, that can help us do that in a quicker time and also a lack of funding cause we need to pay this person . . . there's a lot happening now with the notifiable medical conditions - there's funding for that so that takes priority over other projects. So if there's funding available . . . "

With funding additional staff could also be hired by a department for development of new facilities:

" . . . if I want to develop a system I have to raise funds externally to employ a developer"

In short, the general opinion is that CDW is short staffed, which results in information requestors being kept waiting and also not using the CDW to its full potential. The staff shortage is partly due to the rate of staff turnaround, which results in loss both of important skills and of relationships that had been established with ex-members of staff. These problems are alleviated somewhat for those who have external funding which can be used to hire new people. Despite the fact that there are so

many positive applications of the CDW data to research, administration and population health, it seems that so much more use could be made of it if better staffing resources were available.

In terms of activity theory, 'Staffing' clearly relates to the division of labour, and therefore is mapped onto the same sub-triangle as 'Relationship with CDW Team', as shown in figure 24. This mapping underscores the claim made earlier that there are factors other than the tools and technology that affect the success of the CDW. The CDW can only be used optimally by decision makers if the environment is conducive to them doing so, and certainly having sufficient staff to address their needs would be part of that.

As mentioned previously, the other frustration expressed by interviewees about the relationship with CDW staff was 'Time Delays', which is the theme to be discussed next.

5.5.3 Time Delays

This theme was placed in three different sub-triangles of the activity system for an information request shown in figures 25, 26, and 27 respectively. This choice will be explained once the data for this theme has been presented.

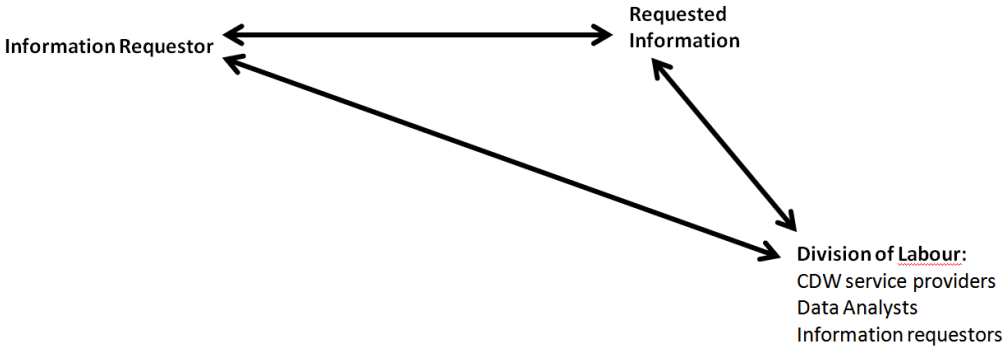


Figure 25: Sub triangle of the activity system assigned to 'Time Delays' (due to Staffing and the Information Request Iterative Cycle)

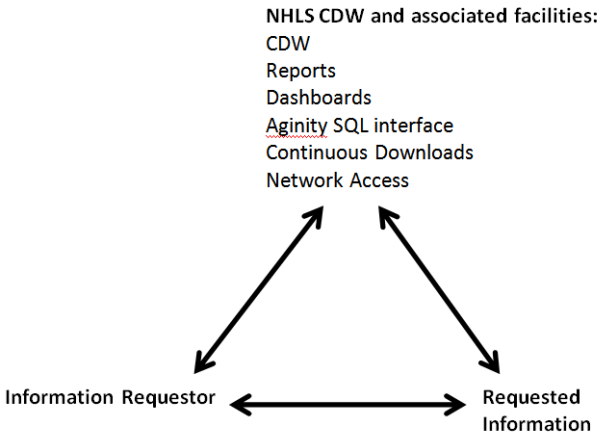


Figure 26: Sub triangle of the activity system assigned to 'Time Delays' (due to Technical Problems)

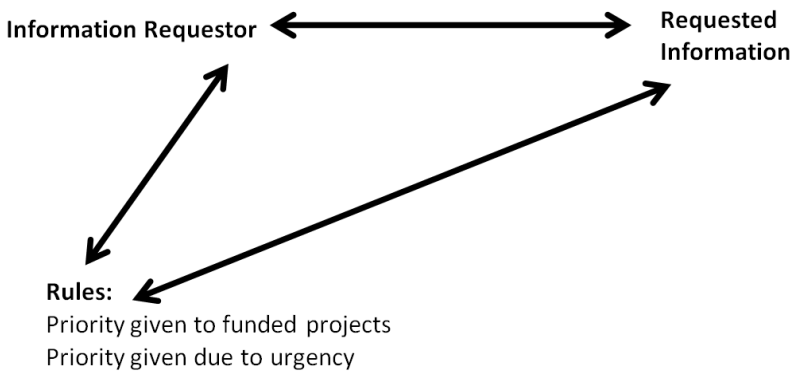


Figure 27: Sub triangle of the activity system assigned to 'Time Delays' (due to Prioritisation)

Staffing constraints and the cycle of iterations required to get an information request right have already been cited as reasons for frustrating time delays. What is interesting, though, is that the degree of time delays differs markedly between the NICD and the NCR on the one hand, and the operational management and the NPP on the other. Whereas NICD and NCR staff complain of time delays of weeks and even months, operational management and NPP staff all say that they have request results within a week, and often less. In fact many of them feel that there is no delay. For this reason they will be discussed separately.

The NPP have two dedicated data analysts who, as discussed earlier, do help to shorten information request turnaround times. Time delays for these requests were only mentioned by two members of NPP staff, who indicated that they were held up by technical problems:

". . . sometimes the system has crashed and . . . not often, but it does happen once in a while and then all our reports will be delayed for the particular period."

"No, the only challenge is when data is not ready and you want to . . . compile your report, or you have a meeting and they're having technical challenges."

The location of the NPP at the Wits Medical School in Braamfontein may be playing a role in these 'technical challenges', since the CDW itself is situated in Sandringham and the connection would be over a wide area network. This may have been what one of the NPP data analysts was referring to when he said:

"There's network issues when we are sitting here and the servers are that side"

Time delays were also experienced by NPP staff in gaining access to information (rather than in getting the information itself). Data analysts have direct access to a subset of the data in the CDW. If a piece of needed information is not in that subset they need to get permission to add it to the subset, which is done by filling in a form, getting appropriate permissions from supervisors, submitting the form to CDW and waiting for access to be granted. Used to quick responses, this additional delay is noticeable but not serious:

" at some point we will have to wait for 2 days to get access to specific data."

Perhaps the relationship with CDW fostered by the data analysts serves to limit this delay to two days.

The operational manager interviewed felt there was no real problem with time delays, saying:

"But in normal circumstances there isn't really a delay. . . . you will log it today, and they will dictate to you how long will it take. Because normally we will want to know the time lines. Sometimes they will tell you no – because of their schedule . . . We agree, in terms of saying I can get it Wednesday or Thursday depending on the day in the week, so then they will give it to me within the agreed time. . . . Mostly in the week you get it"

This contrasts the feelings of frustration expressed by other users, where a member of staff is frustrated because:

" . . . your requests gets prioritised and the time – there's no agreed upon turnaround time." ,

and: "*I have to, you know, without it being an overreaction, I have to say it is absolutely inefficient."*

For this user group this is particularly problematic, because they have not been able to meet a key performance indicator target as a result. The situation is explained:

"Key Performance Indicator for this year was . . . to get 2014 data out by December . . . in order to be where we wanted to be was a three year lag time . . . It is something we promised the Department of Health that we would do and we will not achieve that, either this year or next year"

Staff at NICD also experienced long time delays. Asked about issues and challenges experienced getting information from the CDW, one interviewee said:

"I think it's mainly the time it takes for requests to come."

And another interviewee remarked:

"It takes time to get the data, so I know that when I first requested it took a few weeks to actually get the data, and then again, data had to be re-extracted because not all the data elements were included."

This interviewee did also point out that expected request turnaround time is often about a week, and the urgency of an information request could be escalated in the event of a disease outbreak:

"If it's not urgent, it's about a week or so – the turnaround time is about a week – but for an outbreak situation, the turnaround time is about 2 or 3 days.

There was also the feeling that time delays resulted in the CDW not being used to its full potential:

"I think they can do so much more with the data. I think you need the resources – at the moment there's very little resources that can help us do what we need to do in the time that we need to do it."

" . . . because it's quite a lengthy process you need to request data, sometimes one feels I'd rather not go that route."

Since the CDW is set in the context of the different NHLS divisions that it serves, the extent of time delay experienced for individual information requests seems to depend on some institutional rules about priorities for whose request gets responded to first. There was mention of priorities in one of the quotes above, as well as of the elevation of a request's priority in the event of a disease outbreak. Request priority was also more specifically alluded to with the following comment:

"*there's funding for that so that takes priority over other projects*"

It seems that time delays are affected by various factors, and therefore belong in several sub-triangles on the activity system. Firstly, staffing and the iterative nature of getting information requests right involve the division of labour between the information requestor, CDW service representative and data analyst, and therefore fit into the sub-triangle indicated in figure 25.

Secondly, time delays due to problems with accessing the CDW for network or any other technical reasons fall under the gamut of facilities provided by the CDW department, and therefore form part of the CDW 'toolset'. This type of time delay therefore fits under the 'tools' sub-triangle as shown in figure 26.

Lastly, the existence of a prioritisation structure governing the urgency with which information requests are attended to results from the fact that the request is made within the context of the community of information requestors, and as such must conform to the rules of that community. In this case, the rules dictate the priority with which requests are handled, and thus affect the time delays. As a result, this type of time delay has been assigned to the sub-triangle shown in 27.

The lens of Activity Theory makes it clear that the reasons for time delays are widespread, involving not only the facilities provided by CDW, but also the rules and workings of the organisational context that the information request activity is situated in. In terms of organisational context, it is also interesting that the NICD and NCR had far worse experience with time delays than the NPP and operational management.

Both the 'Staffing' and 'Time Delays' themes are related to another theme called 'Outsourcing', which will be discussed next.

5.5.4 Outsourcing

This theme was placed in the sub-triangles of the activity system for an information request shown in figure 28. This choice will be explained once the data for this theme has been presented.

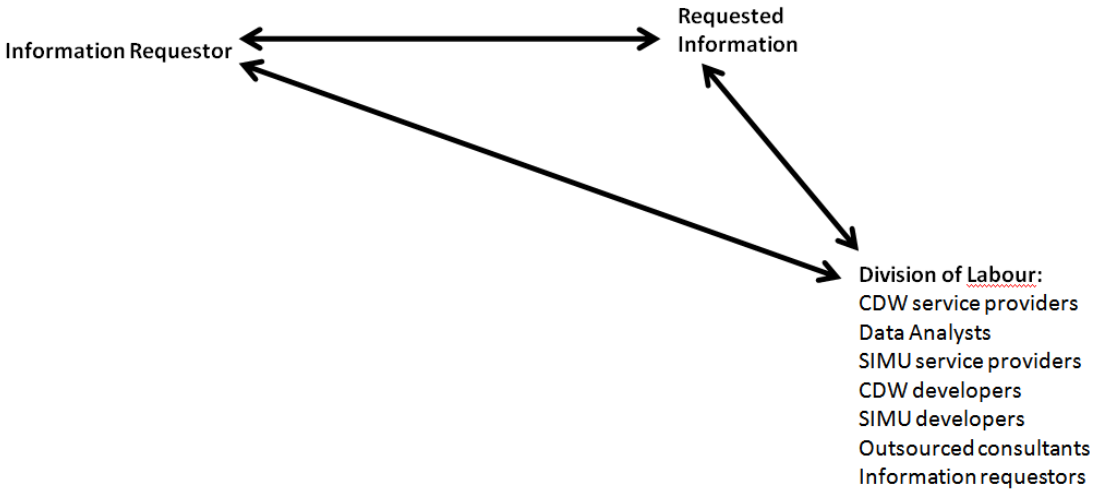


Figure 28: Sub triangle of the activity system assigned to 'Outsourcing'

One way of either alleviating pressures on insufficient staff or overcoming a lack of specific skills, and consequently bringing down time delays, is to outsource some of the work to outside companies. Outsourcing work that was discussed during interviews and the preliminary session fell into three categories – one concerning assistance with the right-time infrastructure, one dealing with a department itself hiring outside contractors, and one regarding the relatively new Surveillance Information Management Unit (SIMU).

In the first category, CDW staff mentioned working together with an outside company to provide the right-time data warehousing infrastructure needed for sending the DoH the information about notifiable diseases within the allocated 24 hour time period:

> *"It's a company. . . . they wrote a program to send us data. We also wrote a procedure on our side. So . . . they run our procedure on their side to send us data the whole day – even if you check now there's data coming. So we then use Informatica to fetch it from landing to our target database, which is Neteza. On Neteza, that's where our ODS tables are."*

In the second category, sheer frustration with time delays caused one of the interviewees to hire outside help to extract data from CDW:

> *"I've had to hire consultants previously . . . to write code to draw the data out of the CDW. Because they didn't have the expertise, and I needed the data and I couldn't wait for them to eventually get to it, so I had to hire a consultant to do that for me"*

And another interviewee talks about hiring a consultant through the CDW:

> *"If we need extra things that will help us improving the data, we source out within CDW and say, OK, can we bring in the consultant who can draw up something like this as a report. So we'll be paying for it in our own budget. . . . CDW, yes. They manage that."*

Also in the second category, a need for missing data caused NHLS staff to turn to academic experts for helping in working out what the data should have been:

> *"80% of our public sector data does not have population group attached, so we impute that. . . . We have a system we developed with one of the mathematics departments at Wits. We developed a hot deck imputation method, where we use surnames to impute race."*

The third category concerns the SIMU, a relatively new initiative spearheaded by the same well-respected person who used to be responsible for the CDW. It was intimated by one of the interviewees that this person and her team are now employed by an outside BI company and that all the work done by the SIMU is actually outsourced:

> *" . . . in terms of the dashboard updates we go through the SIMU, because they've outsourced a company, to get the data or collate the data into development (or whatever they do) . . . So if there's an issue . . . we will communicate with the SIMU directly. . . . So CDW doesn't create our dashboards – they have company that . . . does the dashboards"*

It seems that although the CDW handles the bulk of information requests, in certain scenarios the load is alleviated and external skills capitalised on by outsourcing work to other companies. Where used, there were no apparent problems and frustrations with outsourcing, probably because the CDW takes on the bulk of the responsibility and the situation is not that everything is left up to an outside company.

Outsourcing is an alternative way in which labour can be apportioned to different parties, so therefore it fits into the same sub-triangle of the activity system for an information request as 'Staffing' and 'Relationship with CDW team', as shown in figure 28. Notice that SIMU staff and outsourced consultants have been added to those between whom the labour is divided.

Research findings have now been presented for a group of related themes, 'Relationship with CDW Team', 'Staffing', 'TimeDelays' and 'Outsourcing'. Attention will now turn to two themes unrelated to this group but related to each other, 'Data Quality' and 'Patient-oriented View'.

5.5.5 Data Quality

This theme was placed in two different sub-triangles of the activity system for an information request shown in figures 29 and 30 respectively. This choice will be explained once the data for this theme has been presented.

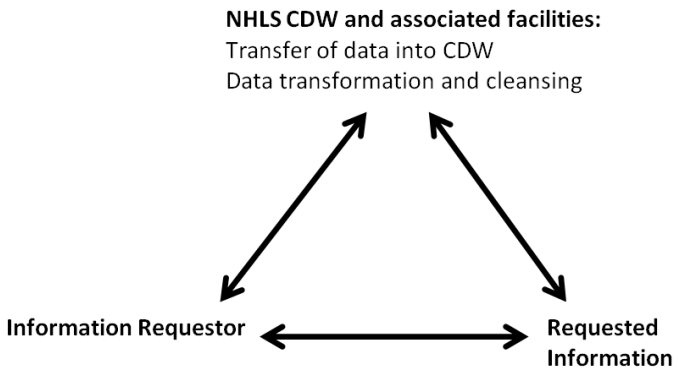


Figure 29: Sub triangle of the activity system for 'Data Quality' (within CDW responsibility)

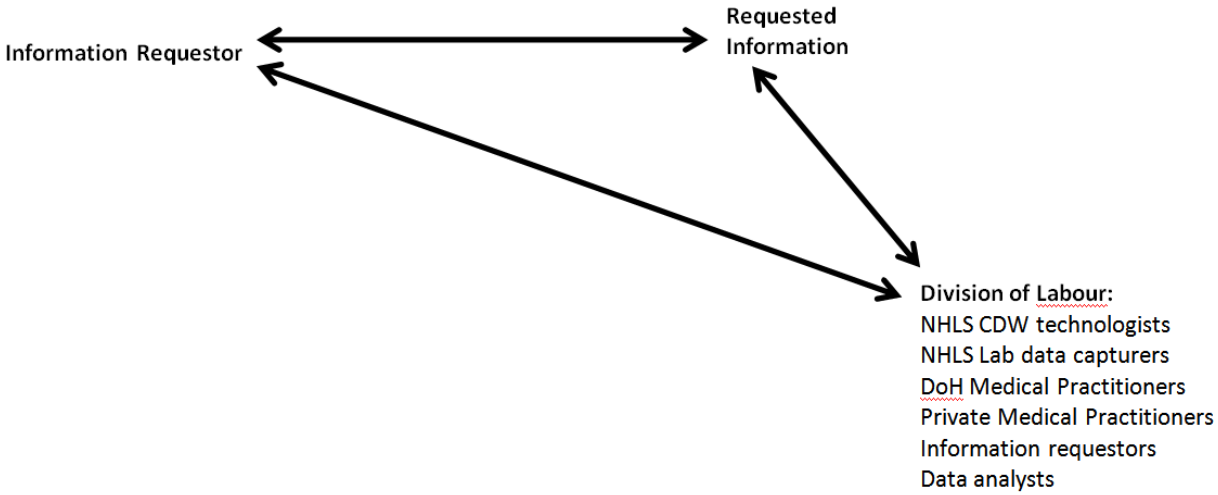


Figure 30: Sub triangle of the activity system for 'Data Quality' (Outside of CDW responsibility)

Decisions based on the data in the NHLS CDW for research, administration and public health tracking rely on the quality of that data. Decisions based on bad quality data are unlikely to produce the desired results. Ensuring that data is of the right quality is a major challenge, particularly since the NHLS CDW processes about 300 million raw data records a month.[32] Even after CDW has performed "extensive data cleansing and transformation"[33], information requestors experience problems with data quality.

[32] According to the "Data Warehouse Overview" presentation received during the preliminary session with CDW staff

[33] Also according to the "Data Warehouse Overview" presentation

The most extensive and impactful data quality issue reported was that of the difficulty experienced in obtaining a patient-oriented view of the data. This is due to the lack of a well-enforced unique patient identifier. Because of the importance and extent of this issue, it is regarded as a theme in its own right called 'Patient-oriented View', which will be discussed after this theme has been completed.

Asked whether they trust CDW data, interviewee responses fell once again into two marked categories. Operational and NPP staff had between 95% and 99.9% trust in the data. NICD and NCR staff, on the other hand, spoke of less trust in CDW data, with the maximum amount of trust in the data being 80%. All divisions, however, spoke about the issue of data quality in some sense. The general feeling was that the bulk of the data was of good quality, but there was cause for concern for a percentage of the data. Understanding which aspects of the data were found to be challenging was done through questioning information requestor perceptions of different dimensions of data quality. Wand and Wang (1996) list twenty five data quality dimensions, of which the most frequently mentioned in the general literature are 'accuracy', 'completeness', 'consistency' and 'timeliness'. Interviewees were questioned on these four dimensions, as well as 'relevance' – also part of Wand and Wang (1996) 's list.

All interviewees found that information received from CDW was sufficiently up to date, or 'timely', for them and relevant to their work, so 'timeliness' and 'relevance' were not found to be issues.

Consistency within the data was also not reported to be an issue. Unlike many data warehouses, where the data comes from many disparate sources, the bulk of the data in the CDW comes from the laboratories, all of which run the same laboratory information system (LIS), TrakCare. Therefore most of the data comes in an expected standard format and it is not surprising that this type of consistency is not considered to be a problem.

A different type of inconsistency was reported, however. Confusion arises when requests for the same data turn up different datasets:

> *"I can request the same dataset three times, I will get a different number of cases in that dataset. I have had that experience, where I asked for colorectal cases between this year and this year and there was a difference of a couple thousand cases every time I requested it."*

> *"We have had instances where the numbers don't balance. So from last month I pulled 20 000 tests, then the following month if I try to pull data from the same period you'll find that there's more tests or there's less tests."*

> *"So if I reported 20 for this December now . . . There's a possibility that when I run the same query I can find 25, when I run it next month, so somehow, somehow it just takes time, but it's something that we have accepted, it's not - around - less than 10%."*

Of course getting a different dataset back may not mean there is any problem with the data – it may actually be that the CDW query retrieving the data is in error. In addition, several of the comments above should be understood in the light of the following explanation, which justifies why requests for information about laboratory tests that occurred in a specific period – December, say – sometimes return different datasets depending on when the request was actually executed:

"The information from CDW is dependent on the labs . . . if a lab somewhere out there has got what you call work that is supposed to be verified. . . . And the supervisor is not checking that verification queue - they don't authorise it. Let's say today is the 30th or the 31st of December, if they don't authorise that work and CDW closes on let's say on 1st January . . . all the work that is still seated at the labs and not yet reviewed won't flow to CDW. . . . But I've found that within my 3 years of staying in the program the labs or the work that the CDW work or data stabilises at most on the third month."

In other words, if CDW takes a snapshot of the data as at 1st January, there may be tests done in December that are not yet recorded as they have not yet been authorised. These tests may later be authorised and included in the CDW as tests done in December. There may also be tests that are recorded for December, but which are later rejected and therefore the test is removed from CDW. Consequently when the dataset for December is requested at a different time, a different number of cases is returned. The interviewer is saying, though, that from his experience after three months all these adjustments are completed and the CDW laboratory test data for that time period becomes stable:

"For me, from the lab manager's perspective, from the client liaison's office and also from the program management perspective I've learnt to say 'You know what, give this data some time, it will balance out', and it always does."

Examination of the last two dimensions of data quality, 'accuracy' and 'completeness', served to highlight the more serious challenges experienced.

A few specific instances of data inaccuracies were referred to, but this is obviously not a comprehensive list of all inaccuracies that may have been found:

" a prostate cancer, and that person is recorded as a female"

"a systematic error in terms of the gender of cases."

"I know there are 52 districts in South Africa, ok, and I'm running a report I'm getting 54 districts. So I had to now go to DoH website and see . . . Sisonke is changed to Harry Gwala."

Probing the 'completeness' dimension revealed issues resulting in both incompleteness and inaccuracies in the data, particularly demographic data. CDW is very often not responsible for this problem, however. To understand how this situation arises, one needs to appreciate where the data comes from, and how the data "*flows*" into the CDW.

The source of the demographic data for patients is the laboratory test request form, as filled in by a medical practitioner. This is where many omissions, and some inaccuracies, originate:

". . . you're relying on the doctor to fill in the variables, doctors only fill in the minimum dataset that is required for them to be able to retrieve the result from the system for themselves"

"And we know which are poorly filled in by the medical practitioners, what is well filled in by the medical practitioners – we've been doing this since 1986. So (laughs) we know – race - 80% of our public sector data does not have population group attached, so we impute that.

". . . this is not a CDW problem. But it is a data problem. . . . a data source problem, in fact."

As a result, there are details about the patient that are never recorded. Information about the disease they have, on the other hand is ample and satisfactory:

"data that is filled in for the pathology reports requests by the doctor is very sketchy. However, having said that, when a pathologist looks at the slide, I mean we get full information from that, because that's their job, you know, they will give detailed description of the cancer etc. . . . Demographic information is where we suffer."

Laboratory test request forms are then sent to the appropriate laboratory, where the next step in the flow of information is for them to be captured into the laboratory information system (LIS). This is the next opportunity for omissions and/or errors to occur:

"Whatever the doctor fills in to a pathology request form SHOULD be captured at the laboratory level, and then SHOULD be pulled through from the laboratory into the CDW"

". . . if somebody from the lab down there does not capture a certain information under that lab information system it does not flow to CDW, so thence and from there I cannot report with pride. I will give you an example. If all my clinics in the correctional services are sending the forms to the lab . . . this person was screened on admission, and somebody at the lab down there does not capture that . . . the report that is generated from CDW does not have that. . . . then the . . . end user of the report does see that there is TB in correctional services, but where is the TB coming from? Is it of all the inmates that have been there for 14 years? Or of the newly incarcerated? Exactly. So yes, it is CDW who houses the lab information, but . . . the labs . . . and the front office staff who are the capturers of data . . . are the ones that are even more important to make sure that information is flowing to CDW."

". . . there are some where a mistake is made at the laboratory level and we know that because . . . we get the lab report and it shows that . . . it was mistakenly captured."

Once the data has been captured into the LIS, yet another step – the transfer of the data from the LIS to the CDW - provides a further opportunity for omissions and/or errors:

"we are aware that there are problems with that pulling through from the laboratory into the CDW as well because (laughs) we check our data. Our last dataset we had a systematic error in terms of the gender of cases. We went back, we looked at what had been wrong from the CDW, we compared it to laboratory reports, and we found that there was an error in the way the information was pulled from the laboratory into the CDW, so there was a mistake in that part of the chain."

Not all of the omissions and/or errors discussed are detected and corrected by CDW's data cleansing and transformation, and many data requestors do their own checking to ensure that the data is adequate for their needs. One data analyst at the NPP does comprehensive checking of the data because he sees it as his role:

". . . because my role here, as a data analyst, is to actually to ensure CDW gives me the complete data, accurate data . . ."

". . . you need to understand, CDW is the only source of data, so you need to work with them to ensure data is complete. . . . After extracting data do your trend, to compare . . . last month versus this month, compare what you've got this month last year versus what you have received. So it's all about doing your controls prior to using the data that you get from CDW"

These checks and corrections are done before the data is passed on to the rest of the NPP staff, which may explain why there is a higher level of trust of the data amongst NPP staff.

Many of the staff who are not data analysts do extensive checks to verify that the data is fit for their own use, sometimes comparing the data against external data sources:

". . . you will be looking your data from, like all the data sources. If I get to a facility I'll extract data from all the registers they have; and have data from the district health information system and data from us. So at times we find that they under report or over report."

". . . sometimes, for example, I've got the data for this particular laboratory, so I then want to try to go back and compare, so I go directly to the lab to the system and do my own individual count, and then you find that 1% or something is missing on my CDW, so on the system I've got more, so I'm not sure where the data got lost."

"I receive monthly information from CDW in a two stream way. Number 1, from our internal data analysts . . . and also from (CDW) who sends me the volumes and statistics, so I validate my own data. . . . maybe to even make it more complex for you (laughs), I can get it from my own guys that's in the labs that are on site. "

"Remember, within NHLS there is CDW and there is, what d'you call, Oracle. . . . "So then the 2 can work back to back as validating tools, one against the other"

"Because we're using almost 3 systems – the laboratory information system . . . TrakCare. And then we've got Oracle which is mainly our financial modules doing the billing and all that. . . . So I can run for volumes, and . . . I can look them at Oracle, or I can look at them as well at the Trak, but mostly I use mainly the two – the Oracle and CDW to compare."

A need for missing data also sometimes results in complex workarounds, as mentioned previously:

"80% of our public sector data does not have population group attached, so we impute that. . . . We have a system we developed with one of the mathematics departments at Wits. We developed a hot deck imputation method, where we use surnames to impute race."

When talking about data quality problems experienced, many of the interviewees made sure to point out that the problems were experienced for only a small section of the data, and that most of the data was fine. CDW was also praised for communicating with requestors when there was a problem with the data, warning them not to use the data until the issue had been resolved.

Three dimensions of data quality – relevance, timeliness and consistency within the data – were not found to be of concern. Inconsistency is experienced in terms of the number of records returned by an information request, but this does not necessarily mean that the data is of bad quality. Issues do exist, however, with the accuracy and completeness of the data, particular patient demographic data, which was often never filled in properly by the medical practitioner in the first place.

Much of the responsibility for the provision of quality data lies with CDW, as the CDW infrastructure allows the transfer of data into the data warehouse and the transformation and cleaning thereof. Therefore part of the issue of data quality can be assigned to the top sub-triangle of the activity system, as shown in figure 29. There were relatively few and less impactful problems in this area.

Getting perspective from the Activity Theory framework, it becomes evident that data quality issues with more serious implications are those that are outside the control of CDW – they are problems originating outside of the NHLS at the source of the data and at various points along the flow of data to the CDW, caused by those responsible for getting the data from the patient and capturing the data into the LIS. Data quality is therefore affected by the division of labour and was thus placed in the sub-triangle of the activity system shown in figure 30.

It is interesting to note that once again there is a contradiction between the way data quality is perceived by NICD and NCR staff on the one hand and NPP and operational staff on the other hand. In this case the more positive perceptions of NPP program managers may be due to the fact that they work through data analysts who take some responsibility for the ensuring the quality of the data before handing it over.

As mentioned previously, the most extensive and impactful data quality issue for the NHLS CDW is the difficulty in obtaining a history of all laboratory tests for a given patient. This is addressed by the 'Patient-oriented View' theme, which will be discussed next.

5.5.6 Patient-oriented View

This theme was placed in the sub-triangles of the activity system for an information request shown in figure 31. This choice will be explained once the data for this theme has been presented.

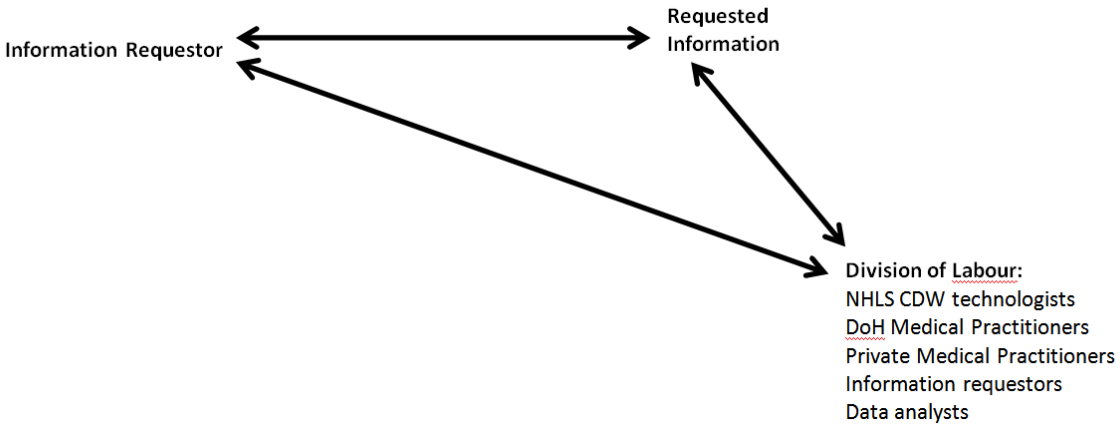


Figure 31: Sub-triangle of the Activity System for 'Patient-oriented View'

'Patient-oriented' refers to the accessibility of all information in the data warehouse pertaining to a particular patient. In the case of the NHLS, this would imply the ability to easily view all of the laboratory tests that have been done for a particular patient. Information per patient is often essential in the management of population health, and should be obtainable from the data warehouse with ease and accuracy 100% of the time. Unfortunately this has proved to be a challenge in the NHLS environment and context, since there is no properly enforced unique patient identifier. This is a problem that the requestors of information from the CDW are aware of:

> *"it's been a challenge, because, you know, we don't have that unique identifier for our patients in South Africa, so it's a challenge that everybody's aware of, and we . . . I think they have understood that it's not easy to get that patient view of information."*

Not all interviewees were interested in viewing the data from a patient point of view. Interviewees focusing on operational monitoring, in particular, required information only from laboratory test perspective. Several other interviewees, however, expressed the need for a unique patient identifier:

> *"Now what you need is to introduce something that will work for the whole country, which is a unique identifier – either the id or the finger print, which everybody will be identified, whether you go to Natal or whatever, you will still be 'Freddy'."*

> *"A unique identifier would make our lives SO much easier"*

> *" Sometimes you just want to link the patient that you had TB, how many went for their HIV test, the viral load testing, CD4 and all those things – you want to link them together, but how can you do that if you don't . . .* [have a unique identifier] *."*

As pointed out by one of the interviewees, without a unique identifier, connecting all the laboratory tests done for a single person "*becomes now a guess work".*

There is a medical record number (MRN) provided by the system which is intended to be a unique identifier for a patient, but this is not properly enforced. The problem, once again, is at the data source, where the MRN is not properly used and patients tend to be allocated a new MRN every time they come for tests:

> *"As we know the NHLS does have . . . the medical record number (MRN) at the lab level . . . after the person will have had their specimen sent to the laboratory . . . and then we can use that number for as long as they live. But I've seen that the clients are not keen to use that number, so . . . They're not capturing it. . . . because they were not taught to use it. They don't quote it in the subsequent specimens that they are collecting. Hence the person will have a new medical record number even though they already have one which is unique."* ('Clients' refers to DoH staff)

This allows people to abuse the system by presenting themselves at more than one facility for the same test:

"someone can do their tests here, and go to the next clinic and change their name and present themself as someone else and do the same tests just to prove that the result that they got it's correct, so . . ."

To add to the confusion, sometimes the same MRN has been found to correspond to different patient details with different tests:

". . . sometimes in the data they'll have probably the same record number, but the details are different. It's not for a big proportion of the records . . ."

It is acknowledged that the lack of a unique identifier is not caused by the CDW:

". . . that's not a CDW issue, that's an NHLS issue. Because CDW is merely the repository."

" So that cannot be drawn back to CDW and saying you are failing, because they didn't create this. They are trying to make this work, no matter how bad it is."

And CDW really are trying to make it work. They have an algorithm that attempts to match up all the laboratory tests for s single patient. As said by a member of CDW staff:

". . . for every test that is done on an individual, it gets linked to the other tests that the individual might do them later on. . . . at the information system we are interested at data at an episode level, or at the event level, so it's not patient level, but in medicine or in health they're interested at the patient level, so we must be able to group all the events."

Figure 32 shows a slide from the CDW overview presentation[34] that concurs with this and indicates the complexity of the task.

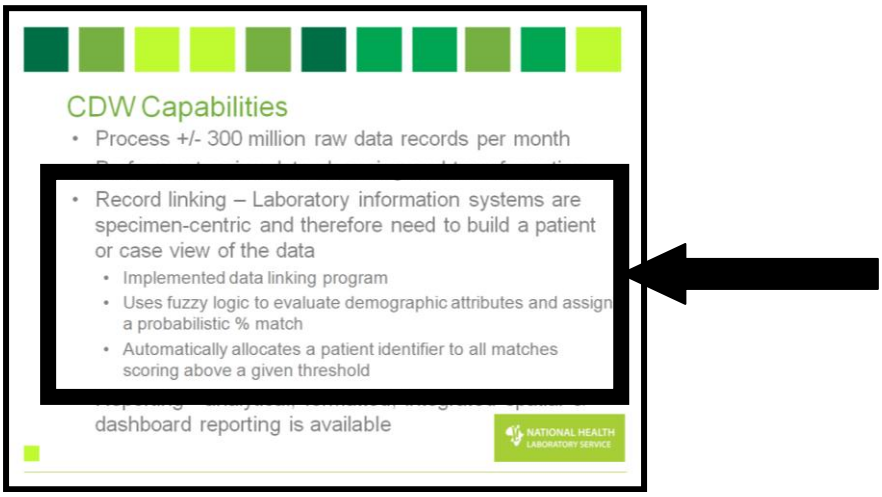


Figure 32: Presentation Slide showing CDW Capabilities

According to this slide, the CDW algorithm uses a specialised fuzzy logic technique, and the matching of records is probabilistic. It will therefore not work 100% of the time, and there will be a certain

[34] Given to me after the preliminary session with the CDW staff

amount of records that will fall through the cracks – i.e. some laboratory tests will not be linked up to the correct patient. Information requestors frequently experience this problem:

"So what they use here is probabilistic, and which can only catch a certain percentage of persons. It will be losing some within – along the way. So for us, this, for now, is all we have, and there's nothing more than that, because this has been tried and tested to give us at least 70 – 75% of real data. . . . 75% well matched in terms of probabilistic matching. So 75% means that even if a patient has been missed somehow, they have tried their best . . ."

Some information requestors resort to taking matters in their own hands and spend a lot of time checking the information manually:

". . . one needs to actually look at the data and clean the data and see whether it's actually suitable, because at times, for example let's say you have patient Joe Soap who comes to Chris Hani Baragwaneth, and then they capture him as male, they don't know his date of birth, but he knows that he's 23 years old. But then that same patient comes back again, but now they've probably spelt his name wrong, maybe by a letter, for example, instead of saying J O E, they have J O E E, but . . . that that might be the same patient, but now do you include that record or do you exclude that record?"

" At some point, the national department were saying to us "NHLS, your figures are high. Are you sure there is no duplication of tests?" And I came down and I took my CDW raw data . . . I was able to link it, because you know we're VLooking[35] it up out – we're able to can see the probabilities of this person being the same person – date of birth, name – no, the probabilities of Lerato Mokwena[36] was born on this day in December of this year, having another person of the same name are very low. . . . I did it manually, yes. I did it manually and there was some time, and I highlighted all those that I was seeing that they were repeated but they were not repeated in the same centre, using the same number"

A more sophisticated approach is preferred by one of the departments, for surveillance:

". . . we go back to our whole database and we remove duplicates . . . we have a program in SAS that does a removal . . . But then we also do a lot of eyeballing of cases because . . . we depend on routine variables filled in by the doctor, so sometimes you will be Smith John, sometimes you will be John Smith[37] - and SAS won't pull that up as an exact duplicate. Sometimes your date of birth will be missing. So, you have to eyeball it and make a decision on those cases where the variables are not exactly the same but . . . a human can see that actually it is the same case . . . they say they have a de-duplication algorithm at the CDW . . . We do our own de-duplication because we – obviously, we deal with a smaller dataset than they deal with, because they have the whole country and all the test cases, and so we are able to de-duplicate with greater reliability on our smaller dataset than the large dataset that they use."

And for research:

". . . for research I would never rely totally on the CDW . . . so we have had to devise very unique methods in order to link patients. We first started on probabilistic record linkage . . . and we found that it wasn't as effective. Now we're using machine learning . . . To link the patients and de-duplicate using machine learning. . . . so this is for research, this is not for the surveillance. . . .

[35] VLookup is a Microsoft Excel function that he was using to look for duplicates in CDW data obtained.
[36] Name changed for anonymity
[37] Name changed for anonymity

it's experimental at the moment, and that's why we're doing it in the research. Once we find that it works for the research project then we will have to look at it for our core surveillance work, but we haven't tested the validity and the reliability for – of it as yet."

Either way – whether manually or using automated techniques – a tremendous amount of effort is being exerted in order to get a patient-oriented view of the data. Even with this effort, there is 100% guarantee that the data will not be 100% right. On the other hand, if there was a well-enforced unique patient identifier, a simple database query would be able to easily satisfy the requirement.

Once again, this could be viewed as a division of labour problem, as it originates at the data source, where the clinician originally fills in the form. This would place the 'Patient-oriented View' theme in the sub-triangle indicated in figure 31.

This concludes the discussion on 'Patient-oriented View'. The next theme to be discussed is 'Privacy and Security'.

5.5.7 Privacy and Security

This theme was placed in the sub-triangles of the activity system for an information request shown in figure 33. This choice will be explained once the data for this theme has been presented.

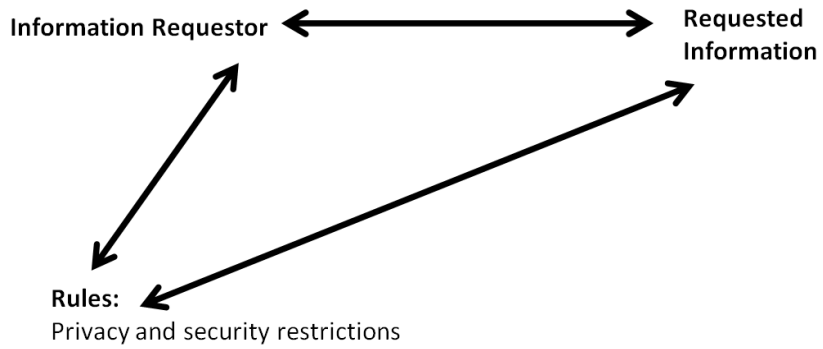


Figure 33: Sub-triangle of the Activity System for 'Privacy and Security'

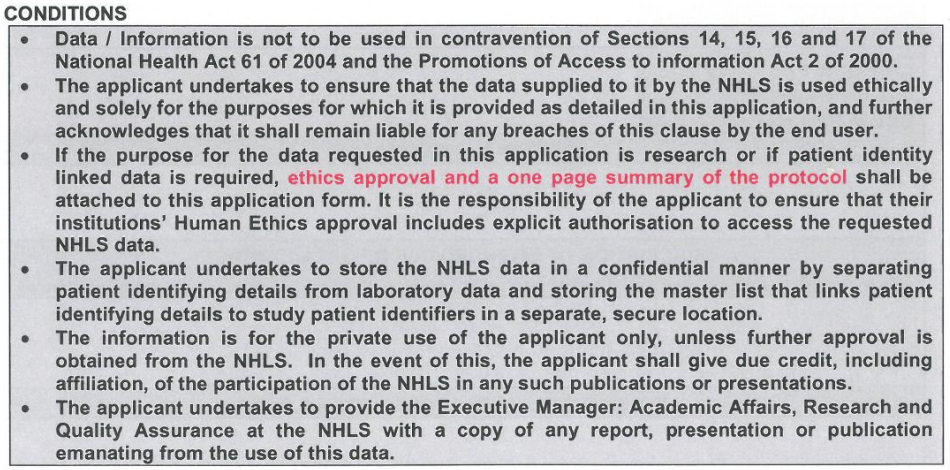
The bulk of the data in the NHLS CDW consists of results of laboratory tests and demographics of the patients who had those tests done, and is in fact exceptionally sensitive medical data. It is therefore of the utmost importance that there is security for the data to protect the privacy of the patients involved.

Interviews with NHLS staff revealed an organisational culture that is very mindful of this issue. Certainly there is an awareness of the sensitivity of patient data:

"So, that data I cannot – the data itself I cannot just give you without – because this is . . . it drills down to patient level, and patient level is identified data."

"Besides we are – that one (sighs) I don't know who can be looking for it because it's not even allowed for us to go into that detail to give it to someone else – you know, the patient data it's very . . ."

The data itself is protected on many levels in different ways for different levels of access. Information requestors are required to fill out an application form whenever requesting data from a CDW service representative. This form stipulates strict conditions that must be accepted by both the representative and his/her supervisor.[38] These conditions are shown in figure 34.



CONDITIONS

- Data / Information is not to be used in contravention of Sections 14, 15, 16 and 17 of the National Health Act 61 of 2004 and the Promotions of Access to information Act 2 of 2000.
- The applicant undertakes to ensure that the data supplied to it by the NHLS is used ethically and solely for the purposes for which it is provided as detailed in this application, and further acknowledges that it shall remain liable for any breaches of this clause by the end user.
- If the purpose for the data requested in this application is research or if patient identity linked data is required, ethics approval and a one page summary of the protocol shall be attached to this application form. It is the responsibility of the applicant to ensure that their institutions' Human Ethics approval includes explicit authorisation to access the requested NHLS data.
- The applicant undertakes to store the NHLS data in a confidential manner by separating patient identifying details from laboratory data and storing the master list that links patient identifying details to study patient identifiers in a separate, secure location.
- The information is for the private use of the applicant only, unless further approval is obtained from the NHLS. In the event of this, the applicant shall give due credit, including affiliation, of the participation of the NHLS in any such publications or presentations.
- The applicant undertakes to provide the Executive Manager: Academic Affairs, Research and Quality Assurance at the NHLS with a copy of any report, presentation or publication emanating from the use of this data.

Figure 34: Conditions Specified on NHLS Request for Data Form

Special attention should be paid to the third and fourth bullet points which both specify additional requirements for patient data, because it is so sensitive. The third bullet point also makes it clear that access to information to be used for research purposes will not be allowed unless ethics clearance for the research has been granted, and that this ethics clearance includes an 'explicit authorisation to access the requested NHLS data'. The process of having to fill out the form and get permission whenever requesting data from a CDW service representative was described by all interviewees making this kind of request. For example:

"Forms have got to be filled, I understand. Remember we've got to also protect the information . . . I really respect that, and I have to abide by that . . ."

"I submitted a request form, and I also dealt with a person"

People requesting data through a customised application do not have to submit an application form every time, but they do need to 'register' and get security clearance in order to get access to the application:

"Yes, there's no request form for the dashboard data. But we have to get permission – so yeah, there is an application form that our heads have to sign off for us to get permission to access the data from the dashboard."

". . . I'm able to, you know like, go and register for . . . that report. . . . I should think when you're looking at the registration, they give the results according to usability. . . . because you have to state who are you, where are you working – you know like, it gives them a sense of what is your job

[38] The full application form is shown in appendix I

profile. So let's take that I'm not somebody that looks after babies. . . . why should you go around giving confidential data to people who don't need it? . . . I know for a fact that they restrict you to your location."

In addition, use of online applications like dashboards is password protected, and reports registered for are sent, password protected directly to the recipient's email address.

Even requests that go through NPP data analysts have to be approved first:

". . . they are not coming direct through to us – we have to go through . . . [our operations manager][39]*, who will . . . give a go ahead . . . or you have to motivate more. . . . Because some people they need it for research and . . . so* [our operations manager] *has to come in before even it comes to us."*

The NPP data analysts themselves, who do have direct access to CDW data, only have access to that subset of the data that they have been given permission for. If their job requires them to access additional CDW data, they need to go through the request process of submitting a form to obtain that additional permission:

". . . remember, NPP . . . it's got access to just a subset, not all of CDW. So if I need more . . . if another request comes for the information that . . . we don't have permissions to access – we have to fill in that form and we indicate why we need that form, what is it gonna be used for, and [our operational manager] *has to sign for it and it goes to CDW, you know they process or they reject, I think some of the, uh, requests they go through the ethics committee . . ."*

"Remember we are not actually requesting data. We are requesting to have access to tables. . . . a researcher . . . will write a request form asking for data. My different project management – NPP management – will come to me and ask data. . . . . If I don't have access to that specific table based on the requirement, then I will request – I will fill in a form and request that particular detail."

". . . even when I am looking for certain fields from our data analysts they will tell me we have got to fill certain forms that have to be signed first by our supervisor that will have then to go to CDW and then CDW would then feed back to us or would then add those fields."

All means of access to the CDW data, therefore, are guarded. Those who have been granted access in any of these ways also have to seek permission before disseminating the information to anyone else:

". . . from our own department, if there's a request, even nationally, it has to go via our operations manager, and he needs to approve whether we provide or do that."

This was experienced personally by the researcher when supporting documents in the form of reports perhaps showing the impact of the population health programs was requested:

"No, on myself I will NOT have those. Because the person who could be able to have that information, if you looking for evidence will be [the head of the NPP]. "

[39] Name removed for anonymity

"Yes, the reports they are available, but the best person that can provide that report to you it's gonna be [our operational manager] *. . . everything that . . . has to go out has to go out through him . . . But the reports are available and if you need them . . ."*

"If you go could back to [the head of the NPP] *and say, you know what, I spoke to* [program manager] *. . . Can we seek first to have the information . . . You understand? . . . I am very sceptical until we follow the right channels of doing this information flow. For me it's all about protecting both you and myself and also our integrity as a national department. . . . But if you satisfy all those processes, I'm very eager to come and share the information."*

When it comes to using CDW data for research, there is an additional concern – the NHLS ensures that the research is ethically conducted by requiring that ethics clearance is granted before access to the data is allowed. As mentioned, this is evident in the conditions on the application form shown in figure 32. This confirms the following comments made, which describe the process:

"First off it depends on the type of request that they are. . . . if ever he or she is a researcher, she has to go to the Academic Research who have to see whether he's got ethics clearance and they have to see whether whatever he is asking is in line with their processes. . . . And then once they have cleared that research then they will send it to us . . ."

". . . for our research projects . . . we go via [the academic affairs person]*, we fill in a data request form, and . . . we have to provide the ethics approval and then we say what years of data we require, the variables we require - and then we give it to somebody in CDW and they are supposed to draw exactly what we have said to them . . ."*

For the sake of the completing the discussion on privacy and security it is worth mentioning again that there is a hierarchical system of security that applies to the access of billing information that is stored on the CDW, which ensures that people only have access to the billing information that is pertinent to them:

"The billing – we also do financial, but the billing part for all the tests that happen . . . We have a BI tool where the DOH can go and check their accounts and their invoices and their CSVs – a breakdown of that reports by transaction. . . .you can only see your own report, your own account, so you can't see other provinces. . . like, for Western Cape we send them billing for those who request billing for the whole of Western Cape, . . . but then certain individuals, even if they are in Western Cape they can only see specific accounts / facilities, and they can't see all of Western Cape."

In summary, NHLS staff was found to be mindful of privacy and security concerns and to adhere strictly to the resulting restrictions. Through sometimes frustrated with the bureaucracy, they ensure that all pathways to the CDW data are guarded.

Privacy and security of the NHLS data can only be adhered to if all members of the CDW 'community' follow the organisational rules or restrictions. The 'Privacy and Security' theme therefore corresponds to the 'rules' sub-triangle of the activity system for an information request, as shown in figure 33.

The discussion on 'Privacy and Security' thus concluded, the next them to be discussed is 'Direct Access'.

5.5.8 Direct Access

This theme was placed in the sub-triangles of the activity system for an information request shown in figure 35. This choice will be explained once the data for this theme has been presented.

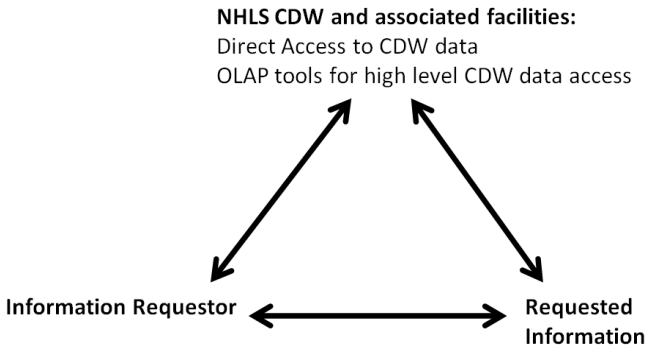


Figure 35: Sub-triangle of the Activity System for 'Direct Access'

Information requestors interviewed – with the exception of the two NPP data analysts – never access the CDW data directly, but rather either go through a customised application or a CDW service representative. Previous literature has shown that personnel in a health planning agency benefited from being given direct access to a data warehouse with an on-line analytical processing (OLAP) tool (Tremblay et al., 2007). Therefore interviewees were questioned on whether they thought they would benefit from having a flexible means of accessing the CDW directly.

The two NPP data analysts do access the CDW directly, but through a low level CDW access tool called SQL, which cannot be used without some technical expertise. Having direct access to CDW gives the data analysts the flexibility to manipulate the data into a form required by people requesting information from them:

> *". . . when they're asking for data they don't really want all the fields that have been specified. They want us to aggregate something based on this, or get the average, or get the mean, or get the difference between the turnaround times. So all those things we have to work around from our side. We don't send a query that side to say 'Give us the in-lab turnaround time' when we know we've got a register date and the review date."*

> *"So you as a researcher, or as maybe, as let's say, one of the doctors, he will write a request form asking for data. My different project management – NPP management – will come to me and ask data. And I'll have expected to get into the CDW and extract that data, analyse it, put it into an appropriate formatting."*

The interviewees serviced by the NPP data analysts are all program managers served by these data analysts and none of them felt that any advantage would be gained for them by personally having direct access to the CDW data. This may be because they are achieving their desired level of flexibility through the direct access that the data analysts have. The interview with the operational manager in a different department also revealed no need for direct access.

Staff at the NICD and NCR, however, felt that direct access to CDW data would be an advantage, provided they received training to enable them to use it effectively:

"Oh if we had a choice of actually going in there and choosing . . . It would be nice, but we need to be trained on actually how to extract the data from the CDW platform that they use or the tables that they use to get the data."

"Definitely. Provided that there were explanations of what data means. So, what does a test code mean? What does a result code mean?"

"Yes. Absolutely! We've been trying to get this with them (deep sigh) for ages. . . . As a medical practitioner, I understand the need for confidentiality. However, the restriction on staff accessing CDW data is absurd. (laughs) . . . why should I have to go through somebody at the CDW as a staff member of NHLS to be able to draw data that I need? Why can't [we] have direct access to the CDW and draw a dataset? . . . all my staff sign a non-disclosure agreement. It has been agreed by the regulation that we have access to that data, but we have to go through a third party in order to pull the data. . . . Like I said, I fully understand patient confidentiality, but this is not about patient confidentiality, this is about an administrative approach to it . . ."

The interviewee who made this last comment is obviously exceptionally keen to get direct access to CDW, and very frustrated that the way in which privacy and security is being managed is preventing this from happening.

Many large organisations make use of 'data marts' to give departments direct access to that part of data warehouse data that is relevant to them (Inmon, 2005). This involves keeping a copy of just that part of the data warehouse data in a separate mini data-warehouse, and giving the staff direct access to this separate copy. Talk of plans to extract data from CDW and build data marts for NICD and NCR was mentioned:

"Ya, well – I think we are moving towards that direction. So we are busy with micro-strategy part of NICD that is pulling data from CDW for us – NICD only. So there what we're going to do is we are going to build our own sub-warehouse from the NICD one that feeds into us, so that we are able to manage ourselves. So if I wake up and say I need this kind of data, I'm able to write a code and draw it up for myself and export, extract it, use it"

". . . they constantly talk about a sandbox – like they can dump our data into a sandbox . . . so we access the data that they give us and we can play around with it . . ."

Staff without technical expertise would need a high level CDW data access tool like OLAP to get to data in the CDW or in a data mart directly. It is surprising that in the NHLS CDW environment OLAP tools are not used by the end users for more flexible access to the data, but rather are only used as part of the highly technical Extract-Transform-Load (ETL) process to bring new data into the data warehouse. The following was the response of a CDW member of staff when asked about the use of OLAP tools with CDW:

"On the ETL part. We do the development for the whole integration."

And from a data analyst from the NICD:

"No, it's part of the BI – Business Intelligence – no we cannot, I cannot use it on my part, because that needs people who are building the warehouse itself, but me, I haven't used it. I analyse data, so that, for me, it's about using statuses and due reporting, so . . ."

The difference between interviewee responses of NICD and NCR staff on the one hand and NPP and operational management is again worth noting. Whereas operational staff and NPP program managers see no need for direct access to CDW data, NICD and NCR staff are keen.

Direct Access to CDW data would enhance the toolset available for information requestors to get the required data. It would therefore belong in the 'tools' sub-triangle of the activity system for an information request, as shown in figure 35.

This concludes the discussion on the 'Direct Access' theme. The next theme to be discussed is 'Metadata'.

5.5.9 Metadata

This theme was placed in the sub-triangles of the activity system for an information request shown in figure 36. This choice will be explained once the data for this theme has been presented.

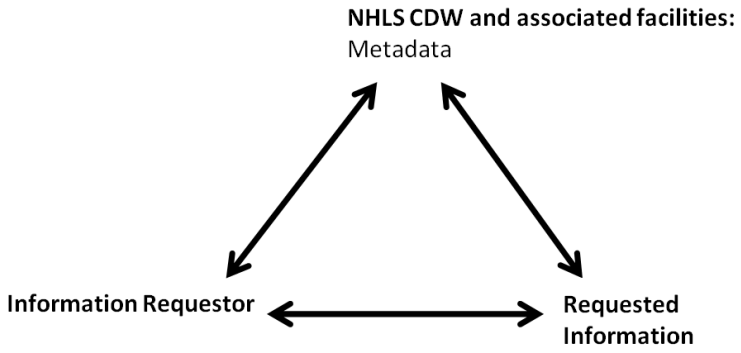


Figure 36: Sub-triangle of the Activity System for 'Direct Access'

The term 'metadata' refers to data about data. In a data warehouse environment, there is a broad range of technical and organisational metadata that documents the structure of the data, the source of the data and how it has been transformed, and the meaning of the data (Inmon, 2005). It is this last element of metadata, the meaning of the data, which is relevant in this study.

The idea of metadata was not originally incorporated in the interview protocol, and was only included after the third information requestor interviewed indicated that difficulty had been experienced in understanding the data. Different laboratory tests done are identified by codes that are not easily understood initially, and time had to be spent on finding out what the different codes meant:

> *" I did find it challenging because I'm not familiar with the codes and things that are used and so I had to first learn about how the coding is done in the laboratory, what tests are being given, what codes when they are entered into the system – to be able to extract, you know, the right data."*

This person was requesting the information through a CDW service representative.

This challenge was not experienced at all by requestors accessing the data via a customised application, as appropriate laboratory tests have been chosen and presented meaningfully for them:

". . . it's clearly stated. Like when you go for HIV reports, they'll tell you exactly – you choose."

A requestor working through a data analyst also expressed a need for information about the meaning of the data in CDW:

"I don't think I'll know from the raw data because it's all very technical, and there's a whole lot of things – they pull time, they pull age, they pull facility or location of the patient. They just pull a whole lot of things but the way they – sometimes they have to explain it to us, because there's also different barcodes that they pull from the clinic, from the hospital and the local barcode internal. . ."

Some requestors felt there was no need for an explanation of the data, as they had brought across an understanding from previous experience at the NHLS:

"I've got previous knowledge – I've worked actually in a lab before, so . . ."

" I was manager in the lab. . . . So when I came to this program management I came with that lab background and understanding . . ."

" I've worked from the receiving office, and that kind of information I expect to go all the way to CDW . . ."

Metadata, it seems, is mainly needed by those employees that are newer and don't have previous experience working in other parts of the NHLS. Providing metadata to new employees will certainly help them, but more significantly, documenting this information is important from an organisational point of view, as pointed out by one of the interviewees who feels it is important to have metadata even if staff do currently understand the data:

". . . if you think about institutional memory, we cannot rely on the individual's history . . . to allow systems to work. . . .we've had that problem, where skills, knowledge, is invested with one person and then that person leaves, all that knowledge and those skills go with that person. So, yes, right now all of us are very well aware of the variables that, but it's not how an organisation should run."

Metadata on the meaning of the data would purely be a service provided by CDW, and therefore would belong in the 'tools' sub-triangle of the activity system for an information system, as shown in figure 36.

That ends the discussion on metadata. The theme 'Ease and Efficiency' will be discussed next.

5.5.10 Ease and Efficiency

This theme was placed in the sub-triangles of the activity system for an information request shown in figure 37. This choice will be explained once the data for this theme has been presented.

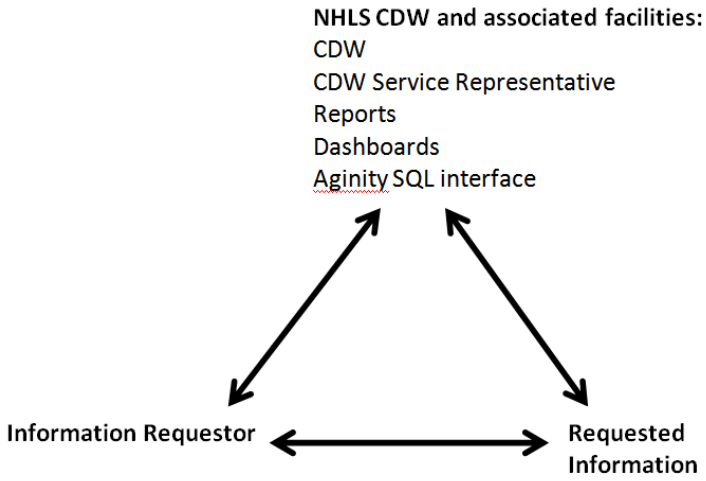


Figure 37: Sub-triangle of the Activity System for 'Ease and Efficiency'

Interaction of staff with operational information system (like TrakCare, the LIS used by NHLS laboratories) is mostly compulsory, since the job cannot be done without entering and retrieving the data required for the operational process. The extent of interaction with a data warehouse, however, may depend on how easy and efficient it is to get the required data out of it. To establish whether this was an issue in the NHLS CDW environment, interviewees were questioned on how easy and efficient they found it to get data from the CDW.

The answers to this question varied according to the way in which the information was being requested. Customised applications are perceived as being very easy to use:

"Well we can always just access it. . . . That one's easy."

"No it is, it does come on to the email, but at least if you ask, if you miss it on the email you can still request it. . . . So I don't really have any challenge in terms of getting the data if I want it."

". . . after registration, before you request a report, already on the side on that web page, they've got all different types of report. . . . And they also . . . give you a sneak preview to say this is how the report looks like. Then you can say mm-mm, no this is not what I want. Then you go to the next one, oh this is exactly what I want."

Requests made through the data analysts are also perceived to be easy, probably because the data analysts take responsibility for ironing out any difficulties:

". . . I would then ask [the data analysts] *to pull data for me, and then . . . ya, I think it works . . ."*

For requests made through a service representative, the process of filling in the form and making the request is perceived to be easy. It is the iterative process of getting the request right, as well as a lack of knowledge of laboratory codes, that interferes with ease and efficiency:

"That process is easy as well, like I said you fill out the CDW request form, and then you submit it to IT . . . say 'Please forward to CDW'. . . and then one of the data technicians will then do the extraction for you and then email you the data. Password protected obviously."

"I submitted a request form . . . I actually repeatedly went to the CDW and sat with that person to try and get the right data extracted for that project. It was difficult in the sense that I had to understand the data I was looking for myself first, and then I had to sit with the person extracting the data and explain to him . . . I did find it challenging because I'm not familiar with the codes . . ."

It seems that the process for some requests is less difficult than others:

"So the outbreak requests, I don't think they are usually as complicated as the meningitis dataset, so . . . the data it would come back to me in an email . . . I would got through it and see if it's not adequate, give feedback by email and then give a revised request."

Aside from the iterative process of getting the request right and the lack of knowledge of laboratory codes, other factors cited as marring ease and efficiency were dependant on the challenges and aspirations of the individual interviewee. One interviewee understands that it is difficult for all new people who need to learn it:

"It's not easy because I am going around the country training people on how it works. I was part of the people who – the team – who developed it. So I developed it based on what I knew, so it can't be difficult for me."

One interviewee expresses frustration with the bureaucracy, but understands that it is a necessary evil:

"It is easy though governed. Governed by policies. Forms have got to be filled, I understand. . . . But it is easy, be it within a certain governed principles . . ."

A data analyst suggests that data quality problems are a stumbling block:

"The process is not difficult . . . There are some challenges . . . Most of them is gonna be mapping of our facilities, because the information that is sitting in CDW is not really linked to the Department of Health . . . So you can find that if a facility changes its name today . . . that won't reflect . . . "

In summary, information requests that are done through a customised application or through a data analyst are perceived as easy and efficient. Even when working through a CDW service consultant it is considered to be relatively straight forward to obtain information from CDW, with only a few stumbling blocks being mentioned. These were the necessity to train new users, a mild frustration with bureaucracy, the iterative process of getting a request right, and a lack of knowledge of laboratory codes.

Ease and efficiency depends on the tools and facilities made available by CDW for accessing the information. The theme 'Ease and efficiency' therefore belongs in the 'tools' sub-triangle of the activity system for an information request, as shown in figure 37.

Having completed the discussion on the theme 'Ease and Efficiency', the theme 'External Data' will be examined next.

5.5.11 External Data

This theme was placed in the sub-triangles of the activity system for an information request shown in figure 38. This choice will be explained once the data for this theme has been presented.

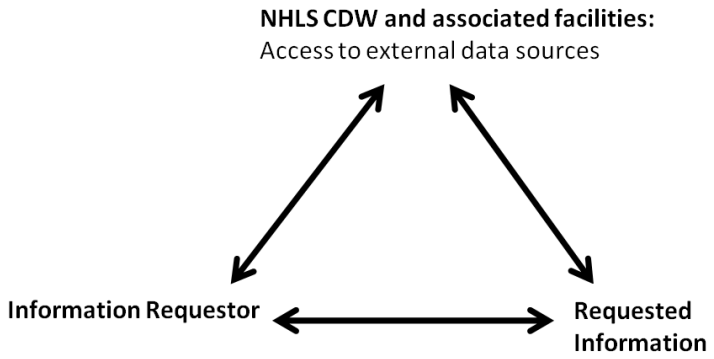


Figure 38: Sub-triangle of the Activity System for 'External Data'

During the course of the interviews it became evident that several information requestors could not fully satisfy their information needs from the CDW and had to consult external sources of data. Once again, there was a discrepancy between the needs of the NICD and the NCR on the one hand and of operational management and the NPP on the other.

As mentioned under the 'Data Quality' theme, operational management and NPP staff tend to use external data sources to verify that the data that they have received from the CDW (and sometimes to verify the data on the alternate data source aswell). They are therefore not seeking out additional data but rather data they can compare against.

In one instance an NPP program manager is using the Oracle Billing system instead of the billing data on the CDW, but there is a good reason for this. CDW data for a given month, as discussed under the 'Data Quality' theme, can take three months to become stable, while laboratory test approval processes are completed and the given month's data continues to be updated. Clients billed for tests done need to refer back to a stable reference point for monies charged. The billing data for a given month on Oracle does not change when new tests are approved, rather the tests are charged for in the following month, and ultimately the Oracle billing and CDW data for that month will be in sync:

". . . because I'm working with international donors . . . for the sake of reporting . . . if we report today and say we spent 900 000 last month and then next month you come and say "Hey, I've got 800 000 for the previous month" , then we lose credibility. So for the sake of me to bill my donors I use the Oracle one . . .CDW will always be updating, updating, updating and even though we're saying say we will update after the third month, to them it felt as if we don't know what we're doing. . . . I can do the VLookups and then show you that at the end of each month and at the end of the third month after we have closed everything, our data will be like – balancing out.

Again, this alternative data source is not additional information, but rather a different source of the same information.

Information requestors in NICD and NCR, however, require additional data to do their jobs properly. Fully understanding the extent and nature of a disease often requires mortality data, hospital and clinic data and data from the private pathology laboratories.

Mortality data makes it possible to establish the outcome of the disease and may even catch previously undiagnosed cases of the disease:

"I mean they don't have death register, so I need to go to the death register, mortuary, whatever you call it, or Stats SA, in order to find out the outcome of these patients because they didn't follow the patient until the death."

". . . the fact that we as a regulated cancer surveillance body do not have access to mortality data is a shame – it's an absurdity, because we cannot capture those patients who may not have been diagnosed during life . . . So everybody asks us 'Do you have mortality data?', and my answer's no, whereas in every other country the cancer registry's automatically linked to the mortality data, so that they get live feeds of when people die . . ."

It was also indicated that certain hospital or clinic information that is not kept in the CDW is necessary to fully understand the nature of infections and disease:

". . . obviously its limitation's it's a laboratory base of data. Often for us we need clinical information as well, for example the date of admission of a patient, and I know that is not in the CDW. So yes, often I have to go and get data from the hospitals themselves or other sources."

"We can't get all the data from CDW, for example patient admission dates and – so we would need that – for example . . . since we're in a centre for hospital acquired infections, AMR and Mycoses, we want to determine . . . whether a patient got the infection – whether it was hospital acquired or community acquired. So using a patient's admission date and a date of specimen collection we will be able to calculate whether it is or it's not. With the CDW that becomes a challenge because we don't have that. . . . we don't have diagnoses, admission dates, etc."

". . . we do enhanced surveillance at certain sites – so we collect additional information on the cases. . . . if you decide if there's an outbreak there and you wanna make that site an enhanced site you collect additional information . . ."

The NCR has implemented a separate population based cancer registry to supplement the pathology based cancer registry that they use the CDW data for. Surveillance officers are actually sent out to the hospitals to collect pertinent data on cancers that may not have been picked up through laboratory tests:

". . . what we've had to do is institute population based cancer registration, where we send our surveillance officers into hospitals and they actually notify the cancers themselves, in addition to our pathology based registration. . . . Because of the deficiencies in the pathology based registry. . . number 1 would be that not everybody would have their cancer biopsied, you know, you might be too sick, it might be a cancer that is not reachable for biopsy – so you wouldn't come through the histology – the laboratory system. So in the population based registry we are able to capture those patients. But secondly demographic information is so poorly filled in, in a pathology based registry, because you're relying on the doctor to fill in the variables . . ."

Hospital data is also externally collected for research purposes:

". . . for research, we'll go directly to the patient and get ethics approval, interview patients and get data. . . . if we are going to patient level, then we will draw the report directly from the patient file, because the pathology report would be in the patient file. . . . for research I would never rely totally on the CDW. . . . Because the information is too patchy and, like I said, the variables are not always completely filled in"

The other external data that is needed to get a full picture of the spread of disease is data from the private pathology laboratories, which test about 20% of the population. There has been some effort to incorporate some of this data into the CDW. Interviewees asserted the need to integrate private lab and CDW data:

". . . you're not getting the true picture of, for example, for AMR, because you're only reporting from public facilities, you're not reporting from private facilities. But now we're trying to get private data integrated as well, for example, we managed, for 2016, to get bloodstream infection data for escaped pathogens onto the CDW – so we're TRYING to get private data. "

". . . I've got some private laboratories, one private laboratory, that has a live feed into the CDW - every time a cancer is reported . . ."

"We request that [private lab] *data and try to match it to ours."*

Private laboratory data cannot be used for research, unfortunately, as ethics clearance has not been obtained for this. At the moment effort is being made to obtain this data from another source:

". . . for research particularly we don't include our private sector data because we won't have ethics for that. What we are looking at is to access data from the council of medical schemes, from other medical schemes, but that would be purely for research purposes."

In summary, sometimes it is necessary for NICD and NCR staff to supplement the CDW data with mortality data, hospital or clinic data, and pathology data from private laboratories. Once again,

requirements for NICD and NCR staff are different to those for operational management and NPP staff.

Access to these external sources of data through the CDW – as is currently the case with some of the private laboratory data – would effectively form part of the tools and facilities offered by the CDW. It therefore belongs in the 'tools' sub-triangle of the activity system for an information request, as shown in figure 38.

This concludes the discussion on the 'External Data' theme, and with that the presentation of all data pertaining to research question 2 is completed. Data concerning research question 3 can now be presented.

5.6 RQ3: What are the individual outcomes for the NHLS CDW data requestor?

The purpose of this research question was to determine whether the individuals requesting data from the NHLS CDW felt that the data thus acquired was helping them to do their jobs. It corresponds to the part of the activity system for an information request shown in figure 39. The 'Information Requested' and now received from the CDW is the 'object' of the activity which leads to an 'outcome'. Research question 3 examines that aspect of the 'outcome' of an information request from the NHLS CDW that concerns the individual information requestor.

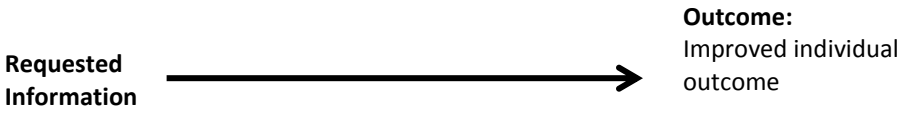


Figure 39: Part of the Activity System for 'Individual Outcome'

In fact, CDW Data was considered by information requestors to be an essential tool in the accomplishment of their work. So much so that written acknowledgements are sometimes sent to CDW staff:

> *". . . you've got other people which maybe write some notes thanking us . . . because we find that the person has got this responsibility to execute in population health, and he solely depend on us . . . and we working hand in hand with them."*

One requestor reminisced about what it was like in the days before the CDW was made available:

> *". . . I've been in the program for too long, and data it might have not been in the CDW, but data has been there. You know like when I first worked . . . data was collected on a manual basis and they were able to . . . give me the numbers . . . Since we have the CDW it's nicer because you don't have to sit there and wait for somebody to give you data. . . . for CDW I can really give them a compliment to say . . . I know where I had to count one for one . . . to now having to go to a click of a button."*

Individual outcomes mentioned by CDW data requestors fell into the same categories that the CDW uses fell into: public health tracking, administration and management, and research. Indications of positive public health tracking outcomes for individuals were:

"it gives you . . . a broad picture, to say that ok this is where we're sitting . . . and how does this impact on the end product, which is the baby."

"In terms of outbreak investigations . . . it is a quick and easy way to see how things have been in the past couple of months, for example, and now with the outbreak to see an increase in the number of cases. So it is an easy way to get your – what we call a baseline"

"For me to know the number of people tested at a centre, you know, and what their values were per any given month helps me to understand the response to the ARVs . . ."

" . . . we can get the total number of . . . specimens that were tested, so we can calculate the positivity rate, . . . and then you'll know. . . whether there has been an increase in the resistance or a decrease in the resistance for a particular antibiotic or whether there has been increase in first line antibiotic resistance or has there been an increase in second line antibiotic resistance."

CDW data is also essential for the management of laboratory operations and national priority programs. Asked whether they benefit from CDW data, a member of the operational staff said:

"Ya, it does, because you can make decisions out of that, what action to take. . . . if you don't get this data you will not know where the performance is. Are we making it or we not making it and what are the challenges . . . and so that you can mitigate. You know it's sometimes embarrassing for you to be told by the client – you should have been aware before they even are aware of that. So . . . it puts you at least to know exactly what's happening in the organisation."

This operational member of staff finds CDW data invaluable in managing laboratory throughput, performance, wastage control and quality control, as well as in other management activities such as budget planning and troubleshooting. (These uses of the CDW for operational management were all discussed under research question 1.)

NPP staff also felt that CDW data helps them to run their programs better:

". . . it's advising me on how to run the program better, so that the people can get best quality results at the end."

". . . everybody is able to identify and to say that "What is it that we're doing right as clinicians, what is it that we're taking for granted, but impacting on the patients' lives"? And it . . . helps to review the program in a very positive way . . ."

Program managers can use the data to determine where the problem areas are so that they can target specific hospitals or clinics with training:

". . . I'm glad though, that with the help of the data analysts they . . . generating . . the rejection lists. . . . that information, to me is what I want, because when I talk to my clients to say, if . . . it's because the specimen is insufficient we need to train you to collect the proper specimen . . ."

For training that is being done, data from the CDW helps to determine the impact that the training has had on the conduct of clinicians:

"And it also helps me when I'm doing training, to say it gives me a baseline. So say this where we started, and how does this impact on the uptake after training. So I'm able to say ok, you were doing 10 tests before we trained, and the next year the same time I want to see the impact."

Factual data from CDW can also be used to refute any denial of inadequate performance on site:

"If I get a challenge and I know for a fact that somebody will say 'No I've been doing well', or you know there's going to be an argument. I request this data and say 'Let's look at the data.'"

Sentiments about using CDW data for research are not quite as positive. The CDW is viewed as a useful resource for research, but one that needs be supplemented :

". . . it is limited to what information is in the laboratory, but I think it is a useful resource."

" More often than not we will go directly, for research, we'll go directly to the patient and get ethics approval, interview patients and get data. . . . then we will draw the report directly from the patient file, because the pathology report would be in the patient file. . . . So for research I would never rely totally on the CDW."

CDW data is particularly important to a member of staff who perceives their livelihood as being dependent on it:

". . . to me as a person, and that the guys that with whom I'm working, because we getting funded by the international donors. We don't get our salary from NHLS directly. . . . without CDW I would be irrelevant and the Global Fund wouldn't see it in me to run with me. (Laughs) . . . for me, indirectly, I need the CDW in order for me to be relevant even going forward."

In summary, aside from the fact that the CDW enables NHLS staff to do their work in the areas of public health tracking and operational and NPP management, it provides support for those doing research and relevance for those relying on international funding for their salaries.

Certainly staff appreciate the value of the information that they receive from the CDW:

"But you see CDW for me, it's a book that I'm reading, and I'm eager preaching as it sends the information to me and it helps me . . ."

"It is very rewarding, you know, to know where we coming from and to where we at now. And for me . . . I used to work with [high level professor]. She used to say to me no no no. This is our record."

Taken together, the findings suggest that the individual impacts can be summarised in terms of the contribution of CDW to improvements in each the following aspects of the individual's working life:

- Productivity
    - "*it is a quick and easy way"*
    - *"you don't have to sit there and wait for somebody to give you data"*

- Decision making effectiveness
    - "*helps me to understand the response to the ARVs"*
    - *"because you can make decisions out of that, what action to take"*
    - *"it's advising me on how to run the program better"*
    - *" because the specimen is insufficient we need to train you"*

- Ability to track progress
    - *"it gives you . . . a broad picture, to say that ok this is where we're sitting"*
    - *"it . . . helps to review the program in a very positive way"*
    - *"So say this where we started, and how does this impact"*

- Sense of meaningfulness of work
    - *"CDW for me, it's a book that I'm reading, and I'm eager preaching"*
    - *"It is very rewarding, you know, to know where we coming from and to where we at now."*

Discussion of all data pertaining to research question 3 is thus completed, and data concerning research question 4 is now presented.

5.7 RQ4: What are the perceived benefits (impacts) of NHLS CDW use for the NHLS as an institution and for population health in general?

Research question 4 was intended to unveil any positive perceived impacts that the NHLS as an institution and that population health in general has experienced as a result of the NHLS CDW. It corresponds to the part of the activity system for an information request shown in figure 40. This part of research question 4 examines that aspect of the 'outcome' of an information request that concerns the NHLS.

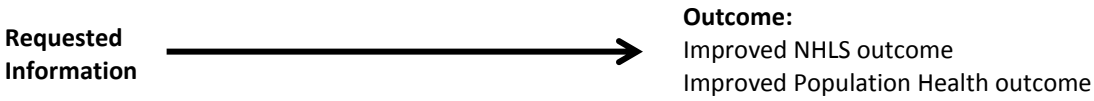


Figure 40: Part of the Activity System for 'NHLS and Population Health Outcome'

The CDW was found to provide support for a range of areas ranging from internal operations to relationships with international donors and the DoH.

It should be noted at the outset that benefits for individuals working for the NHLS discussed under research question 3 mean that they achieve their working objectives more effectively and efficiently, which ultimately benefits the institution as a whole. This is particularly the case with operational management where, as mentioned under research question 3, CDW data is being used to ensure that all 260+ NHLS laboratories around the country run smoothly by managing laboratory throughput, performance, wastage control and quality control.

CDW data also enables the efficient running of the NPP programs:

". . . internally I think it gives us an idea of the work that we do, the volumes, what we are producing."

". . . without CDW I will say (NHLS) *won't be able to perform some monitoring and evaluation."*

Published research based on CDW data upholds the reputation and standing of the organisation as a whole. Certainly many research articles acknowledge CDW as a source of data. One information requestor speaks proudly of research achieved in their department:

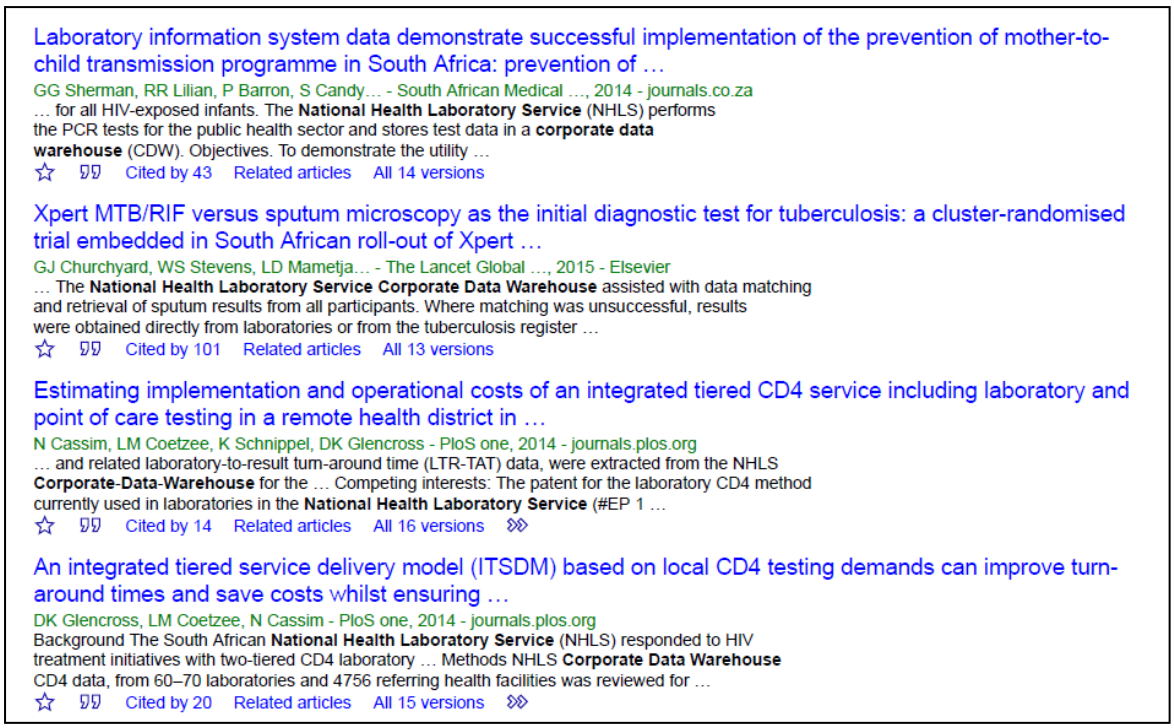
"We've also published papers on breast cancer and prostate cancer. We've published, in Lancet Oncology, looking at our plans for cancer registration and the revitalisation of the NCR. . . . With our HIV Cancer Match study we are creating a virtual cohort of HIV positive patients for research purposes – not just cancer related research, for any research purposes. It will be the largest virtual HIV cohort in the world, because we have the biggest burden of HIV. It is the first time anybody's doing this with CDW data . . ."

"We've got a large research project at the moment that is called the SA HIV Cancer Match study, and we have requested all HIV testing data from the NHLS for a certain time period, and that is in the region of 40 million records."

As mentioned by a member of CDW staff:

". . . some do acknowledge us in terms of what they have achieved – publications . . ."

This can be verified with a simple search on ("national health laboratory service" "corporate data warehouse") in Google scholar. On 24th March 2018 this search returned 87 items. Figure 41 shows the first four results returned. As can be seen from this, there are many research articles acknowledging the NHLS CDW as a source of data. Similar acknowledgements can be found in research articles listed in table 9 as supporting data.


Laboratory information system data demonstrate successful implementation of the prevention of mother-to-child transmission programme in South Africa: prevention of …
GG Sherman, RR Lilian, P Barron, S Candy… - South African Medical …, 2014 - journals.co.za
… for all HIV-exposed infants. The **National Health Laboratory Service** (NHLS) performs the PCR tests for the public health sector and stores test data in a **corporate data warehouse** (CDW). Objectives. To demonstrate the utility …
☆ 99 Cited by 43 Related articles All 14 versions

Xpert MTB/RIF versus sputum microscopy as the initial diagnostic test for tuberculosis: a cluster-randomised trial embedded in South African roll-out of Xpert …
GJ Churchyard, WS Stevens, LD Mametja… - The Lancet Global …, 2015 - Elsevier
… The **National Health Laboratory Service Corporate Data Warehouse** assisted with data matching and retrieval of sputum results from all participants. Where matching was unsuccessful, results were obtained directly from laboratories or from the tuberculosis register …
☆ 99 Cited by 101 Related articles All 13 versions

Estimating implementation and operational costs of an integrated tiered CD4 service including laboratory and point of care testing in a remote health district in …
N Cassim, LM Coetzee, K Schnippel, DK Glencross - PloS one, 2014 - journals.plos.org
… and related laboratory-to-result turn-around time (LTR-TAT) data, were extracted from the NHLS **Corporate-Data-Warehouse** for the … Competing interests: The patent for the laboratory CD4 method currently used in laboratories in the **National Health Laboratory Service** (#EP 1 …
☆ 99 Cited by 14 Related articles All 16 versions »

An integrated tiered service delivery model (ITSDM) based on local CD4 testing demands can improve turn-around times and save costs whilst ensuring …
DK Glencross, LM Coetzee, N Cassim - PloS one, 2014 - journals.plos.org
Background The South African **National Health Laboratory Service** (NHLS) responded to HIV treatment initiatives with two-tiered CD4 laboratory … Methods NHLS **Corporate Data Warehouse** CD4 data, from 60–70 laboratories and 4756 referring health facilities was reviewed for …
☆ 99 Cited by 20 Related articles All 15 versions »

Figure 41: First Four Google Scholar Results for Search on ("National Health Laboratory Service" "Corporate Data Warehouse")

The NHLS's main customer is the DoH, on both national and provincial levels.[40] The contribution that the NHLS makes to the success of DoH is therefore an important measure of NHLS success, and the facilitation of this contribution by the CDW certainly constitutes a benefit to the NHLS. Information requestors were of the opinion that CDW data enables the NHLS to support the DoH to fulfil its responsibilities:

". . . it's not only us as NHLS that is using this data. People from National, they depend on this data to check how the country is doing, because they'll have data that they had from their provinces, but ultimately they will use this data to check – cross-check . . ."

"We are the surveillance arm of the National Department of Health."

"Even some of the reports we send them to the department as well . . . when they want for purposes of operations . . . Or they want . . . to check what is the status of the health. . . So it's formative in that for them to be able to check if their programs are working or what, or patients are complying, so it looks at number of factors across here."

". . . the Department also do request the information to use for their own analysis and also preparation for other, like diseases, to see how - what are we getting."

". . . the outcomes of the data have had a positive impact on . . . the ability of the Department of Health to do its job."

[40] http://www.nhls.ac.za/assets/files/NHLS_Strategic_Plan_2010-15%20for_DOH.pdf

Interviewees also perceive the CDW as helping the DoH to improve its effectiveness. One requestor spoke of the data providing grounds for running targeted campaigns in problematic areas:

". . . so we are aware, and then we also advise the client . . . can view which sites are problematic . . . so they can identify hot spots and do campaigns in those regions."

DoH can also revise their plans as a result of better understanding of the information. One requestor explains how statistics derived from CDW data showed high TB drug resistance rates in a particular area, and how understanding the origin of those statistics helped to identify the fact that most of patients with TB drug resistance were actually coming in from across the Swaziland border:

"We know that the Rifampicin[41] resistance rate on average is about 3 to 4%. . . . we have learnt that there is a place in the Swaziland-South African border where the Rifampicin level has always been up over and around 11%, and we've even visited the place to try and find out – why is it so? . . . Only to find that these are the people that are coming in and going out . . . So it helps the health people to plan around. So . . . we will be able to inform the Department of Health, to say 'Look at the cross border movements - we've seen a lot of TB and Rifampicin resistant TB in this area – in the Ermelo area'."

Another requestor explains how CDW data led to changes in the actual practices of DoH to promote more accurate test results:

". . . we used to do our tests at six weeks. Then . . . with that data, when we went to the technical working groups at National, (colleague) *and I were able to give them this data to say that . . . we're not winning. We need to go to a birth testing. So . . . because now they are looking at data, they can see the numbers, they can see the impact to say that the babies are negative at six weeks but progressively they become positive. Meaning that that we used to give them an anti-retroviral before we test them. So those were false test results. . . . "*

Vulnerable babies are now given the test at birth, which is perceived to have an impact on population health. One of the supporting documents collected and shown in table 9 on page 65, a report on "Early infant diagnosis of HIV infection in South Africa: 2008 to 2010"[42], showed an improvement from 14.6% of babies testing positive in 2008 to 7.9% of babies testing positive in 2010. At this stage tests were still not being done at birth, but it seems to have since further improved, and it is CDW that is allowing them to track the progress:

". . . quite often my report to say mother to child transmission it was sitting at 3% last year at the same time, but this year it is sitting at less than 2%, maybe 1.5. So, at the plans you can see that ok we're making a difference."

One of the requestors sums up the contribution of the NHLS CDW to the DoH as follows:

"Through this data, this is how the minister measures."

[41] Rifampicin is a drug used to treat TB
[42] The CDW was acknowledged in this report as the source of the information

The NHLS CDW also plays an important role in the enhancement of the NHLS's reputation on an international level, which has particular significance given its relationships with international bodies like the WHO, the Global Fund, the CDC and others. CDW data is well received at international meetings, and also helps to refute misconceptions:

". . . when you go to other countries, you know like you tell them about the data from your slides, everybody gets so intrigued and say 'How! You guys, you getting everything on a silver platter!'"

". . . when we stand and we speak in public, when we're out internationally and we're making presentations, we know that the TB positivity rate . . . is not as alarming as we had thought previously, because information is flowing."

New approaches adapted due to decisions based on CDW data are also shared with international teams working on the sites:

". . . I would just say that it's a change in approach of doing business . . . but remember it is being shared amongst us, as NHLS, and the Global Fund teams that . . . it's coming from this because this data here"

The Global Fund also receives NHLS CDW data on a regular basis and uses that information to determine whether the funding has been constructively used and should be renewed:

". . . at the end of each grant, there's what called Requirement Close Up. And Requirement Close Up it summarises all of the work done within the grant period. It incorporates all the data and it talks to what I've been doing in the grant period, and then it then talks about whether we achieved the goals or there is still work that needs to be done. So that has flowed from NHLS . . . on an ongoing monthly basis, to the national TB department of the TB cluster under the Global Fund of the work that we're doing. And then it was then summarised to be a direct 6 months summary of the work that was done, the input and whether we've cracked it, or whether there's still work to be done."

CDW effectively documents NHLS success, and almost guarantees that the relationship with the Global Fund, and consequently the associated funding, will be renewed:

". . . we are the service provider, and we've been the service provider of choice to the Global Fund, . . . and over the years since I've started to work we haven't really failed them. And that is why . . . they never even doubted to say 'You know what, we're running with you guys again because your support to us is of serious value that we can see . . .' . . . in all the reports the word CDW is appearing"

There is a perception that reports from the CDW impact positively on population health through influencing policies, as shown by the following comment:

"I think it does impact in our health as a whole – or the National Department of Health, cos that's how we communicate with them. We give them reports that influence their change in terms of policies . . . Crypto now is part of the casket of care nationally. . . . everything will have to have

stock of Voriconazole in their clinics so that they treat those patients when they get diagnosed, so it's how far we have gone."

This refers to one of the applications of the CDW, "Use of the CDW to Track people with CD4 less than 100 for Cryptococcus Meningitis", discussed under research question 1. As mentioned, HIV/AIDS patients with a CD4 count of less than 100 were found to be susceptible to Cryptococcus Meningitis, which can be deadly if it is not treated in time. Consequently, an "*agreement . . . with NDoH and again with WHO that crypto should fall within the casket of care of HIV and TB"* led to automatic testing for Cryptococcus antigens for any patients with a CD4 count of less than 100, so that patients testing positive could be managed. This was perceived to have had a positive impact on both mortality and morbidity, two important population health indicators:

". . . if CD4 less than 100 automatically that blood is sent for CrAg testing. So all those with positive CrAg will be managed immediately. . . . you are now saving a patient who would be sent back . . . home and . . . can come back within 21 days . . . whereas the patients it's really having underlying stage 4 diseases that are not attended to. So this is saving . . . not only more lives – cost effective in terms of managing the burdens . . . mortality and morbidity . . . because those patients will be coming back and forth from clinic to clinic, ya, until they pass on."

Two government health policies that acknowledge the NCR, whose data is derived from the CDW, for its help in their formulation, are the 'Breast Cancer Prevention and Control Policy' and the 'Cervical Cancer Prevention and Control Policy", both of which are included as supporting documents and shown in table 9 on page 67. The contribution of the NCR to population health is mentioned:

"The National Cancer Registry is the <u>only</u> cancer surveillance organisation in the country. So if we did not perform our function there would be no national statistics for cancer in the country, so one would not be able to do policy, one would not be able to develop programs for cancer management. In fact, the most recent cancer policies, which are the Breast and Cervical Cancer Policy, as well as the National Cancer Control Plan, has utilised the data produced by the cancer registry . . . the data have had a positive impact on – for cancer of the population – for cancer patients, for policy with regards to cancer"

Though perhaps their comments were speaking more to the contribution of the CDW to the NHLS as an institution rather than to population health in general, interviewees all answered "yes" to the question that asked whether their work was benefiting population health – which indeed it would be, if not directly. One of the requestors expressed the sentiment a little more clearly:

"So for me that is very excellent and it's all, all coming from CDW. So you can see it's . . . wealthy - it's got the wealth of information that we are using to make sure that the nation at large is benefitting from that central data warehouse."

In summary, the NHLS CDW is perceived by information requestors to be benefitting population health. Specific examples of how the CDW data led to new practices and policies that were perceived to impact on population health were for mother to child HIV/Aids transmission, the timely

treatment of Aids patients for Cryptococcus Meningitis and the use of the NCR data to inform health policy formulation.
The combination of these findings imply that the perceived benefits to the NHLS can be summarised in terms of the contribution of CDW to improvements in each the following aspects of the NHLS as an institution:

- Management of Internal Operations
    - ensure that all 260+ NHLS laboratories around the country run smoothly
    - *"gives us an idea of the work that we do, the volumes, what we are producing."*
    - *"without CDW . . . won't be able to perform some monitoring and evaluation."*

- Provision of Data for Research
    - *"It will be the largest virtual HIV cohort in the world . . . with CDW data"*
    - *"acknowledge us in terms of what they have achieved – publications"*
    - Google Scholar search revealing research articles acknowledging NHLS CDW

- Fulfilment of National Mandate
    - *"People from National, they depend on this data to check how the country is doing"*
    - *"We are the surveillance arm of the National Department of Health."*
    - *" for them to be able to check if their programs are working . . . or patients are complying"*
    - *"data have had a positive impact on . . . the ability of the Department of Health to do its job."*
    - *"preparation for other, like diseases, to see how - what are we getting"*
    - *"identify hot spots and do campaigns in those regions"*
    - *"it helps the health people to plan around"*
    - *"Through this data, this is how the minister measures."*

- International Credibility
    - *"when you got to other countries . . . tell them about the data from your slides, everybody gets so intrigued"*
    - *"when we're out internationally we know that the TB positivity rate . . . is not as alarming as we had thought previously, because information is flowing"*
    - *"a change in approach of doing business . . . shared amongst the Global Fund teams"*

- Ability to Secure Funding
    - *"Requirement Close Up it summarises all of the work done within the grant period . . . that has flowed from NHLS* (CDW)*"*
    - *"they never even doubted to say 'You know what, we're running with you guys again because your support to us is of serious value. . .' . . . in all the reports the word CDW is appearing"*

- Ability to Make Informed Changes to Practices and the Ability to Formulate Policies.

- *"used to do our tests at six weeks. were able to give them this data to say that . . . we're not winning. We need to go to a birth testing."*
- *"Crypto now is part of the casket of care nationally"*
- Breast Cancer Prevention and Control Policy
- Cervical Cancer Prevention and Control Policy
- National Cancer Control Plan

Presentation of the findings for research question 4 is now complete.

5.8 Chapter conclusion

This chapter has presented the research findings for the study, which has explored the use and perceived impacts, as well as the issues and challenges, of the NHLS CDW. At the start, CDW uses were listed and categorised. Then findings relating to the eleven challenges in requesting information from the CDW were discussed, with the most serious of these being associated with the themes 'Patient-oriented View', 'Data Quality' and 'Staffing'. Individual impacts of CDW use included productivity, decision making effectiveness, ability to track progress and sense of meaningfulness of work, while institutional and population health outcomes revealed were support in the management of internal operations, provision of data for research, the ability to fulfil the national mandate, enhanced international credibility, the ability to secure funding, and most importantly, the ability to make informed changes to practices and the ability to formulate policies.

Chapter 6 will interpret the research findings in the light of the literature, highlighting the main points and making recommendations where appropriate.

6 Discussion of Results

This study has explored and described the ways in which the NHLS CDW is being used, the issues and challenges encountered in this use, the individual outcomes of this use for individual users and the impact that this use is perceived to be having on population health in South Africa. Findings of the research addressed each of the following research questions:

RQ1: What are the current applications of the NHLS CDW?

RQ2: What are the issues and challenges in the use of the NHLS CDW?

RQ3: What are the individual outcomes for the NHLS CDW user and data requestor?

RQ4: What are the perceived benefits (impacts) of the NHLS CDW use for the NHLS as an institution and for population health in general?

This chapter discusses and interprets the findings that were presented in the previous chapter, supplementing insights with interview data falling under the two remaining themes, 'Full Potential' and 'Data Governance Model'. The discussion is organised by research question, beginning with research question 1.

6.1 RQ1: What are the current applications of the NHLS CDW?

Current applications of the NHLS CDW were found to be many and varied. They were revealed in two ways – either as general applications through stories and anecdotes told during the course of an interview, or as information requests specifically described by the information requestors.

General applications identified but not recognised as information requests were found to be concerned with the DoH, NGOs working with the DoH, and NHLS billing. On closer inspection these fall into two categories: those that deal with public health tracking, and those that deal with administration and management.

Applications identified as information requests are summarised in figure 42 on page 129. They have been classified in three ways – by application purpose, type of application request, and NHLS division of origin. A division abbreviation is shown after each request, where [NICD] is the National Institute for Communicable Diseases, [NCR] is the National Cancer Registry, [NPP] is the National Priority Program, and [O] is Operational Management. Each of these requests is described separately in the presentation of the findings for research question 1 in chapter 5.

	Medical Research	Administrative	Public Health Tracking
Customised Application	*(unlikely that customised application would be developed for research as each research project requires different data)*	Monitor laboratory & regional performance through turnaround[0] Prevent wastage by monitoring rational use of laboratory tests[0] Monitor volumes of laboratory tests done[0] Monitor rejections due to incorrect or insufficient samples[0] (by area manager and program manager, so 2 requests)[0][NPP] Monitor performance of laboratories at correctional services[NPP]	Track people with CD4 less than 100 for Crypto-Meningitis[NICD] Track anti-microbial resistance (by 2 epidemiologists, so 2 requests)[NICD] 'Continuous flow' of data from CDW to NCR for cancer tracking[NCR] Track HIV status of babies of HIV positive mothers[NPP] Monitor missed diagnostic opportunities[NPP] Track response to ARVs in correctional facilities[NPP] Track TB positivity and drug resistance in correctional facilities[NPP]
Service Representative	Request information required for Mycology research[NICD] Request Meningitis tests data for a research project[NICD] Request data for SA HIV Cancer Match Study[NCR] Request data for other research on cancer[NCR]	Get more detail on monthly report data to find problem[0] Request projections for planning next year's budget[0] Pinpoint slow episodes at correctional services laboratories[NPP]	Get baseline data on normal disease incidence to verify outbreak[NICD] Investigate disease outbreaks[NICD]
Data Analyst	*(This does occur, but was not included in the sample)*	Monitor performance of laboratories serving correctional services[NPP]	Request EID information for specific time period for trend analysis[NPP] Request viral load test results for women of childbearing age[NPP] Monitor TB positivity and drug resistance rates for GeneXpert tests[NPP]
High level CDW Access Tool	*(High level access tools like data warehouse browsers and OLAP were not found to be used directly against the CDW to satisfy information requests)*	*(High level access tools like data warehouse browsers and OLAP were not found to be used directly against the CDW to satisfy information requests)*	*(High level access tools like data warehouse browsers and OLAP were not found to be used directly against the CDW to satisfy information requests)*
Low level CDW Access Tool	Request data for medical research done by NPP team (by 2 Data Analysts, so seen as 2 requests sampled)[NPP]	Request data for admistration as required by NPP team (by 2 Data Analysts, so seen as 2 requests sampled)[NPP]	Request data for public health tracking as required by NPP team (by 2 Data Analysts, so seen as 2 requests sampled)[NPP]

Figure 42: Information Requests Categorised by Application Purpose, Information Request Type, and Division of Origin

6.1.1 NHLS CDW Applications Classified by Purpose

Information request purposes were classified as Public Health Tracking (12 applications), Administration and Management (10 applications), and Medical Research (4 applications). Applications pertaining to each of these will be discussed in turn.

6.1.1.1 NHLS CDW Applications for Public Health Tracking

The application purpose with the highest number of information requests assigned to it was public health tracking, which stands to reason given the aims of the NHLS to contribute to a healthy nation by detecting, managing and preventing disease. This corresponds with findings of the literature review, where the majority of applications found dealt with surveillance. Surveillance applications identified in the literature were predominantly concerned with early detection of communicable disease outbreaks, eg. Rizi and Roudsari (2013), Tsui et al. (2003), Silva et al. (2013), often reflecting concern over the threat of biological warfare with agents like anthrax, plague, tularemia, and smallpox, as mentioned by Tsui et al. (2003). The bulk of NHLS CDW surveillance applications discovered, however, are focused on the detection and management of HIV/Aids and TB, which is not surprising given the local priority of dealing with the very high incidence of these diseases in South Africa.

Albeit not its main function, the CDW is involved in the detection of disease outbreaks in a few ways. Firstly, it is responsible for ensuring that the DoH is alerted within 24 hours to laboratory test results indicating incidence of any of the category 1 notifiable diseases. Information requestors also identified applications that use CDW data to investigate outbreaks reported by medical facilities or practitioners. Though not part of the scope of this research, other means of surveillance are done by departments that were not interviewed. A notable example is GERMS-SA, which does surveillance through clinical microbiology laboratories, and makes use of the CDW.[43] An examination of all public health surveillance done under the NHLS umbrella could well be a subject for further research.

The two epidemiologists interviewed both reported on applications that track anti-microbial resistance at the different healthcare facilities around the country – this is akin, perhaps, to the CARP application found in the literature that combines data from three hospitals to report trends in hospital infections due to anti-microbial use (Wisniewski et al., 2003). In addition, though not specifically identified as an information request, reference was made by one of the epidemiologists to the contribution of CDW data to the Global Antimicrobial Surveillance System (GLASS).

One application found in the literature, the Comprehensive Assessment for Tracking Community Health (CATCH) tracks community health in general through collecting hundreds of health and social indicators for each local community in Florida, USA (Berndt et al., 2000, Berndt et al., 2001, Berndt et al., 2003). Where the NHLS CDW is unique is that it has automatic access to all public laboratory test results, which accounts for 80% of the population of South Africa – ample information without the collection of other health indicators.

[43] http://www.nicd.ac.za/index.php/germs-sa/

It is interesting to note that none of the surveillance systems in the literature used laboratory test results as the source of the data for their disease tracking. Rather there were a variety of data sources, like the indicators discussed above, hospital data (Wisniewski et al., 2003, Silva et al., 2013, Tsui et al., 2003) and even over the counter retail sales data (Wagner et al., 2003, Xiaohui et al., 2004).

The other public health tracking application in the literature was a disease registry for end stage renal disease, the REIN ESRD registry (Ben Said et al., 2005, Jean-Baptiste et al., 2005, Richard et al., 2005). The National Cancer Registry (NCR) could be likened to this, as it is a disease registry – the only disease registry that was part of the research findings. An extensive number of cancer registries that exist outside of South Africa are documented in the literature (Coebergh et al., 2015), but they were excluded from the literature review as they did not involve BI use.

Evidently, the NHLS CDW is used in a variety of surveillance activities, many of which are similar to those described in the literature. Distinguishing features of the NHLS CDW are firstly that it provides access to laboratory test results as a source of surveillance data, that this data applies to 80% of the population of the whole country, and that the focus of public health tracking activities discovered during this research is around addressing the high incidence of diseases like HIV/Aids and TB.

6.1.1.2 NHLS CDW Applications for Administration and Management

Uses of CDW data found for administration and management by NHLS was comprehensive. While operational management uses regular reports that monitor performance and quality of work done in the laboratories, the NPP division uses CDW data to ensure that the special programs dealing with HIV/Aids and TB are operating effectively. The DoH also uses CDW data to monitor operational performance and the success of its projects and campaigns.

Only three of the applications found in the literature mentioned the use of BI data for administrative purposes. Cancer Care Ontario (CCO) uses the data to minimise wait times for surgical procedures (Garay et al., 2015), and the Oregan Community Health Information Network (OCHIN) use the data to compare the performance of clinics and identify best practices (Devoe and Sears, 2013). These are similar functions to what NHLS and DoH are performing with CDW data, albeit with different end objectives.

The other application in the literature using data warehouse data for administration and management is the REIN ESRD registry, which uses the data to better understand the demand and supply of renal care (Ben Said et al., 2005, Jean-Baptiste et al., 2005, Richard et al., 2005). No mention was made during data collection of the existence of this type of application of the NHLS CDW, but one of the interviewees did suggest it as a possible means of extending the CDW to satisfy its full potential:

> *". . . if the Department of Health wanted a crude estimate of what cancer looked like for 2017 I could be able to give it to them . . . it gives you ideas of the use of services. . . every time somebody comes for a biopsy it costs the Department of Health money. So . . . it tells the Department of Health . . . I need a facility in Northern Cape, for example – we need a facility in the North-West,*

because if this person . . . lives in Northern Cape but had to go to Free State to have their biopsy done."

CDW data is being used comprehensively for administrative purposes, but the discrepancy between the purpose of their administrative applications and the purpose of the administrative applications in the literature could usefully suggest potential future applications of this type.

6.1.1.3 NHLS CDW Applications for Medical Research

Only four NHLS CDW applications for medical research were identified, which would make it seem that very little medical research makes use of CDW data, which is not at all true. In explanation of this, it should first be noted that not all of the interviewees were involved in research. In particular none of the NPP staff interviewed were involved in research themselves, so they made no information requests concerning research that could be identified as applications. There is research being done in the NPP division, however, because the NPP data analysts stated they are often required to retrieve data from the CDW for NPP research purposes. This is can be confirmed by a Google Scholar search on ("national health laboratory service" "corporate data warehouse"), whose results include a number of published NPP research article acknowledging NHLS CDW data. On 24th March 2018 this search retuned 87 results, attesting to the fact that there is more research activity using CDW data than is reflected in the sampled applications.

In addition it should be noted that due to only very broad information being discussed in interviews, two of the identified research applications are very general, being "Request information required for Mycology research" and "Request data for other research on Cancer". Each one of these represents more than one, and probably several, applications of CDW data for research.

According to one of the members of staff interviewed, with 80% of the population's laboratory test results on record, and the high caliber of researchers employed by the NHLS, there is potential to produce a lot more valuable research:

". I think there is so much – the NHLS is the repository for every single lab test that is done for every single patient in the public sector. . . . the research that could come out of that particular system would be phenomenal. . . . it's a treasure trove of data. . . . they're sitting on a goldmine of data. It's there. They have the researchers within the organisation who are fully capable of utilising that data. We have some of the best researchers – we've got world renowned researchers in the NHLS."

None of the applications found in the literature mentioned the use of data warehouse data for doing medical research, which would imply a need for more documented cases of medical research based on CDW data. This study makes a contribution towards documenting such uses.

6.1.2 NHLS CDW Applications Classified by Request Type

As suggested by the literature and discovered in the interviews it was considered to be possible to make requests for information from CDW in each of the following ways: Through a customised application (Berndt, 2001, Berndt et al., 2000, Nambiar et al., 2013) , through a CDW service

representative (Tremblay et al., 2009), through a Data analyst, through a High level CDW data access tool (Xiaohui et al., 2004, Berndt, 2001, Berndt et al., 2000) and through a Low level CDW data access tool (Liyanage et al., 2014, Berndt et al., 2000). Each of these will be discussed in turn.

6.1.2.1 Information Requests made through a Customised Application

Most of the requests identified were made in this way, which makes sense because once a customised application has been created it can be used with a minimum of fuss. Of course this request type will be suited to information that is required repeatedly, usually on a regular basis. Aspects of BI used for these applications include reports, charts and dashboards, where reports are usually distributed in Microsoft Excel format and charts are usually displayed on dashboards. A Graphical Information System (GIS) is not used as a tool, but geographical maps do appear on dashboards. Information requestors often create their own charts in Excel once they have received the data. Five of the applications found in the literature used reports, three used charts and only one used dashboards. No CDW use of alerts was mentioned, whereas three of the surveillance applications in the literature used alerts.

6.1.2.2 Information Requests made through a CDW Service Representative

Information requestors not having access to a data analyst in their department were found to request data through a CDW service representative when they had unusual requirements that were not met by regularly received reports. The BI aspect used by the service representatives is always SQL, which is a low level CDW access tool, and once the data has been retrieved it is sent to the information requestor in the form of a Microsoft Excel spreadsheet. The literature spoke of only one application that allowed for access to data warehouse data using SQL (Berndt et al., 2000).

6.1.2.3 Information Requests made through a Data Analyst

Those lucky enough to have access to a data analyst in their department submit their ad-hoc requests to their data analyst, who also uses the low level CDW access tool SQL to extract the data from the CDW for them. Results are then sent to the requestor in the form of a Microsoft Excel spreadsheet. NPP data analysts also then load the data onto their own SQL Server database, where they can then further understand it using analytical tools like Microsoft Powerpivot, which gives some OLAP facilities, and STATA which allows for statistical analysis. Overall, though, the BI aspect of analytics was not used often in applications found in the NHLS CDW environment. Interestingly, complex analytical techniques like fuzzy logic and machine learning are being applied, not to understand the data better, but rather to link all the laboratory tests done for a single patient. The literature revealed a much more in-depth use of analytics, which were used in eight out of the eleven applications found (Berndt et al., 2003, Ben Said et al., 2005, Wisniewski et al., 2003, Rizi and Roudsari, 2013, Xiaohui et al., 2004, Devoe and Sears, 2013, Walji et al., 2014).

6.1.2.4 Information Requests made through a High level CDW access tool

There were no requests identified that were made using a High level CDW access tool like OLAP or a data warehouse browser. In the NHLS CDW environment OLAP tools are only used in the ETL process of getting data into the data warehouse. In the literature, an empirical study showed that there were substantial benefits in the use of an OLAP tool to get data from the CATCH data warehouse (Tremblay et al., 2007). The use of a data warehouse browser was also reported for one of the applications. The REIN ESRD data warehouse is accessed via a GIS interface that provides OLAP-like facilities (Richard et al., 2005). Once CDW information requestors have the data in Excel format, they can use pivot tables, which is a form of OLAP, a facility also provided by the FluDM surveillance application found in the literature (Rizi and Roudsari, 2013).

6.1.2.5 Information Requests made through a Low level CDW Access Tool

This type of request is only made by the data analysts who are using SQL to request information on behalf of other people in their departments and has been discussed. (SQL is also used by CDW service representatives but this is not classed as an information request.)

In general, BI aspects identified in the literature review that are being used with the CDW include reports, charts, dashboards and SQL. Analytics are seldom employed, and data warehouse browsers, OLAP and Alerts are not being used at all by information requestors.

Of the different types of request, customised applications appear to be the easiest to use and the least resource-intensive for CDW, because they are pre-prepared and just need to be registered for and called up by the information requestor. Dashboards, in particular, appear to be very useful and popular, and several interviewees expressed the wish for more of them. For example, asked whether the CDW was perceived as satisfying its full potential, one requestor replied:

> *"For now I would say not yet. You see the question you asked about the dashboard – that's what we are trying to work on to build it because that will be able to give us more. Because . . . we want to have it on day to day."*

And another pointed out:

> *". . . and what about the routine data we could be producing from tests - like the TB dashboards and the HIV dashboards. Those are for particular priority diseases, but we could be doing that for Diabetes, we could be doing that for Hypertension, we could be doing it for a whole lot of other priority diseases."*

The other type of request that is of interest is the use of High level CDW access tools, which apparently is not available for information requestors. This ties up with the theme 'Direct Access', which will be discussed under the second research question.

6.1.3 NHLS CDW Applications Classified by Originating Division

The different divisions of the NHLS interviewed were the NICD, the NCR, the NPP, and Operational management. There was a distinct difference between the type and the number of requests for information for the NICD and the NCR on the one hand, and the NPP and Operational management on the other.

There were a total of nine applications for NICD and NCR, five for public health tracking and four for research, and none for administration. As mentioned before, two of the applications for research are general, and actually correspond to several research projects. The focus of these divisions is thus on public health tracking and research, which is not surprising, since three of the interviewees in these divisions were highly qualified medical professionals, two being epidemiologists.

The total number of nine requests for the NICD and NCR divisions is substantially lower than the total number of eighteen requests[44] for NPP and Operational management. This may be partially due to the fact that a total of four people from NICD and NCR were interviewed and a total of six were interviewed from NPP and Operational management, but the difference is too great to be attributed to this alone. Findings for research question 2 imply that there is global funding for NPP and they are thus given priority , which may be having an impact. Also the full number of NICD and NCR research requests is not fully represented, as mentioned earlier.

For NPP and Operational management there are seven requests for public health tracking, eleven requests for administration, and no requests for research. Seven of the eleven administration requests are for Operational Management, who make no public health tracking requests. This is understandable because their function in the NHLS is purely administrative. The focus of NPP staff interviewed on public health tracking and administration is also understandable, since three of the interviewees were program managers responsible for the success of various population health programs. These program managers are involved both in the operational aspects of health tests that concern their programs, and the impact that the programs are having on the population.

Any attempt to further improve services offered by the CDW would do well to consider differences in the data needs of NICD and NCR versus those of NPP and Operational Management.

The literature did not contain any discussion on the relative information needs of different departments within a single organisation.

In closing, it should be pointed out that though the NHLS CDW is a single application, it provides a wide gamut of facilities that cover aspects of many of the applications found in the literature. It is possible that not all the functionality of applications found in the literature were documented, but from what is recorded, none of the applications in the literature are as extensively or as diversely used as the NHLS CDW. This is of particular significance considering that the NHLS is situated in a country that is considered part of the developing world, whereas all applications found in the literature were part of the developed world.

[44] Though a total of 23 requests were sampled for NPP and Operational management, 6 of them were made by data analysts on behalf of NPP staff and did not constitute independent applications

This completes the discussion of research findings for research question 1. Research question 2 will be discussed next.

6.2 RQ2: What are the issues and challenges in the use of the NHLS CDW?

Examination of the issues and challenges in the use of the NHLS CDW was done through eleven of the themes that emerged during the thematic analysis of the data. These were 'Relationship with CDW Team', 'Staffing', 'Time Delays', 'Outsourcing', 'Data Quality', 'Patient-oriented View', 'Privacy and Security', 'Direct Access', 'Metadata', 'Ease and Efficiency', ,and 'External Data'.

The study was underpinned by activity theory as elucidated by Engestrom (Engestrom, 1987) and its applications in prior IS work (e.g. (Johnson, 2004), (Lee, 2010), (DeBay, 2013)). According to activity theory, activities are situated in the context of a community. The activity of a subject to attain an object is mediated by the tools being used, the rules of the community, and the division of labour amongst members of the community. RQ2 aimed to identify and explore issues and challenges in the NHLS CDW that could be mapped onto the activity system, thus illustrating the utility of activity theory for understanding why such issues and challenges arise. Identified themes mentioned above were thus mapped onto the activity system for an information request as shown in figure 43.

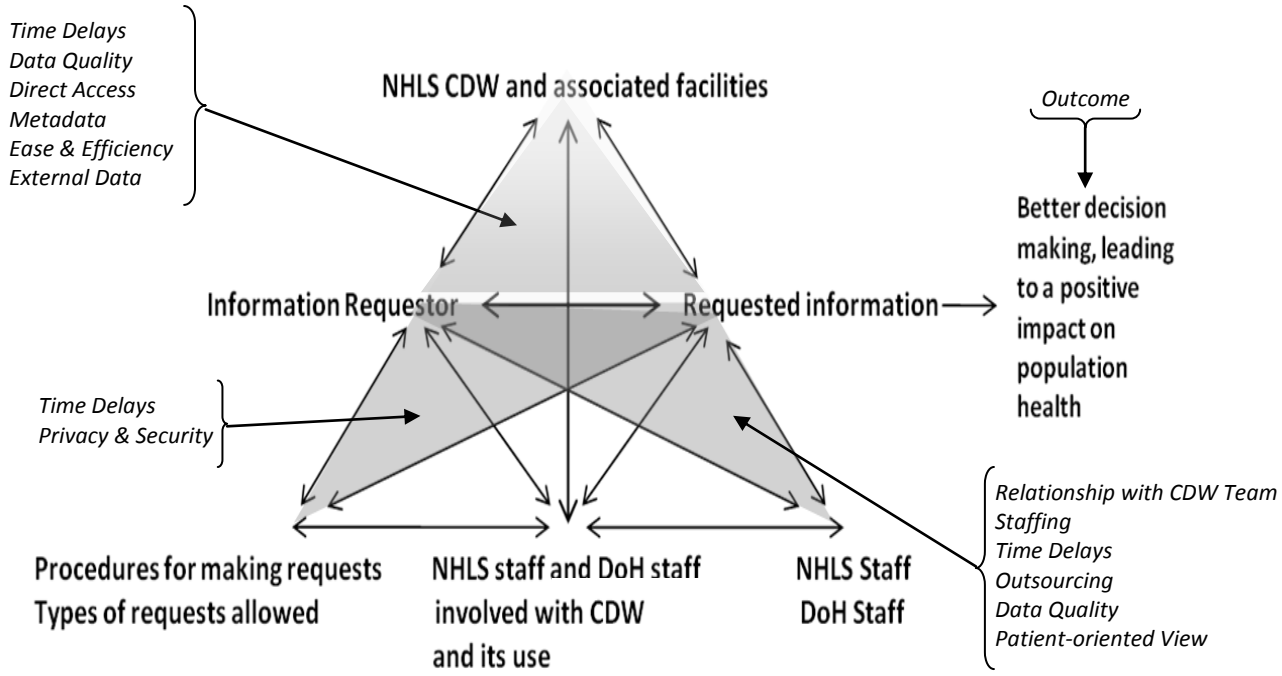


Figure 43: Allocation of Themes to Activity System for an Information Request

Issues and challenges experienced by applications reviewed in the literature were found to fit into three major categories – 'Privacy and Security',' Data Quality' , and 'Implementing BI for Population Health'. The 'Implementing BI for Population Health' category was not relevant to this study, which has focused information requestors and not on implementation issues. The other two categories

were found to be completely relevant to the case of the NHLS CDW and in fact were matched with identified themes of the same names. For this reason the themes 'Privacy and Security' and 'Data Quality' will be discussed first. Also, since there is such close relationship between the 'Data Quality', 'Data Governance Model' and 'Patient-oriented View' themes, they will be discussed together.

6.2.1 'Privacy and Security'

Given the sensitive nature of personal health data, it is hardly surprising that privacy and security was identified in the literature review as the most important issue for BI applications found. In particular, patient details should be kept private, and several of the reviewed applications de-identified patient data before using it for BI purposes (Berndt et al., 2000, Tsui et al., 2003, Walji et al., 2014).

Privacy and security is also an important issue for the NHLS CDW, and most of the staff were found to be extremely mindful of this, with particular concern expressed over the sensitivity of patient data. Though they might find the bureaucracy of it frustrating at times, information requestors adhere strictly to the NHLS regulations in this regard. The NHLS CDW cannot de-identify patient data, because patient details are needed for many of the applications. All means of access to the CDW data are guarded carefully, however, as discussed in chapter 5, with specific attention being paid to the protection of patient details. According to the information request form in appendix I, ethics clearance is required before access to patient identifying data is granted.

Research using NHLS CDW data is only permitted once ethics clearance, which includes permission to access the data, has been granted. The NHLS in fact has a special Academic Affairs department to deal with this research clearance.

Most applications reviewed referred to compliance with government regulations as one of the major challenges encountered, eg. Wisniewski et al. (2003), Wagner et al. (2003), Ben Said et al. (2005). Government legislation in the USA, Canada and France were mentioned. Information requestors at the NHLS may have been aware of government regulations in this regard, but reference to the South African POPI act was made by only one interviewee.

Privacy and security is a certainly a challenge for the NHLS CDW, but it is one that has been well addressed, and therefore does not currently constitute a serious issue.

The discussion of 'Privacy and Security' is now concluded. 'Data Quality' and 'Patient-oriented View' will be discussed next.

6.2.2 'Data Quality', 'Data Governance' and 'Patient-oriented View'

Decisions concerning population health need to be based on good quality data in order to be reliable and effective. Ensuring that data is of good quality is indeed a challenge, especially when dealing

with the large volumes of data found in data warehouses and the multiple ways in which data quality problems can be introduced.

When questioned about their level of trust in NHLS CDW data, all information requestors said that although the bulk of the data was fine, there was a percentage that exhibited problems and that this was an issue. In the literature, too, data quality was identified as an issue for most of the applications, eg. Jean-Baptiste et al. (2005), Berndt et al. (2003), Walji et al. (2014).

To assess where the NHLS CDW data quality issues might lie, interviewees were questioned on five of the standard data quality dimensions: Relevance, Timeliness, Consistency, Accuracy and Completeness (Wand and Wang, 1996). Relevance and Timeliness were not considered to be problems at all.

There was also no problem with Consistency within the data. This is unusual for a data warehouse, which generally combines inconsistent data from multiple disparate sources. Berndt et al. (2003) and Wisniewski et al. (2003) both discuss the difficulty in obtaining consistent data when combining healthcare data from different sources that store the data in different formats. What works in the NHLS CDW's favour, in this respect, is that all the laboratories use the same laboratory information system (LIS), TrakCare, so that much of the data coming in is in a standardised format.

Another type of inconsistency was reported by NHLS users, however. Different record sets can be returned by the same query at different times. As pointed out in chapter 5, this could be a fault in the query, or it could be due to laboratory processes that may cause a lag in the time it takes for data to get to the CDW. If information requestors are made aware of the possibility of this lag time, which means it may take up to three months for CDW data to balance out correctly, they may be more understanding of the inconsistency in result sets returned.

The data quality dimensions of Accuracy and Completeness were the problematic ones, with lack of completeness being the more serious of the two. It was discovered that problems with these dimensions are introduced at various points of the flow of data from the source of the data, to the entry of the data into the LIS, and to the transfer of the data from the LIS into the CDW. The source of the data, where the medical practitioner fills in the test request form, was found to be the most troublesome point, due to important information being omitted. The source, therefore, is the place where these problems need to be attended to. Rizi and Roudsari (2013) and Walji et al. (2014) are also of the view that data quality problems should be remediated at source.

Neglecting to enter all necessary information at source is also responsible for the difficulties experienced by NHLS CDW information requestors in obtaining a patient-oriented view of the data. 'Patient-oriented' refers to the ability to access all information in the data warehouse that relates to a specific patient, in other words to easily view all of the laboratory tests that have been done for that patient. Information per patient is often essential in the management of population health, and should be obtainable from the data warehouse with ease and accuracy 100% of the time. This would be no problem if there was a properly enforced unique patient identifier.

There is a medical reference number (MRN) that is supposed to be a unique patient identifier, but this is not properly entered at source, often resulting in patients possessing more than one MRN,

which defeats the purpose. This allows for the potential abuse of laboratory facilities, where patients present themselves at more than one facility so that they can repeat the same tests to verify the results for themselves. It also necessitates a tremendous amount of effort and expense on the part of those who need a patient-oriented view of the data. Fancy algorithms, developed by CDW and others, achieve this to an extent, but are never correct 100% of the time. A substantial amount of time is also spent manually adjusting the data simply to achieve a patient-oriented view. If only there was a unique patient identifier, this result could be achieved with 100% accuracy by using a simple database query.

Although patient identifiers were discussed in the literature (Berndt et al., 2000, Tsui et al., 2003, Walji et al., 2014), none of the applications reviewed discussed the difficulty of obtaining a patient-oriented view of the data as an issue.

Resolving the challenge of the patient-oriented view is not the only cause for complex workarounds and manual checks that data received is OK. All information requestors reported doing this to some extent on data received. Data analysts take particular responsibility to 'clean' the data before disseminating it to others. The REIN ESRD application uses regional clinical research assistants to verify data warehouse data against patient files (Jean-Baptiste et al., 2005, Richard et al., 2005). Other applications found in the literature also use various ways to ensure the accuracy of their data, eg. Wisniewski et al. (2003), Walji et al. (2014), Wagner et al. (2003).

The only way to remove the need for complex workarounds and hours of manual checking due to inaccurate and incomplete information, is to resolve this problem at the data source. Perhaps a unique patient identifier is so important that it should be enforced by the NHLS. One of the interviewees made this suggestion:

> *". . . for the NHLS itself it's not a problem, because they have the new system called TrakCare . . . So if Trak then allocated a unique identifier to a patient that could be used nationally, because Trak is a centralised system . . So with that progress, why can't we implement a unique identifier for NHLS patients? . . . You could use ID number, and the NHLS could say we are not processing this unless you have an ID or a passport number implemented. So the NHLS can be the gatekeeper of that . . ."*

The early infant diagnosis (EID) program experiences an additional twist to the 'Patient-oriented View' problem, in that new-born babies do not initially have a name, and when they get a name it may change during the first few months of their lives. Without a name it is even more difficult to match up different tests done on the same baby. The EID program has developed a particularly innovative response to this:

> *". . .that one has been created for me. Yeesss. . . . remember we don't have a unique identifier. . . . to keep you in the loop – most of the babies when they are born they don't have names. So it would be "Baby to Tswakane". Then by three days or when they go to the facility the baby has got a name now, and you might find that the baby has got a name that is even the surname that is different from mine – ok - because it's an African thing. Maybe over the weekend they came, they paid lobola and renamed the baby and the baby took the daddy's surname. Then, now, somewhere down the line mommy gets married to somebody, so it is very likely that this baby will*

also change the name and surname. So now what the department – what the paediatric people have been trying to pilot . . . is that the booklet of the baby at the back it has unique stickers . . . that speaks only to this baby. A number. So every time you do a test you'll have to peel a sticker and place it onto the NHLS request form. So it will require attitude from the clinicians also. Meaning that when we see this number we know that it belongs to this booklet. So whether the baby changes names or whatever, we know this is still the same baby."

Both of these approaches require co-operation from external clinicians, whether they work for the DoH or not. The 'rules' that govern the community, which includes the NHLS, the DoH and all private practitioners, would have to change. Whike enforcing a unique patient identifier, measures can be taken to ensure all data is entered completely and accurately. 'Data Quality' and 'Patient-oriented View' would then belong to the sub-triangle of the activity system where rules mediate the quality of the data obtained by the information requestor, as shown in figure 44.

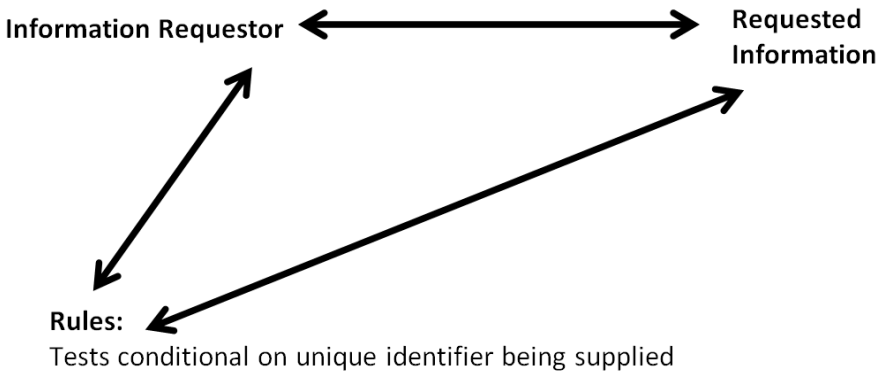


Figure 44: Sub-triangle of the Activity System for 'Patient-oriented View' – Altered Perspective

There was no indication by any of the interviewees of a formal data quality framework that is being applied to the NHLS CDW. There are various methodologies for ensuring data quality that have been suggested and tried. These all involve some kind of procedural quality assessment and improvement cycle (Lee et al., 2002, Batini et al., 2009). Rizi and Roudsari (2013) recommend the early development of a data governance model to ensure the success of a data warehouse. A data governance model would include measures to ensure the usability, integrity and security of the data in the warehouse. The 'Data Governance Model' theme was used to record interviewee responses to the question of whether they were aware is a data governance model for the CDW. Most interviewees answered with a definite "No", or gave a very unsure answer. Content recorded in this theme was so minimal that it has not been considered a CDW challenge in it's own right. It is therefore not likely that the CDW has instituted a data governance model.

This completes the discussion on the themes 'Data Quality' and 'Patient-oriented View'.

Of the remaining themes, the one that is considered to pose the greatest challenge to the NHLS CDW is 'Staffing'. The 'Staffing' theme is related to the 'Relationship to CDW Team' team. It is also related to the 'Time Delays' theme, in that staffing problems are the main cause for time delays, and

to the 'Outsourcing' and 'Direct Access' themes which have the potential to alleviate staffing problems. These themes will therefore all be discussed together.

6.2.3 'Staffing', 'Relationship with CDW Team', 'Time Delays', 'Outsourcing' and 'Direct Access'

The CDW team is tasked with the immense responsibility of maintaining the complex data warehouse infrastructure as well as answering to the information needs of a myriad of different information requestors both inside the NHLS and beyond. All this is accomplished by a relatively small team of people, which, in the words of one of the interviewees, is "*stretched to the limit*".

Information requestors have no real contact with CDW technical staff, but rather come into contact mostly with those members of CDW staff who are acting as CDW service representatives. The iterative process of getting information requests right requires a good relationship between information requestors and the CDW team, and data classified under the 'Relationship with CDW Team' theme revealed this to indeed be the case. NICD and NCR staff did indicate that they felt CDW representatives to be 'limited', but this was referring to the fact that IT staff does not have the medical know-how to fully understand the data.

The problem perceived by information requestors is that "*there aren't many staff working there and they have got numerous projects*". Staff constraints result in frustration and restrict constructive use of the CDW. Requestors complain that "*there's very little resources that can help us to do what we need to do in the time that we need to do it*".

The only reason given for the staff shortage was high CDW staff turnaround, which is not unusual in the IT industry. Wisniewski et al. (2003) cites technical expertise as being one of the challenges involved in implementing a data warehouse. Staff turnaround creates problems for the CDW and its users as relationships that had been formed between information requestors and CDW service representatives are lost and need to be reformed with new members of staff, important expertise is often lost and in one case even an important piece of programming code was lost and has to be re-created. "*To be person-centric it's very dangerous*" and CDW do not appear to have contingency plans whereby knowledge is shared so that work can continue if a member of staff leaves.

Aside from restricting constructive use of the CDW, time delays resulting from staff shortages impact on work being done. NPP and Operational management report minimal time delays that do not seriously impact their work, but long time delays for NICD and NCR staff sometimes result in serious consequences, like the inability to achieve key performance indicator targets. NICD and NCR staff did complain that there is "*no agreed upon turnaround time*" and "*there's funding for that so that takes priority over other projects*", which could account for some of the discrepancy. An additional cause of relatively minor time delays experienced by NPP was technical networking problems, probably experienced because of they are situated on a site that is remote from the CDW.

Adding to the range of available customised applications, external funding and the use of data analysts, outsourcing, and allowing information requestors direct access to CDW data could all possibly alleviate the negative effects of the staff shortage.

Customised applications can be used to retrieve data from the CDW without the intervention of any staff, so it stands to reason that if more of these applications were available there would be less of a load on existing CDW staff.

External funding impacts substantially on the constructive use of CDW data in two main ways. Firstly, it appears that information requests from funded projects take priority, as mentioned above. Secondly, funding enables the division in question to hire additional personnel. Extra staff could assist in various ways, including in the development of new applications, but the most effective new member of staff to acquire would be a data analyst or "*a dedicated person . . . that can help us*". Tremblay et al. (2007) speaks of 'health planners' serving as similar intermediaries between information requestors and the CATCH data warehouse. Information requestors dealing with the CDW team through data analysts seemed considerably more satisfied with having their requirements understood, as well as with turnaround times for information requests and with the quality of data received. Since they are employed full-time to serve the needs of their department, data analysts become familiar with departmental information needs and the medical area that they pertain to and are thus more able to understand requestor requirements. Their data access skills are more advanced than those of the information requestors, so, when granted permission, they can use SQL to access the CDW directly, thus avoiding the lengthy process of filling in an information request form, etc. Data analysts also take some responsibility for data quality, carefully verifying the correctness of data before it is disseminated to requestors. Because they are dedicated to the department, they are always readily accessible to requestors, and because they deal regularly with the CDW team they tend to develop strong relationships with the CDW team, and thus receive willing and prompt service. The deployment of a data analyst seems to be the most effective way of alleviating, or even avoiding, the departmental implications of CDW staff shortages.

Outsourcing is another way in which burdens on insufficient staff can be alleviated, particularly when there is a lack of specific needed skills. The NHLS is not averse to outsourcing, and it has been made use of in various areas, including assistance with the development of technical infrastructure, extracting data from CDW, creation of new customised applications, and the use of experts to create complex algorithms to impute missing data. Requestors felt that outsourcing was successful and had not experienced problems with its use.

Allowing selected information requestors direct access to sections of CDW data, obviously with the requisite permissions, may be another way of relieving the burden on the CDW team. Currently, the NPP data analysts do have direct access to that portion of CDW data that is relevant to them. They are empowered by being able to access this data using SQL, which enables them to explore and understand the data better. NPP project managers have their needs met by these data analysts and have no wish for direct access to CDW data themselves. Operational management also expressed no need for direct access. Medical professionals in NICD and NCR, however, felt that they would greatly benefit from having direct access to the data. It would enable them to explore the data freely and perhaps use their medical knowledge to greater elucidate the information found, and to know what information to look for in future searches. It would also remove the necessity for the iterative cycle of getting a request right. Privacy and security would be an obvious concern, but once a department has been tasked to use the data, and especially if all staff in the department signs a non-disclosure

agreement, access to the appropriate data could be granted without excessive and repetitious bureaucracy.

There was talk of a data mart, or copy of a portion of the CDW data, being created for NICD and NCR to access – as part of the NICD micro-strategy. This may be an alternative way of enabling more direct access to CDW data for these departments.

Information requestors that have direct access to CDW data would be further empowered with a high level CDW access tool, like OLAP, to view that data through. Eckerson (2006) refers to on-line analytical processing (OLAP) as 'spreadsheets on steroids', which allow users to query the data in the way they think about their daily business, by permitting them to easily select the data that they wish to view and summarise by. An empirical study found the use of OLAP with the CATCH data warehouse by knowledge workers in a regional health planning agency to be a highly advantageous way of satisfying information requests (Tremblay et al., 2007).

In summary, information requestors are affected by time delays due to a shortage of CDW staff, by problems caused by rapid CDW staff turnaround, and also by the difficulty CDW staff experience in the understanding of medical data. There are, however, several ways of addressing these problems, which include adding to the range of available customised applications, external funding and the use of data analysts, outsourcing, and allowing information requestors direct access to CDW data.

This completes the discussion of the 'Staffing', 'Relationship with CDW Team', 'Time Delays', 'Outsourcing' and 'Direct Access' themes.

The 'Metadata' and 'Ease and Efficiency' themes are also deemed to be related, since the main purpose of the metadata would be to facilitate the ease and efficiency with which CDW data is extracted. These themes will be discussed next.

6.2.4 'Metadata' and 'Ease and Efficiency'

Interviewees were questioned about how easy and efficient it is to get the required data out of the NHLS CDW, as it was felt that difficulty in access would discourage CDW use.

Requesting information through a customised application or through a data analyst was found to be both easy and efficient. For requests made through a CDW service consultant, the part of the process where the form is filled in is easy, it is the iterative process of getting an information request right, as well as a lack of requestor knowledge of what individual test codes mean that present difficulties and time delays.

Requestors who had been working at the NHLS for an extended period of time, particularly if they had experience working in the laboratories, had no problem understanding what the coded names of the laboratory tests mean. New employees, however, feel the need for some kind of 'lookup' facility, to give them the meaning of the codes. The term 'metadata' refers to data about data. Keeping several types of metadata about several aspects of the data is deemed beneficial in a data warehouse environment (Inmon, 2005). In the case of the NHLS CDW, the implementation of

metadata about the meaning of the data would provide the desired 'lookup' facility. One of the interviewees also validly pointed out that "*we cannot rely on the individual's history*" and that metadata would enhance the "*institutional memory*" of the NHLS.

Other reasons cited for a lack of ease and efficiency of retrieving data varied from person to person. One expresses a mild frustration with bureaucracy , but this does not merit being referred to as a real 'Ease and efficiency' issue. Other issues mentioned have been covered earlier under other themes: the iterative process of getting a request right under 'Relationship with CDW Team'; and problems with data quality under 'Data Quality'. It seems that when all of the other bigger issues are sorted out, the process of obtaining data from the CDW is easy and efficient.

The 'External Data' theme will be discussed next.

6.2.5 'External Data'

The necessity of information requestors supplementing CDW data with information from other sources was discovered during the interview process. NPP and Operational management had very different needs from NICD and NCR in this respect.

NPP and Operational management were mainly seeking out alternative sources of the same data that they could check their own data against for verification.

NICD and NCR, on the other hand, claim that mortality information, hospital and clinic information, and laboratory information, are necessary for them to be able to do their jobs properly. Without mortality information, it is not possible to determine what the outcome of a disease is, and the opportunity to capture those sick people who were not diagnosed while they were alive is lost. In understanding the spread of a disease, hospital and clinic data like date of admission are useful in determining where an infection could have been contracted. Hospital and clinic data is also important for other things, like capturing diseases diagnosed by other means than laboratory tests, and even for research. A full epidemiological picture of a disease can also not be fully obtained without including test results from private laboratories.

The inclusion of some of the private laboratory data in the CDW was mentioned in the interviews, but there was no indication of any connection with mortality or hospital data. In a perfect world, it would be nice if there were interfaces between the given external data sources and the CDW, but this is easier said than done. Aside from the obvious expense and effort that would be involved in the development and maintenance of such interfaces, data sharing agreements would need to be negotiated and signed by participating parties. The CARP data warehouse combined data from only three hospitals and yet cited getting institutional approval as one of the greatest challenges (Wisniewski et al., 2003). Data sharing agreement challenges were also experienced by the RODS surveillance system (Tsui et al., 2003) and the National Retail Data Monitor (NRDM) (Wagner et al., 2003). A CDW interface with hospitals and clinics would also necessitate the existence of an electronic record keeping system at the clinics and hospitals, a luxury that many South African facilities do not possess.

6.2.6 Most Serious Challenges

Of the eleven possible challenges to the NHLS CDW examined, three, 'Data Quality', 'Patient-Oriented View' and 'Staffing', have been highlighted by this discussion as being the most serious challenges to be addressed, and others were seen as lesser challenges. Possible ways of dealing with these challenges that emerged from the research are:

- The 'Patient-Oriented View' and many of the 'Data Quality' issues originate at the data source, where data is not being filled in properly by medical practitioners, and therefore these problems can only be corrected at the data source.

- Several ways of mitigating the staffing problem were discussed, some of which require funding, but some of which could be accomplished with a change of perspective and direction, like allowing authorised information requestors direct access to a limited subset of CDW data.

- 'Time Delays' were found to be problem, but these were mainly due to staffing problems, and thus would be expected to improve if staffing issues are resolved.

- 'Privacy and Security' is an issue in a data warehouse of this nature, but one that NHLS and the CDW seem to have well under control.

- 'External Data' sources are required and it would be nice if interfaces were built to incorporate mortality data, hospital and clinic data and private laboratory data into the CDW, but there would be serious challenges associated with doing this. (There has already been some effort to include private laboratory data.)

- 'Ease and Efficiency' was not found to be a challenge in its own right, but rather to arise from other challenges that have not been resolved. Ease and efficiency in getting CDW would be improved with the addition of metadata to clarify the meaning of the data.

That completes the discussion of findings for research question 2. Research question 3 and research question 4 are closely related, and therefore will be discussed together.

6.3 RQ3: What are the individual outcomes for the NHLS CDW user and data requestor? and RQ4: What are the perceived benefits (impacts) of the NHLS CDW use for the NHLS as an institution and for population health in general?

NHLS CDW data was considered by individual data requestors to be an essential tool in the accomplishment of their jobs, whether those jobs involved public health tracking, administration or research. Their comments in interviews also suggested important positive impacts on productivity,

decision making effectiveness, ability to track progress in their work and a sense of meaningfulness at work.

Inmon (2005), also known as "The father of data warehousing" (Watson et al., 2001: 44), cites the lack of productivity experienced when there were no data warehouses as a reason for his recommendation of the adoption of a data warehouse strategy. The use of a data warehouse would thus be expected to improve personal productivity, as has been confirmed by this study.

Improved decision making is the ultimate objective of making large amounts of data available in a data warehouse, and is in fact stated in the original definition of a data warehouse as a "an integrated, subject-oriented, time-variant, non-volatile collection of data that provides support for decision making" (Inmon, 2005). In fact Kimball et al. (1998), also known as "The father of Dimensional Modelling"[45] (Chaki, 2015: 21) declares that the only true output of a data warehouse is the decisions that are made once the evidence in the warehouse has been considered. Individual impacts of the CDW included an improved ability to make decisions, which therefore confirms these ideas.

A sense of meaningfulness at work may not have been expected by the original 'fathers' of data warehousing and dimensional modelling, but it did come through in this study. Health care workers requesting data from the CATCH data warehouse also intimated that better access to the information empowered them, and in fact enabled a role change from data collector to data consultant (Tremblay et al., 2007).

The literature reviewed did not make reference to a the use of data warehouse data to track of progress being made with work being done on a given project, but this is certainly one of the major purposes it is being used for at the NHLS, particularly for the NPP programs.

Benefits of the NHLS CDW data warehouse to information requestors are therefore substantial, and thus enable them to achieve their stated objectives more competently. This of course means that the contribution they ultimately make to the NHLS as an institution and to population health in general is improved by the CDW.

By the same token, any benefit of the CDW to the NHLS as an institution will ultimately benefit population health, particularly since the NHLS aims to "build a healthy nation" (NHLS, 2015). Similarly, the NHLS has a national mandate to support the DoH, the government body that serves the population's health needs, and thus any benefit to the DoH will ultimately benefit both the NHLS and population health.

Evidence found in this study of benefits to the NHLS as an organisation and to population health in general were the facilitation of the management of internal operations, the provision of data for research, the ability to fulfil the national mandate, the enhancement of international credibility, the ability to secure funding, the ability to make informed changes to practices and the ability to formulate policies.

[45] Dimensional modeling is a data design technique used to design data for data warehouses

It is unusual for a data warehouse to be used for the management of internal operations – typically this is done by operational systems. There was no mention in the literature reviewed of the use of a data warehouse for this purpose. The CDW is, however, the central repository for information from all the different laboratories around the country, and thus a convenient central point from which to view and analyse information from an operational perspective.

The provision of data for medical research was also not reported on in any of the applications reviewed. The prevalence of the research use of the NHLS could be due to the nature of the organisation, and the existence of many "*world renowned researchers in the NHLS*". Certainly medical research done by NHLS researchers that is based on CDW data does enhance the standing and reputation of the NHLS as an institution.

The national mandate of the NHLS is to serve the DoH, which it regards as its major 'client'. CDW data enables this and is in fact often used by the DoH itself. Evidence showed how CDW data helped to improve DoH effectiveness, and assisted them in understanding health dynamics. DoH relies on CDW data, as "*Through this data, this is how the minister measures*". The public health reporting data warehouse described by Rizi and Roudsari (2013) was also built to support a governmental health organisation – the British Columbia Centre for Disease Control in Canada.

International credibility was not cited as being important in any of the literature reviewed. Perhaps this is because international credibility is less important to organisations that are not operating in a developing country. NHLS employees felt they could present data from CDW at international meetings with pride, and even found that other delegates complimented the standard of the data presented, saying that NHLS staff was "*getting everything on a silver platter*".

CDW data helps the NHLS to secure international funding from the Global Fund and other international organisations, as the data makes it possible to track the progress of the projects being funded. Again, no reference was made to this benefit in the literature, perhaps because developed companies are not candidates for international funding.

Most importantly, CDW data informs the DoH in terms of changing current population health practices and formulating new population health policies. Berndt (2001) describes how the aim in establishing the CATCH data warehouse was to provide high quality information to be used in health care planning and policy formulation. Examples of this benefit encountered by NHLS staff include earlier HIV/Aids testing for infants, inclusion of Cryptococcus Meningitis in the national casket of care, recording laboratory test dates to enable tracking of when an infection was contracted, the new Breast Cancer Prevention and Control policy, the new Cervical Cancer Prevention and Control policy and the National Cancer Control Plan. All of these new practices and policies were based on decisions made using CDW data.

Positive outcomes of the NHLS CDW are therefore many. From the individual impacts on productivity, decision making effectiveness, ability to track progress in their work and a sense of meaningfulness at work; to the institutional and population health impacts of the facilitation of the management of internal operations, the provision of data for research, the ability to fulfil the national mandate, the enhancement of international credibility, the ability to secure funding, the

ability to make informed changes to practices and the ability to formulate policies; the NHLS CDW certainly makes a difference.

6.4 Chapter Conclusion

This research study explored the use and impact of the NHLS CDW, the issues and challenges encountered in this use, the individual outcomes of this use for individual users and the impact that this use is perceived to be having on population health in South Africa. This chapter has discussed and interpreted the findings that were presented in the chapter 5. The themes 'Full Potential' and 'Data Governance Model' that had not previously been discussed were used to supplement the insights gained.

Applications of the NHLS CDW were found to be many and varied, and although the NHLS CDW is a single implementation, its diverse applications span what is available from several of the implementations found in the literature taken together. While the biggest issue identified in the literature was privacy and security, the NHLS seem to have risen to the challenge in this regard and ample attention is paid to it. Data quality, on the other hand, was identified to be a challenge both in the literature and for the NHLS CDW. Staff shortages emerged as a serious challenge for the NHLS CDW, though mentioned only in passing in one of the applications found in the literature. The individual impacts of improved decision making, productivity, sense of meaningfulness at work, and the ability to track work progress contribute to organisational success. Instiutrional and population health contributions to fulfilling the national mandate and to the formulation of population health policies correspond with impacts mentioned in the literature. What seems specific to the NHLS, however, are the contribution CDW data makes to internal operations, research, international credibility and the ability to secure funding.

The next chapter will conclude the report by summarising the research, outlining the contributions of the study, making recommendations for the future, and discussing the limitations of the study.

7 Conclusion

7.1 Summary of the Study

This study explored the potential and application of the Corporate Data Warehouse (CDW) and BI facilities of the National Health Laboratory Services (NHLS), as an example of an application of BI to population health. The National Health Laboratory Services (NHLS) is responsible for laboratory tests done at public health institutions throughout South Africa, and as such keep records for 80% of the population, all of which are accessible from the NHLS Corporate Data Warehouse (CDW). With the realisation that the NHLS CDW can only make a difference if the information it contains is utilised by people who base their decisions on it, Activity Theory was used as a lens through which to explore and describe the ways in which the NHLS CDW is being used, the issues and challenges being encountered in its use, and the impact that this use is having on population health in South Africa.

To this end, the following research questions were addressed:

RQ1: What are the current applications of the NHLS CDW?
RQ2: What are the issues and challenges in the use of the NHLS CDW?
RQ3: What are the individual outcomes for the NHLS CDW user and data requestor?
RQ4: What are the perceived benefits (impacts) of the NHLS CDW use for the NHLS as an institution and for population health in general?

Data was collected during a preliminary session with the staff of the CDW department, and ten interviews with members of staff from the NICD, NCR, NPP and Operational management divisions of the NHLS.

Findings from RQ1 showed the NHLS to be an exceptionally large organisation involved in many areas of population health. A broad range of applications of the NHLS CDW were identified in areas as diverse as population health tracking, medical research and administration of laboratory performance.

Findings from RQ2 identified the most important challenges to be due to difficulties obtaining a patient-oriented view of the data, data quality, and staffing issues. Time delays were also found to be problem, but these were mainly due to staffing problems, and thus would be expected to improve if staffing issues are resolved. Privacy and security is always an issue in a data warehouse of a medical nature, but one that NHLS and the CDW seem to have well under control. External data sources are required and it would be nice if interfaces were built to incorporate mortality data, hospital and clinic data and private laboratory data into the CDW, but there would be serious challenges associated with doing this. (There has already been some effort to include private laboratory data.) The ease and efficiency of getting requested data was not found to be a challenge in its own right, but rather to arise from other challenges that have not been resolved. Ease and efficiency in getting CDW would be improved with the addition of metadata to clarify the meaning of the data.

Findings from RQ3 illustrated the positive impacts of CDW data on individual information requestors as being an improvement in productivity, decision making effectiveness, ability to track progress in their work and a sense of meaningfulness at work.

Findings of RQ4 revealed that benefits of the CDW to the NHLS as an organisation and to population health in general were the facilitation of the management of internal operations, the provision of data for research, the ability to fulfil the national mandate, the enhancement of international credibility, the ability to secure funding, the ability to make informed changes to practices and the ability to formulate policies.

The general perception is that the CDW is invaluable to the NHLS as an organisation and for the improvement in population health in general. As always, however, there were clear indications that there is still so much that can be achieved with the NHLS CDW.

The contributions of the study, recommendations for the future, and limitations of the study are presented next.

7.2 Contributions of the Study

Contributions to theory have been made by this study in various ways. Firstly, reporting on a data warehouse implemented in the health sector and used for the benefit of population health has added to the body of existing knowledge on data warehouses and business intelligence technologies, the sectors they are implemented and their uses.

Secondly, the empirical analysis of issues and challenges encountered while making an information request has contributed to the body of evidence on data warehouse use for population health – the literature review revealed only one empirical study in this area, and that was not focused on the data warehouse itself, but rather on OLAP, which is a tool for accessing the data warehouse (Tremblay et al., 2007).

Thirdly, although studies using activity theory as a lens through which to study information systems were found, none of them focused specifically on data warehouses. The application of Activity Theory to this study situated the activity of requesting information from the NHLS CDW in the context of the national health laboratory services as an institution and thus helped to elucidate the issues and challenges. A contribution has thus been made by extending activity theory into the study of data warehouses as information systems.

Furthermore, a methodological contribution was made concerning the approach to using Activity Theory, where Johnson's approach of enumerating participating entities of each activity system element was found to be useful in understanding the complex context of the NHLS CDW, and Lee's method of elucidating issues by assigning themes to activity system triangles resulted in a gradual exposition of the true issues.

Finally, this study has contributed to the knowledge on data warehouses implemented in the South African context.

In terms of contributions to practice, this research could be used by the NHLS to better understand the dynamics around their CDW and its use and inform them about issues and challenges that should be addressed, the most important of which were highlighted by this study as being the difficulty obtaining a patient-oriented view of the data, data quality and CDW staffing. This research could also enlighten officials concerned with population health about the benefits of data warehousing, which may result in the construction of other similar facilities for public benefit. Organisations that currently have a data warehouse may also benefit from the experiences of the NHLS CDW implementation and the issues and challenges identified, and organisations considering building a data warehouse may utilise this study's result as an input in making a decision.

7.3 Recommendations for the Future

South African health organisations should take heed of the NHLS CDW successes and work towards establishing other repositories of useful health data. In particular, it was indicated that hospital and clinic information is needed by the NHLS to perform some of its core functions. It would be extremely beneficial to the population health of the country if a standard electronic means of data capture was instituted at all public hospitals and clinics, and then (at a later stage) if this information could be collected into a central data warehouse.

The NHLS itself would do well to pay attention to the more serious issues and challenges identified by this study as being experienced by requestors of information from the NHLS CDW. The difficulty in obtaining a patient-oriented view of the data due to the absence of a unique patient identifier, and many of the data quality problems, originate at the source of the data which is outside of the control of the CDW. These problems are causing a lot of expense and inconvenience and even then results being obtained are not 100%. NHLS management need to address this problem at its source and insist on complete and correct data is entered by medical practitioners. The other issue of serious importance is the CDW staff shortage, which is interfering with work that needs to be done and inhibiting use of the CDW data. Several options for alleviating this crisis were suggested, and should be seriously considered by NHLS and CDW management. A need for the supplementation of CDW data with mortality data, hospital and clinic data and private laboratory data was discussed, but implementation of interfaces with these data sources may be challenging.

Attention should also be paid to the very different perceptions and needs of the NICD and NCR on the one hand, and the NPP and Operational management on the other. As opposed to the NPP and Operational management, NICD and NCR perceived CDW service representatives to be 'limited' because of their difficulty understanding medical data, experienced very much longer time delays, experienced data quality issues to be much worse, were really keen to use metadata and have direct access to CDW data, and had a serious need for the use of external data. Introspection on these differences leads to the realisation that these two types of departments have very different purposes and are peopled by very different individuals. NPP and Operational management activities are mainly administrative in nature with public health tracking being done for operational purposes, and interviewees from that department were managers. NICD and NCR deal more with population health surveillance and interviewees there were medical experts. Consideration should be given to

meeting the needs of NICD and NCR by providing the desired metadata and direct access to CDW data as this may help to alleviate time delays and other problems experienced.

From a research point of view, Activity Theory was found to be an excellent way of understanding the issues and challenges of making an information request within the organisational context. All information systems are situated in some kind of context, and it is recommended that more studies about information systems use Activity Theory as a theoretical framework. This study also demonstrated the value of an empirical study on the use of BI in the population health space, and it is recommended that more empirical studies are done in this area, particularly since only one empirical study was found in the literature. There is a lot of potential for further research of this type at the NHLS. A more focused study on the use of the NHLS CDW specifically for medical research would probably give interesting results. Research concentrating on how the actual information from the CDW could be utilised for decision making to improve population health could also be done. Further investigation into the relative information needs of different types of information requestors from different departments in a single organisation would also be useful. In the broader public health arena, a more detailed look at resolving some of the issues identified in this research, particularly those concerning data quality and staffing, would also be useful. Future work may wish to consider data quality using different dimensions. Since the need has been highlighted for a uniform hospital information system in public health facilities, perhaps research could be conducted into the feasibility of a project of this nature.

7.4 Limitations of the Study

Because this is a case study, generalisability is not a major objective. Rather, this study was done on the specific implementation of a corporate data warehouse at the NHLS. Notwithstanding the lack of intention for results to be generalisable to other situations, there may be useful lessons for other similar implementations of data warehousing.

Due to restrictions of scope, only ten information requestors were interviewed out of only four of the divisions of the NHLS. The NHLS is a very large organisation and there are many more people that request information than were interviewed. If a different subset of ten information requestors were interviewed the results may have been different. It may also be possible that each interviewee did not mention every request for information that they make. Organisations outside of the NHLS that use CDW data, like the DoH and NGOs that work with the DoH also could not be reached because they were not included in the ethics clearance granted. It is therefore feasible that vital elements of the picture are missing, with many applications of the CDW not having been identified. It follows that there is ample opportunity for further research on the NHLS CDW.

This being a qualitative study, a limitation could be that bias may have been introduced due to the fact that the interpretation of the data depends on the researcher's perceptions, as well as on their ability and experience in interpreting qualitative data, particularly since the researcher is the only person coding the data and there will be no opportunity to formally determine inter-coder reliability.

7.5 Conclusion

This study focused on the corporate data warehouse (CDW) of the National Health Laboratory Services (NHLS) in South Africa. As a result of this study, several uses of the NHLS CDW have been identified in the areas of public health tracking, administration and medical research. Challenges and issues encountered while requesting information from the CDW have been explored, with the most important ones being identified as the difficulties obtaining a patient-oriented view of the data, data quality, and staffing issues. Time delays, privacy and security and the need for data from external data sources were identified as issues of a less serious nature. Individual impacts were found to be improved productivity, decision making effectiveness, ability to track progress in their work and a sense of meaningfulness at work, and institutional and population health impacts were found to be the facilitation of the management of internal operations, the provision of data for research, the ability to fulfil the national mandate, the enhancement of international credibility, the ability to secure funding, the ability to make informed changes to practices and the ability to formulate policies. Notwithstanding all these impacts there is still tremendous potential for further use of CDW data. In the words of one of the information requestors interviewed:

". . . it is just the tip of the iceberg of what we could be doing with the data that comes out of CDW. . . . it's a treasure trove of data. . . a goldmine of data"

References

AGARWAL, P. 2014. Benefits and Issues Surrounding Data Mining and its Application in the Retail Industry. *International Journal of Scientific and Research Publications,* 4.

ANDERSON-LEHMAN, R., WATSON, H. J., WIXOM, B. H. & HOFFER, J. A. 2004. Continental Airlines flies high with real-time business intelligence. *MIS Quarterly Executive,* 3**,** 163-176.

ARCHER, B. N., CENGIMBO, A., DE JONG, G. M., KEDDY, K. H., SMITH, A. M., SOOKA, A., NTSHOE, G. & BLUMBERG, L. 2009. Cholera outbreak in South Africa: preliminary descriptive epidemiology on laboratory-confirmed cases, 15 November 2008 to 30 April 2009. *Communicable Diseases Surveillance Bulletin,* 7**,** 3-8.

ASOKAN, G. V. & ASOKAN, V. 2015. Leveraging "big data" to enhance the effectiveness of "one health" in an era of health informatics. *Journal of Epidemiology and Global Health,* 5**,** 311-314.

BARAKAT, S., ALI AL-ZU'BI, H. & AL-ZEGAIER, H. 2013. The role of business intelligence in knowledge sharing: a Case Study at Al-Hikma Pharmaceutical Manufacturing Company. *European Journal of Business and Management,* 5.

BATINI, C., CAPPIELLO, C., FRANCALANCI, C. & MAURINO, A. 2009. Methodologies for Data Quality Assessment and Improvement. *ACM Computing Surveys,* 41.

BEN SAID, M., LE MIGNOT, L., MUGNIER, C., RICHARD, J. B., LE BIHAN-BENJAMIN, C., JAIS, J. P., GUILLON, D., SIMONET, A., SIMONET, M. & LANDAIS, P. 2005. A Multi-Source Information System via the Internet for End-Stage Renal Disease: Scalability and Data Quality. *Stud Health Technol Inform,* 116**,** 994-9.

BENBASAT, I., GOLDSTEIN, D. & MEAD, M. 1987. The Case Research Strategy in Studies in Information Systems. *MIS Quarterly,* 11**,** 369-386.

BERNDT, D. J. Consumer decision support systems: a health care case study. System Sciences, 2001. Proceedings of the 34th Annual Hawaii International Conference on, 6-6 Jan. 2001 2001. 10 pp.

BERNDT, D. J., FISHER, J. W., HEVNER, A. R. & STUDNICKI, J. 2001. Healthcare data warehousing and quality assurance. *Computer,* 34**,** 56-65.

BERNDT, D. J., HEVNER, A. R. & STUDNICKI, J. Hospital discharge transactions: a data warehouse component. System Sciences, 2000. Proceedings of the 33rd Annual Hawaii International Conference on, 4-7 Jan. 2000 2000. 10 pp. vol.1.

BERNDT, D. J., HEVNER, A. R. & STUDNICKI, J. 2003. The Catch data warehouse: support for community health care decision-making. *Decision Support Systems,* 35**,** 367-384.

BHATTACHERJEE, A. 2012. Social science research: principles, methods, and practices.

BRITZ, E., PEROVIC, O., VON MOLLENDORF, C., VON GOTTBERG, A., IYALOO, S., QUAN, V., CHETTY, V., SRIRUTTAN, C., ISMAIL, N. A. & NANOO, A. 2016. The epidemiology of meningitis among adults in a South African province with a high HIV prevalence, 2009-2012. *PloS one,* 11**,** e0163036.

CATALYST, H. 2016. Interview with Stanford CIO, Carolyn Byerly: Launching A Clinical Data Warehouse in Months. *Health Catalyst* [Online].

CHAKI, S. 2015. Enterprise Information Management in Practice. Springer.

COEBERGH, J. W., VAN DEN HURK, C., ROSSO, S., COMBER, H., STORM, H., ZANETTI, R., SACCHETTO, L., JANSSEN-HEIJNEN, M., THONG, M., SIESLING, S. & VAN DEN EIJNDEN-VAN RAAIJ, J. 2015. EUROCOURSE lessons learned from and for population-based cancer registries in Europe and their programme owners: Improving performance by research programming for public health and clinical evaluation. *European Journal of Cancer,* 51**,** 997-1017.

CRAWFORD, K. & HASAN, H. 2006. Demonstrations of the Activity Theory Framework for research in Information Systems. *Australsian Journal of Information Systems,* 13.

DA SILVA, J., DE OLIVEIRA, A. G., FIDALGO, R. N., SALGADO, A. C. & TIMES, V. C. 2010. Modelling and querying geographical data warehouses. *Information Systems,* 35**,** 592-614.

DEBAY, D. 2013. *Mathematical self-efficacy and understanding: Using Geographical Information Systems to mediate urban high school studentss' real-world problem solving.* Doctor of Philosophy, Boston College: Lynch School of Educations.

DEVOE, J. E. & SEARS, A. 2013. The OCHIN community information network: bringing together community health centers, information technology, and data to support a patient-centered medical village. *J Am Board Fam Med,* 26**,** 271-8.

DLAMINI-MVELASE, N. R., WERNER, L., PHILI, R., CELE, L. P. & MLISANA, K. P. 2014. Effects of introducing Xpert MTB/RIF test on multi-drug resistant tuberculosis diagnosis in KwaZulu-Natal South Africa. *BMC infectious diseases,* 14**,** 442.

ECKERSON, W. 2006. *Performance Dashboards: Measuring, Monitoring and Managing Your Business,* New Jersey, Wiley.

ENGESTROM, Y. 1987. *Learning by expanding: An activity-theoretical approach to developmental research*.

ENGESTROM, Y. 2001. Expansive Learning at Work: toward an activity theoretical reconceptualisation. *Journal of Education and Work,* 14.

ERASMUS, L., KOORNHOF, H. & COETZEE, G. 2008. Multi-drug resistant and extensively drug resistant tuberculosis in South Africa from data extracted from the NHLS corporate data warehouse. *Communicable Diseases Surveillance Bulletin,* 6**,** 8-13.

FINK, A. S. The role of the researcher in the qualitative research process. A potential barrier to archiving qualitative data. Forum Qualitative Sozialforschung/Forum: Qualitative Social Research, 2000.

FOX, C., LEVITIN, A. & REDMAN, T. 1994. The notion of data and its quality dimensions. *Information Processing and Management,* 30**,** 9-19.

GANGADHARAN, G. R. & SWAMI, S. N. Business intelligence systems: design and implementation strategies. Information Technology Interfaces, 2004. 26th International Conference on, 2004. IEEE, 139-144.

GARAY, J., CARTAGENA, R., ESENSOY, A. V., HANDA, K., KANE, E., KAW, N. & SADAT, S. 2015. Strategic analytics: towards fully embedding evidence in healthcare decision-making. *Healthc Q,* 17 Spec No**,** 23-7.

GOODHUE, D. L. & THOMPSON, R. L. 1995. Task-Technology Fit and Individual Performance. *MIS Quarterly,* 19**,** 213-236.

HASAN, H. 1999. Integrating IS and HCI using Activity Theory as a philosophical and theoretical basis. *Australasian Journal of Information Systems,* 6**,** 44-55.

IGIRA, F. T. 2008. The situatedness of work practices and organizational culture: implications for information systems innovation uptake. *Journal of Information Technology,* 23**,** 79-88.

INMON, W. H. 1995. What is a data warehouse? *Prism Tech Topic,* 1**,** 1.

INMON, W. H. 2005. *Building the Data Warehouse,* Indianapolis, Wiley.

JEAN-BAPTISTE, R., TOUBIANA, L., LE MIGNOT, L., BEN SAID, M., MUGNIER, C., LE BIHAN-BENJAMIN, C., JAIS, J. P. & LANDAIS, P. 2005. A Web-based GIS for health care decision-support. *AMIA Annu Symp Proc***,** 365-9.

JOHNSON, N. A. 2004. *The dance of sustaining technology innovations in teacher education: Case of an award-winning program.* Doctor of Philosophy, Iowa State University.

JOSHI, S., BARRETT, M., WALSHAM, G. & CAPPLEMAN, S. 2007. Balancing Local Knowledge Within Global Organisations Through Computer-Based Systems: An Activity Theory Approach. *Journal of Global Information Management,* 15.

KAPTELININ, V. 1996. Activity Theory: Implications for Human-Computer Interactions. *In:* NARDI, B. A. (ed.) *Context and Consciousness: Activity Theory and Human-Computer Interaction.* Cambridge: MIT Press.

KIMBALL, R., REEVES, L., ROSS, M. & THORNTHWAITE, W. 1998. *The Data Warehouse Lifecycle Toolkit: Expert Methods for Designing, Developing and Deploying Data Warehouses,* USA, Wiley.

KINDIG, D. 2007. Understanding Population Health Terminology. *The Milbank Quarterly,* 85**,** 139-161.

KLEIN, H. K. & MYERS, M. D. 1999. A Set of Principles for Conducting and Evaluating Interpretive Field Studies in Information Systems. *MIS Quarterly,* 23**,** 67-93.

KRUMHOLZ, H. M. 2014. Big data and new knowledge in medicine: the thinking, training, and tools needed for a learning health system. *Health Aff (Millwood),* 33**,** 1163-70.

KULKARNI, S. S., CHAUDHARI, N., MULLA, J., KAMBLE, A. & SOFI, B. 2017. Review on Business Intelligence solutions in Healthcare sector. *International Journal of Advanced Research in Computer and Communication Engineering,* 6.

KUUTTI, K. 1996. Activity Theory as a potential framework for human-computer interaction research. *Context and Consciousness: Activity Theory and Human Computer Interaction ????*

LEE, J. C. 2010. *Tensions in a Nepali Telecenter: An ethnographic look at progress using Activity Theory.* Doctor of Education in Educational Technology, Pepperdine University.

LEE, Y. W., STRONG, D. M., KAHN, B. K. & WANG, R. Y. 2002. AIMQ: a methodology for information quality assessment. *Information & Management 40 (2002) 133–146,* 40**,** 133-146.

LIYANAGE, H., DE LUSIGNAN, S., LIAW, S. T., KUZIEMSKY, C. E., MOLD, F., KRAUSE, P., FLEMING, D. & JONES, S. 2014. Big Data Usage Patterns in the Health Care Domain: A Use Case Driven Approach Applied to the Assessment of Vaccination Benefits and Risks. Contribution of the IMIA Primary Healthcare Working Group. *Yearb Med Inform,* 9**,** 27-35.

MAZANDERANI, A. H., MOYO, F. & SHERMAN, G. G. 2017. Missed diagnostic opportunities within South Africa's early infant diagnosis program, 2010–2015. *PloS one,* 12**,** e0177173.

MERCY, J. A., IKEDA, R. & POWELL, K. E. 1998. Firearm-related injury surveillance: An overview of progress and the challenges ahead. *American Journal of Preventive Medicine,* 15**,** 6-16.

MOHER, D., SHAMSEER, L., CLARKE, M., GHERSI, D., LIBERATI, A., PETTICREW, M., SHEKELLE, P., STEWART, L. A. & GROUP, P.-P. 2015. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic Reviews,* 4.

MOONEY, S. J., WESTREICH, D. J. & EL-SAYED, A. M. 2015. Commentary: Epidemiology in the era of big data. *Epidemiology,* 26**,** 390-4.

MUSA, G. J., CHIANG, P. H., SYLK, T., BAVLEY, R., KEATING, W., LAKEW, B., TSOU, H. C. & HOVEN, C. W. 2013. Use of GIS Mapping as a Public Health Tool-From Cholera to Cancer. *Health Serv Insights,* 6**,** 111-6.

NAMBIAR, R., BHARDWAJ, R., SETHI, A. & VARGHEESE, R. A look at challenges and opportunities of Big Data analytics in healthcare. Big Data, 2013 IEEE International Conference on, 6-9 Oct. 2013 2013. 17-22.

NANOO, A., IZU, A., ISMAIL, N. A., IHEKWEAZU, C., ABUBAKAR, I., MAMETJA, D. & MADHI, S. A. 2015. Nationwide and regional incidence of microbiologically confirmed pulmonary tuberculosis in South Africa, 2004–12: a time series analysis. *The Lancet Infectious Diseases,* 15**,** 1066-1076.

NARDI, B. A. 1996. Activity Theory and Human-Computer Interaction. *In:* NARDI, B. A. (ed.) *Context and consciousness: activity theory and human-computer interaction.* Cambridge: MIT Press.

NEGASH, S. 2004. Business Intelligence. *Communications of the Association for Information Systems,* 13.

NEWS@ORCHARD 2015. Crystal Run Healthcare Using Data Analytics as an Integral Part of Improving Patient Care. *news@Orchard.* Indiana, USA: Orchard Software Corporation.

NHLS 2015. Annual Report 2014/2015 - National Health Laboratory Service.

NHLS. 2017. *National Health Laboratory Service Web Site* [Online]. [Accessed 2018].

OATES, B. J. 2006. *Researching Information Systems and Computing,* London, SAGE Publications.

OLSZAK, C. M. & BATKO, K. The use of business intelligence systems in healthcare organizations in Poland. Computer Science and Information Systems (FedCSIS), 2012 Federated Conference on, 9-12 Sept. 2012 2012. 969-976.

PARKER, S. T. 1979. General and theoretical: mind in society: the development of higher psychological processes. ls vygotsky. *American Anthropologist,* 81**,** 956-957.

PFEIFFER, D. U. & STEVENS, K. B. 2015. Spatial and temporal epidemiological analysis in the Big Data era. *Prev Vet Med*.

PHAN, D. D. & VOGEL, D. R. 2010. A model of customer relationship management and business intelligence systems
for catalogue and online retailers. *Information & Management 47,* 47**,** 69-77.

POSTHUMUS, R. & BANK, S. 2008. GIS as a tool in business intelligence. *Position IT*.

PRABDIAL-SING, N., PUREN, A. & SCHOUB, B. 2013. The status of hepatitis c–the slient "volcano"–in South Africa. *National Institute or Communicable Diseaess*.

RICHARD, J. B., TOUBIANA, L., LE MIGNOT, L., BEN SAID, M., MUGNIER, C., LE BIHAN-BENJAMIN, C., JAIS, J. P. & LANDAIS, P. 2005. SIGNe: A Geographic Information System on the Web for End-Stage Renal Disease. *Stud Health Technol Inform,* 116**,** 713-8.

RIZI, S. A. & ROUDSARI, A. 2013. Development of a public health reporting data warehouse: lessons learned. *Stud Health Technol Inform,* 192**,** 861-5.

ROGERS, E. 2010. *Diffusion of Innovations*, Simon and Schuster.

SAUNDERS, M., LEWIS, P. & THORNHILL, A. 2009. *Research methods for business students,* Italy, Prentice Hall.

SHANI, M. 2000. The impact of information on medical thinking and health care policy. *International Journal of Medical Informatics,* 58–59**,** 3-10.

SHERMAN, G. & LILIAN, R. 2011. Early infant diagnosis of HIV infection in South Africa: 2008 to 2010. NHLS.

SHERMAN, G., LILIAN, R., BARRON, P., CANDY, S. & BHARDWAJ, S. 2014a. Laboratory information system data demonstrate successful implementation of the prevention of mother-to-child transmission programme in South Africa: prevention of mother-to-child transmission-Progress towards the Millennium Development Goals. *South African Medical Journal,* 104**,** 235-238.

SHERMAN, G. G., MAZANDERANI, A. H., BARRON, P., BHARDWAJ, S., NIIT, R., OKOBI, M., PUREN, A., JACKSON, D. J. & GOGA, A. E. 2017. Toward elimination of mother–to–child transmission of HIV in South Africa: how best to monitor early infant infections within the Prevention of Mother–to–Child Transmission Program. *Journal of global health,* 7.

SHERMAN, G. G., R., L. R., BARRON, P., CANDY, S. & BHARDWAJ, S. 2014b. Laboratory information system data demonstrate successful implementation of the prevention of mother-to-child transmission programme in South Africa : prevention of mother-to-child transmission - Progress towards the Millennium Development Goals. *South African Medical Journal,* 104**,** 235-238.

SHOLLO, A. & KAUTZ, K. Towards an Understanding of Business Intelligence. ACIS, 2010.

SIESLING, S., LOUWMAN, W. J., KWAST, A., VAN DEN HURK, C., O'CALLAGHAN, M., ROSSO, S., ZANETTI, R., STORM, H., COMBER, H., STELIAROVA-FOUCHER, E. & COEBERGH, J. W. 2015. Uses of cancer registries for public health and clinical research in Europe: Results of the European Network of Cancer Registries survey among 161 population-based cancer registries during 2010–2012. *European Journal of Cancer,* 51**,** 1039-1049.

SILVA, J. C., SHAH, S. C., RUMORO, D. P., BAYRAM, J. D., HALLOCK, M. M., GIBBS, G. S. & WADDELL, M. J. 2013. Comparing the accuracy of syndrome surveillance systems in detecting influenza-like illness: GUARDIAN vs. RODS vs. electronic medical record reports. *Artificial Intelligence in Medicine,* 59**,** 169-174.

SINGH, E., RUFF, P., BABB, C., SENGAYI, M., BEERY, M., KHOALI, L., KELLETT, P. & UNDERWOOD, J. M. 2015. Establishment of a cancer surveillance programme: the South African experience. *The Lancet Oncology,* 16**,** e414-e421.

SINGH, E., SENGAYI, M., URBAN, M., BABB, C., KELLETT, P. & RUFF, P. 2013. The South African National Cancer Registry: an update. *Health*.

TREMBLAY, M. C., BERNDT, D. J. & HEVNER, A. R. 2009. Measuring information volatility in a health care information supply chain. *Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology.* Philadelphia, Pennsylvania: ACM.

TREMBLAY, M. C., FULLER, R., BERNDT, D. & STUDNICKI, J. 2007. Doing more with more information: Changing healthcare planning with OLAP tools. *Decision Support Systems,* 43**,** 1305-1320.

TSUI, F. C., ESPINO, J. U., DATO, V. M., GESTELAND, P. H., HUTMAN, J. & WAGNER, M. M. 2003. Technical description of RODS: a real-time public health surveillance system. *J Am Med Inform Assoc,* 10**,** 399-408.

WAGNER, M. M., ROBINSON, J. M., TSUI, F.-C., ESPINO, J. U. & HOGAN, W. R. 2003. Design of a national retail data monitor for public health surveillance. *Journal of the American Medical Informatics Association,* 10**,** 409-418.

WALJI, M. F., KALENDERIAN, E., STARK, P. C., WHITE, J. M., KOOKAL, K. K., PHAN, D., TRAN, D., BERNSTAM, E. V. & RAMONI, R. 2014. BigMouth: a multi-institutional dental data repository. *J Am Med Inform Assoc,* 21**,** 1136-40.

WALSHAM, G. 1995. Interpretive case studies in IS research: nature and method. *European Journal of information systems,* 4**,** 74-81.

WALSHAM, G. 2006. Doing interpretive research. *European Journal of Information Systems,* 15**,** 320-330.

WAND, Y. & WANG, R. Y. 1996. Anchoring Data Quality Dimensions in Ontological Foundations. *Communications of the ACM,* 39.

WATSON, H., ARIYACHANDRA, T. & MATYSKA, R. J. 2001. Data warehousing stages of growth. *Information Systems Management,* 18**,** 42-50.

WATSON, H. J. 2009. Tutorial: Business Intelligence – Past, Present, and Future. *Communications of the Association for Information Systems,* 25.

WATSON, H. J. 2010. BI-based Organizations. *Business Intelligence Journal,* 15**,** 4-6.

WATSON, H. J., WIXOM, B. H., HOFFER, J. A., ANDERSON-LEHMAN, R. & REYNOLDS, A. M. 2006. Real-time business intelligence: Best practices at Continental Airlines. *Information Systems Management,* 23**,** 7.

WHITTEMORE, R., CHASE, S. K. & MANDLE, C. L. 2001. Validity in Qualitative Research. *QUALITATIVE HEALTH RESEARCH, Vol. 11 No. 4, July 2001 522-537,* 11**,** 522-537.

WISNIEWSKI, M. F., KIESZKOWSKI, P., ZAGORSKI, B. M., TRICK, W. E., SOMMERS, M., WEINSTEIN, R. A. & FOR THE CHICAGO ANTIMICROBIAL RESISTANCE, P. 2003. Development of a clinical data warehouse for hospital infection control. *Journal of the American Medical Informatics Association,* 10**,** 454-462.

WIXOM, B. H. & WATSON, H. J. 2001. An Empirical Investigation of the Factors Affecting Data Warehousing Success. *MIS Quarterly*.

WIXOM, B. H., WATSON, H. J. & WERNER, T. 2011. Developing an Enterprise Business Intelligence Capability: The Norfolk Southern Journey. *MIS Quarterly Executive,* 10**,** 67-71.

XIAOHUI, Z., FIEDLER, R. & POPOVICH, M. 2004. A biointelligence system for identifying potential disease outbreaks. *Engineering in Medicine and Biology Magazine, IEEE,* 23**,** 58-64.

YEOH, W. & KORONIOS, A. 2010. Critical Success Factors for Business Intelligence Systems. *Journal of computer information systems,* 50**,** 23-32.

YIN, R. K. 2003. *Case Study Research: Design and Methods,* USA, SAGE Publications.

APPENDIX A: Interview Protocol

Demographic and General Questions
1. What institution are you employed by?
2. What is your position at your institution?
3. What is your role at your institution?
4. How long have you been performing this role?
5. How often do you request information from the NHLS Corporate Data Warehouse(CDW)?
6. What requests for information from the NHLS CDW do you make? (Obtain a list)

(PTO)

The following questions must be repeated for each different information request		
Object and Outcome – Impact of this request on Population Health	**Source**	**AT Element(s)**

7. Can you explain for what purpose you request this information from the NHLS CDW?
• Can you explain:
  • why you are requesting this information?
  • in what ways this information is helping you?
  • in what ways this information is helping the NHLS?
  • in what ways you feel that the information may be improving population health?
		Outcome

8. Have you traced or recorded any of the benefits of this information for population health? • Could you provide me with any of the evidence? • Are you aware of any other documentation or indication of the benefits of this information to population health?		Outcome
9. Are you aware of any changes to population health policy that have resulted from information requested from the NHLS CDW? • Can you explain what they are/ provide me with evidence?		Outcome

Type of Request		
10. Can you please describe how you access information from the NHLS CDW?		Subject, Tools, Rules, Community
• Through a person to whom you submit an information request?	(Tremblay et al., 2007)	
• Through a special program (app) / dashboard /report that has been created for your use?	(Berndt, 2001, Berndt et al., 2000, Nambiar et al., 2013) and Conversation with NHLS staff member	
• Through general purpose information requesting software made available for use with the CDW (eg. Data warehouse browsing facility, OLAP)?	(Xiaohui et al., 2004, Berndt, 2001, Berndt et al., 2000)	
• Through using SQL to retrieve information from the CDW or using advanced analytics?	(Liyanage et al., 2014, Berndt et al., 2000)	

Issues and Challenges for this request		
The following questions only apply if the request is made through a person		
11. Can you please describe how easy and efficient it is for you to request the information you need? • How would you rate the service of the person who helps you? • How quickly do you get results?		Subject, Community, Division of Labour
12. Is sufficient assistance and expertise available to you to guide you as to <u>what</u> information should be requested to best suit your purpose?	(Tremblay et al., 2007)	Subject, Community, Division of Labour
The following questions only apply if the request is NOT made through a person		
13. Can you please describe how easy it is for you to make the request for the information you need? • Describe any training you may have received for the facilities (program /app/ dashboard / report / other) that have been provided for making information requests? • Describe how easy you find these facilities to use • Describe the how intuitive you find the facilities to use	(Berndt, 2001, Rizi and Roudsari, 2013)	Subject, Tools
14. Discuss what assistance and expertise is available to you • To show you <u>how</u> to make the information request • To guide you as to <u>what</u> information should be requested to best suit your purpose • To make sure the technology is working and the CDW and associated facilities are available	(Tremblay et al., 2007)	Subject, Community, Division of Labour
15. Is this assistance available from staff of the NHLS or do you need to request it from another company? • In your opinion, does requesting assistance from another company have implications?	(Wisniewski et al., 2003, Rizi and Roudsari, 2013)	Subject, Community, Division of Labour
The following questions apply whether the request is made through a person or not		
16. Can you explain how you know what information is available in the CDW and how to select the parts of it that you want. • Would it help to have a lookup facility that would clarify the meanings of field names and field content?		Subject, Tools, Object
17. Are you able to request <u>only</u> that precise part of the information that is required?	(Rizi and Roudsari, 2013)	Subject, Tools
18. Are you satisfied with the degree to which you can specify exactly which information is requested? - Are you able to choose and modify the fields to view, filter by and summarise by? - Would it be useful to have a software tool that would enable you to access the CDW and choose the fields to view, filter by and summarise by?	(Berndt et al., 2000, Rizi and Roudsari, 2013)	Subject, Tools

Object and Outcome for this request		
19. To what extent do you trust the information you get from the NHLS CDW? • Can you explain how accurate (correct) you feel the data is? • Can you explain any problems you may have experienced with missing data? • Can you explain how consistent you have found the information to be? • Can you explain how up to date you find the information to be? • Can you explain the extent to which these aspects of the data affect your proposed outcome?	(Berndt et al., 2000, Rizi and Roudsari, 2013, Coebergh et al., 2015, Mercy et al., 1998, Fox et al., 1994, Siesling et al., 2015)	Object, Tools
20. Do you find the information received from the NHLS CDW to be relevant? • Does it provide appropriate assistance in achieving your current objective?	(Tremblay et al., 2007)	Object, Tools
21. Have you been able to obtain all the data you need from the NHLS CDW, or could your objective only be achieved by obtaining information from other sources as well? • Are you aware of any data sharing agreements that the NHLS has with other organisations? • Would your proposed outcome benefit from the inclusion of mortality information from the death registry?	Previous interview	Object, Tools
22. Can you explain to what extent to which the outcome of your request requires a patient-oriented view of the data? • Were any difficulties encountered in acquiring this view of the data?	Previous interview	Object, Tools, Outcome

The following questions must be asked once, after all requests are dealt with		
Issues and Challenges	**Source**	**AT Element(s)**
23. Have privacy and security restrictions on the CDW prevented you from getting to the information that you need? • Has anything been done to 'anonymise' the data and make this less of an issue? • Are you aware of any security and privacy policies or laws that may prevent access to the data? • Is there anything that you have done to get around this problem?	(Berndt et al., 2000, Garay et al., 2015, Rizi and Roudsari, 2013, Wisniewski et al., 2003, Coebergh et al., 2015)	Rules, Object
24. Do you feel you are using the NHLS CDW and associated facilities to their full potential? • If not, what is stopping you from doing so?	(Kimball et al., 1998)	Subject, Tools, Object
25. Are you aware whether the NHLS has a CDW data governance model? • What would you consider to be the advantages and disadvantages of having a governance model?	(Rizi and Roudsari, 2013)	Rules
26. Are there any other issues and challenges that you have experienced when making a request for information from the NHLS CDW?		Subject, Tools, Object, Rules, Community, Division of Labour
General		
27. Are there any comments, compliments or criticisms that you would like to make regarding the information that you have received from the NHLS CDW?		Subject, Tools, Object, Rules, Community, Division of Labour
28. Do you have any documents or other evidence of what we have discussed that you can give to me?		Outcome

APPENDIX B: First Page of Template for Interview Transcription

	dd Month yyyy time **Participant Name (Participant Type)** **at Participant Location**	
Interviewer		
Participant		
Interviewer		
Participant		
Interviewer		
Participant		
Interviewer		
Participant		
Interviewer		
Participant		
Interviewer		
Participant		
Interviewer		
Participant		
Interviewer		
Participant		
Interviewer		
Participant		
Interviewer		
Participant		
Interviewer		
Participant		
Interviewer		
Participant		
Interviewer		
Participant		
Interviewer		

APPENDIX C: Illustration of Unitising Transcribed Interviews and Colour Coding for Themes

Interviewer	Do you find it efficient to get to the information? It just comes to you in an email, so I suppose that's very efficient for you?	**Ease and Efficiency**
Participant	No it is, it does come on to the email, but at least if you ask, if you miss it on the email you can still request it. Because sometimes, you know, electronic – sometimes it doesn't work as you will expect. But you can request it and they will resubmit it to you. So I don't really have any challenge in terms of getting the data if I want it.	
Interviewer	Can you explain - well, this doesn't really apply because you're getting them in a report - but if you ever need to change the report, do you ever decide that you want information in the report that is not in the report, and you need to have it changed?	**Outcomes and Impact**
Participant	It's seldom; it doesn't happen a lot of time. It's very seldom, like for example if I've found that um . . . Like let's take one report. I'm not happy with what performance is projecting, or one of my manager objects to that, then we will go into raw data, and then I will be able to interrogate the raw data as it is to see, in terms of what the outcome was on the summarised report – is it doing that, or to find data where the gaps are, in terms of saying, is it in receiving? Is it in testing? Is it . . . You know I try to analyse.	**Direct Access (OLAP-type tools /other)** **Use**
Interviewer	So in other words you want to drill down into the data.	Through: Service Rep
Participant	Yes, ya, and then I can get that.	
Interviewer	And if you want to drill down into the data, if you put a request in . .	(For finding problems through drilling down into raw data)
Participant	I put a request for them to send me the raw data.	
Interviewer	Ok, and the request goes on a form?	
Participant	Normally I can request it – sometimes, ya, they can give me a form to complete to ask for that data, because it's not normally a standard report which is being done on regular basis, it's on request.	
Interviewer	But if you just want information do you just sometimes just phone them and then they give it to you?	
Participant	Not necessarily – they will require some documentation for records so that they can possibly – I think it's for them to know who took which data and all that, so they do request – it's sort of controlled.	
Interviewer	And, when you're in that position you find that there's help for you from them to get the correct data?	**Relationship with CDW Staff**
Participant	Yes, I did.	
Interviewer	To what extent do you trust the information that you get from them? Do you find the information to be accurate?	**Data Quality**
Participant	For me so far I don't have any reason to doubt that. Because, like the report that	

	I've just alluded to, that if I'm in doubt of the other, I have – because I was also sceptical about the data, so once I get the raw data, I have managed to say there's no – I mean I've managed to confirm what their report has said – except where . . . There's not really a time where I can say I had objected to any data because they have always proved that the data is correct, so I don't have any . . .	
Interviewer	So you find the data to be accurate.	
Participant	Ya, it is.	
Interviewer	Complete – always all there? Consistent? And up to date?	
Participant	Ya, it is up to date.	
Interviewer	For you it needs to be up to date, because you needing it on an operational basis. And do you find the information to be relevant to you?	
Participant	Ya, for the specific information I request it is.	
Interviewer	And it helps you with your job.	
Participant	Ya, it helps me with my job, ya.	
Interviewer	Have you been able to get all the data you need from the CDW, or have you had to get data from other sources as well?	**External Data**
Participant	No . . . um . . . there is, I can, there is another data source we have. We're running on – in terms of data I can extract from another – I can extract from there to compare if the numbers are tallying. Because we've got Oracle, which is also now the CDW extract – I don't know how the technical issues that – but the information moves from Oracle into CDW. Because we're using almost 3 systems – the laboratory information system -	
Interviewer	That's TrakCare.	
Participant	Ya, that's TrakCare. And then we've got Oracle which is mainly our financial modules doing the billing and all that. And then you get CDW which is mainly for data, but it links with this to extract some of the data into that. So I can run for volumes, and if I look to look I can look them at Oracle, or I can look at them as well at the Trak, but mostly I use mainly the two – the Oracle and CDW to compare.	
Interviewer	So the only time you really use the Oracle is when you want to verify what you've found. Do you find that most of the time . . .	**Data Quality**
Participant	Most of the time – I mean it's neglible – you know there are room for error. It's always there because it will never be – because we are aware of it, that what do we expect it to be averaging. Because they're dependant on the time of the cut off – because they are cut off when you cut off the closing, because you can't leave it so they . . .	
Interviewer	So if you have different closing times you have different values.	
Participant	Ya, so you will have - but it's neglible, you can compare and see that it's . . .	

APPENDIX D: Participant Information Sheet

EXPLORING AND DESCRIBING THE USE AND IMPACT OF THE NHLS CDW FOR POPULATION HEALTH IN SOUTH AFRICA

Good Day

I am a researcher conducting a study on the use of the National Health Laboratory Services (NHLS) Corporate Data Warehouse (CDW), and the possible impact of this use on population health.
As a person who makes requests for information from the NHLS CDW, you are **invited** to take part in a face-to-face interview. The purpose of the interview is to better understand the information request process, to establish the quality and value of the information obtained from the NHLS CDW, and to explore any positive impact that this information may have on population health. I have purposefully chosen to invite you to participate because of your knowledge and experience with the services available through the NHLS CDW.

Why is this being done?

With the prevalence in South Africa of communicable diseases like AIDS and Tuberculosis, the increase in the incidence of non-communicable diseases like Cardiovascular disease and Diabetes, as well as health care access, management and insurance issues, it is clear that the area of population health in South Africa is in need of urgent attention. Initiatives in this regard benefit from the use of trustworthy and relevant information about population health that decisions can be based on. Business Intelligence is a form of information technology that enables the collection, extraction and analysis of information that has been captured as a result of computerisation. Data warehouses, like the NHLS CDW, are one of the components of a business intelligence infrastructure. They gather information from different sources within an organisation and outside of the organisation, creating an exceptionally useful pool of information to be analysed and used for decision making. The NHLS is responsible for all blood tests done at public health institutions throughout South Africa, and as such keep blood test records for 80% of the population, all of which are accessible from the NHLS CDW. This is an extremely valuable and useful pool of information that could contribute to the betterment of population health in South Africa. This study will explore and describe the ways in which the NHLS CDW is being used, and the impact that this use is perceived to be having on population health in South Africa.

What is expected of participants?

You are invited to a one hour interview where questions about the NHLS CDW information request process will be asked.

- How you use the NHLS CDW
- The nature of assistance and facilities available to you in your use of the NHLS CDW
- Your trust and satisfaction with the information received from the NHLS CDW
- Restrictions to and/or challenges with the use of the NHLS CDW

May I withdraw from the study?

Your participation is completely voluntary. You are free to decide whether or not you wish to be interviewed. There will be no consequences to you whether you participate or not. Note that you may withdraw from the interview at any stage. If you do not wish to answer any particular question or if you wish to leave the interview at any point then we will respect and honour your decision.

What about confidentiality?

Your response is important and there are no right or wrong answers. Your name and affiliation will be removed from the transcription of the interview. Thus confidentiality of your details will be ensured in transcriptions and reporting of

findings. You may withdraw from the interview at any time and are not obliged to respond to any particular question. You may choose to leave the room at any time during the interview without the need to give any reason for leaving.

What about audio recording?

The interview will be audiotaped. A separate consent form is included to provide for this. The audio recordings will only be available to the researchers and will not be shared with any other parties. The recordings will be stored securely for a period of up to five years before being destroyed.

Risks and Benefits

There are no risks associated with participation. Participation in this research is entirely voluntary and as a result no financial reward or incentive will be given. Although there will be no direct benefit to yourself, the study hopes to improve the use of the NHLS CDW and its resulting impact on population health by the lessons learnt.

Permission

Official permission to conduct this research has been granted by the NHLS. Ethical permission to conduct this research has been granted by the Human Research Ethics Committee (Medical) of the University of the Witwatersrand. Should you require any information regarding your rights as a research respondent, or have any complaints regarding this study, you may contact the HREC Chairperson, Professor Peter Cleaton-Jones, Tel: 0117172301; email: peter.cleaton-jones1@wits.ac.za or Ms Zanele Ndlovu, Secretariat, Tel: 0117171252/1234/2700, email: Zanele.ndlovu@wits.ac.za.
If you are willing to participate, we will request you to sign the formal consent form for both the participation and the recording of the interview. Copies of consent forms will be provided to you.

Who should I contact if I have any questions?

Should you have any questions, or should you wish to obtain a copy of the results of the interview, please contact Lyrice Cohen on +2711-717-8153 lyrice.cohen@wits.ac.za
Thank you for considering participating.
Sincerely

_____
Lyrice Cohen, University of the Witwatersrand, Johannesburg

APPENDIX E: Interview Consent Form



CONSENT FORM - INTERVIEW

Concerning participation in the Research Project:

EXPLORING AND DESCRIBING THE USE AND IMPACT OF THE NHLS CDW FOR POPULATION HEALTH IN SOUTH AFRICA

Investigator: Lyrice Cohen
University of the Witwatersrand, Johannesburg

I understand that I have been invited to participate in an interview.
I have heard the aims and objectives of the Research Project that are proposed. I was given opportunity to ask questions and was also given enough time to think about this Research Project. I have not been forced or pushed in any way to take part. I am clear about the aims of the Research Project.
I understand that taking part in this Research Project is completely voluntary i.e. of my own choice. I know that I may withdraw from it at any time without giving any reasons.
I know that the results of this Research Project will be used for scientific and educational purposes and that may include being published. I agree to this, provided any identifying data of region and name are removed.
I hereby agree to participate in the interview in this Research Project.

.....	.....	.....
Name of participant	Signature of participant	Witness

.....	.....
Place	Date

_____

Statement by the interviewer:
I have given written and oral information regarding this Research Project to the participant.
I agree to answer any future questions concerning the Project as best as I am able.
I will adhere to the protocol as it has been approved.

…… . ……. ……. …….
Name of interviewer Signature Date Place

APPENDIX F: Recording Consent Form



CONSENT FORM - AUDIOTAPING

Concerning audiotaping:

EXPLORING AND DESCRIBING THE USE AND IMPACT OF THE NHLS CDW FOR POPULATION HEALTH IN SOUTH AFRICA

Investigator: Lyrice Cohen
University of the Witwatersrand, Johannesburg

I understand that I have been invited to participate in an interview.
I have heard the aims and objectives of the Research Project that are proposed. I was given opportunity to ask questions and was also given enough time to think about this Research Project. I have not been forced or pushed in any way to take part. I am clear about the aims of the Research Project.
I understand that taking part in this Research Project is completely voluntary i.e. of my own choice. I know that I may withdraw from it at any time without giving any reasons.
I do understand that the interview will be audiotaped, and that this recording will be used only for the purposes of analysing as data in this Research Project. I have been told that only the researchers above will be able to access the tape recording. I have also been told that when this analysis is complete these recordings will be destroyed. I agree to participate in the interview with me being audio-taped.
I know that the results of this Research Project will be used for scientific and educational purposes and that may include being published. I agree to this, provided any identifying data of region and name are removed.
I hereby agree to being audiotaped in the interview in this Research Project.

.....	.....	.....
Name of participant	Signature of participant	Witness

.....	.....
Place	Date

_____

Statement by the interviewer:
I have given written and oral information regarding this Research Project to the participant.
I agree to answer any future questions concerning the Project as best as I am able.
I will adhere to the protocol as it has been approved.

.....	.....	.....	.....
Name of interviewer	Signature	Date	Place

APPENDIX G: Letter of Permission from the NHLS

NATIONAL HEALTH LABORATORY SERVICE

Academic Affairs and Research
Modderfontein Road, Sandringham, 2031
Tel: +27 (0)11 386 6142
Fax: +27 (0)11 386 6296
Email: babatyi.kgokong@nhls.ac.za
Web: www.nhls.ac.za

13 September 2017

Applicant: Ms Lyrice Cohen
Institution: University of the Witwatersrand
Faculty: School of Economic and Business Sciences
Department: Information System Division
Email: Lyrice.Cohen@wits.ac.za

Re: Provisional Approval to access National Health Laboratory Service (NHLS) Data

Your application to undertake a research project "**The Use and Impacts of a Corporate Data Warehouse: The Case of the National Health Laboratory Service**" using data from the NHLS database has been reviewed. This letter serves to advise that the application has been provisionally approved.

Please note that final approval will be granted on your compliance with the NHLS conditions of service and that the study can only be undertaken provided that the following conditions have been met.

- Ethics approval is obtained from a recognised SA Health Research Ethics Committee.
- Processes are discussed with the relevant NHLS departments (i.e. Information Management Unit and Operations Office) and are agreed upon.
- Confidentiality is maintained at participant and institutional level and there is no disclosure of personal information or confidential information as described by the NHLS policy.
- A final report of the research study and any published paper resulting from this study are submitted and addressed to the NHLS Academic Affairs and Research office and the NHLS has been acknowledged appropriately.
- NHLS Data cannot be used to track patients as no pre-approval/consent is obtained from Patients.

Please note that this letter constitutes approval by the NHLS Academic Affairs and Research Office. Any data related queries may be directed to NHLS Corporate Data Warehouse, contact number: 011 386 6074 email: zarina.sabat@nhls.ac.za

Dr Babatyi Malope-Kgokong
National Manager: Academic Affairs and Research

APPENDIX H: Final Ethics Clearance Certificate

R14/49 Ms L Cohen

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)
CLEARANCE CERTIFICATE NO. M170642

NAME: (Principal Investigator)	Ms L Cohen
DEPARTMENT:	School of Economics and Business Sciences Information Systems Division
PROJECT TITLE:	The use and impacts of a corporate data warehouse: the case of the National Health Laboratory Service
DATE CONSIDERED:	30/06/2017
DECISION:	Approved unconditionally
CONDITIONS:	
SUPERVISOR:	Professor J Cohen
APPROVED BY:	Professor PE Cleaton-Jones, Chairperson, HREC (Medical)
DATE OF APPROVAL:	12/10/2017

This clearance certificate is valid for 5 years from date of approval. Extension may be applied for.

DECLARATION OF INVESTIGATORS

To be completed in duplicate and **ONE COPY** returned to the Research Office Secretary on 3rd floor, Phillip V Tobias Building, Parktown, University of the Witwatersrand, Johannesburg.

I/We fully understand the conditions under which I am/we are authorised to carry out the above-mentioned research and I/we undertake to ensure compliance with these conditions. Should any departure be contemplated from the research protocol as approved, I/we undertake to resubmit to the Committee. I agree to submit a yearly progress report. The date for annual re-certification will be one year after the date of convened meeting where the study was initially reviewed. In this case, the study was initially reviewed in June and will therefore be due in the month of June each year. Unreported changes to the application may invalidate the clearance given by the HREC (Medical).

Principal Investigator Signature

October 2017
Date

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES

APPENDIX I: NHLS CDW Information Request Form

NATIONAL HEALTH LABORATORY SERVICE HELPDESK
Tel: (011) 386-6125/6/7/9 Fax: (011) 386-6308 email: helpdesk1@nhls.ac.za

APPLICATION FOR DATA FROM NHLS INFORMATION SYSTEMS (Q-Pulse FMI0069)

Each application will be approved or rejected subject to the ability to extract this data and the availability of the data, and subject to the intended usage of the requested data. Applications that are incomplete and/or do not contain supporting documentation, will be rejected.

APPLICANT'S DETAILS

Applicant Name		Tel No		Email	
Business Role / Designation					
Laboratory / Department / Branch / Region (Internal applicants)					
Organisation (External applicants)					
Supervisor Name		Tel No	( )	Email	
Supervisor Designation					

CONDITIONS

- Data / Information is not to be used in contravention of Sections 14, 15, 16 and 17 of the National Health Act 61 of 2004 and the Promotions of Access to information Act 2 of 2000.
- The applicant undertakes to ensure that the data supplied to it by the NHLS is used ethically and solely for the purposes for which it is provided as detailed in this application, and further acknowledges that it shall remain liable for any breaches of this clause by the end user.
- If the purpose for the data requested in this application is research or if patient identity linked data is required, ethics approval and a one page summary of the protocol shall be attached to this application form. It is the responsibility of the applicant to ensure that their institutions' Human Ethics approval includes explicit authorisation to access the requested NHLS data.
- The applicant undertakes to store the NHLS data in a confidential manner by separating patient identifying details from laboratory data and storing the master list that links patient identifying details to study patient identifiers in a separate, secure location.
- The information is for the private use of the applicant only, unless further approval is obtained from the NHLS. In the event of this, the applicant shall give due credit, including affiliation, of the participation of the NHLS in any such publications or presentations.
- The applicant undertakes to provide the Executive Manager: Academic Affairs, Research and Quality Assurance at the NHLS with a copy of any report, presentation or publication emanating from the use of this data.

ACCEPTANCE OF CONDITIONS

By signing this document we accept the conditions stated above.			
Applicant Signature		Date	/ /20
Supervisor Signature		Date	/ /20

In the event of a dispute concerning this document, the electronic version stored on Q-Pulse will be deemed to be the correct version

National Health Laboratory Service- All rights reserved

NATIONAL HEALTH LABORATORY SERVICE HELPDESK
Tel: (011) 386-6125/6/7/9 Fax: (011) 386-6308 email: helpdesk1@nhls.ac.za

APPLICATION FOR DATA FROM NHLS INFORMATION SYSTEMS (Q-Pulse FMI0069)

All fields in this section must be completed

DATA REQUEST DETAILS

Request Type (Tick)	☐ New ☐ Modify	**Data Format (Tick)**	☐ Excel ☐ CSV	**Data Delivery (Tick)**	☐ CD / DVD ☐ Email
Frequency of Extract (Tick)	☐ Once ☐ Repeat	**If Repeat, specify frequency**	☐ Daily ☐ Weekly		☐ Monthly ☐ Annually

DESCRIPTION OF REQUIRED DATA EXTRACT	
Data required	
Region (for data extract, e.g. Province or Laboratory	
Date range of extract	
Fields required (e.g. Patient name, Date of Birth, etc)	

ADDITIONAL INFORMATION

DESCRIPTION OF INTENDED USE OF DATA EXTRACT (e.g. research, epidemiology study, cost analysis of service, drug effectiveness, disease surveillance)

LIST WHO WILL HAVE ACCESS TO THIS DATA

PROJECT NAME AND REGISTRATION NUMBER (if data is required for a registered research project. Please attach the Ethics Approval.)

NATIONAL HEALTH LABORATORY SERVICE HELPDESK
Tel: (011) 386-6125/6/7/9 Fax: (011) 386-6308 email: helpdesk1@nhls.ac.za

APPLICATION FOR DATA FROM NHLS INFORMATION SYSTEMS (Q-Pulse FMI0069)

NHLS RESPONSIBILITIES

The NHLS will:

- Ascertain if it is possible to extract the required data.
- Register the application and issue a registration number.
- Only release the requested data to the applicant whose name is specified on this application form.

After this application has been completed and approved, please raise a service request with the NHLS IT Service Desk (Contact Number: (011) 386-6125/6/7/9):

- Send an email to helpdesk1@nhls.ac.za, and cc the CDW Manager (sue.candy@nhls.ac.za)
- Scan this application form and attach it to the email, or fax it to (011) 386-6308.

FOR OFFICE USE					
APPROVAL BY BUSINESS (Approval will be obtained by the CDW Manager)					
INFORMATION MANAGEMENT UNIT APPROVAL (required for external requests and patient identifying data)					
Check list for external applicants	☐ Signed by Supervisor ☐ Ethics Approval attached, if applicable				
Executive Manager: Academic Affairs, Research and Quality Assurance		Signature		Date	/ /20
CEO APPROVAL (required for sensitive data requests)					
Chief Executive Officer		Signature		Date	/ /20
APPROVAL BY IT					
CDW Manager		Signature		Date	/ /20
REQUEST TRACKING					
Service Request Number					
Request Commence Date	/ /20				

In the event of a dispute concerning this document, the electronic version stored on Q-Pulse will be deemed to be the correct version

National Health Laboratory Service- All rights reserved

[i] http://www.hsrc.ac.za/uploads/pageNews/72/SANHANES-launch%20edition%20(online%20version).pdf