UNIVERSITY OF THE WITWATERSRAND

FACULTY OF HEALTH SCIENCES

SCHOOL OF PUBLIC HEALTH

Circumstances associated with injury-related deaths and changes in causes of injury-related deaths over three decades in rural northeast South Africa: Insights from text mining of verbal autopsy narratives from Agincourt

By

Zamangema Mlungwana

Student Number – 673023

A Research Report submitted to the University of Witwatersrand in partial fulfillment to the degree, Master of Science Epidemiology

Public Health Informatics

Supervisor/s: Daniel Ohene-Kwofie

: Chodziwadziwa Kabudula

June 2025

DECLARATION

I Zamangema Nokulunga Mlungwana, hereby declare that the work presented in this dissertation is my work (except where acknowledgments indicate otherwise) and that neither the whole work nor any part of it has been, is being, or shall be submitted for another degree at this or any other university, institution for tertiary education or examining body.

Signature: ____________ **Date: 17/06/2025**

Dedication

This piece goes to my late mom Ms NE Nsele in spirit I know I had your full support. To my beautiful children Zamile, Siyamthanda, and Singabakho. Finally, to My husband, Mr V, your support and strengths kept me going.

Acknowledgments

The completion of this dissertation would not have been possible without the support and encouragement of many individuals. I take this opportunity to express my deepest gratitude to all those who contributed to this journey.

First and foremost, I extend my heartfelt appreciation to my supervisors, Mr. Daniel Ohene-Kwofie and Dr. Chodziwadziwa Kabudula, for their unwavering guidance, insightful advice, and continuous encouragement. With no prior programming experience, you patiently guided me, providing the technical support I needed to navigate this process. Your mentorship has been invaluable—thank you.

I am also grateful to the MRC/Wits Agincourt team for granting me access to their data. This research would not have been possible without their generosity and assistance in providing the dataset.

A special thank you to my manager, Mr. S.S. Dlamini, for his understanding and support as I balanced my academic and professional responsibilities. Your patience, particularly during the times when my workload and duties were affected, did not go unnoticed. I also deeply appreciate your technical guidance, especially in data analysis with STATA it played a crucial role in the success of this work.

To my family, thank you for your patience and understanding during this journey. A special mention to Ms. K. Zikhali, my domestic helper, for taking over my motherly duties and allowing me to focus on my studies without worry.

Lastly, to my husband, who embarked on this journey with me you not only supported me but also made the effort to understand the concept entirely unfamiliar to you, just so I could grasp it better. Your unwavering belief in me, your support, and your words of encouragement have meant everything. Thank you, Myeni Wam. This accomplishment is as much yours as it is mine.

Abstract

Injuries remain an underprioritized public health issue despite contributing to approximately 5.3 million deaths annually and leaving millions with temporary or permanent disabilities. Low-to-middle-income countries bear the greatest burden from injuries. However, due to the lack of comprehensive medical certification of deaths occurring outside health facilities, most injury-related deaths may remain unreported. Verbal autopsy (VA) serves as a crucial method for assigning causes of death in such settings, offering valuable insights into the sequence of events leading to death. However, VA narratives consist of highly unstructured text, making analysis a challenge.

This retrospective longitudinal study employed natural language processing (NLP) techniques, including n-gram analysis, Latent Dirichlet Allocation (LDA), and BERTopic, to extract meaningful patterns from VA narratives. Data were drawn from the Agincourt Health and Demographic Surveillance System, covering deaths recorded between January 1, 1993, and December 31, 2023. The text-mining workflow involved five key steps: data acquisition, pre-processing, feature extraction, topic modeling, and knowledge discovery.

Results revealed that road traffic accidents, homicide, and suicide were the predominant causes of injury-related deaths over the 30-year period (1993-2023) while drowning and fall-related fatalities remained consistently low. Road accidents primarily involved vehicle collisions and pedestrian fatalities, with most victims dying at the scene and only a few accessing medical care. Homicide was largely driven by gun and knife violence, often influenced by alcohol, particularly in stabbing-related incidents. Suicide cases were primarily due to hanging, with a smaller proportion involving poison ingestion. Additionally, falls were more prevalent among the elderly, while drowning incidents disproportionately affected younger individuals.

This study provides critical insights into the circumstances surrounding injury-related deaths, the possible risk factors contributing to injuries, and the most prevalent types of injuries by analyzing unstructured textual data from verbal autopsy narratives. The findings underscore the effectiveness of automated text-mining techniques in extracting meaningful patterns and enhancing injury surveillance. These results can be used to strengthen injury prevention efforts through targeted, information-driven interventions aimed at reducing the overall injury burden.

Table of Contents

List of Figures

DEFINITION OF TERMS

1. Verbal Autopsy (VA)

Verbal autopsy is a structured method used to determine causes of death through interviews with caregivers, witnesses, relatives, or friends of the deceased in settings where medical certification of death is limited or unavailable (WHO 2016).

2. External Death

These are deaths that are due to external factors such as accidents, injuries, poisoning, homicide, or suicide, other than natural diseases. They are normally classified under injury-related mortality and often require forensic investigation (GDB 2020).

3. Non-External Deaths

These are deaths due to natural causes such as chronic illnesses (communicable and con communicable diseases), infections, or genetic conditions (GDB 2020).

4. Text Mining

Text mining is extracting useful information and patterns from large volumes of textual data using natural language processing (NLP), machine learning, and statistical analysis to uncover useful relationships and trends in unstructured text (Feldman & Sanger 2007).

5. Topic Modeling

Topic modeling is an unsupervised machine-learning technique used to discover hidden thematic structures within large collections of text. It groups words with similar meanings into topics and analyses the semantic relationships between text data (Blei et al. 2003).

6. Natural Language Processing (NLP)

Natural Language Processing (NLP) is a field of artificial intelligence that allows computers to understand, interpret, and generate human language.

7. Algorithm

An algorithm is a step-by-step procedure designed to perform a specific task or solve a problem. It is commonly used in computing for data processing, problem-solving, and automating tasks in machine learning and artificial intelligence applications (Cormen et al. 2009).

8. Injury

Injury is a physical harm or damage to the body due to external force such as accidents, falls, violence, or burns. Injuries can be classified into intentional (e.g., homicide, suicide) and unintentional (e.g., road traffic accidents, falls) Peden et al. (2002).

9. N- gram

An **n-gram** is a contiguous sequence of *n* items (words, characters, or symbols) extracted from text or speech data. In natural language processing (NLP), n-grams capture contextual information and patterns within text, making them valuable for text analysis (Nasser et al., 2021)

ABBREVIATIONS

1.	BERT	Bidirectional Encoder Representations from Transformers
2.	HDSS	Health Demographic Surveillance System
3.	LDA	Latent Dirichlet Allocation
4.	LMIC	Low-Middle-Income Country
5.	NLP	Natural Language Processing
6.	MIC	Middle-Income Country
7.	RTI	Road Traffic Injuries
8.	VA	Verbal Autopsy
9.	WHO	World Health Organization

CHAPTER ONE: INTRODUCTION

1.1.Chapter Introduction

This introductory chapter lays the foundation for the research report by presenting the background and context of the study. It begins by outlining the global and regional burden of injury-related mortality, highlighting its significant impact on public healthparticularly in low- and middle-income countries. The chapter then introduces verbal autopsy as a vital tool for determining probable causes of death in settings where formal death certification systems are inadequate or absent.

Subsequently, the chapter outlines the problem statement, articulates the rationale for the study, and clearly defines the study's aims and objectives. The research questions guiding the investigation are also presented. Finally, the chapter provides an overview of the structure of the dissertation, offering a roadmap of the content covered in the subsequent chapters

1.2.Background

Injuries are a public health challenge in low-to-middle-income countries, and yet they remain underprioritized (Hardcastle et al.2016). It is estimated that approximately 5.3 million people lose their lives due to injuries annually, while tens of millions suffer either temporary or permanent disabilities (WHO, 2021).

South Africa, a middle-income country (MIC), faces a significant burden of injuries. According to Norman et al. (2007), injury-related deaths are part of the quadruple burden of disease affecting the country. Injuries account for at least 13% of annual deaths in South Africa (Matzopolous et al. 2015). The young, healthy population with low socio-economic status is the most affected by injuries (Hardcastle et al. 2016). Of all injury-related deaths, 40% could have been prevented if adequate and immediate medical intervention had been sought (Matzopolous et al., 2015).

Norman et al. (2007) further highlight that injuries significantly contribute to premature deaths and disabilities. However, many low-to-middle-income countries struggle with unreliable injury statistics, making it difficult to measure their impact accurately (Norman et al. 2007). Verbal autopsy (VA) narratives have been widely used in such settings as an alternative method for ascertaining causes of death, including injuries (Edem et al. 2019). Injuries represent a significant burden in terms of preventable mortality and physical and emotional suffering, making it crucial to prioritize their prevention (Seedat *et al.,* 2009). Understanding the circumstances surrounding injuries can aid in developing effective injury prevention strategies. VA narratives contain rich textual information that is useful for determining causes of death and capturing detailed summaries of events leading to and following an injury (Edem et al. 2019).

1.3.Problem Statement

Injuries are responsible for a significant proportion of deaths worldwide, leaving many survivors with long-term disabilities. Despite this, injuries remain a neglected public health problem (Hardcastle et al. 2016). According to WHO (2021), more than 40% of injury-related fatalities are preventable, emphasizing the need for improved responses to this avoidable mortality burden. Unfortunately, injury statistics remain unreliable globally, particularly in low-to-middle-income countries, which bear the highest injury burden yet have the weakest injury surveillance and reporting systems. Effective injury prevention interventions require a deep understanding of the circumstances surrounding injury deaths.

1.4.Study Rationale

VA narratives contain rich free-text summaries of information regarding events leading to an injury and the post-injury sequence before death (Hardcastle et al. 2016). According to Garrib et al. (2011), injury-related deaths follow a well-defined sequence of events that can be reconstructed using VA data. Understanding these events is crucial for developing targeted injury prevention strategies. However, analysing VA data remains problematic due to the large volume of unstructured narrative text.

While substantial research has been conducted using medical datasets from emergency rooms and postmortem reports to classify injury deaths, limited studies have focused on classifying the circumstances surrounding injury deaths using VA data. This study aims to analyse VA narratives to determine the circumstances associated with injury deaths, thereby contributing to injury prevention policies and interventions. The study findings will enhance knowledge about injury-related death circumstances and support the development of data-driven prevention strategies.

1.5.Research Questions

1. What are the circumstances associated with injury-related deaths in rural Northeast South Africa?
2. How have the circumstances associated with injury-related deaths changed over time?

1.6.Study Aims and Objectives

1.6.1. Study Aim

This study aims to use text-mining techniques to determine the circumstances associated with injury-related deaths and how they have changed over a 30-year period (1993-2023) in the rural Agincourt sub-district of Bushbuckridge Municipality, Mpumalanga Province, South Africa.

1.6.2. Study Objectives

2. To classify injury-related deaths in Agincourt between 1993 and 2023 into different cause-of-injury death categories.

3. To determine the circumstances associated with each cause-of-injury death category.

4. To analyse trends in injury-related deaths over time for each category.

4.1.Dissertation Organization

This dissertation is structured into several chapters, each focusing on different aspects of the study to provide a comprehensive analysis of injury deaths and circumstances surrounding these deaths in the Agincourt area over the past 30 years. The organization is as follows:

1. **Chapter 1: Introduction** – This chapter provides an overview of the research background, problem statement, objectives, research questions, and significance of the study.

2. **Chapter 2: Literature Review** –This chapter will review relevant related work that has been done on text mining of verbal autopsies. It will also discuss theoretical background on text mining uses of VA narratives and injury-related deaths focusing more on rural settings.

3. **Chapter 3: Methodology** – The chapter will present the description of the research methodology employed in this study. It will further describe the analytic procedure or techniques used for the analysis.

4. **Chapter 4: Results** – The chapter presents the findings of the study in detail using figures and tables where necessary.

5. **Chapter 5: Discussion** – This chapter interprets and contextualizes the results by comparing them with existing literature. It also highlights the implications of the findings for public health, social policies, and future research. It further provides the study limitations, conclusion summarizing key findings and recommendations for policy planning and future research.

4.2.Summary

This chapter stated the purpose of the current dissertation which is to determine circumstances surrounding injury deaths in Agincourt over the past 30 years. This chapter has also outlined research questions as well as objectives for the study. Finally, the chapter provided an outline of the chapters in this dissertation.

The following chapter will provide a review of the literature related to text mining, verbal autopsy, and injury-related mortality

CHAPTER TWO: LITERATURE REVIEW

2.1.Chapter Introduction

This chapter reviews relevant literature on injury-related deaths and their possible causes, with a particular focus on rural South Africa. It begins with an overview of injury-related fatalities, emphasizing their occurrence in rural areas. The review then explores different types of injury-related deaths and their underlying causes. Additionally, it examines the use of verbal autopsy (VA) methods to enhance death classification in resource-limited settings, highlighting the challenges associated with this approach. Furthermore, the chapter discusses text mining techniques in health data analysis, specifically focusing on topic modeling and insights from related studies. Finally, it identifies gaps in the existing literature.

2.2.Overview of Injury-Related Deaths

According to the World Health Organisation (WHO) (2024), globally, injury-related deaths result from accident-caused causalities which include road traffic accidents, falls, suicide, homicides, and drowning. WHO reported that injuries result in over 5 million annual deaths which exceeds all other types of mortality worldwide (WHO, 2024). Ahmed et al. (2023) found that road traffic injuries persist in both urban and rural areas as high-incidence injuries because of inadequate infrastructure, poor vehicle conditions, and insufficient road safety understanding among residents. Violence including domestic violence and physical assaults are significant types of injuries that also lead to injury-related deaths. The occurrence of violence demonstrated major differences since young men and other specific population groups face higher rates of assault (Khurana et al. 2025). Phillips et al. (2020) highlighted that self-inflicted injuries such as suicides and self-harm injuries occupy the third significant area for death resulting from injuries.

The combination of mental health problems coupled with substance abuse frequently leads people to self-harm resulting in their death. The other injuries that lead to death as a result include poisoning through drugs or weapons as well as accidental overdoses of drugs (Merrill, 2019). Inadequate health systems in LMICs lead to serious challenges when managing cases of fatalities due to injuries (Mehmood et al. 2019). The injury-related mortality burden in Africa is worsened by a lack of access to emergency services and incomplete road safety programs. This results in premature deaths as victims of injury may delay getting necessary medical attention. The Agincourt area together with other rural areas in South Africa experienced unusually high numbers of injury fatalities because of bad healthcare infrastructure, insufficient resources, and local community traditions (D'Ambruoso et *al.,* 2022).

2.3.Injury-Related Deaths in Rural South Africa

The situation of injuries leading to fatalities becomes serious throughout rural South African regions because social economic and healthcare together with infrastructure deficiencies create greater risks (Prinsloo et al. 2024). The mortality rates from injuries exceed those of urban areas in both rural communities because of distinct regional factors. Prinsloo (2019) found that limited healthcare access stands as a main reason behind the issue because rural regions typically experience medical facility shortages along with shortages of emergency services and healthcare workers. The unavailability of immediate medical help after an injury increases the chances of dying. Rural populations suffer from the combined challenges of social detachment and economic scarcity together with little access to psychological care (Dixon et al. 2024).

2.4.Circumstances Associated with Injury-Related Deaths

The South African rural regions face injury-related deaths that emanate from multiple interconnected socioeconomic and environmental elements. The injury-related deaths differ based on the circumstances (such as accidents, violence and self-inflicted wounds) which triggered injury casualties (Merrill, 2019). Ahmed et al. (2023) highlighted that road traffic accidents rank among the major injury-related death contributors. This is because poor infrastructure becomes dangerous when combined with insufficient street lighting, limited pedestrian areas, and dangerous driving practices. Giummarra et al. (2021) noted that over-speeding or driving under alcohol influence alongside weak road safety regulation enforcement leads to further increases in injury-related death rates. Lack of formal transportation access in rural areas forces the local population to use dangerous transport options (broken carriages and overcrowded buses) which raises their risk of suffering fatal accidents.

Savitsky et al. (2020) found that social and economic problems (gender bias and lack of employment) create conditions that cause both violent deaths and homicides. This is because the social and economic differences result in poverty, high levels of unemployment, and substance abuse. The absence of law enforcement officers leads to an increased incidence of alcohol-related violence in both informal settlements and rural communities (Purcell et al. 2021). Domestic violence is another significant factor that predominantly harms women and their children. The occurrence of homicide crimes, disputes between individuals and communities, and violent robberies in rural areas lead to a high number of homicide-related injury deaths (Purcell et al. 2021). Since, rural areas fail to provide sufficient psychological support services and face high stigma levels, mental health issues lead to high rates of self-inflicted injuries that include suicides. Reduced social interaction, financial trouble, and strained interpersonal are eminent suicide causes (Savitsky et al. 2020). Merrill (2019) identified that injury-related deaths including drowning and firearm injuries take place because of unsafe environments combined with inadequate toxic substance storage and limited firearm control measures. The evaluation of these circumstances through verbal autopsy can provide vital information to create targeted interventions that minimize injury fatalities (Merrill, 2019).

2.5.Trends in Injury-Related Deaths

Trends in injury-related deaths have shifted throughout history because of changes in socio-economic, political as well as environmental changes. Owolabi et al. (2023) found that road traffic injuries led to major injury-related deaths, however, the reasons for these incidents varied over time. During the 1990s up to the early 2000s, faulty roads, and insufficient traffic law enforcement aided as the main causes. In the last two decades, advanced vehicle possession rates and unsafe driving habits comprising speed violations and drunk driving continued to produce elevated death tolls related to road accidents. Despite that, road safety developments including better infrastructure and enhanced traffic law enforcement have succeeded in lowering death rates across certain areas (Fraser et al. 2020).

During the late 1990s and early 2000s, South Africa observed increased mortality rates from violent deaths and homicides because of post-democracy political and socioeconomic disorder. Community safety programs and improved police presence have contributed to decreasing numerous types of violent actions in recent times (Chersich et al. 2019). Deaths from domestic violence and alcohol-inflicted assaults continue to cause injury-related fatalities, particularly in rural locations (Merrill, 2019; Chersich et al. 2019). Kootbodiem et al. (2020) reported that the suicide rate has increased steadily, especially among young male individuals. The main elements forcing people toward suicide comprise economic troubles together with unemployment and social solitude. Limited access to mental health services in rural areas creates a huge obstacle that prevents early medical help from reaching those who need it (Kootbodiem et al. 2020).

2.6.Verbal Autopsy

Verbal autopsy is an important tool for determining the cause of death in areas with limited medical records or health care services (Hupkova et al. 2014). VA involves the use of structured interviews with a family member or next of kin or witness of the deceased where they will provide a narrative explanation about the circumstances or the events that occurred before or after an injury has occurred which at a later stage will be analysed for a cause of death determination (Byass et al. 2010)

The use of VA is an important technique used to discover causes of death amongst populations with limited medical certification or without official death documentation (Fraser et al. 2020). The study by Maqungo et al. (2024) conducted in the rural community of Agincourt in South Africa, emphasized problems with formal death certification because most deaths occur outside healthcare centres.

The growth of VA methods took place over several decades to improve death surveillance in settings with limited means. The VA assessment system now aids as a globally accepted mortality verification technique (Edem et al. 2019). Global health organizations including WHO support standardised techniques for VA that protect data consistency throughout different regions (WHO, 2023). Le Vay et al. (2021) stated that medical professionals conducted physician reviews before the advent of automated methods for assessing the causes of death through manual analysis of written reports. Such procedures demanded significant time and fluctuated in reliability because of personal interpretation. Advanced methods have improved the precision as well as the scale capabilities of VA thus making it a useful tool for extensive death monitoring operations (Fraser et al. 2020).

VA serves as an essential tool for injury death investigations because it includes comprehensive details that official hospital records and death registries fail to record (Le Vay *et al.,* 2021). Maqungo *et al.* (2024) found that injuries leading to mortality require witness or family member involvement for an external evaluation to correctly identify the causes leading to death. VA provided researchers with tools to classify sudden deaths into road traffic accidents, murder, suicide, drowning, and poisoning groups (Maqungo *et al.,* 2024). The injury-related deaths connecting road traffic accidents request information from witnesses and relatives for recognizing insecure driving or unacceptable road conditions as possible contributing elements. VA evidence offers details about homicide weapons in addition to conflicts and preceding threats that reveal likely killing motives preceding violent deaths (Fraser *et al.,* 2020).

When inspecting incidents of suicide, the collected evidence might expose past suicide attempts together with pointers of depressive indications and distress issues including financial losses or partnership splits (Purcell *et al.,* 2021). VA allows for the differentiation of unintended events from intentional ones. The investigators can detect hidden elements that official mortality data fails to show via narrative analysis (Edem *et al.,* 2019). Le Vay *et al.* (2021) presented that the cost-effective nature of VA remains beneficial over formal autopsy in mortality surveillance work, especially in areas with limited resources. Through comprehensive narrative retrieval, the VA system enables investigating officers to obtain profound insight regarding variables that play essential roles in injury death occurrence. The information from the VA serves as a foundation for developing targeted public health strategies that cut down preventable death numbers (Le Vay *et al.,* 2021). VA serves an essential role in mortality surveillance because it functions well even though medical certification is insufficient. Advanced automated classification systems together with text mining technologies demonstrate promise to enhance accuracy which would transform VA into a better method for examining rural injury deaths in South Africa (Fraser *et al.,* 2020).

2.7.Challenges for Verbal Autopsy in injury-related death analysis

The use of VA remains important for injury death cause determination despite several issues that affect its accuracy and reliability (Anteneh and Endris, 2020). Groenewald et al. (2023) argued that the participants tend to experience recall bias because they find it hard to remember exact event information when the death occurred long before the interview. Mapundu *et al.* (2024) highlighted that misclassification is another significant challenge in injury-related death analysis while using VA. The separation of purposeful and accidental poisonings as well as homicide from suicide makes classification difficult. Sociocultural pressure and concerns about legal consequences cause families to convert suicide deaths into accidental causes. The threat of persecution or unwanted consequences forces the criminal offenders to omit to report homicide occurrences. In this way, the ability of the VA to accurately measure injury-related death rates decreases when there are reporting mistakes (Mapundu et al. 2024).

Inconsistent data collection processes create a significant restriction during the analysis. Standardized VA questionnaires aim to enhance reliability through standardized procedures, yet inconsistencies arise due to inconsistent interviewer training as well as linguistic obstacles along with varied respondent comprehension of questions (Anteneh and Endris, 2020). The diverse cultural views of injury deaths present challenges for accurate cause-of-death classification in the rural community of Agincourt thus affecting the reliability of death attributions (Nichols et al. 2022). Anteneh and Endris (2020) found that the delay between death occurrence and the VA interview causes information recollection inaccuracies after prolonged periods. Finding a dependable respondent to provide information becomes challenging for public-place injury deaths involving multiple parties.

2.8.Text Mining

Žižka *et al*., (2020) define text mining as "a knowledge-intensive process in which a user interacts with a collection of documents by using analytics tools to identify and explore interesting patterns". Therefore, text mining is an important method of extracting meaningful patterns, trends, and insights from large collections of textual data. Text mining uses both the statistical and/or machine learning approaches and the natural language processing techniques to find knowledge that is hidden in the large number of text (Žižka, et al. 2020). Text mining encompasses a range of techniques, including natural language processing, machine learning, and statistical analysis, which enable the extraction, analysis, and synthesis of information from unstructured text data (Talib et al. 2016). According to Cassim et al (2021), the logical steps for the text mining algorithm are data acquisition, pre-processing, feature extraction, feature value representation, feature selection, information extraction, classification, and discovered knowledge.

2.9.Use of Text Mining in Health Data Analysis

Recent advances in text mining have enabled researchers to analyze large volumes of unstructured textual data from sources like Verbal Autopsies (Feng et al., 2018). Text mining techniques including natural language processing (NLP) have been applied to health data to categorize causes of death, identify common risk factors, and analyse temporal trends (Shah et al. 2019). In rural South Africa, text mining has shown promise in identifying emerging trends in injury deaths and this can be used to inform targeted public health interventions (O'Neill et al. 2017). Through the automation of the process of narrative classification, text mining offers a more scalable and cost-effective approach to understanding the complex mortality data in resource-constrained settings (Vogt *et al.,* 2020)

2.10. Natural language processing

Natural Language Processing (NLP) is a field of artificial intelligence, that allows machines to extract, and interpret, human language with increasing accuracy and sophistication, it bridges the gap between computer science and human language, empowering computers to process and analyze vast amounts of natural language data, unlocking insights and automating tasks that were once exclusively within human capabilities (Lin & Chang, 2021). Since Computers are not programmed to understand the human language that is embedded within the text narratives of verbal autopsies, NLP plays an important role in ensuring that VA narratives are converted into a language they can understand. According to Manaka *et al.,* (2022), NLP has an important role in text mining of extracting important information from unstructured text.

NLP has many applications including machine translation, sentiment analysis, chatbot development, and information retrieval (Bachate & Sharma , 2020). NLP algorithms are designed to analyze in detail text and speech, identifying patterns and relationships that reveal the underlying meaning of the data (Bachate & Sharma , 2020). The use of artificial intelligence technologies in linguistics has promoted the growth of NLP, allowing for the creation of more sophisticated models that can handle the ever-evolving human language (Baburoglu et al. 2019). As such NLP promises to play an increasingly important role in bridging the communication gap between humans and machines, unlocking new possibilities for automation, insight generation, and human-computer interaction (Baburoglu et al. 2019; Dinesh et al. 2021).

The application of NLP in VA analysis is an important improvement in the field of public health, particularly in regions with limited access to comprehensive medical records or trained physicians. The unstructured nature of VA narratives poses a significant challenge for traditional data analysis techniques, often requiring laborious manual coding and interpretation. NLP therefore provides a solution to this challenge by automating the extraction of key information from VA reports, allowing for more efficient and accurate analysis of data (Torfi et al. 2020). NLP can analyze unstructured data more effectively and efficiently compared to the traditional ways of analyzing data and as a result, its algorithms can easily identify specific symptoms, medical conditions sequence of events, and risk factors associated with death through the understanding of the language used in VA interviews (Maniar et al. 2022).

NLP has a bigger and more pivotal role to play in the health research space. Integration of NLP into VA analysis facilitates the process of information extraction, therefore decreasing the time and resources required for the analysis of VA reports. This role would allow public health officials to spend minimal time trying to understand comprehensive mortality trends. According to Yu et al. (2018), NLP can help improve the accuracy of VA coding therefore minimizing the bias currently present and ensuring that data analysis is standardized.

2.11. N-gram natural language processing

An **n-gram** is a contiguous sequence of *n* items (words, characters, or symbols) extracted from text or speech data. In natural language processing (NLP), n-grams capture contextual information and patterns within text, making them valuable for text analysis (Nasser et al., 2021). They serve as a robust form of text categorization, tolerating textual errors while remaining efficient and reliable. Additionally, they can be applied to both short and long documents without requiring prior content or semantic analysis (Nasser et al., 2021).

N-grams are classified as follows:

- **Unigrams (1-gram):** Single words (e.g., "car")
- **Bigrams (2-gram):** Consecutive word pairs (e.g., "car collided")
- **Trigrams (3-gram):** Three-word sequences (e.g., "car collided with")
- **Higher-order n-grams (4-gram and above):** Longer sequences capturing more complex linguistic patterns

N-grams play a crucial role in health-related text analysis, particularly in clinical and epidemiological research. Wang et al. (2018) utilized n-grams to categorize clinical documents, electronic health records (EHRs), and medical notes into different medical conditions. Similarly, Kavuluru et al. (2015) found that n-grams improve the accuracy of predictive models for detecting early signs of diseases such as cancer, diabetes, and cardiovascular disorders. Moreover, n-grams facilitate the extraction of key medical terms including disease names, procedures, and prescribed medications from unstructured clinical text. This enables automated extraction of meaningful medical concepts from free-text narratives, improving efficiency in health data analysis (Kavuluru et al., 2015).

2.12. Topic Modelling

Topic Modeling is an unsupervised machine-learning technique that is capable of extracting and identifying themes underlying a collection of documents by analyzing the co-occurrence of the words (Blei et al. 2003). Through topic modeling, researchers are able to discover hidden structures of large text. Topic modeling has emerged as a powerful tool to identify latent themes in medical texts. Latent Dirichlet Allocation (LDA) and BERTopic are two usually used topic modeling approaches (Blei et al., 2003).

Topic models are generative models and can be categorized into different types such as algebraic, fuzzy, probabilistic as well as neural models. Topic modeling is used mainly in information retrieval to discover hidden themes within a collection of documents and therefore provides an automatic way of organizing, understanding as well and summarizing large textual information. Through a topic model, researchers can mine large text by taking a group of documents and uncovering the underlying topics without any supervision. The term topic represents the hidden information that links the words in a vocabulary to the words in the collection of documents (Tong and Zhong, 3023). According to Mapundu et al. (2024), a topic can be made up of a group of words and the word weight in a document, and one word can belong to the different number of topics in a document.

2.13.Topic Modelling using LDA

LDA is a generative probabilistic model that allows hidden information in the textual dataset to be explained by unobserved groups using the pattern of word co-occurrence (Mapundu et al. 2024). This is one of the most common probabilistic topic modelling techniques that has been applied in domains such as the classification of the genomic sequence, social media networks, modelling of health data as well as extracting latent topics from clinical reports.

Primarily, the LDA approach wants to maximize the probability of the word belonging to a particular topic and the probability of the word belonging to a particular topic (Mapundu et al. 2024). The model has a 3-level hierarchical structure whereby each document is modelled as several topics, and each topic is modelled as a set of terms

In healthcare, LDA has been applied in various domains, including disease classification, patient risk stratification, and medical text summarization (Blei et al. 2003). While LDA has been effective in many applications, it has certain limitations, such as difficulty handling short texts and lack of interpretability in some cases (Blei et al. 2003)

2.14.Topic Modelling using BERTopic

BERTopic is an advanced topic modeling technique that simplifies the modeling process by utilizing embedding techniques and class-based Term Frequency-Inverse Document Frequency (c-TF-IDF) to create clusters. This approach enhances topic interpretability while ensuring that the most significant words remain in the topic description. Unlike traditional methods such as Latent Dirichlet Allocation (LDA), which rely on a probabilistic framework, BERTopic leverages transformer-based models to generate dense vector representations of words and documents, improving the coherence and relevance of extracted topics (Clarke et al. 2021).

BERTopic is able to extract topics with the most relevant keywords with minimal preprocessing. It incorporates contextual word embeddings, and captures semantic similarities between words, allowing for more accurate topic discovery even in complex and highly unstructured text datasets (Reimers & Gurevych 2019). Additionally, BERTopic adapts to variations in topic granularity, making it particularly useful for analyzing large-scale textual data such as verbal autopsy (VA) narratives, social media data, and clinical records (Mao et al. 2021).

Another advantage of BERTopic over traditional topic modeling approaches is its robustness in handling short texts, which is a well-known limitation of models like LDA. Since short texts often lack sufficient word co-occurrence patterns, BERTopic compensates by leveraging pre-trained language models such as BERT (Devlin et al. 2019) and SBERT (Sentence-BERT) (Reimers & Gurevych, 2019) to generate semantically meaningful topic representations.

Furthermore, BERTopic allows for interactive visualization of topics, providing researchers with greater flexibility in interpreting and refining topic clusters. Its integration with dimensionality reduction techniques such as Uniform Manifold Approximation and Projection (UMAP) further enhances its ability to uncover latent structures within textual data (McInnes et al., 2018).

2.15.Related Work

Recent advances in text mining have allowed researchers to analyze large volumes of unstructured textual data from sources like verbal autopsies (Feng et al. 2018). Text mining techniques including natural language processing have been applied to health data to categorize the cause of death, identify common risk factors associated with the cause of death or disease, and analyse trends (Shah et al. 2019). Text mining offers a more scalable and cost-effective approach to understanding mortality data in settings with limited resources (Vogt et al. 2020).

Some work has been done on the application of text mining in the analysis of verbal autopsy narratives globally, though it remains fairly minimal. It is worth noting that the VA remains a manual method used to collect relevant information about the cause of death from relatives, friends, or caregivers through structured questionnaires. Information collected through VAs is rich in textual information that requires time and skills to be analyzed. Applying text mining techniques to these textual data has proven to be very useful in the few reported studies as it can automatically extract structured data from unstructured text allowing for the discovery of patterns, trends, and relationships that could be very difficult to discover when using manual methods (Hassani et al. 2020).

In a recent study by Mapundu et al. (2024), they were able to successfully apply text mining techniques in VA analysis using n-gram natural language processing, LDA, and BERTopic modeling techniques to extract the most prevalent disease-causing deaths in Agincourt. They were further able to identify common symptoms associated with each disease. Gao & Sazana (2023); Nowacki (2021) both applied text mining in VA narratives to extract causes of death and circumstances associated with each cause of death. They reported that Topic modeling provides a powerful model for identifying prevalent themes and patterns within textual data to uncover any hidden relationships between medical history, symptoms, and possible causes of death.

Cassim et al. (2021) also used text mining methods to extract prostate cancer predictive information (Gleason score). They found that their text mining algorithm could handle different reporting formats and could successfully extract the Gleason Score directly from laboratory information. They further concluded that based on the performance of the text mining algorithm used, the same can be applied to different types of cancers to extract useful clinical information.

While there is limited work done on verbal autopsy text mining, there is enough work that has been conducted on the application of text mining in health care research. One study by Cerrito and Cerrito (2005) used text mining to analyse electronic medical records from the hospital emergency department to classify patients being attended according to emergency need. Another study on text-mining clinical medical records was conducted by Brunekreef et al. (2021) where they applied text-mining techniques to identify patients with Systemic Lupus Erythematosus and they found that their text-mining algorithm can accurately give the diagnoses compared to physician specialists. They further discovered that their text-mining algorithm was able to detect life-threatening symptoms with high specificity and good sensitivity.

2.16.Literature Gap Identified.

Even though there have been some advancements in text mining and natural language processing (NLP), the application of these techniques to VAs for determining injury-related deaths in South Africa remains under researched. Existing studies on VA analysis have primarily focused on cause-of-death classification using machine learning algorithms but have not fully employed topic modeling or deep learning-based text mining methods to uncover underlying circumstances leading to injury-related deaths. Moreover, most studies rely on structured VA data, neglecting the rich unstructured narratives that provide crucial contextual information about deaths and events leading to deaths. The lack of text mining models trained on South African VA data further limits the accuracy and interpretability of automated cause-of-death assignment. Additionally, challenges such as language diversity, limited annotated datasets, and inconsistencies in VA reporting prevent the effective deployment of text-mining approaches in VA narratives. Addressing these gaps can improve mortality surveillance, improve injury prevention strategies, and strengthen public health interventions in South Africa.

2.17.Chapter Summary

The literature reviewed provided insights into the use of VA to determine the cause of death in settings where there is limited access to medical certifications. Furthermore, reviewed studies provided a growing need for an automated approach into analysing VA narratives. Furthermore, background on text mining and its application to analysis of VA narratives was reviewed.

CHAPTER THREE: RESEARCH METHODOLOGY

3.1. Chapter introduction

This chapter will discuss the design of the current research highlighting the study setting, source of data as well as the rationale for the study design employed in this study. This chapter will also provide a detailed explanation of the text mining techniques used for data extraction and the data analysis procedures used to interpret extracted data.

3.2. Study Setting

This study used data from the Agincourt Health and Demographic Surveillance System (HDSS) from the Bushbuckridge rural sub-district within the Ehlanzeni District of Mpumalanga Province South Africa. Agincourt is a surveillance site that provides evidence-based monitoring to improve health priorities and inform policy (Mapundu et al., 2022)

Agincourt HDSS started operation in 1992 across 450 km² of land containing 31 villages and reached a population total of 120,000 people (Mabetha et al. 2023). High poverty conditions along with limited job opportunities have forced residents to use old-age pensions combined with child support grants as their main income sources (Kahn et al, 2012). Even though there are primary and secondary schools in each village, educational accomplishment remained low owing to long periods of academic delay (Kahn et al, 2012). There are a total of nine healthcare facilities in the subdistrict comprising of six clinics, two health centres, and three district hospitals which are 25-60 km apart (Kahn et al, 2012). Since the area is at the border of South Africa and Mozambique, about a third of the population is of Mozambique origin.

3.3.Data Source

The study used secondary data acquired from injury VA reports that ran from 1992 to 2023 in the Agincourt HDSS region. Agincourt HDSS is a surveillance site that seeks to improve health priorities and practice through evidence-based monitoring (Mapundu et al. 2022). VA narratives were routinely collected by trained fieldworkers using a WHO VA data collection tool. Hussain-Alkhateeb et al. (2016) defined the VA narratives data collection method as the one that involved trained fieldworkers interviewing the caretaker present to obtain information on signs, symptoms, and health circumstances related to the death of the deceased. Those interviews worked to reconstruct the death journey patterns because they grouped critical information about how situations and injuries developed during these events.

A data request form (Annexure A) was sent to Wits/MRC Agincourt to request extraction of relevant data fields. The requested data fields include, Individual ID, village ID, date of death, date of birth, type of injury and VA Narratives. Data extracted contained VA narratives collected between 1 Jan 1993 and

31 December 2023. Two datasets were received, containing data on observed individuals with 310333 rows and 15 columns, and the second one with injury deaths having 1732 rows and 6 columns. The observed individual dataset provided the overall individuals population observed between 1 Jan 1993 to 31 Dec 2023. We used this dataset for population demographic analysis and overall mortality trends in Agincourt. The injury deaths dataset was used for the analysis of injury related mortality and circumstances surrounding injuries.

Even though coding of injuries was done according to the VA tool guide by WHO, there were some inconsistencies in the coding of the injuries which we found not to align with the guide. According to Anteneh and Endris, (2020), inconsistency in data collected through Vas is common due to the inconsistency in trainings of data collectors as well as diversity of the languages. Based on the context of each narrative, we further refined and manually grouped types of injuries according to six groups, presented in table 1 below.

Table 1: Types of injuries in Agincourt dataset

Injury type from the dataset	Type of injury grouped.
Traffic (488), accidental (132)	Road accidents
Stabbed (11), stab (9), strangled (7), assault (469) firearm (28)	Homicide
Poison (124),	Poison
self (114), inflict (69), suicide (106)	Suicide
Falls (22)	Falls
Drowns (51)	Drowns
Other (5), struck (3), animal (3), electricity (3), Vermon (2), fire (81)	Other

3.4.Study Design

This study followed a retrospective longitudinal design, utilizing secondary data from VA injury narratives collected over the past 3 decades (1993–2023) in the Agincourt HDSS. A retrospective approach allows for the analysis of pre-existing data without the need for new data collection (Creswell & Cresswell 2017). On the other hand, the longitudinal nature of the study enabled the examination of trends and patterns in incidents and occurring over time (Wang et al. 2017). This study design was appropriate for examining injury-related mortality trends because it enabled researchers to detect substantial changes in death causes alongside corresponding demographic and situational patterns.

3.5.Data Processing and Analysis

3.5.1. Data Acquisition

The current study used VA narratives that were collected in Agincourt HDSS between 1993 and 2023. Data was extracted from Agincourt HDSS in an Excel format.. VA narratives are being widely used across LMIC to represent the context and causes of death, particularly in the areas that lack cause-of-death certification by medical personnel (Edem et al. 2019). Previous studies acknowledged VA data as a reliable source of cause-of-death determination in settings where there is no medical certification of death (Danso e*t al.,* 2013; Mapundu *et al.,* 2022).

3.5.2. Analytic Techniques

For this study, text mining techniques were used on verbal autopsy narratives to categorize death according to injury types, to determine common circumstances surrounding injury-related death, and to explore temporal trends in injury-related mortality. Text mining is a computational technique that makes it possible to extract useful information and patterns from unstructured textual data to identify patterns that are difficult to extract through manual inspection, it integrates NLP with statistical and machine learning methods (Allahyari et al., 2017; Žižka et al. 2019). The text mining algorithm used followed five basic text mining steps, depicted by figure 1 below.

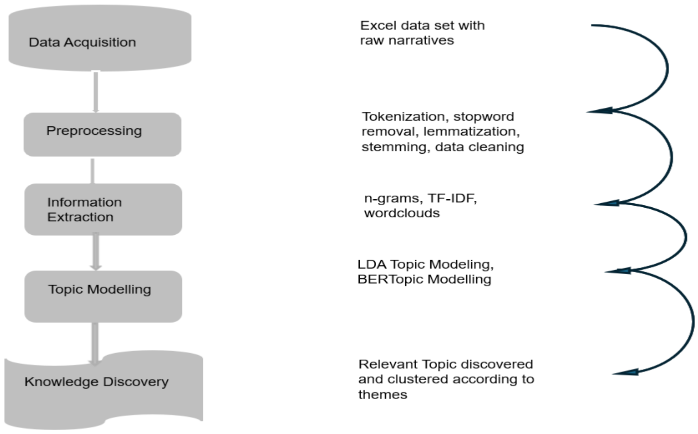

Figure 1: Text mining process.

3.5.3. Data Cleaning and Pre-processing

The narratives were pre-processed to clean and normalize the data before analysis can be conducted. Cassim *e*t al. (2021) defines data preprocessing as the process whereby text narratives are converted into data formats that can be read by machines. We first cleaned the unstructured text by converting all text into lowercase and removing punctuations, spaces, numbers, and special characters.

We removed uninformative, irrelevant words using the NLTK library of English stop words. In **NLP**, **stopwords** are commonly used words (such as "the," "is," "in," "and," etc.) that have no significant meaning in a document. These were the words that we found to be irrelevant for the analysis. Tokenization was also performed to break down long texts into single work tokens. Through the use of Python Spacy package, we performed lemmatization to convert words into their base form (i.e. convert falling → fall; fallen → fall ect.). According to Mapundu et al. (2017) the Python Spacy Library, uses a dictionary of already known forms of different words. Thus, it considers the role that the word has in the sentence and then extract its normal/original form.

3.5.4. Data Set Visualisation

In order to understand the distribution of deaths caused by injuries as well as the trends in all-cause mortality rates for the past 30 years, bar charts to show common terms related to injury-related deaths, word clouds and word frequencies of the data revealed where injuries were most commonly caused, and what the leading causes and reasons were for injuries.

3.5.5. Feature Engineering

Feature engineering to determine the common representative features that we used in the algorithm was performed. According to Zaki & Meira (2019), feature engineering is a process that is used to create a new set of features from existing ones. There are three main steps that we followed, feature extraction, feature selection, and feature value presentation.

3.5.5.1. Feature Extraction

Feature extraction was achieved using Matrix Factorisation, particularly the Non-Negative Matrix Factorisation (NNMF) approach. As a result, this approach was used to discover latent themes within injury narratives by reducing counts of textual data items to lower-dimensional representations (Chen et al. 2015). We extracted features using n-grams natural language processing and they were extracted as tokens of words. We extracted unigrams n=1 such as "*stab*", bigrams n = 2 such as "stab chest" and trigrams of n = 3 such as "taken hospital admitted". The n-grams were extracted based on their frequency in the corpus. When the token had higher frequency, it was deemed as having more context in relation to the document and thus extracted. It was these n-gram models that provided us with a better understanding of the narrative corpus and showed us the common circumstances that surround the injury deaths.

3.5.5.2. Feature Value Presentation

Term Frequency–Inverse Document Frequency (TF-IDF) is a statistical measure used to evaluate the importance of a word in a document relative to a collection (or corpus) of documents. It assigns higher weights to words that are frequent in a specific document but rare across the entire corpus, thereby reducing the impact of commonly occurring but less informative words (Danso et al., 2013). In this study, TF-IDF was used to vectorise the verbal autopsy (VA) narratives, converting them into a high-dimensional sparse matrix where each document was represented as a vector of weighted term values. This process allowed for the extraction of keywords that summarized the content of each narrative. As noted by Mapundu et al. (2021), such keyword extraction is an essential text analysis technique that automatically identifies the most relevant terms within a document, capturing key themes while ignoring redundant or generic words.

3.5.5.3. Feature Selection

To improve computational efficiency and increase the semantic representation of the data, Singular Value Decomposition (SVD) was applied to the TF-IDF matrix. SVD reduces dimensionality by projecting the original high-dimensional feature space into a lower-dimensional latent semantic space. This transformation retains the most informative patterns while eliminating less meaningful features (Leskovec et al., 2020). By selecting the most representative components, SVD served as a critical feature selection step, making the dataset more manageable and improving the performance of subsequent topic modelling.

3.6. Topic Modelling Using LDA and BERTopic

Topic modelling was used to identify and cluster common circumstances surrounding injury-related deaths. Two different approaches were implemented: Latent Dirichlet Allocation (LDA) **and** BERTopic**.**

For LDA, a **term-document matrix** was used as the input, providing a structured representation of the textual data. While LDA typically relies on raw token counts, in our case, the matrix generated from the TF-IDF and subsequently reduced via SVD was used to enhance the quality and coherence of the topics extracted. This preprocessing helped the LDA algorithm identify semantically coherent word groupings that correspond to real-world circumstances described in the VA narratives (Leskovec et al., 2020).

Topic Modelling Using LDA: LDA is a probabilistic generative model in which each word in a document is generated from one of the topics. The goal is to discover the hidden topic distribution of each document and the word distribution for each topic. According to Diouf et al. (2023) the LDA process attempts to maximize both the probability of a document belonging to a certain topic and the probability of a word belonging to a topic. We used the hyperparameter α to optimize the topic distribution for each document.

Document-to-Topic Assignment: the LDA algorithm assigns documents to topics based on probabilistic inference. Each document is assumed to be a mixture of topics, and each topic is characterized by a distribution of words. The result is a meaningful categorization of documents into topics, aiding in the analysis of injury death circumstances.

3.7. Topic Modelling Using BERTopic

In addition to LDA, we implemented BERTopic, a modern topic modelling technique that leverages transformer-based language models for creating rich semantic embeddings of textual data. Although BERTopic does not directly use the TF-IDF and SVD matrices generated in earlier stages, the previously extracted n-gram features and keywords informed the interpretation and labelling of the resulting topics, contributing to a more nuanced understanding of the textual content.

The BERTopic modelling process involved four key steps:

Document Embeddings: The first step involved generating document embeddings using a pre-trained transformer-based language model. These embeddings encode semantic information by representing entire documents as dense vectors in a high-dimensional space. Documents with similar themes are placed closer together in this vector space, facilitating effective clustering. As noted by Buenano-Fernandez et al. (2023), these embeddings enhance the semantic understanding of the data, though they are not directly used in topic generation.

Dimensionality Reduction with UMAP: To make the embedding space more manageable and visually interpretable, we applied Uniform Manifold Approximation and Projection (UMAP). UMAP reduces the high-dimensional vector space into a lower-dimensional space while preserving both local neighbourhood structures and the global topology of the data. This step is crucial for enhancing the quality of clustering in the next phase.

Clustering Using HDBSCAN: Next, we applied Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) to identify dense regions in the UMAP-reduced space. HDBSCAN effectively discovers natural groupings within the data and treats sparse or irrelevant points as noise. According to Buenano-Fernandez et al. (2023), this improves topic coherence by ensuring that only semantically similar documents are grouped together, while unrelated narratives remain unclustered.

Topic Assignment with c-TF-IDF: The final step involved assigning topics to each cluster using class-based Term Frequency–Inverse Document Frequency (c-TF-IDF). Unlike standard TF-IDF, which considers individual documents, c-TF-IDF calculates term importance across the entire cluster, enabling the extraction of representative keywords and phrases for each topic. This method enhances both the interpretability and semantic consistency of the resulting topics.

By combining LDA and BERTopic, we achieved a robust topic modeling approach that enabled a more delicate exploration of the circumstances surrounding injury-related deaths as described in verbal autopsy narratives.

3.8. Text clustering and Categorization

We used Latent Dirichlet Allocation (LDA) and BERTopic modelling for clustering the narratives into common topics. LDA and BERTopic are topic modeling techniques used for clustering text data by identifying hidden structures within large corpora. It applies Bayesian inference to determine word-topic and topic-document distributions, making it effective for discovering underlying themes in textual data. LDA assigns probabilities to each word belonging to different topics, allowing researchers to cluster documents based on dominant themes. Despite its effectiveness, LDA struggles with short texts, requires careful tuning of topic numbers, and lacks interpretability when topics overlap significantly (Blei et al., 2003).

In contrast, BERTopic, is an advanced topic modelling approach that leverages Bidirectional Encoder Representations from Transformers (BERT) along with clustering techniques such as UMAP (Uniform Manifold Approximation and Projection) and HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise). Unlike LDA, which relies on term frequency and topic probability distributions, BERTopic extracts contextual word embeddings using BERT or other transformer-based models, reducing dimensionality through UMAP and forming topic clusters with HDBSCAN. This pipeline enables BERTopic to capture semantic nuances in text and group documents into more coherent and interpretable topic clusters than LDA. It is especially useful for short-text clustering, such as tweets or customer reviews, where traditional LDA often underperforms (Grootendorst, 2020).

Comparative studies highlight that BERTopic outperforms LDA in modern NLP tasks, particularly when dealing with short and noisy text data. Grootendorst (2022) demonstrated that BERTopic could extract more granular and semantically meaningful topics compared to LDA, which often produced overlapping and less interpretable results. However, LDA remains a robust choice for large corpora where computational efficiency and probabilistic topic distribution analysis are priorities. The choice between LDA and BERTopic depends on the dataset, with LDA being better suited for large document collections (e.g., academic papers, news articles) and BERTopic excelling in dynamic, real-world text streams (e.g., social media, customer feedback). Future advancements may combine LDA's probabilistic topic modeling with BERTopic's deep learning capabilities for even more effective text clustering and topic discovery (Grootendorst, 2022; Blei et al. 2003).

3.9.Model Assessment Metrics

To evaluate and compare the performance of the Latent Dirichlet Allocation (LDA) and BERTopic models, several model assessment metrics were employed. For LDA, the coherence score was the primary metric used to determine the optimal number of topics. Topic coherence assesses the semantic similarity between high scoring words within a topic, offering a quantitative measure of topic interpretability (Mimmo et al.,2011). Higher coherence values indicate that the words grouped together are more likely to be understood as a meaningful theme by human evaluators (Röder, Both and Hinneburg, 2015). Coherence was computed over a number of topics, and the model with the highest coherence before saturation or decline was selected as optimal.

In the case of BERTopic, which uses transformer-based embeddings and clustering techniques, topic diversity and inter -topic distance mapping alongside coherence was used. Topic diversity which is the proportion of unique top words across all topics was used to evaluate redundancy and ensure model generalizability. Furthermore, inter-topic distance mapping through UMAP and HDBSCAN algorithms was utilized to visually inspect the distribution and separation of topics in the embedding space (Grootendorst, 2022). These visualizations helped identify overlapping or noisy topic clusters, particularly when the number of topics increased.

Through combining quantitative (coherence and diversity) and qualitative (visualization) evaluation approaches, the most interpretable and meaningful topic structure was selected for further analysis.

3.10. Knowledge Discovery

Knowledge discovery is a crucial component of the text-mining process, enabling the extraction of meaningful and useful information from large volumes of unstructured text (Zhu et al. 2013). According to Mapundu et al. (2024), the knowledge uncovered through this process often includes descriptive patterns that contribute to a deeper understanding of a specific domain. This extracted knowledge can serve as supplementary information for further analysis, aiding in the identification of additional trends and associations within the dataset.

In this study, we applied knowledge discovery techniques to mine VA narratives and uncover key insights into the circumstances surrounding injury-related deaths in Agincourt. By employing text-mining methods like natural language processing (NLP) and topic modeling.

3.11. Ethical Considerations

This research study protected participant information in accordance with ethical principles. Ethical clearance was received from the University of the Witwatersrand Human Research Ethics Committee (WHREC). Furthermore, Permission was obtained from the MRC/WITS Agincourt Research Unit to use VA narratives collected between 1993 and 2023. The received dataset was anonymized with no personal identities like names and ID numbers listed. Only the individual study ID, household ID, and village ID as used to identify the entries in the dataset. All received dataset was password protected.

3.12. Summary

This chapter provided a clear description of the methodology that was employed in this study. Furthermore, a description of study site, study design and data source as well as the data analysis technique was provided.

CHAPTER FOUR (4): RESULTS

4.1. Introduction

The previous chapter focused on the research design and methodology, discussing how the study was conducted. This current chapter provides a detailed analysis and explanation of the findings.

4.2.Demographic analysis

4.2.1. Overall Population Distribution

Between January 1, 1992, and December 31, 2023, there were 310,294 individuals recorded in the Agincourt HDSS database. Females constituted an overall higher proportion (55%) than males (45%). Figure 2 below illustrates the population distribution by sex and age group across different time periods over the past 30 years.

From 1993 to 1997, the Agincourt population exhibited an expanding structure, as seen in the population pyramids with broad bases. These periods were characterized by high birth rates and elevated mortality rates among adults resulting in a narrow peak at the top of the pyramids.

Between 2018 and 2023, a notable demographic shift occurred, marked by a decline in the population of children under the age of five years, indicating a decreased birth rate. This trend resulted in the gradual narrowing of the base of the population pyramid, while the upper segments expanded due to an increase in the population aged five years and above. This transformation indicates a stabilization of the population structure, as the younger cohorts aged into older groups and adult mortality rates declined, contributing to a growing proportion of elderly individuals. Over the past three decades, females have consistently outnumbered males across the population, a trend that has remained steady throughout the observed period.

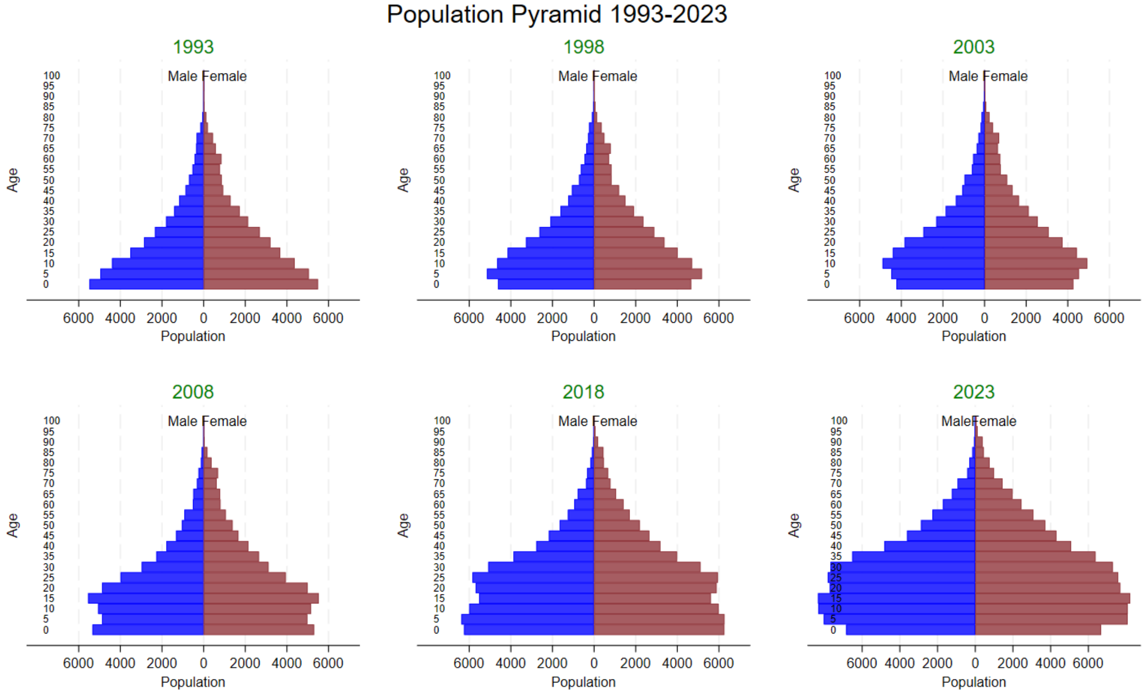

Figure 2: Population distribution according to sex and age in Agincourt over the past 30 years.

4.2.2. Mortality rates

There was a total of 21129 deaths that were recorded in Agincourt over the period 1993 to 2023 (30 years) caused by different causes. There was a sharp increase in the mortality rates for both males and females between 1993 and 2004. In these years male mortality rates remained higher than that of females. In the years 2005- 2008, male mortality rates reached a peak of about 13.3% while females reached their mortality rate peak of 11.9% in 2005. From 2008, mortality rate started to decline for both males and females with a decline in females being more noticeable compared to males. At about the year 2020, there was a spike in the mortality rate for both males and females and this higher mortality rate remained for about 2 years then it started declining again, with a narrowed difference noted between mortality rates for both genders in the year 2023.

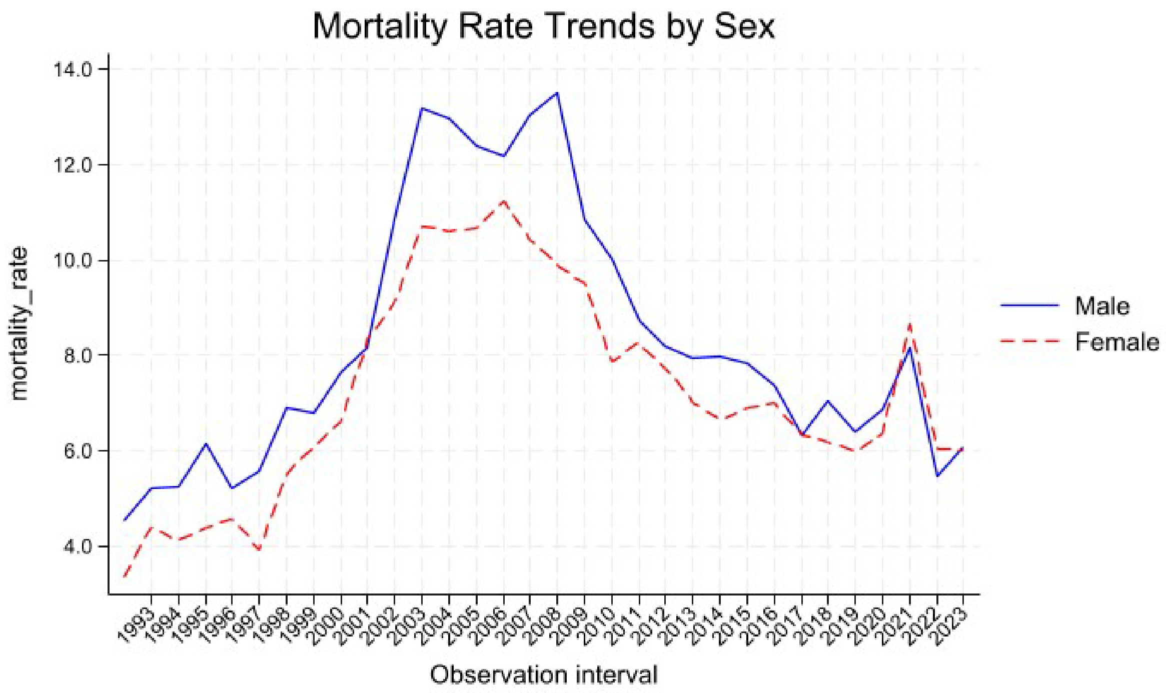

Figure 3: Overall, All-cause Mortality rate according to Sex

4.2.3. Mortality rates according to age and sex.

We further zoomed into mortality rates for different age groups and sex. Mortality rates of all different age groups followed a similar trend, a trend that was noted for the overall mortality according to gender in Figure 3 above. The mortality rate for infants (0-4 years) increased from around 1992 and reached the peak of 80 per 1000 person-years in males and 50 per 1000 person-years. From the year 2020, there was a decline in mortality rates for both males and females. The mortality rate in children aged 5-14 years was characterized by fluctuations throughout the 30-year period. There were some years (1997, 2003) where females had higher mortality rates than males. Between 2018 and 2023, there was a noticeable difference in mortality rates between males and females with males remaining consistently higher than females of the same age group. The trend noted in age groups 15-49 and 50-64 is almost similar with the exception that the difference in mortality rates is more pronounceable in age groups 50-64 years. For both age groups, the mortality rate peaked at 120 per 1000 person-years for males in the year 2005 followed by a decline. Females on the other side had a different picture where in age groups 5-49 mortality peaked at 100 per 1000 person-years in year 2004, followed by a decline, and females aged 50-64 years had mortality rate peak at 70 per 1000 person-years in the year 2007 and remained slightly at that point for about 3 years then a sharp decline in 2011 was noticed. Trends in mortality rate in individuals aged 65 years and above, resemble that of children aged 5-14 years characterized by consistent fluctuations throughout the period, but noticeable was that males remained higher than females.

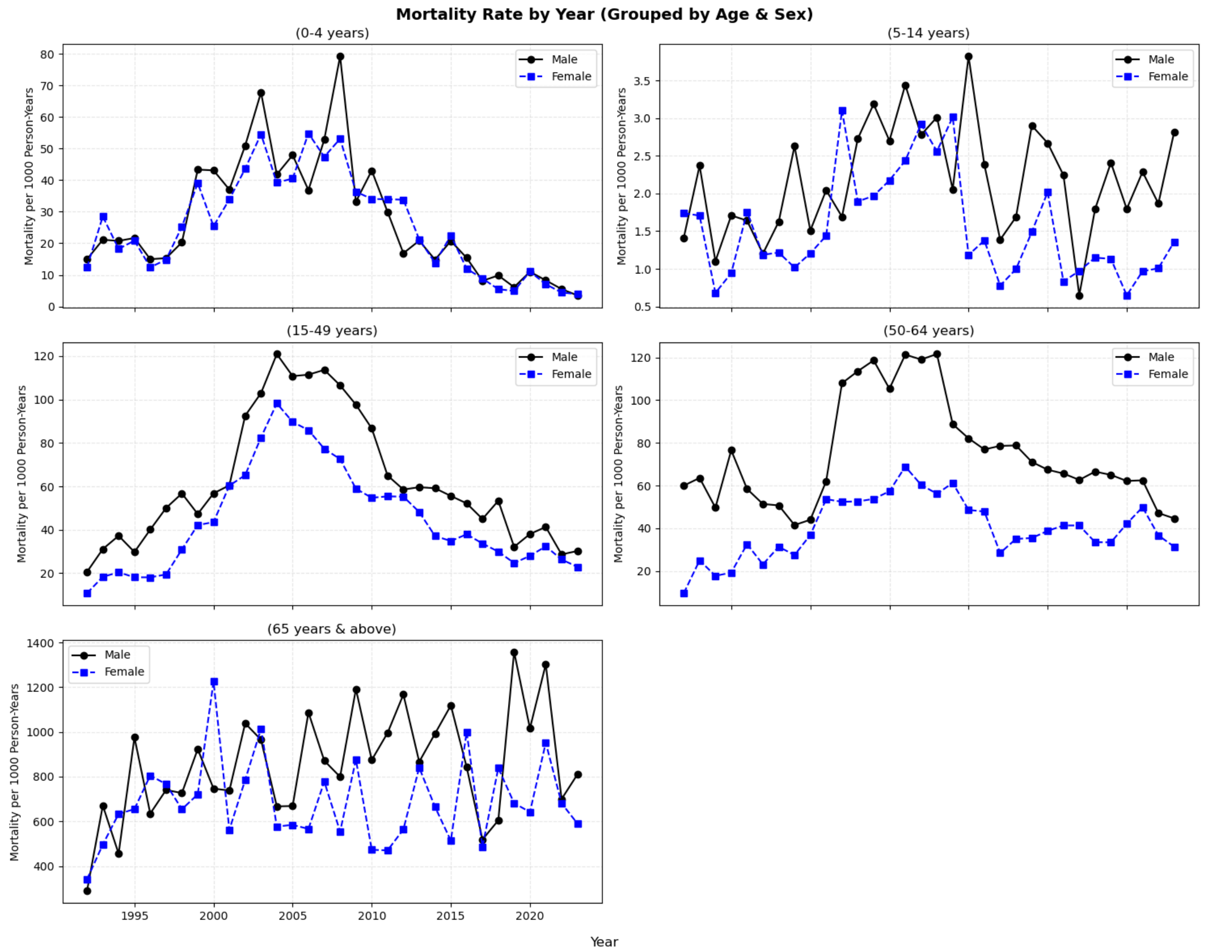

Figure 4: Mortality by sex and age in Agincourt over the past 30 years.

4.2.4. Cause-specific mortality rates.

A total of 1732 deaths that occurred in Agincourt over the past 30 years were due to injuries. This accounted for 8% of the total deaths. There were more males (77.2%) compared to females (22.8%) who died due to injuries. Causes of death were categorized as either non-external causes of death (inclusive of all diseases and any underlying medical conditions) or external causes of death (which are deaths due to homicides, suicides, violence, road accidents as well as any other type of accidents). Figure 5 indicates the trends in cause-specific mortality rates (mortality due to a specified cause of death). It can be noted that non external causes of death had higher mortality rates compared to external causes of death (referred hereto as injuries in this study). Mortality rates due to external causes remained fairly low for the past 30 years with some minimal fluctuations noted. There was a slight increase in injury deaths between 2018 and 2021 then declined again to it all times low rate.

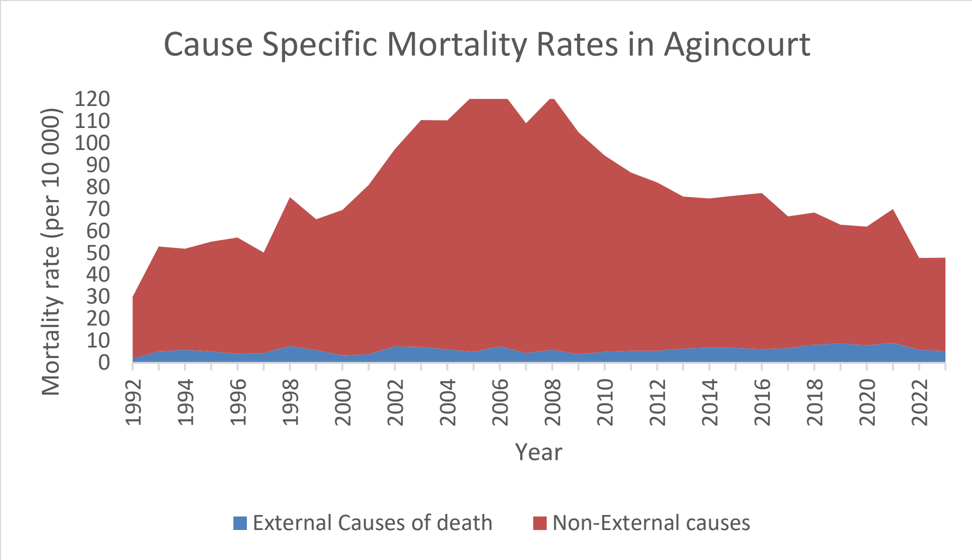

Figure 5: Mortality rates according to age group and gender

Although injury-related deaths represent a smaller proportion of total mortality compared to non-external deaths, the current study focuses specifically on injury deaths due to their significant public health implications and potential for prevention. Figure 6 illustrates the distribution of injury-related deaths in Agincourt, as determined through verbal autopsy (VA) interviews. Road traffic accidents emerged as the leading cause, accounting for 620 deaths, followed by homicide (524) and suicide (413). Falls (22) and drowning (51) contributed the fewest deaths. Less common but collectively impactful injuries such as those caused by animal bites, electrocution, lightning strikes, and other machinery were grouped under the category “other,” which accounted for 100 deaths, surpassing the combined toll of drowning and fall

From Figure 7, Over the past 30 years, road traffic injuries have remained the leading cause of injury-related deaths, persisting as a major public health and safety concern with little fluctuation. Homicide was the second leading cause of injury-related deaths, maintaining a consistent trend over the years. A slight decrease was observed between 2013 and 2017, but rates increased again between 2018 and 2023. Although suicide-related deaths accounted for less than 10% of injury-related deaths for the first decade (1992–2002), there was an increase between 2003–2007 and again from 2018–2023. Poison-related deaths also saw a notable rise beginning in 2013 and continuing through 2023.

While drowning remained a relatively less common cause of injury-related deaths, it was consistently present throughout the 30-year period. Fall-related deaths, however, emerged as a new concern starting in 2013 and continued through 2023, having been previously unreported. Deaths classified under "other" causes remained consistently present but minimal, suggesting that injuries from burns, animal bites, lightning strikes, and similar incidents contributed to a small proportion of overall injury-related fatalities.

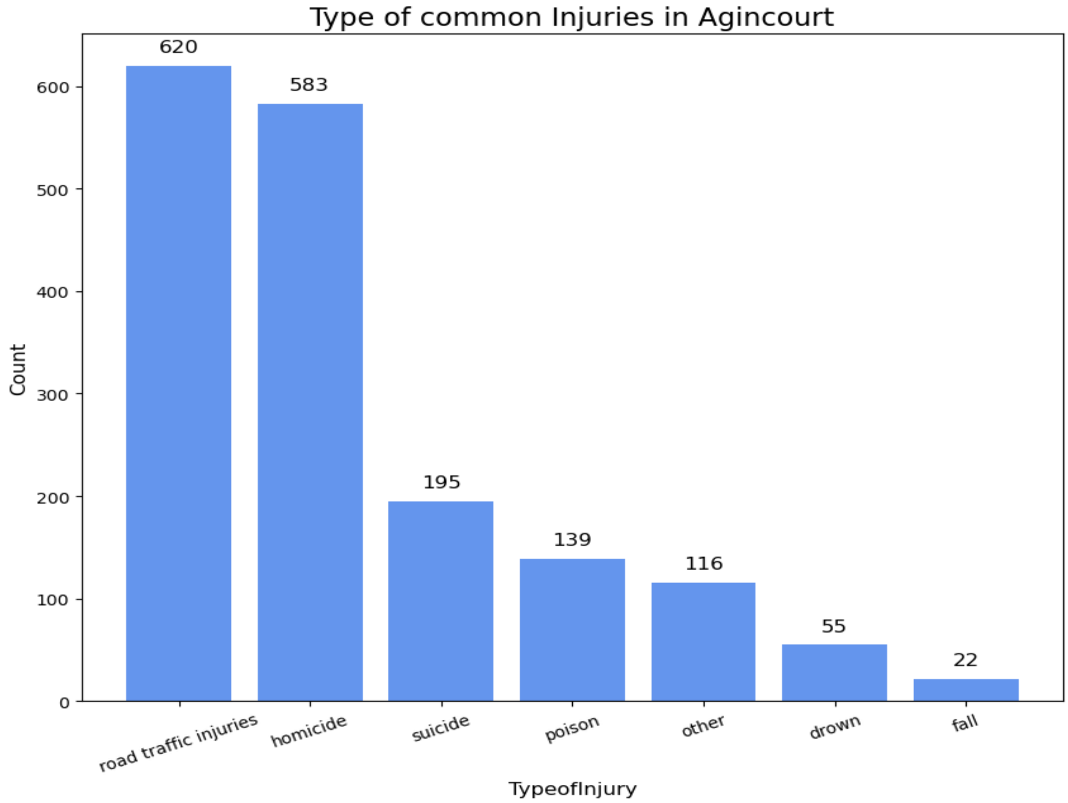

Figure 6: Total number of deaths due to each type of injury in Agincourt for the past 30 years

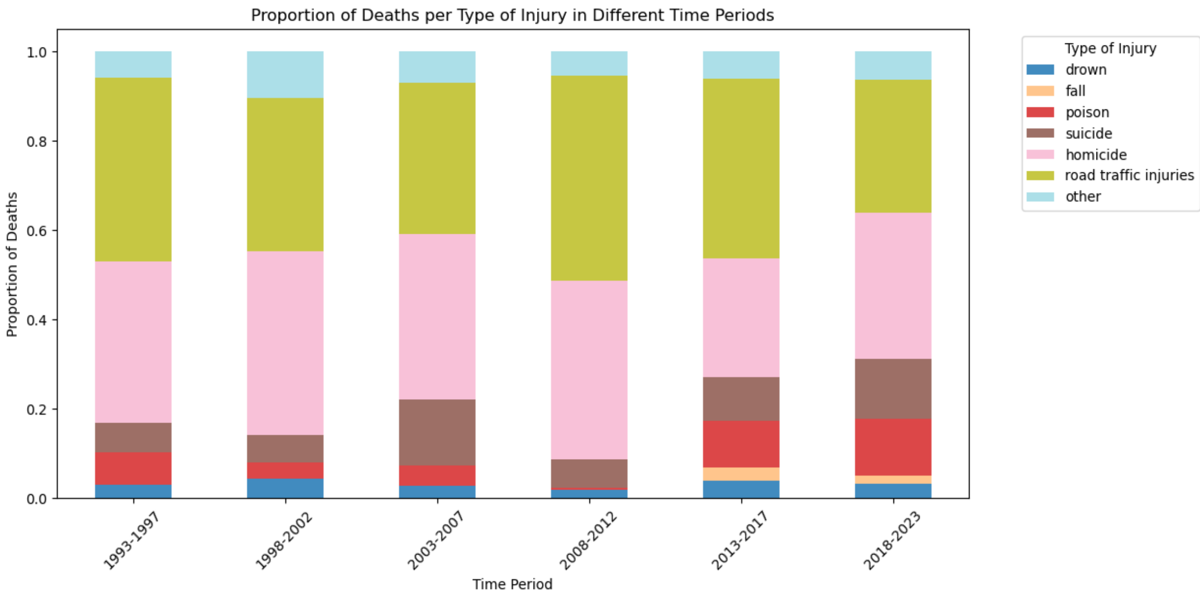

Figure 7: Proportion of deaths due to different types of injuries over time.

Over the past three decades, road traffic injuries have been the leading cause of injury-related deaths among females aged 0-4 years, accounting for more than 50% of cases in most periods. However, in the early years of the Agincourt Health and Demographic Surveillance System (AHDSS), poisoning was a major concern, contributing to over 70% of injury deaths in this group. In later years, drowning incidents became more prominent, particularly between 2013 and 2017, where they peaked at nearly 50%. Among males in the same age group, road traffic injuries remained the dominant cause of death, while drowning-related deaths saw a significant increase between 2008 and 2017, reaching around 25% in some years. Homicide-related deaths were recorded between 2003 and 2017 but remained relatively low, ranging between 10% and 30%. In more recent years (2018-2023), falls-related deaths began to emerge, accounting for approximately 5% of male injury deaths.

For children aged 5-14 years, road traffic injuries continued to be the primary cause of death in both genders, contributing to more than 50% of cases in most periods. However, from 2008 onwards, there was a decline in road accident-related deaths, while drowning-related deaths increased, particularly among males, where they reached approximately 40% in certain years. Poisoning-related deaths were rare in males but accounted for over 25% of injury deaths in females around 2013-2017. During the same period, falls-related deaths emerged, making up about 10% of deaths among females and less than 5% among males.

Among individuals aged 15-49 years, road traffic injuries and homicide consistently accounted for over 60% of injury-related deaths over the past 30 years. While poisoning and homicide remained persistent throughout this period, a slight increase in these causes was observed between 2018 and 2023 in both males and females. Drowning and falls were recorded, but they remained rare, each contributing to less than 1% of total injury deaths.

In the 50-64 age group, homicide-related deaths increased significantly among males, reaching around 30% in some periods. In contrast, for females in this age group, suicide emerged as the leading cause of injury deaths, accounting for over 35% in recent years. Additionally, poisoning-related deaths became more prominent among females, peaking at over 15% between 2013 and 2017. Drowning and falls-related deaths were also noted in this age group, affecting females more frequently than males.

For individuals aged 65 years and above, homicide and suicide were the major contributors to injury-related deaths among females, with suicide showing an increasing trend and reaching approximately 40% in recent years. Meanwhile, road traffic injuries remained the leading cause of injury-related deaths among males in this age group, consistently accounting for over 40% of cases. Fall-related deaths also showed an upward trend, with fatalities due to falls reaching about 3.5% among females between 2013 and 2017 and approximately 2.8% among males between 2018 and 2023.

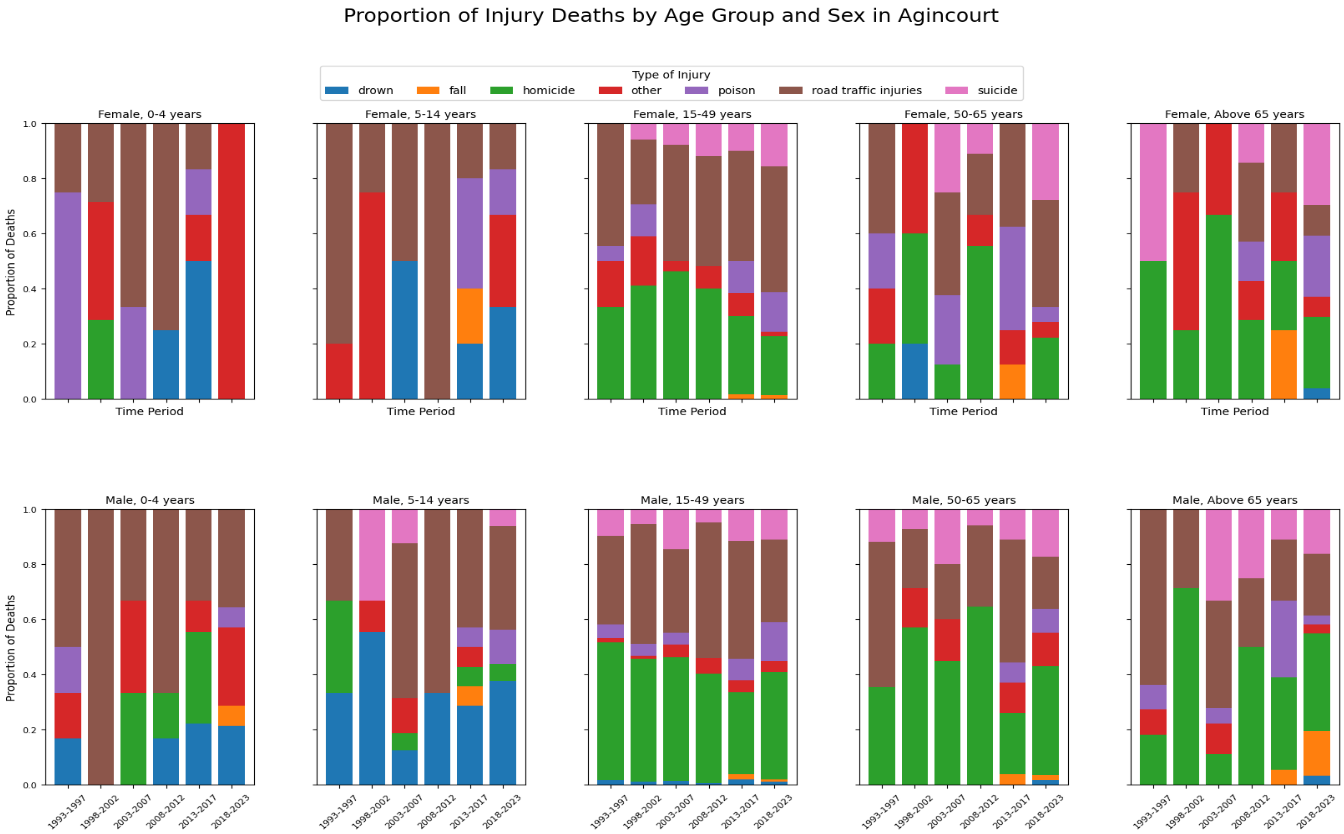

Figure 8: Proportion of deaths due to injuries for different age groups by sex

4.3. Text Mining

4.3.1. Term Frequencies Using Word Cloud

We extracted common terms in the VA narratives and used the word cloud to illustrate their importance in the corpus. Unigrams, bigrams, and trigrams were plotted in 3 different word clouds shown in Figures 9 (a), (b), and (c). Terms like "*spot, hospital, home, hit, hand, shot, car"* were most frequently occurring unigrams whilst terms like "t*aken hospital", "hang spot", "found dead", "found scene", "hit car", "car collided"* were most commonly occurring bi grams and terms like "t*aken hospital admitted", "collided another car", "another car spot", "passenger light vehicle", "car hit another", "spot hang spot", "stabbed knife chest"* were the most commonly" occurring trigrams in the VA corpus

Figure 9(a). Word Cloud for the commonly occurring Unigrams in the VA Narratives

Figure 9(b). Word Cloud for the commonly occurring bigrams in the VA corpus.

Figure 9(c). Commonly Occurring Trigrams in VA corpus.

4.3.2. Term frequencies using TF-IDF

Table 1 below shows the top 20-word frequencies using TF-IDF. Based on the table, top 10 unigrams were found to be “*spot, car, hospital, hit, hang, way, taken, shot, head, scene*”. Additionally top 10 bigrams included terms like “*hang spot, taken hospital, car spot, hospital admitted, shot gun, car collided, car hit, water drip, stabbed knife*” and to 10 trigrams included terms like “ *hit car spot, taken hospital admitted, car collided car, hospital admitted doctor, shot gun spot, light vehicle car, collided car spot, water drip injection, stabbed knife chest, water drip given*.” These were all the terms that had higher frequencies and were the most important terms in our document.

Table 2: Frequently occurring terms using TF-IDF

Unigrams		Bigrams		Trigrams	
Term	TF-IDF Score	Terms	TF-IDF Score	Terms	TF-IDF score
spot	168.05	hang spot	63.69	hit car spot	43.72
car	140.69	taken hospital	51.55	taken hospital admitted	29.04
hospital	97.65	car spot	49.96	car collided car	18.07
Hit	76.36	hit car	47.49	hospital admitted doctor	19.98
hang	69.54	hospital admitted	31.97	shot gun spot	16.60
way	64.41	shot gun	29.64	light vehicle car	14.37
taken	58.86	car collided	26.03	collided car spot	12.63
Shot	55.96	car hit	25.89	water drip injection	11.43
head	55.11	water drip	25.34	stabbed knife chest	11.10
scene	48.14	stabbed knife	23.71	water drip given	11.06
dead	47.24	private car	21.16	admitted doctor said	10.63
given	45.93	chest spot	21.00	way taken hospital	9.87
stabbed	40.62	gunshot spot	20.00	car hit car	9.73
deceased	40.26	treatment given	19.84	collided truck spot	9.51
chest	39.57	light vehicle	19.44	hanged rope spot	9.00
gun	39.49	injured head	18.94	car capsized spot	8.76
collided	37.90	car accident	18.10	passenger light vehicle	8.73
deceased	36.93	collided car	17.92	treatment given changed	8.65
admitted	36.70	died scene	17.41	pedestrian hit car	8.33
chest	36.03	admitted doctor	16.70	driving private car	7.93

4.3.3. Term Frequencies using n- grams

To further extract possible consequences surrounding injury deaths in Agincourt, we extracted the most commonly occurring terms using n-gram, bigrams and trigrams. Figure 10 below represents the unigrams, bigrams, and trigrams. Similarly, with n-grams data extraction, we could discover similar frequent terms as with the word cloud and TF-IDF. For the unigrams, terms like " *hospital, car, spot, taken, way, given, head, hit*, were discovered as frequently occurring alongside the following bigrams ; " t*aken hospital", "hospital admitted", "water drip", "hit car", "treatment given", "car spot"* and trigrams like " *taken hospital admitted", "hit car spot", "water drip injection", "light vehicle car", "car collided spot"*. The extracted terms using TF-IDF, n- grams as well as the representation in wordcloud, it was clear which circumstances surrounds injury deaths in Agincourt. For a clear understanding of these circumstances, n-grams specifically bigrams and trigrams were extracted per injury type.

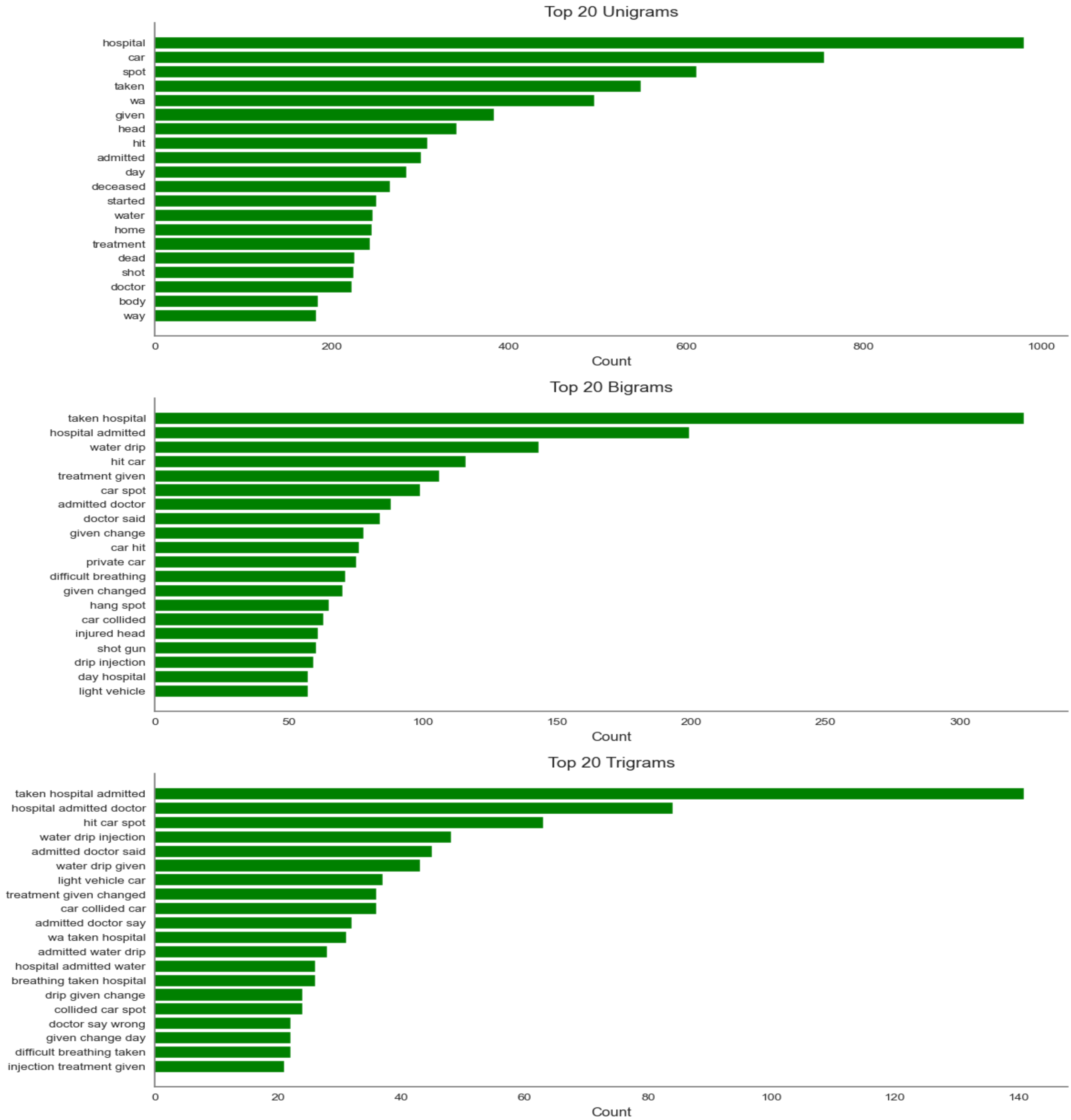

Figure: 10: Unigrams, bigrams, and trigrams extracted from the VA narratives using n-grams processing

4.3.4. Term Frequencies using n-grams per type of injury

Figures 11 and 12 below shows commonly occurring bigrams and trigrams per each type of injury. It was clear that most road accidents involved cars colliding with one another as well as pedestrians being hit by cars as indicated by the presence of bigrams like *"hit car", "car spot", "hospital admitted", "private car", "car collided"* and trigrams like "*hit car spot", "taken hospital admitted", "light vehicle car", "car collided car*". Furthermore, presence of terms like "*hospital admitted, taken hospital*" indicates that more often when these injuries occur people were taken to hospital to seek medical help. The most frequently occurring bigrams in homicide related injuries were "shot *gun, stabbed knife, hospital admitted, chest spot, treatment given*" and the trigrams were "taken *hospital admitted, stabbed knife chest, water drip given, shot gun spot"*. Based on these we could see that the common to circumstances surrounding homicide-related injuries were commonly violent whereby the individual is either stabbed in the chest or shot with a gun. Also, there was evidence that medical help was usually sought after an injury had occurred. Medical intervention was common in suicide-related injuries as indicated by the presence of medical-related terms in both the bigrams and trigrams. Common bigrams from poison-specific verbal autopsies included *"taken hospital," "hang spot," "hospital admitted," "water drip," and "treatment given,*" with corresponding trigrams such as *"taken hospital admitted,"* "*water drip given," "started stomach pain," "home taken hospital,"* and *"stomach pain started."* The presence of phrases like *"started stomach pain*" suggests that many poison-related deaths were likely due to ingesting substances that caused severe stomach pain. Additionally, the frequent mention of medical interventions, such as hospital admission and treatment with a water drip, indicates that many victims sought medical help after poisoning.

In suicide related deaths, hanging was the most common method, as evidenced by bigrams like "*hang spot," "tree hanging," "hanged rope,"* and *"inside house,"* along with trigrams such as *"hang rope spot," "hanging tree rope," "morning hanging roof," "hanging roof using,"* and *"dead hanging tree."* These patterns suggest that most suicide-related deaths occurred either by hanging from a tree or inside a house, often using a rope tied to the roof

With drowns, the common bigram found were words like "*inside water", "dead inside", "drowned water", "drowned river", "drowned dam"* and trigrams included terms like "*drown damn spot*", d*rink lot water", "boy drowned river", "dam died spot", "search boy inside"*. These trigrams and bigrams indicated that most frequently circumstances surrounding drowning incidents included drowning in freshwater settings and the presence of terms with word "*boy*" also indicates that common drowns occur in younger males. Unlike drowns where there was a minimal indication of any form of medical intervention or health-seeking behaviours, with the falls commonly occurring bi grams were words like "*water drip", "taken hospital", "hospital admitted", "taken doctor",* alongside trigrams like "*water drip given", "taken hospital admitted", "admitted doctor hospital" and "head injury scene"*. These terms indicate that upon a fall-related injury victim is taken to the hospital and in most cases, they get admitted. Also, the presence of a "*head injury scene*" depicts that they suffer head injuries and commonly so, may die in the scene. For other types of injuries, there were common terms like *"taken hospital, hospital admitted, water drip, inside the house, badly burnt, difficult breathing*" and trigrams included *"water drip given, admitted doctor said, the way taken the hospital, admitted water drip*". The presence of terms like *"badly burnt, inside house* as well as *difficult breathing"* may indicate burns-related injuries that would have occurred inside the house.

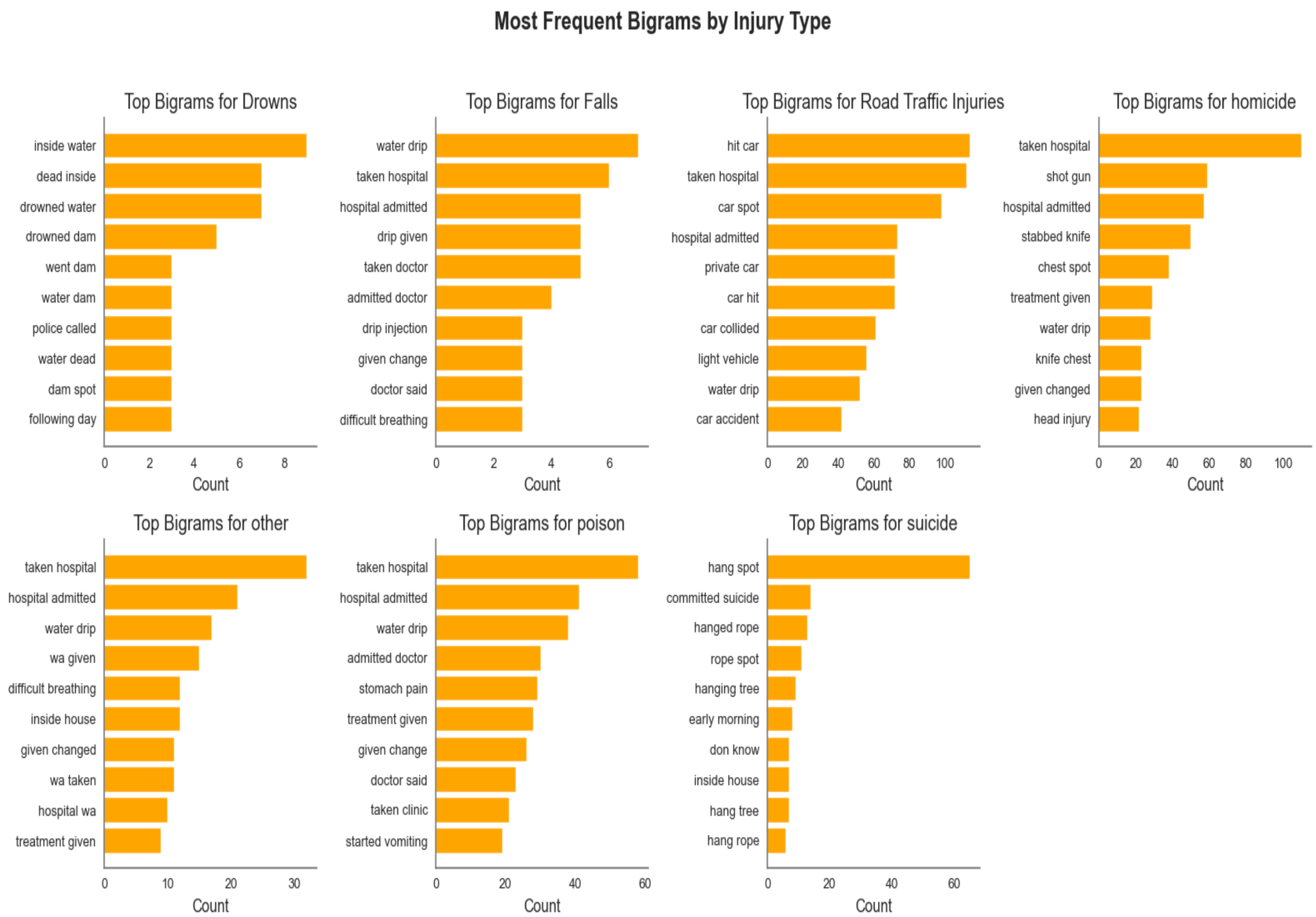

Figure 11: list of common Bi-grams per injury type

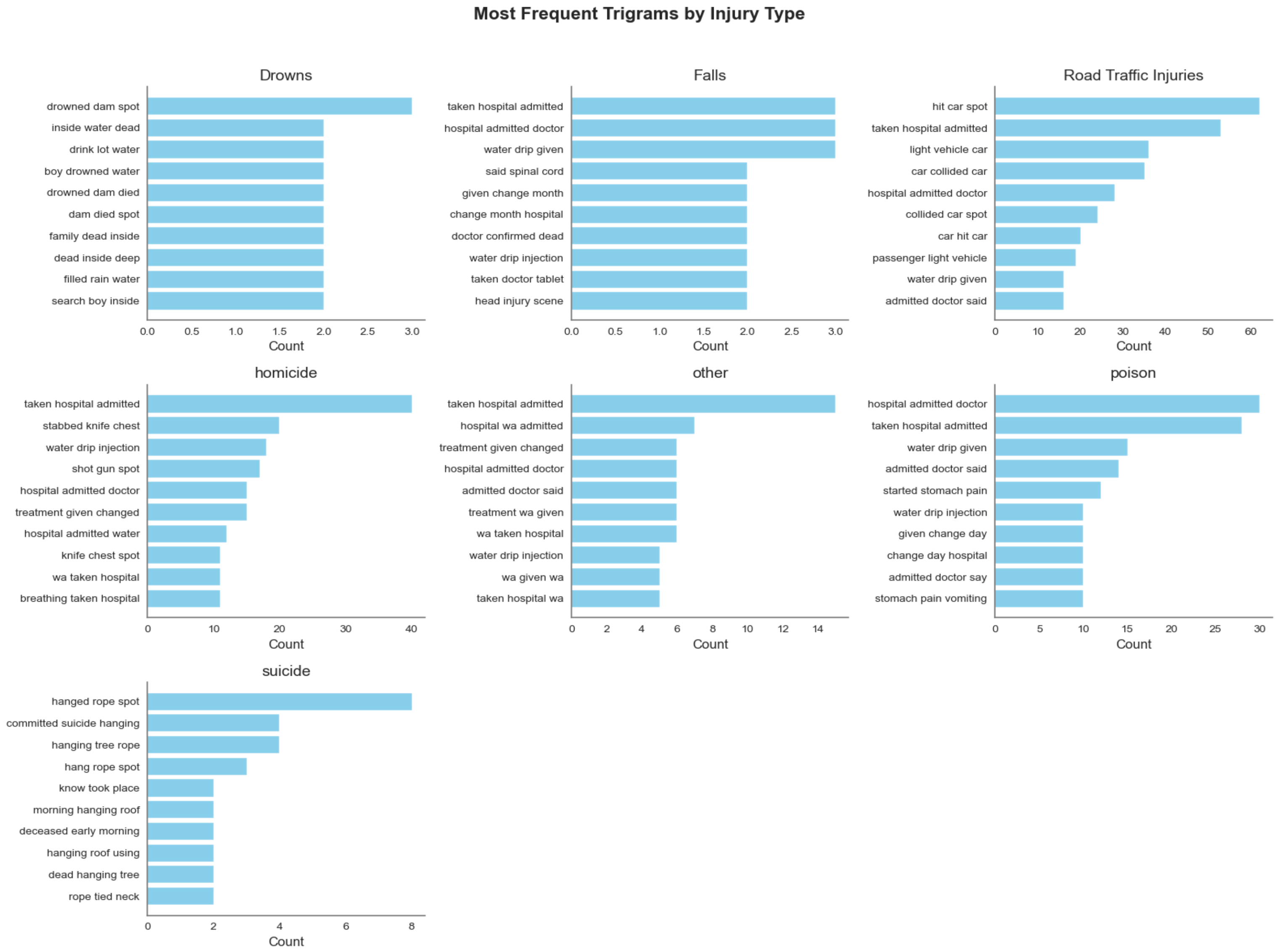

Figure 12: List of trigrams per injury type

4.4. Topic Modelling

4.4.1. Topic Modelling Using LDA

To gain a better understanding of the circumstances surrounding common injuries in Agincourt, we extracted the most prevalent themes from our text corpus using Topic Modelling techniques. We used both Latent Dirichlet Allocation (LDA) and BERTopic to identify and visualize recurring patterns in the data. Figure 13 illustrates the relationships among the top 35 topics identified in the model. The overall coherence score of the model was 0.55, suggesting a moderate level of semantic consistency across topics. Based on the Modelling process, the coherence scores generally improved as more topics were added into the model, but the model performed optimally when limited to 9 topics as indicated by the first peak in the coherence score curve. When more topics beyond 9 were added, the numerical coherence score continued to increase but based on the intertopic distance mapping shown in figure 14, it was clear that there was a reduced distinctiveness of the tops due to a significant overlap in topics. These overlapping "nuisance topics" introduced ambiguity and undermined the clarity of thematic boundaries. Therefore, the 9-topic configuration was selected as the most meaningful and interpretable model for understanding the injury-related narratives in the corpus.

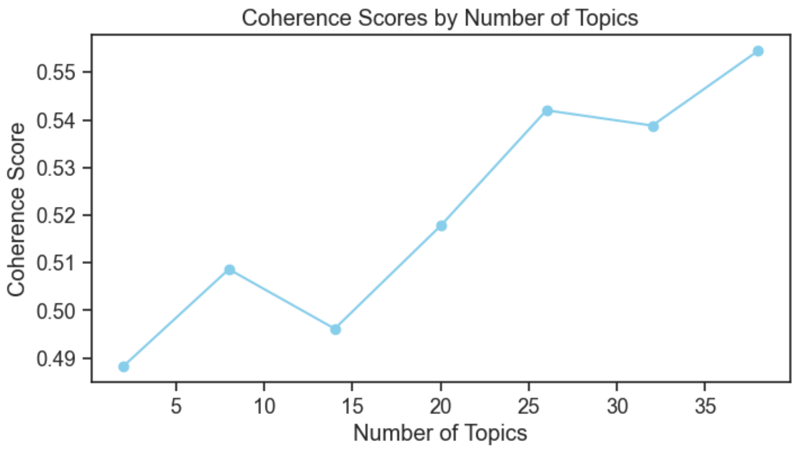

Figure 13: Hyper-parameter optimization of the number of topics in LDA topic modelling

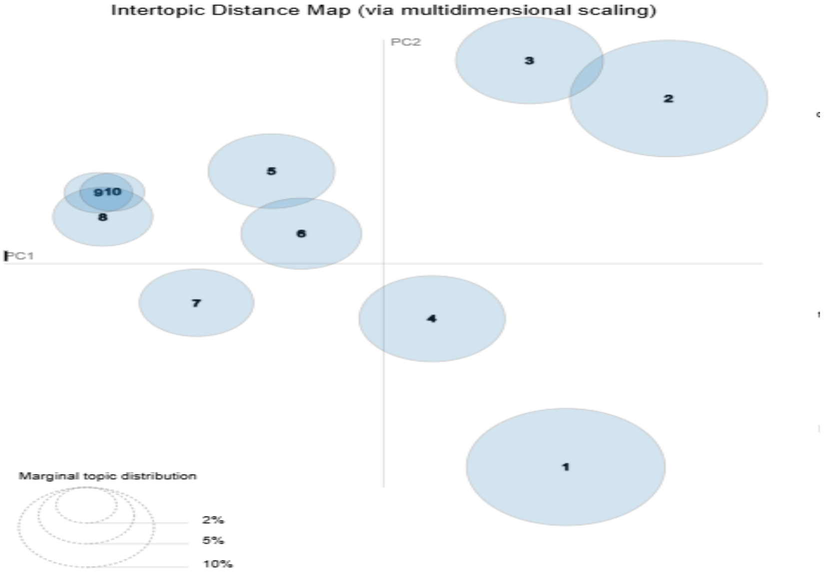

Figure 14: LDA inter-topic distance map using multi-dimensional scaling

Table 3 below shows the results of the common 20 topics that we mined from the VA narratives which is the summary of the common 20 circumstances that surround injury deaths in Agincourt. We found that traffic accidents were the major circumstances surrounding injury deaths as denoted by the highest score of topic number 4. There was also the prevalence of violence-related circumstances as denoted by the prevalence of topics 6,8,11, and 13 that had key circumstances of *"shooting, gunshot, stabbing, and hanging"*. We further noted that drowns and falls were less dominant in Agincourt compared to traffic accidents and violence.

Table 3: Most common circumstances identified using LDA.

Dominant Topic Number	Key Circumstances	Keywords	Topic Contribution
4	Car, spot, hit	car, spot, hit, passenger, collide, driver, truck, light vehicle, accident, injure	0.76
2	hospital, take, give	hospital, take, give, day, start, doctor, admit, treatment, pain, vomit	0.73
6	find, go, dead	find, go, dead, come, shoot, friend, home, fight, take, morning	0.71
13	shot, respondent, know	shot, respondent, know, find, say, dead, gun, day, death, take	0.71
8	spot, gunshot, head	spot, gunshot, head, beat, bleed, injury, burn, scene, stone, fighting	0.70
11	spot, hang, stab	spot, hang, stab, chest, knife, find, dead, gun, rope, head	0.70
10	child, road, bus	child, road, bus, water, drown, train, find, come, go, dead	0.70

1	burn, fire, get	burn, fire, get, hospital, take, day, body, come, house, decease	0.67
5	find, dead, water	find, dead, water, dam, friend, go, come, day, river, drown	0.66
0	hospital, take, day	hospital, take, day, injure, head, get, say, week, family, give	0.65
7	car, scene, way	car, scene, way, drive, taxi, bakkie, hit, truck, accident, come	0.65
3	deceased, police, criminal	deceased, police, criminal, worker, hand, strike, say, stranger, work, license	0.65
9	hospital, take, head	hospital, take, head, road, injure, morning, talk, way, bleed, find	0.64

Based on the similarity matrix represented by a heat map in Figure 15, we were able to note that most of the topics had a distinct relationship as indicated by the diagonal line of 1. There were however some topics that showed some overlap like "*weak route gives*" and "*bandage difficulty, route"*

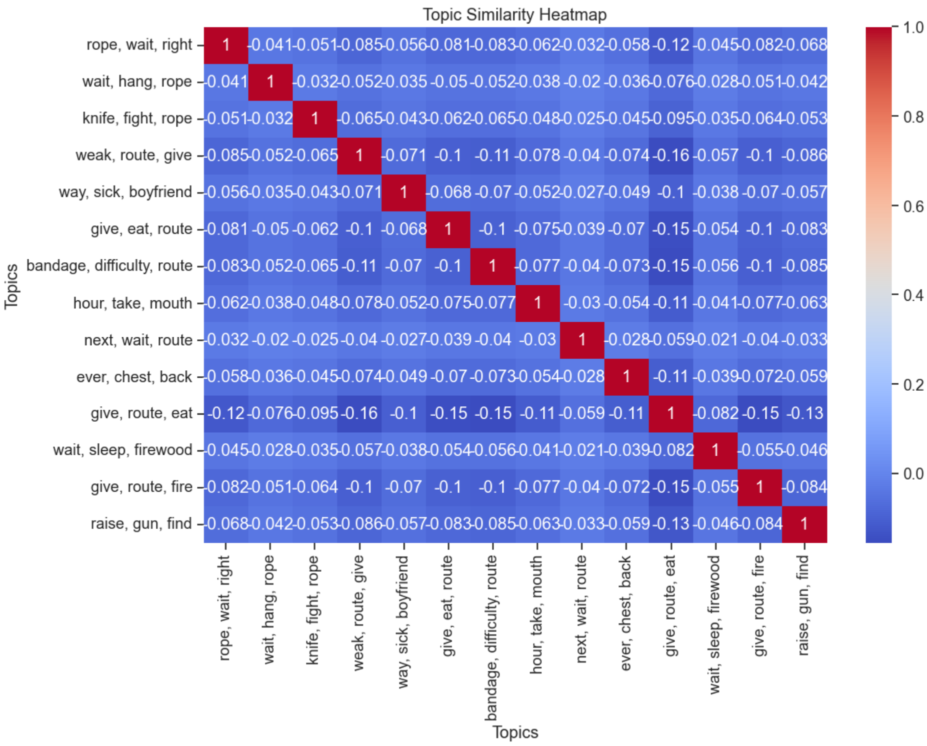

Figure 15: Heatmap of similarity matrix for the topics

4.4.2. Topic Modelling using BERTopic

Clustering VA narratives using BERTopic approach successfully captured different themes, giving insights into the circumstances surrounding injury deaths in Agincourt. Based on the generated topics, we were able to come up with bigger themes for the circumstances surrounding injury deaths. Topic 9, Topic 10, Topic 13, Topic 15 ("car*, truck, accident, collided, passenger, and driving*") are about road traffic accidents.

There were more topics related to homicide, and these were Topic 4 & Topic 6: with words like "*stabbed, knife, chest, fighting*", Topic 12: "*dead, found, knows, stabbed"* suggest violent injuries, related to homicides, and Topic 19: "*beaten, mob, injuries, sticks, internal*" suggest mob violence which is part of homicide-related injuries.

Topic 0 had words like "*started, given, vomiting, taken, and treatment"* which suggest illness-related circumstance, possibly cases of poisoning or medical conditions. Topic 3 & Topic 18: has words like the *hospital, admitted, oxygen, drip"* which indicate topics related to hospitalization and medical treatment. Topic 4 & Topic 6: Words like "*stabbed, knife, chest, fighting*", Topic 12: "*dead, found, knows, and stabbed* suggest violent injuries, related to homicides, and Topic 19: "*beaten, mob, injuries, sticks, internal*" suggest mob violence which is part of homicide-related injuries. Topic 2 & Topic 14: Words like *gun, shot, and shoot* strongly indicate firearm-related incidents, possibly homicides or suicides. Topic 16: Words like *gunshot, scene, and spot* further reinforce shooting incidents. Topic 7 & Topic 8: Words like *hanged, rope, hanging, and university* suggest suicide-related cases, potentially within university or academic settings. Topic 17 has a mixture of circumstances indicating self-inflicted as depicted by the presence of words like *hang, machine, fridge, and operating.* Topic 5 & Topic 10 with the presence of words like "*dam, drowned, river, and water"* indicates drowning incidents, that occurred in natural disasters. Topic 11 had the word *lightning* which suggests a connection to natural causes, possibly electrocution or lightning strikes.

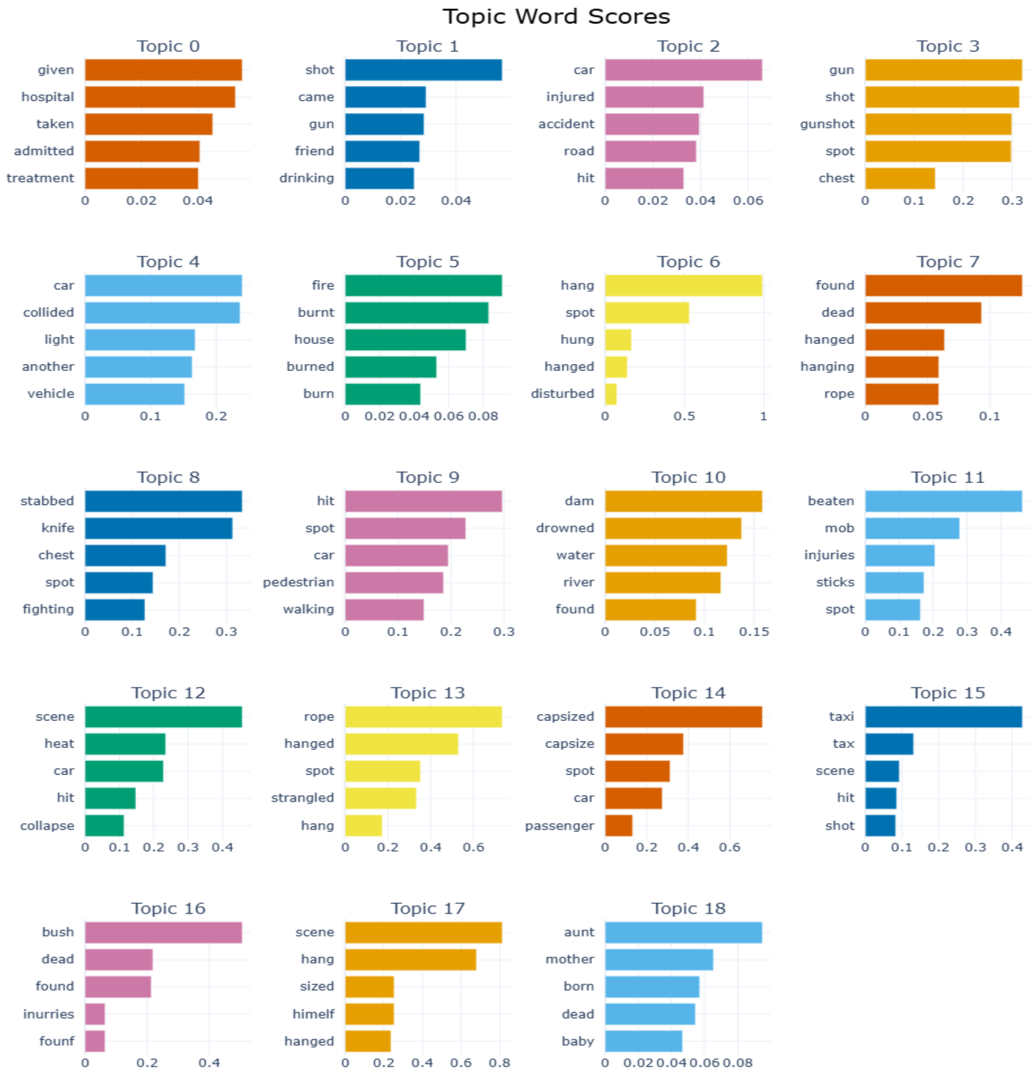

Figure 16: Top 20 topics generated using BERTopic modelling

Figure 17 presents the top five dominant clusters identified from the verbal autopsy (VA) narratives using an optimal clustering algorithm applied to the topic modelling outputs. These clusters represent sets of thematically similar topics that commonly co-occur in descriptions of injury-related deaths, providing insight into the recurring circumstances surrounding such incidents in the Agincourt population. **Cluster 1**, composed of topics (2, 14, 9, 4, 12, and 15), was the most prominent and is strongly associated with road traffic accidents. The narratives grouped within this cluster commonly describe events involving motor vehicles, high-impact collisions, and roadside fatalities. **Cluster 2**, which includes topics (8, 3, and 11), captures scenarios involving interpersonal violence, including the use of firearms and knives. Additionally, the cluster features references to mob justice, supporting its interpretation as representing homicide-related deaths.

Cluster 3 consists of topics (5 and 0), which relate primarily to burn injuries and subsequent medical care. Topic 0 is notably dominant, indicating its strong presence across the dataset. The clustering of burns with medical intervention suggests that burn-related deaths often occurred after hospital admission, rather than at the scene of the incident. **Cluster 4**, comprising topics (7, 10, 16, and 18), reflects narratives involving drowning and other cases where the deceased was found dead under unclear or poorly understood circumstances. This cluster indicates deaths where the mechanism of injury may not have been immediately apparent. **Cluster 5**, including topics (13, 6, and 17), also pertains to road traffic incidents but differs from Cluster 1 in that it specifically involves car-to-car collisions and pedestrian injuries.

Figure 17: Inter-topic distance map for 5 common clusters

The similarity matrix displayed in Figure 18 provides semantic similarities between different topics using heatmaps. The X-axis and Y-axis represent topic labels, which describe different circumstances surrounding injury deaths. Each cell in the matrix represents a similarity score between two topics, with darker blue shades indicating higher similarity and lighter green shades indicating lower similarity. For this report, we will group our topics based on the similarity level.

(A) High-Similarity Topics (Strongly Related Topics)

Topic 2 (shot, gun, shoot) and *Topic 14 (gun, shot, spot)* exhibit strong similarities, suggesting they represent similar firearm-related deaths. *Topic 16 (gunshot, spot, scene)* and *Topic 30 (gunshot, chest, times)* also show a close relationship, further confirming a common cluster for gun-related violence. *Topic 4 (stabbed, knife, chest)* and *Topic 36 (stabbed, knife, chest)* are almost the same, confirming the repeat clustering of similar stabbing incidents. *Topic 6 (drinking, stabbing, fighting)* may also relate to violent confrontations but introduces alcohol-related aggression, which now brings in another circumstance meaning that some of these stabbings occurred due to alcohol intoxication. *Topic 10 (light, vehicle, collided)* and *Topic 26 (collided, another, car)* suggest vehicle crashes. *Topic 22 (road, crossing, hit)* and *Topic 34 (pedestrian, hit, spot)* relate to pedestrian accidents, with impact injuries on roads.

(B) Moderate Similarity Topics (Partially Related Clusters)

Topic 28 (capsized, capsize, spot) and *Topic 40 (collision, buckie, capsized)* suggest water-related accidents, but Topic 40 may involve vehicle-related incidents leading to drowning. Topic 12 *(dead, found, knows)* and Topic 32 *(bush, dead, found)* show moderate similarity, possibly covering external death that is unknown as the deceased was found already dead in different settings. Topic 8 *(hang, spot, cape)* and *Topic 7 (hanged, found, rope)* indicate hanging-related suicides. Topic 17 *(hang, spot, machine, fridge, operating)* may include other mechanical suffocation-related suicides.

(C) Low-Similarity Topics (Distinct Categories)

Topic 0 *(started, given, vomiting)*, which may represent medical-related deaths, has low similarity with violence- or accident-related topics. Topic 18 *(he, admitted, hospital)* likely focuses on hospitalization events before death, which differs from sudden deaths from trauma or suicide. Topic 38 *(bakkie, driver, control)* appears distinct, suggesting a unique type of vehicle-related incident.

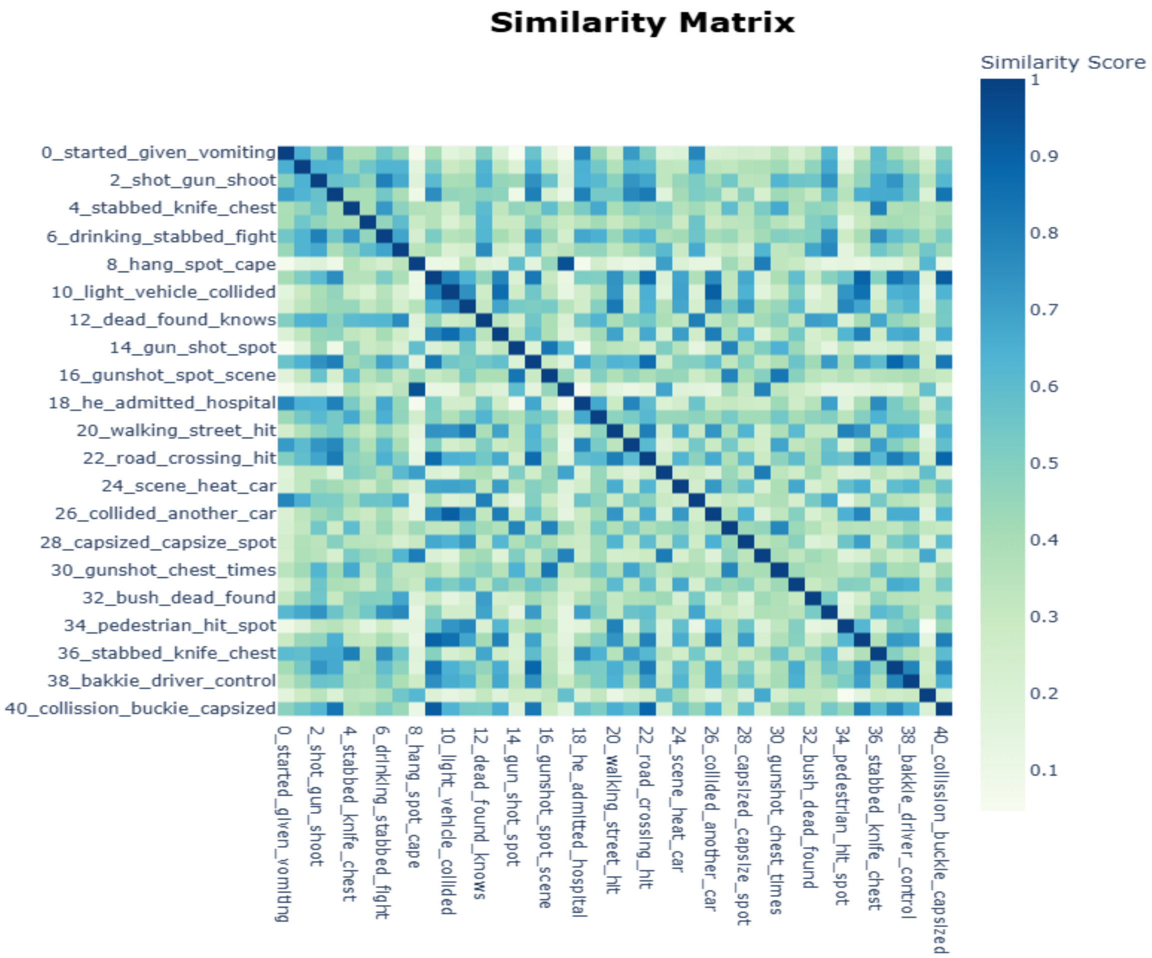

Figure 18: Heat map of similarity matrix

We further clustered the key circumstances using a hierarchical clustering dendrogram in order to visualize common clusters based on their similarities and based on the dendrogram depicted in Figure 19 below. Based on the dendrogram, we were able to cluster common circumstances into four main clusters based on the similarity of the topics. The first cluster (top left) can be clustered under medical-related terms as it can be seen by the presence of words like *"vomiting, hospital, treatment,* and *bleeding*". The second cluster (upper middle) identified relates to violence, as depicted by the presence of terms like *"gunshot, "shot, chest, head, stabbed and scene"*. These terms indicate violence-related circumstance that involves mostly guns and stabbing. The presence of the word *"scene"* further indicates that most individuals die on the scene.

The upper middle cluster relates to suicide and homicide circumstances as words like *"hanged, rope, strangled, and disturbed*" indicate hanging circumstances with the rope being used for hanging while the presence of *"strangled"* clearly indicates that the deceased may have been strangled to death. The final cluster (lower middle) relates to road accidents as words like *"car accident, collided, hit, vehicle,* *and taxi"* were seen. These accidents involved both car collisions as well as pedestrians being hit by cars.

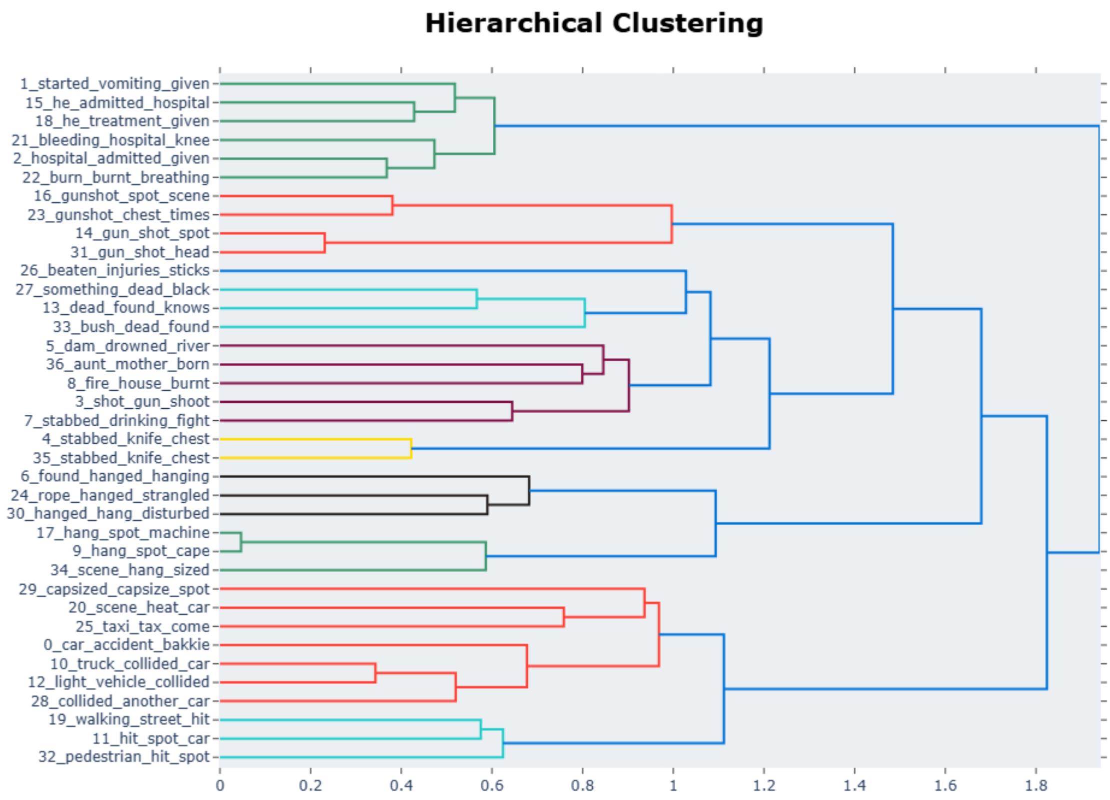

Figure 19: Hierarchical clusters of similar topics.

4.5. Results from Knowledge Discovery process: Identified Topics

Our analysis suggests that most injury deaths in Agincourt were due to road accidents, homicide, and suicide. To a lesser extent drowns and falls were also found. Topic clusters of T10, T26, T22, and T34 indicate the prevalence of road accidents. The presence of keywords like vehicle collision as well as pedestrian hit indicates that these road accidents involved vehicles colliding as well as pedestrians being hit by cars. The presence of T28 and T40 were drowns-related injuries but with an element of vehicle capsizing, therefore indicating a form of car accident that resulted in the passengers drowning.

Clusters of T2, T14, T16, T30, T4, T36, and T6 showed a very strong correlation, and these topics related to homicide or violence-related incidences. From these topics, we were able to further deduce that common violence involved guns and knives. Cluster T4, T36, and T6 was another interesting discovery, where it brought an element of alcohol consumption alongside stabbing, indicating that the majority of the stabbing incidences occurred after alcohol intoxication.

Another interesting discovery was found in clusters T12, and T32 where these were deaths recorded as injury deaths but the type of injury that caused such death is not known as it was indicated that the deceased was found dead. These correlate with previous literature where it is argued that the use of VA narratives for the determination of the cause of death is limited and, in most cases, an expert opinion remains an important aspect in coding death.

4.6. Chapter Summary

In this chapter, the results of the study were presented, and VA narratives were analyzed in detail using text-mining techniques to uncover the circumstances surrounding injury deaths. Mortality analysis was also done to assess the changes in mortality rates over time.

In the next chapter, the presented results will be discussed in detail, implications of the findings, and recommendations will be outlined. Limitations and avenues, for future research will be discussed.

CHAPTER FIVE (5): DISCUSSION

5.1. Chapter Introduction

The results presented in the previous section provide valuable insights into population changes, mortality trends, common causes of injury-related deaths, and circumstances surrounding injury deaths in the form of commonly occurring terms and topic clusters. Through patterns in trends of injury-related mortality over time, and circumstances leading to injury deaths this study highlights key areas that may have significant public health implications.

In the following chapter, these findings are interpreted with existing literature, theoretical frameworks, and socio-economic contexts. The implications for public health policies interventions, and future research directions are explored to provide a comprehensive understanding of injury-related mortality.

5.2.Population Growth and Mortality Trends

Our analysis of the data from the Agincourt HDSS over a 30-year period indicates a steady increase in population size between 1993 and 2003. During the earlier years, population distribution by age revealed a higher proportion of younger individuals compared to the elderly, signalling an expansive population growth pattern. According to Saroha (2018), an expansive population pyramid is characteristic of high fertility rates, high mortality rates, and low life expectancy at birth. However, from 2004 onward, we observed a shift in the population structure, with the population pyramid showing a narrower base. Khan et al. (2007) reported that this change was driven primarily by a decline in fertility rates, likely due to evolving fertility preferences, alongside an increase in child mortality. Additionally, Khan et al. (2007) identified two major contributors to population decline in Agincourt as decreasing fertility rates and out-migration, as individuals moved toward urban areas in search of better economic opportunities. Similarly, Garenne et al. (2013) found that in rural South Africa, declining fertility rates have been strongly associated with increased access to family planning services and shifts in socio-cultural attitudes toward childbearing.

These findings align with global demographic transition trends, as described by Caldwell (2006), who explained the role of socioeconomic development, education, and healthcare improvements in fertility and mortality patterns.

5.3.All-Cause Mortality Trends

The analysis of all-cause mortality rates revealed an overall increasing trend since 1993, with mortality rates peaking at 13.8 deaths per 1,000 person-years in 2008 for males and 11.7 deaths per 1,000 person-years in 2006 for females. These findings align with those of Kabudula et al. (2017), who reported that overall mortality rates in Agincourt began rising in 1998, reaching a peak of 13.3 per 1,000 person-years in 2008 for males and 11.3 per 1,000 person-years in 2006 for females.

This trend can be partially attributed to the HIV/AIDS epidemic, which significantly increased mortality rates in the early 2000s (Herbst et al. 2009; Kabudula *et al.,* 2017). The introduction of antiretroviral therapy (ART) in South Africa around 2009 contributed to a decline and stabilization of mortality rates (Bor et al. 2013; Tanser et al. 2013). While HIV/AIDS-related mortality peaked in the 2000s and later declined, injury-related mortality, particularly from RTIs, suicides, and homicides, has become increasingly significant.

The age and gender-based analysis showed that males were disproportionately affected across all age groups. This observation is consistent with global findings on male vulnerability to mortality, particularly in rural settings (Wang *et al.,* 2021). Studies suggest that men in rural areas tend to engage in high-risk behaviours, including excessive alcohol consumption, unsafe driving, and hazardous occupational activities, leading to higher mortality rates (Peltzer & Phaswana-Mafuya, 2012; GBD 2019 Risk Factor Collaborators, 2020).

5.4.Injury-Related Mortality Trends

Our study specifically examined injury-related mortality in Agincourt over the 30-year period. Although injury-related deaths remained relatively low, a slight increase was observed between 2008 and 2010, contributing 8% of total deaths.

5.4.1. Road Traffic Injuries (RTIs)

Road traffic injuries (RTIs) were the leading cause of injury-related deaths, showing an increasing trend from 2008 to 2018 before declining between 2018 and 2023. This pattern is comparable to findings by Nantulya & Reich (2002), who highlighted that RTIs remain a major cause of mortality in low- and middle-income countries (LMICs), particularly in rural regions where road infrastructure and traffic regulations may be inadequate.

Males aged 15-65 years were most affected by RTIs', with pedestrian accidents and vehicle collisions being the most common causes. Although the proportion of RTI-related deaths in females was lower, it remained relatively consistent over time (Peden et al. 2004).

5.4.2. Homicide and Suicide Trends

While RTI-related deaths declined post-2018, an increase in suicide-related deaths and, to a lesser extent, homicide-related deaths were observed. Homicide-related deaths were frequently associated with violent incidents, such as gunshot wounds and stabbings, mirroring broader crime trends in South Africa (Seedat et al. 2009). These findings are consistent with reports from Norman et al. (2010), who argue that interpersonal violence is a leading cause of death, particularly among young adult males in South Africa. There was high homicide-related mortality in economically active males (15-64 years) and extracted narratives showed that common circumstances surrounding homicide-related injuries were gun-related as well as stabbing. This pattern aligns with Abrahams et al. (2009), who noted that middle-aged men in South Africa often experience socioeconomic stressors, alcohol abuse, and interpersonal conflicts that increase their risk of violent encounters. Similarly, Prinsloo (2019) found that firearms played a significant role in homicides, with urban and peri-urban settings showing the highest prevalence of violent deaths. In contrast, among females, homicide was most prevalent in the 15-49 age group, supporting previous findings by Mathews *et al.* (2008)**,** which highlighted intimate partner violence (IPV) as a major contributor to female homicide rates in the country.

Both males and females have experienced an increase in suicide-related deaths in recent years, but the impact has been disproportionately higher among females, particularly in older age groups. Among individuals aged 15-49 years, suicide has remained a persistent cause of injury-related deaths, though it has generally been overshadowed by road traffic injuries and homicide. However, the rising number of suicide cases in this demographic indicates a growing public health concern that requires further investigation into the underlying socio-economic, psychological, and environmental stressors contributing to self-harm (World Health Organization, 2019).

Moreover, the increase in suicide cases among older age groups, as observed in females aged 50 years and above, aligns with global trends recognizing mental health disorders as a significant and escalating public health crisis (Patel et al., 2018). This trend underscores the urgent need for targeted mental health interventions in vulnerable populations (WHO, 2019). Bantjes et al. (2016) also reported a rising suicide rate in South Africa, particularly among younger individuals, attributing it to factors such as untreated depression, economic distress, and the stigma surrounding mental health care.

Additionally, studies indicate that older women face distinct risk factors that may contribute to suicidal behaviours, including chronic illness, the loss of family members, social isolation, and inadequate access to mental health support (Conejero et al., 2018). Addressing these challenges through comprehensive mental health programs, improved healthcare accessibility, and community-based support systems is crucial in reversing the growing burden of suicide, particularly among aging female populations.

5.4.3. Drowning and Falls

Drowning and fall-related deaths remained consistently low over the past 30 years. However, there were some patterns found in both categories. Drowning accidents were mostly reported among males aged 5-14 years, with most cases occurring in natural water bodies such as dams and rivers. This was an expected trend since Agincourt is a rural area and according to World Health Organization (2014) drowning incidences in rural areas commonly occur in natural water bodies. This aligns with global and South African studies showing that children, particularly boys, are at an increased risk of drowning due to inadequate swimming skills, lack of supervision, and exposure to open water sources (Peden et al., 2019; Saunders et al., 2019). Research further highlights that drowning rates tend to be higher in low-resource settings where access to swimming education and safety measures is limited (WHO, 2021).

There were a minimal number of injury deaths that could be attributed to falls in the dataset. Also, our analysis could not show any strong cluster of topics related to falls. According to Pillay et al. (2020) falls are a significant cause of non-fatal injuries in South Africa, particularly among older adults and individuals working in hazardous occupations. This may further support our findings that falls-related injuries do happen but to a lesser extent result in death. Furthermore, what we found was that there was a strong correlation between falls and medical interventions. This indicates that in most cases individuals who may have suffered a fall usually seek medical help. Fall-related deaths increased among females aged 15-49 years between 2013-2017, while a notable rise was observed among males over 65 years from 2018-2023. This pattern aligns with broader trends in age-related injury risks. Falls among older adults are often linked to age-related frailty, decreased mobility, and underlying health conditions (James et al. 2020; Naidoo et al. 2021). Peltzer & Phaswana-Mafuya, 2017 argues that older individuals, particularly in rural areas, face an increased risk of fall-related injuries due to poor infrastructure, uneven terrain, and inadequate access to assistive devices

5.5.Natural Language Processing (NLP) and Circumstantial Analysis

Our n-gram analysis using natural language processing (NLP) confirmed that road traffic accidents (RTAs) were the most common type of injury, as evidenced by the frequent occurrence of terms such as *"vehicle collision"* and *"pedestrian hit car."* This finding aligns with global and regional studies indicating that RTAs remain a leading cause of injury-related mortality in South Africa. Research suggests that South Africa has one of the highest RTA fatality rates globally, which is almost four times the global average (Matzopoulos *et al.,* 2015; WHO, 2018). The high burden of pedestrian fatalities, in particular, is well-documented in studies highlighting poor road infrastructure, inadequate traffic regulations, and high-risk driving behaviors in low- and middle-income countries (Nantulya & Reich, 2002; Peden et al. 2004).

Additionally, our homicide and suicide analysis revealed that homicide-related injuries were predominantly associated with gunshots and stabbings, often leading to immediate death, as indicated by frequent terms such as *"died on the spot."* This observation is in line with studies showing that South Africa is one of the countries with the highest homicide rates in the world, with firearms and sharp weapons being the most common methods (Seedat et al. 2009; Abrahams et al. 2013). Research also suggests that men, particularly those in lower socioeconomic settings, are disproportionately affected by violent deaths (Norman et al. 2007), further confirming our findings.

We further discovered that suicide-related injuries, drownings, and falls were more likely to involve medical intervention, as reflected in frequent terms such as *"taken to hospital," "admitted,"* and *"treatment given."* This trend aligns with studies showing that while suicide attempts often result in hospitalization, firearm-related suicides tend to have high fatality rates due to limited medical intervention opportunities (WHO, 2019; Patel et al. 2018). In South Africa, suicide rates are estimated to be nearly four times higher than the global average, with hanging and poisoning, being the most commonly reported methods (Mars et al. 2014). This is in line with our findings wherein we found that the most common terms associated with suicide included mostly “hanging”.

Furthermore, Latent Dirichlet Allocation (LDA) modelling identified drowning incidents as being particularly prevalent among younger males and often occurring in natural water bodies such as rivers and dams. This finding aligns with global research indicating that boys, particularly those in low-resource settings, are at greater risk of drowning due to limited supervision and engagement in risky water-related activities (Peden *et al.,* 2019; WHO, 2021). South African studies have similarly highlighted environmental hazards, inadequate access to swimming education, and the lack of safety measures in rural areas as key contributors to child drowning incidents (Saunders *et al.,* 2019).

5.6.Study Limitations

The research encountered several limitations because of the specific features of VA information. The VA narratives are summarised information that is obtained from interviews of individuals who were close to the deceased during the time of death, which can be a relative, caregiver, friend, or next of kin. There is a possibility that the injury death sequence of events might have been inaccurately reported since these accounts relied on memory rather than factual records. Caregivers might overlook or inaccurately describe information during interviews because recall bias is mostly influential for distant incidents. Also, they may omit some important events due to their cultural beliefs or fear of possible criminality surrounding the death. Despite these limitations, the results obtained from this current study still stand.

5.7.Recommendations

The findings of this study provide important insights into the circumstances surrounding or events leading to injury deaths using verbal autopsy (VA) narratives. Our results highlighted the need to improve the accuracy and consistency of cause of death assignment within VA reporting. Additionally, further research is required to uncover the underlying contextual factors contributing to prevalent causes of injury-related mortality, which is essential for targeted prevention strategies.

5.8.Conclusion

This study utilized unsupervised text mining techniques to extract valuable information about the circumstances and events leading to injury-related deaths. By applying this approach, we established a robust framework for analysing and understanding the contextual factors surrounding such fatalities. The findings of this study contribute to public health research by introducing an interpretable and scalable methodology for extracting key information from VA narratives. Additionally, this automated text mining approach can complement traditional human annotation, improving the accuracy and efficiency of identifying causes of injury deaths and tracking evolving mortality patterns

REFERENCES

1. Abrahams, N., Mathews, S., Martin, L. J., Lombard, C., & Jewkes, R. (2013). Intimate partner femicide in South Africa in 1999 and 2009. *PLoS Medicine, 10*(4), e1001412.
2. Ahmed, S.K., Mohammed, M.G., Abdulqadir, S.O., El-Kader, R.G.A., El-Shall, N.A., Chandran, D., Rehman, M.E.U. and Dhama, K. (2023) 'Road traffic accidental injuries and deaths: A neglected global health issue', *Health science reports*, *6*(5), p.e1240.
3. Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B. and Kochut, K. (2017) 'A brief survey of text mining: Classification, clustering, and extraction techniques', *arXiv preprint arXiv:1707.02919*.
4. Anteneh, A. and Endris, B.S. (2020) 'Injury related adult deaths in Addis Ababa, Ethiopia: analysis of data from verbal autopsy', *BMC public health*, *20*, pp.1-8.
5. Baburoglu, B., Tekerek, A. and Tekerek, M., 2019. Development of Deep Learning Based Natural Language Processing Model for Turkish. *arXiv preprint arXiv:1905.05699*.
6. Bachani, A. M., Gielen, A. C., & Mooney, S. J. (2017). "Road safety in LMICs: Challenges and opportunities." Injury Prevention, 23(5), 364-365.
7. Bachate, R.P. and Sharma, A., 2020. Acquaintance with natural language processing for building smart society. In *E3S Web of Conferences* (Vol. 170, p. 02006). EDP Sciences.
8. Bantjes, J., Kagee, A., McGowan, T., & Steel, H. (2016). "Symptoms of posttraumatic stress, depression, and anxiety as predictors of suicidal ideation among South African university students." Journal of American College Health, 64(6), 429-437.
9. Bauni, E., Ndila, C., Mochamah, G., Nyaguara, A., Kihara, J., Williams, T.N. and Newton, C.R., 2011. Validating physician-certified verbal autopsy and probabilistic modeling against hospital cause of death: a case study of childhood mortality in Kenya. *Population Health Metrics*, **9**(1), pp.1-9.
10. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). *Latent Dirichlet Allocation.* Journal of Machine Learning Research, 3, 993-1022.
11. Bor, J., Herbst, A. J., Newell, M. L., & Bärnighausen, T. (2013). Increases in adult life expectancy in rural South Africa: Valuing the scale-up of HIV treatment. *Science, 339*(6122), 961-965.
12. Brunekreef TE, Otten HG, van den Bosch SC, Hoefer IE, van Laar JM, Limper M, Haitjema S. Text Mining of Electronic Health Records Can Accurately Identify and Characterize Patients With Systemic Lupus Erythematosus. ACR Open Rheumatol. 2021 Feb;3(2):65-71. doi: 10.1002/acr2.11211. Epub 2021 Jan 12. PMID: 33434395; PMCID: PMC7882527.

13. Buenano-Fernandez D, Gonzalez M, Gil D, Luja´n-Mora S. Text mining of open-ended questions in self assessment of university teachers: An LDA topic modeling approach. IEEE Access. 2020; 8:35318–35330. https://doi.org/10.1109/ACCESS.2020.2974983
14. Byass, P. (2010). *The Usefulness of Verbal Autopsy Data in Health Systems: A Review.* Public Health Reports, 125(5), 548-555.
15. Caldwell, J.C., 2006. Demographic transition theory. *Springer Science & Business Media.*
16. Cassim, N., Mapundu, M., Olago, V., Celik, T., George, J. A. and Glencross, D. K. (2021) 'Using text mining techniques to extract prostate cancer predictive information (Gleason score) from semi-structured narrative laboratory reports in the Gauteng province, South Africa', *BMC Medical Informatics and Decision Making*, **21(1)**, pp. 1-11.
17. Cerrito, P. B., & Cerrito, J. A. (2005). Data mining electronic medical records using text mining to discover patterns in patient histories. *Proceedings of the 20th Annual SAS Users Group International Conference (SUGI 20)*, 1-12.
18. Chen, L., Vallmuur, K. and Nayak, R. (2015) 'Injury narrative text classification using factorization model', *BMC Medical Informatics and Decision Making*, **15(1)**, pp. 1-12.
19. Chersich, M.F., Swift, C.P., Edelstein, I., Breetzke, G., Scorgie, F., Schutte, F. and Wright, C.Y. (2019) 'Violence in hot weather: Will climate change exacerbate rates of violence in South Africa?', *South African Medical Journal*, *109*(7), pp.447-449.
20. Clarke, R., Vickers, W., & Humphries, S. (2021). Exploring the Utility of BERTopic for Unsupervised Topic Modeling of Clinical Text. *International Conference on Healthcare Informatics*.
21. Conejero, I., Olié, E., Courtet, P., & Calati, R. (2018). Suicide in Older Adults: Current Perspectives. *Clinical Interventions in Aging*, *13*, 691-699
22. Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2009). *Introduction to Algorithms.* MIT Press.
23. Creswell, J. W., & Creswell, J. D. (2017). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches.* Sage Publications.
24. D'Ambruoso, L., Price, J., Cowan, E., Goosen, G., Fottrell, E., Herbst, K., van der Merwe, M., Sigudla, J., Davies, J. and Kahn, K. (2022) 'Refining circumstances of mortality categories (COMCAT)',https://pureoai.bham.ac.uk/ws/portalfiles/portal/169176051/Refining_circumstances_of_mortality_categories_COMCAT_a_verbal_autopsy_model_connecting_circumstances_of_deaths_with_outcomes_for_public_health.pdf [Accessed on 19 Feb. 2025]
25. Danso, S., Atwell, E. and Johnson, O. (2013) 'Linguistic and statistically derived features for cause of death prediction from verbal autopsy text', *Language Processing and Knowledge in the Web*, **25**, pp. 47-60.

26. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171–4186.
27. Dinesh, P.M., Kumar, J.H., Sekar, S.C., Swarnambiga, M. and Ranjani, S., 2021, March. A REVIEW ON NATURAL LANGUAGE PROCESSING Report: Statistics and Facts. In *IOP Conference Series: Materials Science and Engineering* (Vol. 1084, No. 1, p. 012037). IOP Publishing.
28. DIOUF M, THIAM M, ROCHE M. New approach to discover meaningful terms to specify cause of death from narratives verbal autopsy using TF-IDF and the LDA topic model. In: IEEE EUROCON 2023-20th International Conference on Smart Technologies. IEEE; 2023. p. 502–507.
29. Dixon, J., de Vries, S., Fleischer, C., Bhaumik, S., Dymond, C., Jones, A., Ross, M., Finn, J., Geduld, H., Steyn, E. and Lategan, H. (2024) 'Preventable trauma deaths in the Western Cape of South Africa: A consensus-based panel review', *PLOS global public health*, *4*(5), p.e0003122.
30. Edem, I. J., Dare, A. J., Byass, P., D'Ambruoso, L., Kahn, K., Leather, A. J., Tollman, S., Whitaker, J. and Davies, J. (2019) 'External injuries, trauma, and avoidable deaths in Agincourt, South Africa: A retrospective observational and qualitative study', *BMJ Open*, **9(6)**, pp. e027576.
31. Feldman, R., & Sanger, J. (2007). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data.* Cambridge University Press.
32. Feng, Y., Zhang, M., & Li, Y. (2018). Text mining for causal relations from unstructured data in biomedical literature. *Briefings in Bioinformatics*, 19(5), 888-900.
33. Franklin, R. C., Peden, A. E., Hamilton, E. B., et al. (2015). "The burden of drowning: global, regional and national estimates of mortality." Injury Prevention, 21(2), 81-85.
34. Fraser, A., Le Vay, J.N., Byass, P., Tollman, S., Kahn, K., D'Ambruoso, L. and Davies, J.I. (2020) 'Time-critical conditions: assessment of burden and access to care using verbal autopsy in Agincourt, South Africa', *BMJ global health*, *5*(4), p.e002289.
35. Gao, Y., & Sazana, A. (2023). New approach to discover meaningful terms to specify cause of death from narratives verbal autopsy using TF-IDF and the LDA topic model. *Journal of Biomedical Informatics*, 137, 104218.
36. Garenne, M., Tollman, S., Kahn, K., Collins, T. and Ngwenya, S., 2013. Fertility trends and net reproduction in Agincourt, rural South Africa, 1992–2004. *Scandinavian Journal of Public Health*, **41**(8), pp.68-74.

37. Garrib, A., Herbst, A.J., Hosegood, V. and Newell, M.L., 2011. Injury mortality in rural South Africa 2000–2007: rates and associated factors. *Tropical Medicine & International Health*, *16*(4), pp.439-446.
38. GBD 2019 Risk Factor Collaborators. (2020). Global burden of disease study 2019: Risk factor estimates for 204 countries and territories, 1990-2019. *The Lancet, 396*(10258), 1223-1249.
39. Giummarra, M.J., Beck, B. and Gabbe, B.J. (2021) 'Classification of road traffic injury collision characteristics using text mining analysis: Implications for road injury prevention', *PloS one*, *16*(1), p.e0245636.
40. Groenewald P, Kallis N, Holmgren C, Glass T, Anthony A, Maud P, Akhalwaya Y, Afonso E, Niewoudt I, Martin LJ, De Vaal C, Cheyip M, Morof D, Prinsloo M, Matzopoulos R, Bradshaw D. Further evidence of misclassification of the injury deaths in South Africa: When will the barriers to accurate injury death statistics be removed? S Afr Med J. 2023 Sep 4;113(9):30-35. doi: 10.7196/SAMJ. 2023.v113i9.836. PMID: 37882130; PMCID: PMC11017197.
41. Groenewald, P., Thomas, J., Clark, S.J., Morof, D., Joubert, J.D., Kabudula, C., Li, Z. and Bradshaw, D. (2023) 'Agreement between the cause of death assignment by computer-coded verbal autopsy methods and physician coding of verbal autopsy interviews in South Africa', *Global Health Action*, *16*(1), p.2285105.
42. Grootendorst, M. (2020). BERTopic: Leveraging BERT and c-TF-IDF to create easily interpretable topics. *GitHub repository.*
43. Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794.*
44. Hardcastle, T.C., Oosthuizen, G., Clarke, D. and Lutge, E., 2016. Trauma, a preventable burden of disease in South Africa: review of the evidence, with a focus on KwaZulu-Natal. *South African Health Review*, *2016*(1), pp.179-189.
45. Hassani, H., Beneki, C., & Unger, S. (2020). Text Mining in Big Data Analytics. *Algorithms*, 13(1),
46. Herbst, A. J., Cooke, G. S., Bärnighausen, T., KanyKany, E., Tanser, F., & Newell, M. L. (2009). Adult mortality and antiretroviral treatment roll-out in rural KwaZulu-Natal, South Africa. *Bulletin of the World Health Organization, 87*(10), 754-762.
47. James, S. L., Lucchesi, L. R., Bisignano, C., Castle, C. D., Dingels, Z. V., Fox, J. T., & Mokdad, A. H. (2020). The global burden of falls: global, regional and national estimates of morbidity and mortality from the Global Burden of Disease Study 2017. *Injury Prevention, 26*(Supplement 1), i3-i11.
48. Jeblee, S., Gomes, M., Jha, P., Rudzicz, F. and Hirst, G. (2019) 'Automatically determining the cause of death from verbal autopsy narratives', *BMC Medical Informatics and Decision Making*, **19**, pp. 1-13.

49. Kabudula, C. W., Houle, B., Collinson, M. A., Kahn, K., Tollman, S., & Clark, S. J. (2017). Progression of the epidemiological transition in a rural South African setting: findings from population surveillance in Agincourt, 1993–2013. *BMC Public Health, 17*(1), 424.
50. Kahn, K., Collinson, M.A., Gómez-Olivé, F.X., Mokoena, O., Twine, R., Mee, P., Afolabi, S.A., Clark, B.D., Kabudula, C.W., Khosa, A. and Khoza, S. (2012) 'Profile: Agincourt health and socio-demographic surveillance system', *International journal of epidemiology*, 41(4), pp.988-1001.
51. Khan, K., Collinson,M., Tollman S., and Garenne,M., 2007. Mortality trends and migration in South Africa: Evidence from Agincourt. *African Journal of Health Sciences*, 14(2), pp.88-95.
52. Khurana, B., Bayne, H.N., Temple, J., Andover, P. and Loder, R. (2025) 'Comparative analysis of injuries related to self-harm, assault, and intimate partner violence: insights from US Emergency Departments (2005–2021)', *Injury Prevention*.
53. Kootbodien, T., Naicker, N., Wilson, K.S., Ramesar, R. and London, L. (2020) 'Trends in suicide mortality in South Africa, 1997 to 2016', *International journal of environmental research and public health*, *17*(6), p.1850.
54. Le Vay, J.N., Fraser, A., Byass, P., Tollman, S., Kahn, K., D'Ambruoso, L. and Davies, J.I. (2021) 'Mortality trends and access to care for cardiovascular diseases in Agincourt, rural South Africa: a mixed-methods analysis of verbal autopsy data', *BMJ open*, *11*(6), p.e048592.
55. Leskovec J, Rajaraman A, Ullman JD. Mining of massive data sets. Cambridge university press; 2020
56. Lin, J.W. and Chang, R.G., 2021. Optimizing Chinese Story Generation Based on Multi-channel Word Embedding and Frequent Pattern Tree Structure.
57. Mabetha, D., Ojewola, T., Van Der Merwe, M., Mabika, R., Goosen, G., Sigudla, J., Hove, J., Witter, S., D'Ambruoso, L. and On behalf in collab the Verbal Autopsy with Participatory Action Research (VAPAR)/Wits/Mpumalanga Department of Health Learning Platform (2023) 'Realising radical potential: building community power in primary health care through Participatory Action Research', *International Journal for Equity in Health*, 22(1), p.94.
58. Manaka, T., van Zyl, T. and Kar, D., 2022, November. Improving cause-of-death classification from verbal autopsy reports. In *Southern African Conference for Artificial Intelligence Research* (pp. 46-59). Cham: Springer Nature Switzerland.
59. Maniar, K., Haque, S. and Ramzan, K., 2022. Improving clinical efficiency and reducing medical errors through NLP-enabled diagnosis of health conditions from transcription reports. *arXiv preprint arXiv:2206.13516*.
60. Mao, K., Sun, Q., & Zhao, Y. (2021). An Improved Topic Modeling Approach for Short Texts Using Pre-Trained Language Models. *Journal of Artificial Intelligence Research, 70*, 1025–1054.

61. Mapundu, M. T., Kabudula, C. W., Musenge, E., Olago, V. and Celik, T. (2022) 'Performance evaluation of machine learning and Computer Coded Verbal Autopsy (CCVA) algorithms for cause of death determination: A comparative analysis of data from rural South Africa', *Frontiers in Public Health*.
62. Mapundu, M.T., Kabudula, C.W., Musenge, E., Olago, V. and Celik, T. (2024) 'Text mining of verbal autopsy narratives to extract mortality causes and most prevalent diseases using natural language processing', *Plos one*, *19*(9), p.e0308452.
63. Mapundu, M.T., Kabudula, C.W., Musenge, E., Olago, V. and Celik, T., 2021. Robust Application of Supervised Machine Learning Techniques for Cause of Death Determination from Verbal Autopsies.
64. Maqungo, M., Nannan, N., Nojilana, B., Nichols, E., Morof, D., Cheyip, M., Rao, C., Lombard, C., Price, J., Kahn, K. and Martin, L.J. (2024) 'Can verbal autopsies be used on a national scale? Key findings and lessons from South Africa's national cause-of-death validation study', *Global Health Action*, *17*(1), p.2399413.
65. Mars, B., Burrows, S., Hjelmeland, H., & Gunnell, D. (2014). "Suicidal behaviour across the African continent: A review of the literature." BMC Public Health, 14(1), 606.
66. Mathews, S., Abrahams, N., Jewkes, R., Martin, L. J., & Lombard, C. (2008). Intimate femicide-suicide in South Africa: a cross-sectional study. *Bulletin of the World Health Organization, 86*(7), 552-558.
67. Matzopoulos, R., Prinsloo, M., Wyk, V. P., et al. (2020). "The injury mortality survey: a national study of injury mortality levels and causes in South Africa in 2017." South African Medical Journal, 110(10), 920-927.
68. Matzopoulos, R., Prinsloo, M., Wyk, V., & Bradshaw, D. (2015). The injury mortality survey: A national study of fatal injuries in South Africa. *South African Medical Journal, 105*(8), 635-642.
69. McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv preprint arXiv:1802.03426*.
70. Mee, P., Collinson, M.A., Madhavan, S., Kabudula, C., Gomez-Olive, F.X., Kahn, K., Tollman, S.M., Hargreaves, J. and Byass, P. (2014) 'Determinants of the risk of dying of HIV/AIDS in a rural South African community over the period of the decentralised roll-out of antiretroviral therapy: a longitudinal study', *Global health action*, 7(1), p.24826.
71. Mehmood, A., Hung, Y.W., He, H., Ali, S. and Bachani, A.M. (2019) 'Performance of injury severity measures in trauma research: a literature review and validation analysis of studies from low-income and middle-income countries', *BMJ open*, *9*(1), p.e023161.
72. Merrill, R.M. (2019) 'Injury-Related Deaths according to Environmental, Demographic, and Lifestyle Factors', *Journal of environmental and public health*, *2019*(1), p.6942787.

73. Naidoo, D., Van Aardt, C. J., & Joubert, G. (2021). Falls in older adults: The burden and distribution of risk factors in a South African hospital. *African Health Sciences, 21*(1), 446-454.
74. Nantulya, V. M., & Reich, M. R. (2002). "The neglected epidemic: Road traffic injuries in developing countries." BMJ, 324(7346), 1139-1141.
75. Norman, R., Matzopoulos, R., Groenewald, P., & Bradshaw, D. (2007). The high burden of injuries in South Africa. *Bulletin of the World Health Organization, 85*(9), 695-702.
76. Nowacki, T. (2021). Overview of Statistical and Machine Learning Techniques for Determining Causes of Death from Verbal Autopsies: A Systematic Literature Review. *Journal of Health Informatics in Developing Countries*, 15(2).
77. Owolabi, E.O., Ferreira, K., Nyamathe, S., Ignatowicz, A., Odland, M.L., Abdul-Latif, A.M., Byiringiro, J.C., Davies, J.I. and Chu, K.M. (2023) 'Social determinants of seeking and reaching injury care in South Africa: a community-based qualitative study', *Annals of global health*, *89*(1).
78. Patel, B., & Ray, B. (2018). Three leading suicide methods in the United States, 2017–2019.
79. Patel, V., Saxena, S., Lund, C., Thornicroft, G., Baingana, F., Bolton, P., ... & Unützer, J. (2018). The Lancet Commission on Global Mental Health and Sustainable Development. *The Lancet*, *392*(10157), 1553-1598.
80. Peden, A. E., Franklin, R. C., & Mahony, A. (2019). The burden of drowning: Global, regional, and national estimates of mortality. *Injury Prevention, 25*(6), 507-514.
81. Peden, M., McGee, K., & Krug, E. (2002). *Injury: A Leading Cause of the Global Burden of Disease.* World Health Organization.
82. Peden, M., Scurfield, R., Sleet, D., Mohan, D., Hyder, A. A., Jarawan, E., & Mathers, C. (2004). World report on road traffic injury prevention. *World Health Organization*.
83. Peltzer, K., & Phaswana-Mafuya, N. (2012). Problem drinking and associated factors in older adults in South Africa. *African Journal of Psychiatry, 15*(5), 352-359.
84. Phillips, D., Lidón-Moyano, C., Cerdá, M., Gruenewald, P. and Goldman-Mellor, S. (2020) 'Association between unintentional injuries and self-harm among adolescent emergency department patients', *General hospital psychiatry*, *64*, pp.87-92.
85. Pillay, J., Riva, J.J., Tessier, L.A., Colquhoun, H., Lang, E., Moore, A.E., Thombs, B.D., Wilson, B.J., Tzenov, A., Donnelly, C. and Émond, M., 2021. Fall prevention interventions for older community-dwelling adults: systematic reviews on benefits, harms, and patient values and preferences. *Systematic reviews*, *10*, pp.1-18.
86. Prinsloo, M, Bradshaw, D, Joubert, J, Matzopoulos, R, & Groenewald, P. (2017). South Africa's vital statistics are currently not suitable for monitoring progress towards injury and violence Sustainable Development Goals. *SAMJ: South African Medical Journal*, *107*(6), 470-471.

87. Prinsloo, M. (2019) ‘Estimating injury mortality in South Africa and identifying urban-rural differences’,https://open.uct.ac.za/bitstreams/ba4c7b04-80f8-4032-bd8c-bad885b7dc57/download [Accessed on 19 Feb. 2025]
88. Prinsloo, M., Mhlongo, S., Roomaney, R.A., Marineau, L., Mamashela, T.A., Dekel, B., Bradshaw, D., Martin, L.J., Lombard, C., Jewkes, R. and Abrahams, N. (2024) ‘Injury mortality in South Africa: a 2009 and 2017 comparison to track progress to meeting sustainable development goal targets’, *Global Health Action*, *17*(1), p.2377828.
89. Purcell, L.N., Ellis, D., Reid, T., Mabedi, C., Maine, R. and Charles, A. (2021) ‘In-home interpersonal violence: Sex based prevalence and outcomes’, *African journal of emergency medicine*, *11*(1), pp.93-97.
90. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982–399
91. Saroha, E., 2018. Population pyramids and demographic transition: A visual analysis. *Journal of Population Research*, **35**(3), pp.211-225.
92. Saunders, C. J., Sewduth, D., & Naidoo, N. (2019). "Drowning prevention strategies in South Africa: An overview." International Journal of Aquatic Research and Education, 11(2), 1-9.
93. Savitsky, B., Radomislensky, I., Goldman, S., Gitelson, N., Frid, Z. and Peleg, K. (2020) ‘Socio-economic disparities and returning to work following an injury’, *Israel journal of health policy research*, *9*, pp.1-22.
94. Seedat, M., Van Niekerk, A., Jewkes, R., Suffla, S. and Ratele, K., 2009. Violence and injuries in South Africa: prioritising an agenda for prevention. *The Lancet*, *374*(9694), pp.1011-1022.
95. Shah, N. H., Milstein, A., & Bagley, S. C. (2019). Making machine learning models clinically useful. *JAMA*, 322(14), 1351-1352.
96. Tanser, F., Bärnighausen, T., Grapsa, E., Zaidi, J., & Newell, M. L. (2013). High coverage of ART associated with decline in risk of HIV acquisition in rural KwaZulu-Natal, South Africa. *Science, 339*(6122), 966-971.
97. Torfi, A., Shirvani, R.A., Keneshloo, Y., Tavaf, N. and Fox, E.A., 2020. Natural language processing advancements by deep learning: A survey. *arXiv preprint arXiv:2003.01200*.
98. Vogt, F., Schwappach, D. L. B., & Bridges, J. F. P. (2020). Navigating the ethics of text mining for healthcare research: A comprehensive review. *Health Policy*, 124(5), 519-526.
99. Wang, H., Abbas, K. M., Abbasifard, M., Abbasi-Kangevari, M., Abbastabar, H., Abd-Allah, F., & Murray, C. J. L. (2021). Global age-sex-specific fertility, mortality, healthy life expectancy (HALE), and population estimates in 204 countries and territories, 1950–2019: a

comprehensive demographic analysis for the Global Burden of Disease Study 2019. *The Lancet, 396*(10258), 1160-1203.

100. WHO (2016). *Verbal Autopsy Standards: The 2016 WHO Verbal Autopsy Instrument.*
101. WHO (2023). *Verbal autopsy standards: ascertaining and attributing causes of death tool*. [online] www.who.int. Available at: https://www.who.int/standards/classifications/other-classifications/verbal-autopsy-standards-ascertaining-and-attributing-causes-of-death-tool.
102. WHO. (2018, 2019, 2021). "Global status report on road safety; Preventing suicide; Drowning prevention." World Health Organization.
103. WHO. (2019). Preventing suicide: A global imperative. *World Health Organization*.
104. World Health Organisation (WHO) (2024). *Injuries and violence*. [online] www.who.int. Available at: https://www.who.int/news-room/fact-sheets/detail/injuries-and-violence.
105. World Health Organization (WHO), 2016. *Verbal Autopsy Standards: The 2016 WHO Verbal Autopsy Instrument.*
106. World Health Organization. (2019). Suicide prevention. *WHO Fact Sheets*.
107. Yu, K.H., Beam, A.L. and Kohane, I.S., 2018. Artificial intelligence in healthcare. *Nature biomedical engineering*, *2*(10), pp.719-731.
108. Zaki MJ, Meira Jr W. Data Mining and Machine Learning: Fundamental Concepts and Algorithms. Cambridge: Cambridge University Press (2019)
109. Zhu, F., Patumcharoenpol, P., Zhang, C., Shen, B., & Hong, J. (2013). Biomedical text mining and its applications in cancer research. *Journal of Biomedical Informatics, 46*(2), 200-211.
110. Žižka, J., Dařena, F. and Svoboda, A., 2019. *Text mining with machine learning: principles and techniques*.

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG	HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)

Office of the Deputy Vice-Chancellor (Research and Innovation)

TO: Ms Z Mlungwana
School of Public Health
Faculty of Health Sciences

E-mail: 673023@students.wits.ac.za

CC: Supervisor: Dr C Kabudula and Mr D Ohene-Kwofie
<Chodziwadziwa.Kabudula@wits.ac.za>
and <HREC-Medical Research Office@wits.ac.za>

FROM: Mr Iain Burns
Human Research Ethics Committee (Medical)
Tel: 011 717 1252

E-mail: Iain.Burns@wits.ac.za

DATE: 2024/10/22

REF: R14/49

PROTOCOL NO: **M2409105** (This is your ethics application reference number. Please quote it in all enquiries, oral or written, relating to this study.)

PROJECT TITLE: *Circumstances associated with injury-related deaths and changes in cause of injury-related deaths over three decades in rural north-east South Africa: insights from text mining of verbal autopsy narratives from Agincourt*

Please find attached the Clearance Certificate for the above project. I hope it goes well and that an article in a recognized publication comes out of it. This will reflect well on your professional standing and contribute to Government funding of the University.

MSWorks2000/Iain0007/Clearscan.wps

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

R49 Ms Z Mlungwana

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)
CLEARANCE CERTIFICATE NO. M2409105

NAME: **(Principal Investigator)** Ms Z Mlungwana **DEGREE:** MSc

DEPARTMENT: School of Public Health
Faculty of Health Sciences

PROJECT TITLE: *Circumstances associated with injury-related deaths and changes in cause of injury-related deaths over three decades in rural north-east South Africa: insights from text mining of verbal autopsy narratives from Agincourt*

DATE CONSIDERED: Ad hoc

DECISION: Approved unconditionally

CONDITIONS: Retrospective record review

NOTE: If contact information regarding student study participants is required, please contact the Registrar's office - <Nicoleen.Potgieter@wits.ac.za>

SUPERVISOR: Dr C Kabudula and Mr D Ohene-Kwofie

APPROVED BY: ______________________
Ms M van Niekerk, Co-Chairperson, HREC (Medical)

DATE OF APPROVAL: 2024/10/22 **EXPIRY DATE:** 2029/10/21

This Clearance Certificate is valid for 5 years from the date of approval. An extension may be applied for.

DECLARATION OF INVESTIGATORS

To be completed in duplicate and **ONE COPY** returned to the Research Office secretariat on the 3rd floor, Phillip Tobias Building, Parktown, University of the Witwatersrand, Johannesburg.

I/we fully understand the conditions under which I am/we are authorized 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 submit details to the Committee. **I agree to submit a yearly progress report** in the format available at https://wits.ac.za/research/researcher-support/research-ethics/ethics-committees/. This report is due annually, for the duration of the Clearance Certificate, beginning on the first anniversary of Date of Approval above. Unreported changes to the study may invalidate the clearance given by the HREC (Medical).

______________________ ______________
Signature of Principal Investigator Date

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

FACULTY OF HEALTH SCIENCES

PLAGIARISM DECLARATION TO BE SIGNED BY ALL HIGHER DEGREE STUDENTS

SENATE PLAGIARISM POLICY: APPENDIX ONE

I Zamangema Mlungwana (Student number: 673023) am a student registered for the degree of MSc. Epidemiology in the academic year 2025

I hereby declare the following:

I am aware that plagiarism (the use of someone else's work without their permission and/or without acknowledging the original source) is wrong.

I confirm that the work submitted for assessment for the above degree is my own unaided work except where I have explicitly indicated otherwise.

I have followed the required conventions in referencing the thoughts and ideas of others.

I understand that the University of the Witwatersrand may take disciplinary action against me if there is a belief that this is not my own unaided work or that I have failed to acknowledge the source of the ideas or words in my writing.

I have included as an appendix a report from "Turnitin" (or other approved plagiarism detection) software indicating the level of plagiarism in my research document.

Signature: ____________ Date: 17/06/2025

MlungwanaZN Turnitin-3.docx 17/06/2025

ORIGINALITY REPORT

10%	5%	6%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to University of Witwatersrand Student Paper	3%
2	www.ncbi.nlm.nih.gov Internet Source	1%
3	Michael Tonderai Mapundu, Chodziwadziwa Whiteson Kabudula, Eustasius Musenge, Victor Olago, Turgay Celik. "Text mining of verbal autopsy narratives to extract mortality causes and most prevalent diseases using natural language processing", PLOS ONE, 2024 Publication	<1%
4	Ndebele, Sikhuphukile Gillian. "Clustering of Child and Adult Mortality During Pre and Post ART Rollout Eras at Agincourt and Dikgale Health and Demographic Surveillance Systems in South Africa", University of the Witwatersrand, Johannesburg (South Africa), 2025 Publication	<1%
5	fastercapital.com Internet Source	<1%
6	Injury Research, 2012. Publication	<1%
7	wri-india.org Internet Source	<1%