University of East London

School of Architecture, Computing and Engineering

Title

Analysis of Heterogeneous Data Sources for Veterinary Syndromic Surveillance to Improve Public Health Response and Aid Decision Making

By

Victor Adejola

A Thesis submitted in partial fulfilment of the requirement of the University of East London for the degree of Professional Doctorate in Data Science.

September 2021

Abstract

The standard technique of implementing veterinary syndromic surveillance (VSyS) is the detection of temporal or spatial anomalies in the occurrence of health incidents above a set threshold in an observed population using the Frequentist modelling approach. Most implementation of this technique also requires the removal of historical outbreaks from the datasets to construct baselines. Unfortunately, some challenges exist, such as data scarcity, delayed reporting of health incidents, and variable data availability from sources, which make the VSyS implementation and alarm interpretation difficult, particularly when quantifying surveillance risk with associated uncertainties. This problem indicates that alternate or improved techniques are required to interpret alarms when incorporating uncertainties and previous knowledge of health incidents into the model to inform decision-making. Such methods must be capable of retaining historical outbreaks to assess surveillance risk.

In this research work, the Stochastic Quantitative Risk Assessment (SQRA) model was proposed and developed for detecting and quantifying the risk of disease outbreaks with associated uncertainties using the Bayesian probabilistic approach in PyMC3. A systematic and comparative evaluation of the available techniques was used to select the most appropriate method and software packages based on flexibility, efficiency, usability, ability to retain historical outbreaks, and the ease of developing a model in Python. The social media datasets (Twitter) were first applied to infer a possible disease outbreak incident with associated uncertainties. Then, the inferences were subsequently updated using datasets from the clinical and other healthcare sources to reduce uncertainties in the model and validate the outbreak. Therefore, the proposed SQRA model demonstrates an approach that uses the successive refinement of analysis of different data streams to define a changepoint signalling a disease outbreak.

The SQRA model was tested and validated to show the method's effectiveness and reliability for differentiating and identifying risk regions with corresponding changepoints to interpret an ongoing disease outbreak incident. This demonstrates that a technique such as the SQRA method obtained through this research may aid in overcoming some of the difficulties identified in VSyS, such as data scarcity, delayed reporting, and variable availability of data from sources, ultimately contributing to science and practice.

Subject Keywords

Veterinary Syndromic Surveillance, Data Scarcity, Syndromic Surveillance, Disease Outbreak Detection, Early Outbreak Detection, Risk, Stochastic Quantitative Risk Assessment, Public Health, Decision Making, PyMC3, Changepoint, Bayesian Methodology, Probabilistic Approach, Markov Chain Monte Carlos, Frequentist Approach, Uncertainty, SAVSNET, Prior Distribution, Posterior Distribution, Poisson Distribution, and Negative Binomial Distribution (NB).

Ethics Approval

-This project required ethical approval from the University of East London Ethics Committee, and confirmation of approval is under the application ID: ETH2122-0019 updated on 20th September 2021. The original application ID is ETH1920-0217. The ethics approval is provided in the Appendix section of this research work.

Acknowledgement

As Nelson Mandela puts it, "It always seems impossible until it is done"; this is true for the more difficult tasks in life. However, the subtle or obvious assistance from you all is the critical component that helped me complete this research work. Therefore, I thank everyone who has crossed my path in the course of undertaking this research work.

I want to express my gratitude to my first supervisor and former director of studies, Professor Allan Brimicombe, for initiating this work with me. He was constantly available to assist; his mentorship, vast knowledge, and experience were extremely beneficial throughout this research work, even after he retired from the school during this study. You have always provided me with the appropriate balance of freedom and support to accomplish my goals, and most importantly, you gave me your trust. I will always remember and appreciate this trust, and I hope to pass it on to other junior researchers who may cross my path. I believe you would consider this a satisfactory recompense, do you agree, Prof?

Furthermore, I thank Dr Yang Li, who took over from Professor Allan, for his supervisory during the final stages of my doctorate programme and for all his guidance, help, and leadership throughout my study journey at UEL. He helped steer me through the challenges when required. I am also thankful to the SAVSNET team at the Institute of Veterinary Science, Faculty of Health and Life Science, University of Liverpool, for their help and support on SAVSNET data access; You made this research work possible by supporting with access to the SAVSNET platform, and for interpreting the data. Also, I am thankful to the entire team of the company where I work, The Irish Equine Centre, for supporting, motivating, and interpreting the syndromic surveillance datasets.

My heartfelt appreciation goes out to my family, Shade, Daniella, Dammy, Esther, and Joshua, who suffered a lack of attention because of my doctorate studies. The odd hours spent on this research work and the general lack of time they all received from me. Special thanks to all my friends who withstood the suffering because I could not give them attention or time throughout this programme. I appreciate your understanding, and I want you all to know how much I owe you for the anguish you endured during the six years of my doctoral studies.

Please note that this acknowledgement does not mention an editor, implying that I am solely responsible for writing, drafting, and editing my thesis. As a result, no portion of this research was written with the assistance of a professional editor.

Contents

List of Figures

List of Tables

List of Acronyms

ARIMA - Autoregressive Integrated Moving Average

EWMA - Exponentially Weighted Moving Averages

MEWMA - Multivariate Exponentially Weighted Moving Average

MCMC - Markov Chain Monte Carlo

MCUSUM - Multivariate Cumulative Sum

NHS - National Health Service, United Kingdom

NLP - Natural Language Processing

PHE - Public Health England

QRA - Quantitative Risk Assessment

SAVSNET - Small Animal Veterinary Surveillance Network

SDT - Signal Detection Theory

SyS – Syndromic Surveillance

TFP - TensorFlow Probability

VSyS – Veterinary Syndromic Surveillance

Chapter 1 – Research Background and Introduction

1.1 Introduction

Traditional disease surveillance methods have been used for many years under the public health function as part of national health programs (Smith et al., 2017). Their purpose is to identify the presence of a particular pathogen or reportable disease in a population by monitoring and analysing clinically confirmed cases and potential outbreaks (Dupuy et al., 2013a). The effectiveness of the method is apparent for the disease they are designed to monitor. However, literature evidence shows that such a surveillance approach often fails to identify new disease outbreaks in real-time or provide a risk assessment (Faverjon et al., 2019a; Dupuy et al., 2013a; Dórea and Vial, 2016). In the current context, potential health threats may be unpredictable as present globalisation, biological terrorism, global warming, human and animal population growth are favourable to the emergence of new diseases (Jones et al., 2008; Noufaily et al., 2019). Even if potential health threats could be identified, there are no sufficient resources to implement dedicated traditional surveillance systems to monitor the risk that each poses to a population.

Also, Syndromic Surveillance (SyS) is a public health practice designed not to replace the traditional disease surveillance but to supplement it with the broad aim of identifying the early, often weak, signal of a health event in the absence of accurate identification of health incident by the traditional surveillance system and healthcare officials (Ziemann et al., 2016). SyS was first implemented in the USA in the late 1990s to detect a bioterrorist attack in real-time. While in Europe, the motivation for SyS grew in 2003 after the heatwave incident, which led thirteen EU states to execute emergency surveillance and automated mortality systems (Dupuy et al., 2013a).

A basic SyS strategy involves collecting, processing, and analysing clinical and non-clinical datasets to identify, understand, and monitor health events before confirming a diagnosis (Centers for Disease Control and Prevention, 2020; Triple- S Project, 2013). Also, to monitor non-specific health indicators that are critically essential for informing decision-making and stimulating timely and appropriate public health action. Therefore, SyS uses data that are routinely collected for other purposes to analyse and detect possible disease outbreaks. The

most notable SyS method relies on alarm-generating statistical models, where the data points above a defined threshold indicate an unusual event and triggers an alert.

Dórea and Vial (2016) argued that the failure of traditional disease surveillance techniques and SyS to recognise new health threats and provide the risk impact on time is a limitation and may sometimes have serious consequences. For example, the three-week delay in detecting the 2001 foot and mouth outbreak in the UK caused the diseases to spread quickly throughout the country undetected, resulting in more than 6.5 million livestock destroyed to control the disease. This health event resulted in **£8 billion** to eradicate (Office NA, 2002). A recent example is the COVID-19 outbreak. Delay in identifying the infection and its potential risks resulted in a global widespread that infected over 223 million people and caused over 4.6 million deaths when writing this research work.

1.2 Identified Challenges

The standard VSyS technique uses statistical algorithms to monitor surveillance datasets and indicate deviation from baselines in the form of alarms (Vial et al., 2016). However, some studies highlighted the reduced sensitivity of the current approach to changes in disease frequency, which often lead to increased false alarms (Fischer et al., 2015; Andersson et al., 2014; Robertson et al., 2011). Also, in practice, all alarms must be investigated to understand the nature and possible cause. Such investigation is called the risk assessment of the public health alarm, and the existing approach appears to be manual driven and frequently requires special training, time, and maintenance of an expertise base to establish which alarms are important public health concerns (Lake et al., 2019).

Although the outputs of the current VSyS methods may be simple to understand, the interpretations are potentially complicated to use, especially when a sudden spike in the dataset is close to the baseline or when there is a slowly increasing outbreak, or the number of cases reported in each time unit is too small to trigger an alarm. Furthermore, it could be difficult to combine alarm outputs with other epidemiological knowledge such as environmental factors, disease seasonality and uncertainties, or any previous knowledge of the health incident to inform public health decisions. Therefore, there is no straightforward approach to measuring risk when the underlying algorithm is based on an alarm threshold method (Veldhuis et al., 2016).

Evidence from the literature review shows that most alarm-generating techniques produce many false alarms when the threshold value is lowered and few alarms when the threshold is increased. This behaviour shows that the current VSyS techniques cannot explain the impacts of each data point when quantifying surveillance risk. Therefore, the alarm-based VSyS may not provide sufficient information to discriminate adequately between low and high-risk events.

A method to quantify disease outbreak occurrence from data signals and provide detailed information to differentiate between high and low-risk periods still needs further research investigations. The review of previous research publications on the Quantitative Risk Assessment (QRA) method indicates that the available techniques are focused on identifying disease introduction into a population using the @Risk Microsoft Excel application. This application is not easily adaptable for implementing a method to quantify the risk associated with SyS alarm due to several challenges identified in this study, including the limitation of the @Risk software. Therefore, quantifying alarms by measuring the magnitude of risks and estimating the associated uncertainty in VSyS requires further investigation. However, earlier research shows that epidemiological datasets are rarely investigated with uncertainty measurement in mind (Smith et al., 2017); this is a knowledge gap in veterinary syndromic surveillance.

Due to the identified challenges, the veterinary public health domain remains a poorly investigated study area. Existing data gathering strategies in VSyS are not as mature as in human medicine, where the NHS and Public Health England (PHE) in the U.K coordinate the collection and analysis of various datasets under a cohesive framework. Datasets used for VSyS are generated from various sources without a standardised approach; this is because veterinary medicine is predominantly a private sector with no identifiable data framework, causing a lack of standardisation among the different systems that generate the datasets (Vanderwaal et al., 2017; Dórea et al., 2019).

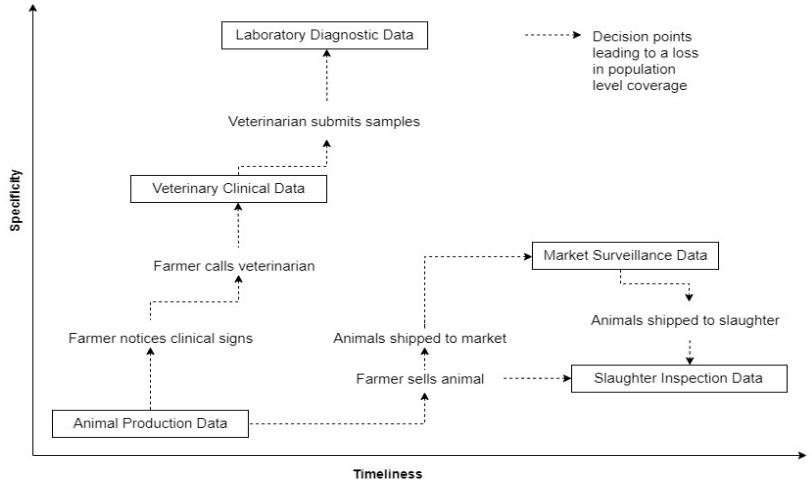

Figure 1: *Veterinary data lifecycle, specificity and timeliness (adapted from Egates et al. (2015))*

Many researchers also identified the data scarcity and under-reporting of animal health incidents as major difficulties for progressing investigative studies in VSyS (Alkhamis et al., 2018; Robertson and Yee, 2016; Egates et al., 2015; Dórea et al., 2011; Dórea and Vial, 2016). Furthermore, the timeliness of the animal health dataset for outbreak detection analysis is highly variable as different sources generate data at relatively different times. For example, Figure 1 presents the data lifecycle and generation path for a typical veterinary scenario. It shows that the availability of the animal production dataset for disease outbreak analysis is near real-time compared to the slaughter inspection or market surveillance dataset. These datasets take much longer to be available in the epidemiological data lifecycle, which is additional complexity.

The social media platforms can generate an abundance of data needed to address the data scarcity in animal medicine (Yousefinaghani et al., 2019), particularly in companion animals. Nowadays, many animal owners can quickly surf the web for information on sick animals or post questions about observed symptoms/signs of unwell animals or livestock to generate useful discussions on different internet platforms. Such activities can contribute to non-clinical datasets for VSyS implementation. However, a method that provides a thorough understanding of probability distribution which is critical to identifying high and low-risk regions in datasets with a clear knowledge of the associated uncertainties for proactive public health strategies is lacking (Faverjon et al., 2017).

Consequently, research into the probability distribution of data is vital for implementing the risk assessment of alarms in VSyS. This may involve investigating and incorporating techniques that addressed data scarcity and delayed reporting. Hence, these challenges call for a new VSyS approach that can kick off by using readily available datasets such as social media datasets, like Twitter or Facebook, to infer a possible outbreak quickly and subsequently update inferences when more datasets are available from clinical and other healthcare sources to quantify risks and validate disease outbreaks.

1.3 Research Aim and Objectives

In light of addressing the challenges and gaps identified in knowledge, the author aims to explore the development of a novel VSyS technique for analysing heterogeneous data sources and assessing surveillance risk while considering existing knowledge of disease outbreaks and associated uncertainties to improve public health response and aid decision-making. The research approach will rely on data science techniques and combine social media datasets in real-time with the routine clinical and laboratory datasets to produce a method with a detailed step-by-step procedure as the major contribution to public health practice and science.

The Bayesian probabilistic modelling approach using the Markov Chain Monte Carlo (MCMC) sampling technique and changepoint analysis is selected as the preferred data science technique. The goal is to demonstrate an approach that uses the successive refinement of analysis of different data streams through probability distribution and uncertainty measures to define a changepoint signalling a disease outbreak.

The objectives of this research are:

1. To consider the concept of risk assessment in public health surveillance and define the risk terminologies in the context of their application to this work. Furthermore, clarify risk analysis, risk assessment and management, and differentiate quantitative from qualitative risk assessment procedures. Evaluate various theoretical frameworks and determine which theories are most appropriate for this work.

2. Conduct a review of the published literature on routinely explored data sources in VSyS development and investigate the data modelling techniques to ascertain whether they

account for associated uncertainties, assumptions, and existing knowledge about health incidents to interpret VSyS alarms and inform decision-making.

3. Present the limitations of the existing VSyS methodologies and identify ways to address them in the context of the aims of this research work.

4. Consider and select the data science techniques, programming platforms, and software packages to deliver the research objectives and address the identified challenges. Also, to indicate the difficulties associated with the proposed technique of analysing heterogeneous data sources in VSyS and how they can be resolved.

5. Use the available datasets from Twitter and Small Animal Veterinary Surveillance Network (SAVSNET) to develop, validate, and test the method. Also, to present the detailed step-by-step procedure of the proposed method as the contribution of this study to knowledge.

1.4 Research Questions

In this study, various sources of information that shed light on the development of VSyS and its associated data modelling techniques were explored. The main research question to deliver the primary aim of this study is "how to develop a procedure in step-by-step detail for analysing heterogeneous data sources and evaluating VSyS alarms to improve public health response and aid decision making?". However, due to the scope of the research work, the following sub-questions based on objective opinion were used to address the primary research question.

1. How is risk knowledge and its components perceived in public health surveillance, and which theory or theories can be used to conduct this research?

2. What are the existing VSyS techniques and data sources? How do they account for uncertainties and previous knowledge of health incidents in a model to evaluate risk? Furthermore, what are the difficulties associated with their use?

3. Is there any published research demonstrating the use of QRA techniques in VSyS?

4. What techniques in data science can be applied to datasets from social media and veterinary medicine to evaluate risks in VSyS to aid public health decision-making?

What is the detailed step-by-step procedure? Furthermore, what are the obstacles to overcome and practical solutions?

1.5 Thesis Structure

The structure of this thesis document is as follows; in chapter two, an investigation of public health risk was conducted from an interdisciplinary perspective, beginning with the concept of risk and the theories that surround it. Then, risk assessment and risk analysis were differentiated, and key terms used throughout this study were defined. In chapter three, a systematic analysis of peer-reviewed publications was conducted to identify the routinely explored data sources and the challenges of using them for veterinary disease surveillance. Furthermore, a literature review of the existing VSyS modelling techniques and their ability to retain historical outbreaks and account for associated uncertainties was critically analysed to highlight research gaps. The datasets used in this study were presented and described in chapter four, including detailed information about their sources and the methods used to obtain them. Also, previous publications that used the datasets or part of the data in their research were critically analysed to identify their contributions and data science techniques. The aim is to ensure that knowledge is not duplicated.

The research methodology was explored in chapters five to seven. The approach for filling the identified gaps using various data science techniques was presented and analysed in chapter five. Different software packages were identified and critically investigated for implementing the proposed method. In chapter six, the datasets acquired for this study were explored and visualised to understand their structures, shapes, attributes, and the accompanying challenges due to the unique characteristics of veterinary surveillance datasets. In chapter seven, different probabilistic programming approaches were investigated, and a suitable technique was selected based on flexibility, efficiency, usability, ability to retain historical outbreaks, and the ease of developing a model in Python language. The proposed method was developed, and a detailed procedure of its application to the Twitter data and the results were presented in chapter eight.

In chapter nine, the author presented the difficulties encountered during the data modelling phase and discussed the solutions to overcome them. Also, the resulting SQRA method was tested and validated using real-life veterinary epidemiological datasets from SAVSNET, and their validity and reliability were discussed. In chapter ten, the author summarised the whole thesis and concluded that the SQRA method could be implemented using the Bayesian

probabilistic modelling approach with the MCMC sampling technique and changepoint analysis to interpret VSyS alarms with uncertainty measures. Furthermore, the author concluded that the approach proposed in this study could improve public health response and support the decision-making process in syndromic surveillance. The study's limitations, further research work and recommendations were also outlined in the final sections of chapter ten.

Chapter 2 – Public Health Surveillance and Risk Concept

2.1 Chapter Introduction

The modern description of public health practice dates back to Charles-Edward Amory Winslow's definition in 1920. Winslow defined public health practice as "the science and art of preventing disease in a population in order to prolong life and promote physical health through specialised community efforts for a clean environment, infection control, individual hygiene education, and the organisation of health care." (Winslow, 1920). Public health surveillance is an aspect of public health practice that has drawn more and more attention in the last decade from multiple research fields. This is because it serves as a tool for estimating the health status and behaviour of a population. It is useful for evaluating the risk of infection and measuring the need for interventions and their effects (Martin-Moreno et al., 2016; Kivits et al., 2019; Ruddell et al., 2014).

In this chapter, public health surveillance and its components are presented. The relationships between disease surveillance and risk concepts are defined. The goal is to build a foundation for the elements of this study and clarify the differences between active and passive surveillance practices and why they are important in this study. The different key terminologies used throughout this work are also explained to eliminate ambiguous interpretations.

A theoretical framework serves as the blueprint that provides the guideline for developing and interpreting a research study (Osanloo and Grant, 2016). As a result, the author investigated different theoretical and conceptual risk models to establish and present the appropriate theoretical framework that best fits the delivery of this research work.

2.2 Public Health Surveillance

Public health surveillance is undertaken to inform the public health practice of health threats to humans and animals. It is an interdisciplinary function that involves “an ongoing systematic collection, analysis and interpretation of health-related data, which are essential for planning, implementing, and evaluating public health practice” (WHO, 2020). The two approaches to public health surveillance identified in the literature are traditional disease surveillance and syndromic surveillance (SyS) (Abat et al., 2016; Vial et al., 2016; Salathé, 2016; Ziemann et al., 2016; Dupuy et al., 2013a; Triple- S Project, 2013). Traditional disease surveillance has been around for decades with a primary focus on identifying a particular disease or pathogen

in a population of interest by tracking potential cases through relevant laboratory testing or medical experts' experiences (Dupuy et al., 2013a). In comparison, Dórea and Vial (2016) highlighted SyS as a new approach under public health surveillance based on real-time tracking and analysis of unspecific clinical and non-clinical datasets to enable the early identification of the impact of potential health threats. SyS aims to inform the decision-making of public health concerns through risk assessment (Smith et al., 2017).

Faverjon (2017) highlighted two main strategies for SyS as active and passive surveillance techniques. Active surveillance indicates the active role of health officials in collecting datasets. The benefit of this approach is that it can identify the disease that is designed to monitor in a population. However, a major disadvantage is that the active surveillance approach needs to be exceptionally large in data sampling to identify rare diseases like a new strain of a disease outbreak. Implementing an extensive active sampling of a population of interest for different disease types could be very expensive and cost-prohibitive.

On the other hand, passive surveillance is one of the most popular surveillance strategies for identifying exotic and rare diseases. The approach implies any passive disease reporting systems, whereby data about symptoms and signs of diseases are collected from routine datasets meant for other purposes. For example, emergency attendance, ambulance calls out, GP visits, pharmaceutical sales, and 111 calls, to name just a few. The passive approach is highlighted as a low-cost strategy for data collection. However, it has been identified that under-reporting of health incidents, especially in veterinary medicine, could be a major disadvantage of the passive approach due to a lack of standardised data collection framework in the veterinary sector (Dórea et al., 2011; Dórea and Vial, 2016; Egates et al., 2015).

Many research literature agreed that the purpose of SyS is to complement traditional disease surveillance under public health functions since health event identification still largely relies on laboratory confirmation of an outbreak (Struchen et al., 2017; Ziemann et al., 2016; Dupuy et al., 2013a). Therefore, more and more recent research work focuses on the efficiency of the SyS technique to identify disease outbreaks within a population of interest in a timely fashion using the most up-to-date technologies (Faverjon et al., 2019b; Hale et al., 2019; Smith et al., 2017; Lake et al., 2019; Bollig et al., 2020). Consequently, leading to the development of various public health surveillance systems (Musa et al., 2018) but with little emphasis on systems that can streamline the assessment of public health risks (Faverjon, 2017; Smith et al., 2017).

2.3 Risk as a Concept

There are various views to risk in the literature; some risk descriptions are focused on probability distribution, expected values or chance, some are based on undesirable events or danger, and others are described from uncertainty measures. Some literature describes risk as subjective and epistemic, which depends on available knowledge; others view risk with an ontological status independent of assessors (Aven, 2012). In a general context, the risk concept may be described as an act or phenomenon that poses a hazard and can produce harm or other undesirable consequences to humans, animals, the environment, or things.

The risk components are; hazard, exposure, and vulnerability (Cardona et al., 2012). Although hazard has been used interchangeably with risk, it has recently been widely accepted as a risk component and not risk itself (Field et al., 2012). The extent of the hazard may be described as the amount of resulting harm, for example, the number of animals, humans or things exposed and the severity of consequence. This means that the risk concept can also quantify hazards by attributing the probability distribution of being realized to each stage of potential harm (National Research Council, 1989).

Aven (2012) highlighted a gradual change in how practitioners now perceive the risk concept from a rather narrow risk perspective focusing on probabilities and expected values to a broader view, including a clear distinction between risk measurements, uncertainties, and risk concept as a whole. Many literature reviews, such as those by Renn (1992), Althaus (2005) and Aven and Renn (2010), revealed the measurements and descriptions of several risk perspectives in different disciplines, showing a classification of how different studies and disciplines look at risks, which means risk differs in its application to different fields. However, Aven (2012) view is that different disciplines must create different methods for defining, assessing, and managing risks.

As a result, public health risk is an event or an action that may likely cause harm to human or animal health or contribute to disease among humans or animals, such as infections, bioterrorism attacks, and germs. It also includes transmittable infectious agents from the environment, water, decomposing biological remains, and harmful substances, to name a few (Queensland Health, 2019). However, the risk concept cannot be discussed without explaining other terminologies used as part of the description of risk in public health practice concerning disease surveillance, such as signal, alarm, alert, and uncertainty.

2.3.1 Signal

Signal refers to the data resulting from a monitored health event occurrence. In public health practice, an event occurrence could be an increased number of visits to a hospital emergency department to treat a specific symptom or syndrome (Smith et al., 2017). A sudden or gradual increase in the signal could be a disease outbreak. Therefore, public health surveillance aims to identify the occurrence of an event such as an epidemic from a signal (Wagner et al., 2001). A SyS processes the signal through a statistical algorithm to produce an output that establishes whether an outbreak is present or not. Heffernan et al. (2004) referred to signal as statistically significant aberration; this means that SyS use various aberration detection techniques to identify syndrome increase above a predetermined threshold. However, a signal corresponding to disease outbreaks could consist of substantial background noise, giving rise to uncertainties in health event occurrence (Henning, 2004). With the more rapidly increasing datasets nowadays, it is essential to differentiate signal from noise and place potentially significant events in context.

2.3.2 Alarm

The signal may lead to an alarm to indicate that a statistical aberration representing an epidemiological signal of interest has occurred. The alarm means that the statistical analysis result may need further investigation because the population under surveillance may be affected by a disease outbreak or an event (Triple- S Project, 2013). Alarms are sometimes used interchangeably with statistical indicators from statistical analysis to reveal a significant deviation from the normal situation or threshold and quantify the degree of such departure from the normal situation. SyS typically produce output signals compared against a threshold to determine whether an alarm has occurred or not. For example, GP in-hours for asthma with a normal baseline of 10 and upper limit of 18, but a signal of 20 on a particular day would result in an alarm (Wagner et al., 2001; Morbey et al., 2015).

2.3.3 Alert

Alert is a notification that an alarm raised requires further consideration for public health action and decision. Launching a public health alert may be divided into two main steps: (1) verification of data quality and confirmation of the statistical alert, and (2) epidemiological verification of the signal that generated the alert (Triple- S Project, 2013). Implementing the risk assessment procedure aims to identify those alarms that need to be prioritized to result in an ‘alert’ as reliably and consistently as possible (Smith et al., 2017). The alert is an important

output of a surveillance system, highlighting a possible emerging health event. The general workflow of a syndromic surveillance investigation typically begins with a statistical aberration of the signal before progressing to an alarm and, finally, an alert (Triple- S Project, 2013).

2.3.4 Uncertainty

Uncertainty has been described as a situation where the available information may not be sufficient to decide the issue under consideration (Fonnesbeck et al., 2018). In other words, uncertainty refers to the likelihood of no accurate impact prediction. The risk concept and uncertainty are closely related, so like risk, uncertainty has many perspectives based on its application in different fields. Similarly, the National Research Council describes uncertainty as the lack of complete information essential for performing risk assessment (National Research Council, 2009). Scientific uncertainty can provide risk assessments with confidence levels about decisions made and evaluate the extent of uncertainty's influence on a specific decision's consequences.

As a result, the probability of realising an event can be estimated using current and historical datasets; this implies that the difference between uncertainty and risk is the probability of estimating an event occurrence, which can be established through risk measurement but not through uncertainty measurement. Cardona et al. (2012) also indicated that the risk consequences could either be negative or positive. Therefore, measuring a risk involves combining data containing consequences and incident frequency, which means that some uncertainties always accompany risk estimation techniques. In cases where the data used for risk estimation contain issues, such condition is described as "knowledge uncertainty", which is when gaps arise in knowledge creation and the data used for creating the knowledge (Berztiss, 2004). Therefore, it is important to manage uncertainty. Though it is impossible to eliminate uncertainty from a knowledge creation process, identifying and managing uncertainty is worthwhile to forestall unfavourable decision-making issues (Li et al., 2012).

However, establishing the nature and magnitude of uncertainties remains a major challenge in risk assessment due to the lack of best practices for characterising and quantifying uncertainties and variability in epidemiological datasets (Burns et al., 2014). Thus, disease surveillance datasets are critical for risk assessment efforts in public health practice, but research evidence shows that epidemiological datasets are rarely investigated with uncertainty measurement in mind (Smith et al., 2017).

2.3.5 Risk Analysis and Risk Assessment

Risk assessment is an approach of investigating potential future events to avoid, reduce or manage risk more efficiently (Wilson and Crouch, 1987). According to the World Health Organization (2012), risk assessment is an organised process for gathering, evaluating, and recording information to determine the relative importance of risks and assign severity to them based on risk analysis. Also, since the risk concept and the notion of uncertainty are related, risk assessment can be viewed as a structured way to address uncertainty using scientific and technical evidence to make judgements on the tolerability of the risk. Layton et al. (2015) considered risk assessment as the foundation for safety regulatory decision-making bound by the policy and legislative requirements for making timely decisions using available resources. In the context of public health, risk assessment provides the foundation for acting, managing, and reducing the consequences of serious public health risks; see the public health risk management cycle in Figure 2. Risk assessment is one of the core elements of a risk management strategy. Additionally, communication within the risk management strategy is seen as a key element of the process (Cardona et al., 2012). Thus, Risk assessments and risk identification must be linked to risk communication types and strategies for effective risk management techniques.

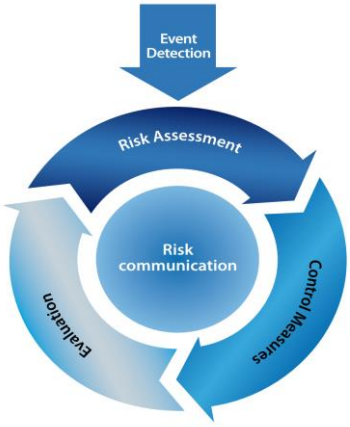

***Figure 2**: The risk management cycle (World Health Organization, 2012)*

However, both terms (risk assessment and risk analysis) have been used interchangeably by non-scholarly sources for a long period of time. In contrast, they mean two different processes within risk management methodology. Risk analysis is a method for identifying and investigating the vulnerabilities of potential threats. Therefore, it is a process that provides a flexible framework within which identified risks of adverse consequences resulting from a

course of action could be evaluated in a systematic and science-based manner (Ezekiel et al., 2011).

In comparison, risk assessment involves many steps and forms the backbone of risk management, providing the mechanisms to identify and prioritise risk so that each identified risk is analysed to understand the root cause. In other words, risk assessment is one of the processes of risk management strategy, while risk analysis is at the core of risk assessment. Thus, Figure 3 illustrates the relationship between risk management, risk assessment, and risk analysis. The risk analysis approach permits a defendable decision on whether a risk posed by a particular action or "hazard" is acceptable and provides the means to evaluate possible ways to reduce the risk from an unacceptable level to one that is acceptable. Therefore, to inform the risk management process, risk assessments must be structured to provide the mechanisms for risk prioritization and cost-effectiveness in addressing the potential consequences of identified hazards through risk analysis, as indicated in Figure 3; this means that risk analysis is embedded within a risk assessment strategy.

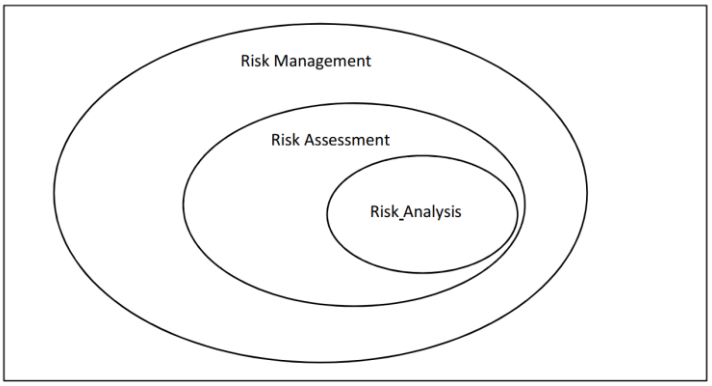

***Figure 3**: Risk management, risk assessment and risk analysis relationships*

2.3.6 Qualitative and Quantitative Risk Assessment

Existing research literature distinguishes between risk assessment as quantitative or qualitative, depending on whether data is used as a strategy to drive the assessment and estimate the impact of risks. Qualitative risk analysis uses descriptive techniques to ascertain the potential impact and likelihood of risk situations and prioritise them. Therefore, qualitative risk assessment methods are modelled with qualitative knowledge to identify several possible risk situations worthy of consideration (Coletti et al., 2020). Bode et al. (2018) argued that qualitative risk

assessment was preferred in risk assessment implementation because the alternative quantitative approach appears rigorous and less practical due to its requirement for quality data.

On the other hand, Souley Kouato et al. (2018) argued that the qualitative risk assessment approach must be thorough to ensure that genuine risk and not false risk perceptions are captured. This is because the qualitative risk assessment technique relies on qualitative knowledge, which might not capture all factors responsible for different risk situations (Coletti et al., 2020). As a result, Souley Kouato et al. (2018) recommended the quantitative risk assessment technique since it allows a numeric estimate of the probability of risk and the magnitude of the consequences in a measurable manner. This approach enables the modelling of uncertainty to determine the effects of random input parameters affecting risk magnitude. Also, it allows discrimination between large and small risk magnitudes. This approach's downside is the complex implementation requirements, extended project time, increased project resources, and the need for high-quality and accurate datasets. Nevertheless, quantitative risk assessment methods have proven useful in the last decades for assessing the magnitude of threats to human and animal health (Opatowski et al., 2020). Therefore, the benefits of quantitative risk assessment techniques may outweigh their downside if the gaps and constraints due to data quality, time, and resources could be closed or reduced efficiently (Coleman and Marks, 1999).

2.4 Theoretical Frameworks

The theoretical framework serves as a conceptual foundation for comprehending, analysing, and developing methods for investigating a study problem (Osanloo and Grant, 2016). As a result, the theoretical framework comprises time-tested hypotheses that incorporate multiple investigations into how phenomena may have occurred. On the other hand, a conceptual framework establishes the researcher's vision for how to approach a research problem. Conceptual frameworks are built on a theoretical framework with a much broader resolution scale.

Therefore, researchers need to define their approach to a research problem and justify their methodology in order for readers to understand the perspective of the researcher on the issues been addressed. Additionally, Kitchel and Ball (2014) defined a theoretical framework as a collection of related concepts, propositions, and definitions that illustrate and explain the relationship between variables. As a result, the conceptual and theoretical frameworks must

specify the research focus, anchor it firmly in theoretical constructs, and enhance the acceptability and significance of the study findings (Anfara Jr and Mertz, 2014). While theoretical and conceptual frameworks appear similar, their approach, style, and application within a study are distinct. As a result, a study requires a theoretical framework, a conceptual framework, and a literature review. These three components lay the groundwork for the study, demonstrate how the research advances knowledge, conceptualises and evaluates the study, and serve as a reference point for interpreting findings. Therefore, the author critically examined various theoretical frameworks in the subsequent sections to identify the most appropriate theory for interpreting the results of this ongoing research study.

Under public health practice, any health risk is expected to be monitored and assessed through the disease surveillance function (Heffernan et al., 2004). Such disease surveillance function produces data that can be used to implement quantitative risk assessment to understand and manage public health risk occurrences. These datasets may be suitable for differentiating high and low-risk impacts if the right data analysis strategy is employed. However, the public health organisations such as "PHE" in England do not instruct individuals or organisations on how to measure the magnitude of risks resulting from public health incidents (either through a qualitative or quantitative risk assessment approach).

2.4.1 Organisational Theory

The organisational theory involves investigating and describing enterprise behaviours and their environments and using the resulting knowledge for decision-making in its operations. Human factors are involved at every level of most organisational activities and are responsible for incidents/accidents. Therefore, more high-risk organisations adopt safety management, emphasising lessons learned from past risk experiences and other similar organisations to generalise safety management methods (Grote, 2012). With a focus on the significance of human behaviour in the industry, some research studies proposed new approaches to support organisational risk assessment in industrial environments (Carpitella et al., 2018; Khan and Burnes, 2007; Trucco et al., 2008). These studies demonstrated efforts to apply theoretical frameworks based on the study of organisations' behaviours to risk management. Enya et al. (2018) reviewed the available evidence of organisational theory as a strategy to manage construction safety in high-reliability organisations. They concluded that high-risk but high-reliability organisations have unique safety management principles entrenched in organisational behaviours, which helps them implement effective risk management policies.

2.4.2 Contingency Theory

Contingency theory suggests that organisational effectiveness results from the "organisation fitting characteristics" such as their structure to contingencies. Contingencies include the organisational size, environment (internal and external), and organisational strategy (Donaldson, 2001). The Contingency theory describes how organisations are shaped through contingencies since enterprises need to "fit" their changing contingencies to avoid performance loss. Some academic studies examined the adoption and impact of Enterprise Risk Management (ERM) and put forward a contingency theory of ERM, identifying factors that may describe observable variation in the ERM mix implemented by different organisations (Mikes and Kaplan, 2014; Woods, 2009; Grötsch et al., 2013). While researchers may hypothesise a relationship between contingent variables such as risk and ERM mix and outcomes such as organisational effectiveness, contingency theory can be thought of as a framework that promotes the principle of "no single optimal strategy" for risk management. As a result, advocating for the implementation of alternate pathways may be more appropriate for an organisation's unique contingency needs. Additionally, because each risk is unique, it must be managed according to its specific characteristics and location in a defined time.

2.4.3 Risk Behaviour Theory

Jiang et al. (2018) applied risk behaviour theory to investigate and interpret the risks of excessive Internet usage among adolescents in China. The authors' research focused on risk assessment techniques and proposed a conceptual model with a theoretical origin in risk behaviour theory. According to Trimpop (1994), risk behaviour is "any consciously or unconsciously controlled behaviour with a perceived uncertainty about its outcome, or its possible benefits, or costs for the physical, economic or psychosocial well-being of oneself or others." Killianova (2013) highlighted several applications of risk behaviour theories depending on the field of research study and the perception of the risk concept. The theory of risk behaviour provides adequate ground for investigating and explaining the relationships between risky human behaviours and risk concepts.

Globocnik (2019) relied on risk behaviour theory to investigate the relationship between risk propensity and employees' secret innovative behaviour and whether risk propensity can foster bootlegging behaviour in organisations. In public health practice, risk behaviour theory is commonly proposed to explain how risky human behaviours may encourage the spread of

serious contagious diseases. For example, Rhodes (1997) proposed risk behaviour theory for investigating and interpreting the problems and relationship between acquired immunodeficiency syndrome (AIDS) and drug addiction. Therefore, risk behaviour theory is a multidisciplinary model for explaining or interpreting phenomena and connecting human behaviours and risk concepts.

2.4.4 System Behaviour Theory

In a typical organisation, different components interact with each other or operate independently as a continuous system. These individual components perform different functions that contribute to an organisation's overall operation. Therefore, their behaviour can be investigated separately. The assumption that underpins the investigation of system behaviours is that the system can be divided into smaller subunits and produce results that can be analysed independently of the entire system (Miller, 1972).

As a result, system behaviour theory defines a system as "interdependent components that cooperate to accomplish a common goal." (Smith, 2010). Several studies proposed the application of system behaviour theory to investigate accidents or incidents in different organisations and industries, operating complex systems to understand their risk management strategies (Rodríguez and Díaz, 2016; Leveson, 2002; Larsson et al., 2010). The need for hazard analysis techniques in system safety was highlighted by Larsson and colleagues (Larsson et al., 2010). Leveson and Stephanopoulos (2013) defined system theory differently as an approach that focuses on the behaviour of a process system in its entirety rather than on its constituent events. They argued that studying system behaviour theory should entail an examination of the system's components and the social, human, legislative, and regulatory frameworks that influence the system's behaviour. As a result, they concluded that applying a system behaviour theory may enable the development of novel types of risk and accident analysis.

2.4.5 Signal Detection Theory (SDT)

SDT is a theoretical modelling framework founded in the mathematical domain for analysing data. It has recently become popular in supporting the decision-making process, offering an approach to dissecting components of risks and uncertainties (Cohen et al., 2020; Lynn et al., 2018). The theoretical purpose of SDT is to understand the perceptual decision-making process in situations of uncertainty. In other words, the ability to distinguish signal or information from

the background noise or random patterns that distract from the information. Inferring what a person is thinking, feeling, or deciding involves perpetual uncertainty and behavioural risk, and such decisions can be described by SDT (Lynn and Barrett, 2014).

SDT serves as a model for optimal decision-making and analytical methods for evaluating decision-making performance on behaviour. SDT is a well-established analytical technique for describing decision-making behaviour in various perceptual and conceptual phenomena. It establishes a theoretical foundation for predicting or explaining behaviour. Therefore, SDT is a predictive technique for simulating the perceptual uncertainty and behavioural risk inherent in various decisions made in and out of the laboratory. Lynn and Barrett (2014) emphasized that the model enables the formulation of novel experimental questions about the computational processes underlying bias, sensitivity, and functional decision-making.

Many literature studies presented evidence of the use of SDT in different disciplines and sciences for risk assessment and decision-making tasks (Weil et al., 2018; Lusted, 1971; McFall and Treat, 1999). For example, audit procedures may not guarantee the freedom of financial statements from management fraud. Karim and Siegel (1998) applied SDT to detecting management fraud by independent auditors using signal detection theory and risk assessments to understand the association between various audit components and organisations' practices. Zhang and Maloney (2012) applied SDT to the probability distribution of the decision-making process under risk and uncertainty conditions to explain why probability distortion is a key factor in describing the peculiar characteristics of experience-based decision-making. Canfield et al. (2016) proposed SDT for measuring the risk of vulnerability to phishing attacks.

The application of SDT can improve decision-making quality, service quality and public health safety. Wagner et al. (2001) highlighted the extreme timeliness of disease detection as a vital requirement for public health surveillance systems. They proposed SDT and Decision theory for identifying and improving the timeliness of disease detection in SyS. However, this theory has been sparingly applied to influence the development of surveillance systems. The reason may be that few studies attempt to look at risk assessment and uncertainty measures under the public health surveillance practice.

2.4.6 Decision Theory

Decision theory is the study of decision-making analysis, and it has two interrelated facets, normative and descriptive. A normative Decision theory is how decisions should be made and are concerned with prescribing courses of action that conform most closely to the decision maker's beliefs and values. A descriptive theory describes how decisions are made, describing beliefs and values and incorporating them into the decision process (Slovic et al., 1977). Modern Decision theory has been around since the middle of the 20th century. It is now considered an academic subject typically pursued by researchers from different fields. Also, it has been widely applied to risk problems, assisting in answering research questions about acting when there is uncertainty and a lack of information.

Gaffney and Ulvila (2000) presented a technique to analyse intrusion detection systems and determine the optimum configuration using a decision analysis approach. Cienfuegos (2012) highlighted the usefulness of Decision theory in risk management and explained the implicit dependency of risk management on rules derived from general knowledge and precepts of Decision theory. Once a risk analysis has passed the assessment phase, a decision must be made about what to do with the identified outcome. Linville et al. (1993) introduced Decision theory to interpret HIV infection and transmission risks, analysing empirical data on decision biases in HIV risks.

Similarly, Decision theory was proposed as a tool for evaluating the effectiveness of public health interventions, applying it to cases where prior beliefs are sufficiently strong to allow decision-makers to assume the direction of change of the intervention's outcome in the context of the transparent and deliberative decision-making process (Fischer et al., 2013; Fischer and Ghelardi, 2016). Wagner et al. (2001) relied on the mathematical foundations of Decision theory and SDT to identify strategies for improving the early warning system timeliness. This is based on the Decision theory's ability to estimate the benefit of true alarms against the cost of false alarms to establish optimal sensitivity, specificity, and timeliness of the application.

2.4.7 Selected Theory for the Ongoing Research

This current research is concerned with providing a novel technique for integrating the QRA approach into VSyS to account for risk magnitudes, associated uncertainties, prior knowledge, and any variability due to the background noise of the input signals. The aim is to produce the probability distribution of outcomes with credible intervals to improve public health response

and aid decision-making. It would be appropriate to use SDT and Decision theory as the theoretical frameworks for this research work since previous studies indicate the ability of the SDT theory to support the decision-making process while explaining the components of risks and uncertainties in the data (Cohen et al., 2020; Lynn and Barrett, 2014; Lynn et al., 2018; Zhang and Maloney, 2012; Canfield et al., 2016).

Decision theory was used to explain risk management's dependency on situations where prior information might be essential to inform choices (Gaffney and Ulvila, 2000; Fischer et al., 2013; Gaube et al., 2019). However, the risk assessment process designed by Smith et al. (2017) relied on SDT and Decision theory for interpreting the risk of a disease outbreak from an alarm-based surveillance system. Their research focused on a qualitative risk assessment technique, so risk measurement was based on the qualitative knowledge of experts. While their techniques' outcomes can be simple to understand, the interpretations may be potentially complicated to implement, especially when the risks of the outbreak are detected close to the alarm threshold. Therefore, their approach appears tedious, unable to discriminate between high and low risks, and may not account for uncertainties and prior information.

2.5 Conclusion

Public health surveillance is a vital function of the public health practice, which provides essential tools for monitoring and evaluating disease outbreaks, interventions and their effects. On the other hand, surveillance risk assessment is an element of public health surveillance that evaluates an ongoing disease outbreak or potential outbreak to stimulate decision-making. Smith et al. (2017) highlighted an increase in the development of different public health surveillance techniques focusing on the efficiency of disease identification. However, little emphasis has been placed on improving how the public health surveillance system interprets alarms and assesses risks quantitatively with uncertainty measures.

There are different views about risk in the literature, but the view by Aven (2012) suggests that different disciplines need to identify suitable methods for assessing and managing risks in their domains. Therefore, the public health risk may be assessed better when the concepts of risk are incorporated as part of the overall methodology of implementing a disease surveillance system. The concepts identified in this chapter are signal, alarm, alert, uncertainty, risk assessment, and risk analysis. Uncertainty always accompanies risk assessment techniques. Therefore, managing uncertainty in risk evaluation is important to avoid issues in decision-making

processes that rely on such risk systems. However, it is worth noting that uncertainty measurement is not a common practice in veterinary disease surveillance. Therefore, posing a challenge for implementing VSyS techniques with risk evaluation components.

For this thesis, the author selected the SDT and Decision theory as the theoretical framework for explaining the quantitative risk assessment process in uncertain situations. Especially in a condition where the datasets might contain background noise with existing knowledge of the health incident and assumptions that may influence the decision-making process. Many research literature indicated the evidence of using SDT or Decision theory individually for research study design and interpretation, but only Smith et al. (2017) combined both theories in their work on qualitative risk assessment techniques for public health disease surveillance. In comparison, this current research work focuses on the quantitative risk assessment approach in veterinary syndromic surveillance.

Chapter 3 – Literature Review

3.1 Chapter Introduction

Technological advancement in animal production and healthcare activities has increased data generation and storage, and most of these data are generated for administrative and economic purposes. Consequently, the need to analyse the data has exponentially increased too. They are also used as passive surveillance models to develop various VSyS techniques. The goal is to monitor non-specific health indicators in real-time for early detection of disease outbreaks and more rapidly characterise them than the traditional notifiable disease methods. Therefore, VSyS can measure the temporal or spatial aberrations in the occurrence of health incidents using algorithms that triggers an alarm when the number of health incidents is above a set threshold of a target population. Unlike traditional disease surveillance, it is worth noting that VSyS does not focus on specific diseases as it is based on data collected for various routine activities before a diagnosis has been confirmed.

Recent research studies on VSyS concentrate on the development and efficiency of detection algorithms, and very little emphasis has been placed on evaluating the surveillance risk they tend to communicate. Many of these research studies demonstrated that different algorithms perform better with different temporal or spatial characteristics of the surveillance data, such as autocorrelations, seasonal trends, or day-of-week effects. Also, the performance of the algorithm can be influenced by different challenges that are peculiar to the datasets generated in veterinary medicine. Therefore, in this systematic review, the various VSyS data sources and techniques making use of animal health data for disease surveillance implementation are investigated to address the following research questions:

a. What are the routinely explored VSyS data sources, and what data modelling techniques do they implement? How do they interpret syndromic alarms and account for uncertainties or previous knowledge of outbreaks to inform decision-making?
b. What are some of the drawbacks to their use?
c. Is there any proof of current published research on using QRA techniques to assess syndromic alarms quantitatively?

By answering the research questions through this systematic literature review, the author aimed to present the current state of the art in the VSyS field and highlight the limitations of the

existing techniques. Furthermore, present the challenges accompanying the analysis of veterinary syndromic surveillance datasets and identify gaps in the literature.

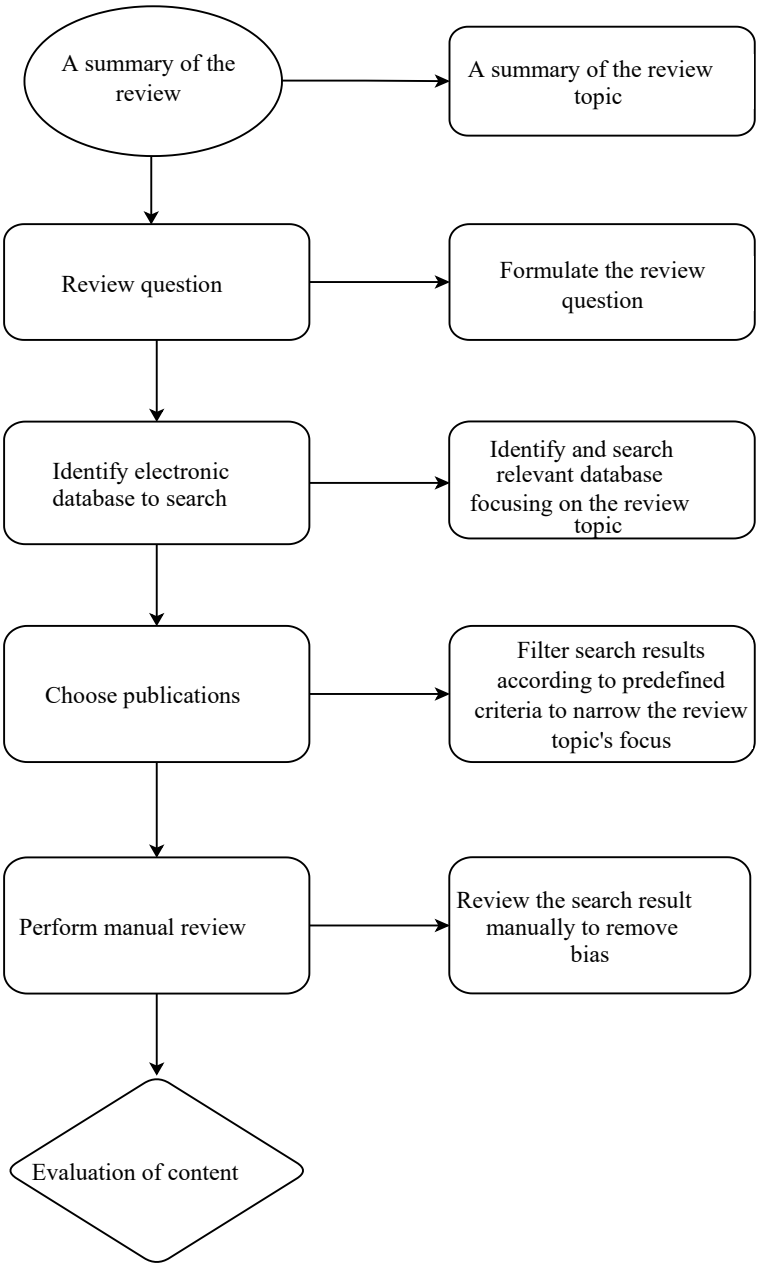

***Figure 4**: The systematic review flow chart*

3.2 Method

In this chapter, the author adopted the systematic approach of searching the academic database, investigating, and synthesising evidence from the literature while ensuring the process can be reproduced (Bettany-Saltikov, 2012). The adopted systematic review is presented in a flow chart diagram in Figure 4. The PRISMA systematic review search strategy was implemented, and the flow diagram of the search strategy is illustrated in Figure 5. It represents the procedure

for identifying relevant research studies that explore the objectives of this research work and help to answer the formulated research questions. The critical appraisal framework approach suggested by Moule and Hek (2011) was used to evaluate each selected research paper and synthesise evidence from the literature. The inclusion and exclusion criteria in section 3.2.2 were applied to select relevant studies for the literature review.

3.2.1 Electronic Literature Search

Scopus is one of the world-leading citation platforms and one of the largest single abstract and indexing databases (Burnham, 2006). It provides the capabilities to search research papers across different scientific disciplines and combines search results from the most popular electronic platforms (Zhu and Liu, 2020). According to Meester et al. (2017), Scopus is the primary research citation platform for top universities and research institutes worldwide. Since Scopus connect different digital databases, it was searched for peer-reviewed journals and articles relevant to the review questions in this work. The results were narrowed down to the most pertinent papers for answering the research questions within the allotted time frame for this project work using predefined filters. The flow chart for the systematic review framework is illustrated in Figure 4. Other processes of the framework are explained in the subsequent sections.

3.2.2 Filters for Inclusion and Exclusion

Before the systematic review was undertaken, it was decided that a time filter for the research publication must be considered. SyS was not in the mainstream academic research publications until 2001 because the initiatives were relatively new (Dupuy et al., 2013a). The uptake of the SyS research acquired momentum after the anthrax attack in 2001 and outbreaks of emerging infectious diseases, such as SARS in 2003. Although their aims are different from the aims of this ongoing literature review, Dorea and colleagues conducted extensive research on the SyS domain in 2011 and 2016, respectively, focusing on papers published from the early 2000s to 2016 on animal health surveillance (Dórea et al., 2011; Dórea and Vial, 2016).

Likewise, Egates et al. (2015) reviewed SyS initiatives to investigate routinely observed data for animal disease surveillance. Their research aims were different from the goal of this present research work. These authors classified the data sources into different data streams routinely collected for syndromic surveillance. Therefore, in this study, we decided that the final search result will be manually reviewed to exclude any publication that does not clearly explain the data source investigated – the purpose is to streamline further the focus of the search result

using the data stream classification in Dorea and their colleagues combined with Egates and colleagues. Also, it was decided that a filter to exclude all publications before 2010 be added to consider only literature published in the last ten years.

Other filters to exclude irrelevant publications include the following:

- The language filter was applied to select papers whose domain language is English.
- Filter for the inclusion of veterinary and animal domains was applied to ensure that literature on human disease surveillance was eliminated. Also, fish and amphibians were excluded.
- A filter was added to ensure that the search result only included peer-reviewed journals and articles.
- The search result was further filtered to remove any patents, blogs, books, or websites from the search result.
- Since the research involves a multidisciplinary approach, the author included many relevant subject domains in the search results. However, irrelevant domains, such as earth and planetary sciences, neuroscience, chemical engineering, health professions, materials science, pharmacology, toxicology, and pharmaceuticals, were filtered out.
- The disease surveillance method must be passive and not active to include a publication in the final search result, as explained in section 2.2.
- The author only considered publications with a clear explanation of data analysis techniques.

3.3 Result

The Scopus database search returned a total of 1317 publications initially. The search string evolved in Table 1 below from the initial search string in item 1 to the final search string in item 5, which returned 142 research publications. The final search string was achieved by applying various exclusion and inclusion filters described earlier. Then, each article was double-checked for bias prior to being manually reviewed. 24 publications were found to have failed the inclusion and exclusion criteria because they refer to disease outbreaks in human beings or are focused on fish and amphibians or non-relevant research domains such as referring to environmental research sciences or genomics and proteomics.

Each remaining publication was checked to establish that the data source and data analysis techniques were clearly explained. Also, the article was checked to establish that the data collection methodology is based on a passive technique described in chapter 2, section 2.2. 44 research publications were discarded due to failing to meet these inclusion criteria, and 9 research articles were excluded for reasons. 65 peer-reviewed research articles were finally selected for the literature review and are listed in Table 2 below. Also, Figure 5 shows the search processes for selecting the publications for the systematic literature review.

Table 1: *Search string evolution*

Item	Search Strings	Publication returned
1	“Syndromic Surveillance” OR “Disease Outbreak System” OR “Disease Detection System”	1,317
2	“Syndromic Surveillance” OR “Disease Outbreak System” OR “Disease Detection System” AND veterinary OR animal	178
3	“Syndromic Surveillance” OR “Disease Outbreak System” OR “Disease Detection System” AND veterinary OR animal AND PUBYEAR > 2009	162
4	"Syndromic Surveillance" OR "Disease Outbreak System" OR "Disease Detection System" AND veterinary OR animal AND PUBYEAR > 2009 AND (LIMIT-TO (SUBJAREA, "VETE") OR LIMIT-TO (SUBJAREA , "AGRI") OR LIMIT-TO (SUBJAREA, "MEDI") OR LIMIT-TO (SUBJAREA, "IMMU") OR LIMIT-TO (SUBJAREA, "BIOC") OR LIMIT-TO (SUBJAREA, "MULT") OR LIMIT-TO (SUBJAREA, "ENVI") OR LIMIT-TO (SUBJAREA, "COMP") OR LIMIT-TO (SUBJAREA, "SOCI") OR LIMIT-TO (SUBJAREA, "ENGI") OR LIMIT-TO (SUBJAREA, "DECI")) AND (LIMIT-TO (LANGUAGE, "English"))	153
5	"Syndromic Surveillance" OR "Disease Outbreak System" OR "Disease Detection System" AND veterinary OR animal AND PUBYEAR > 2009 AND (LIMIT-TO (SUBJAREA, "VETE") OR LIMIT-TO (SUBJAREA, "AGRI") OR LIMIT-TO (SUBJAREA, "MEDI") OR LIMIT-TO (SUBJAREA, "IMMU") OR LIMIT-TO (SUBJAREA, "BIOC") OR LIMIT-TO (SUBJAREA, "MULT") OR LIMIT-TO (SUBJAREA, "ENVI") OR LIMIT-TO (SUBJAREA, "COMP") OR LIMIT-TO (SUBJAREA, "SOCI") OR LIMIT-TO (SUBJAREA, "ENGI") OR LIMIT-TO (SUBJAREA, "DECI")) AND (LIMIT-TO (LANGUAGE, "English")) AND (LIMIT-TO (DOCTYPE, "ar") OR LIMIT-TO (DOCTYPE, "re"))	142

The selected articles for the literature review and the corresponding outcomes of their content analysis are listed in Table 2 below.

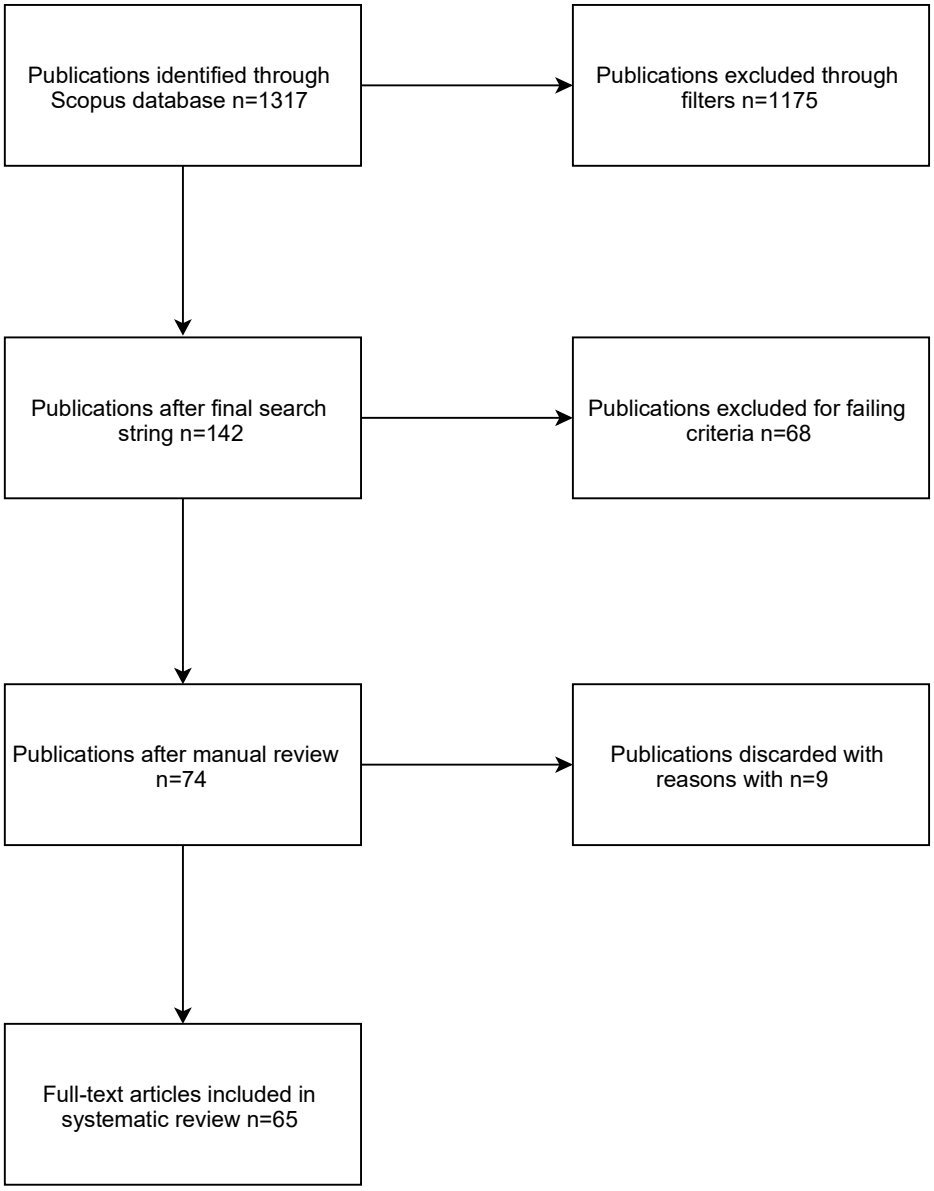

Figure 5: *The PRISMA flow diagram showing the systematic review search strategy applied in this study (Page et al., 2021)*

Table 2: *The articles selected for the ongoing literature review*

Data Stream Category	Authors	Title	Data Analysis Techniques	Risk Analysis / Baseline / Alarm Definition
Animal production and farm activities data stream (24 papers)	(Church et al., 2014)	Influence of environmental factors on infrared eye temperature measurements in cattle	Statistical test (comparison of means)	Data points above the baseline were used for alarm definition
	(Banakar et al., 2016)	An intelligent device for diagnosing avian diseases: Newcastle, infectious bronchitis, avian influenza	Support vector machine (SVM) classifier (classification)	None
	(Veldhuis et al., 2016)	Application of syndromic surveillance on routinely collected cattle reproduction and milk production data for the early detection of outbreaks of Bluetongue and Schmallenberg viruses	A linear mixed model was used to construct the baseline, then prediction with a CUSUM algorithm, and the harmonic linear regression model	Data points above the baseline were used for alarm definition
	(Marceau et al., 2014)	Can routinely recorded reproductive events be used as indicators of disease emergence in dairy cattle? An evaluation of 5 indicators during the emergence of bluetongue virus in France in 2007 and 2008	Logistic harmonic regression was used to predict and construct the baseline in the absence of a major epidemic	Data points above the baseline were used for alarm definition
	(Oyas et al., 2018)	Enhanced surveillance for Rift Valley Fever in livestock during El Niño rains and threat of RVF outbreak, Kenya, 2015-2016	Descriptive and statistical analysis (correlations and associations between data variables)	None
	(Madouasse et al., 2013)	Evaluation of a Continuous Indicator for Syndromic Surveillance through Simulation. Application to Vector Borne Disease Emergence Detection in Cattle Using Milk Yield	A linear mixed model for prediction (considers random and fixed effects)	Log-likelihood ratios (LLR) and cluster detection for alarm definition
	(Madouasse et al., 2014)	Use of monthly collected milk yields for the detection of the emergence of the 2007 French BTV epizootic	Linear mixed model prediction (considers random and fixed effects)	Log-likelihood ratios (LLR) and cluster detection for alarm definition
	(Bronner et al., 2015b)	Devising an indicator to detect mid-term abortions in dairy cattle: A first step towards syndromic surveillance of abortive diseases	Poisson regression model with over-dispersion	The method proposed by Farrington was used in the alarm definition
	(Nusinovici et al., 2016)	Quantification of the increase in the frequency of early calving associated with late exposure to bluetongue virus serotype 8 in dairy cows: Implications for syndromic surveillance	Mixed-logistic regression model	None

	(Mubamba et al., 2018)	Is syndromic data from rural poultry farmers a viable poultry disease reporting tool and means of identifying likely farmer responses to poultry disease incursion?	Logistic regression model	None
	(Bronner et al., 2015a)	Syndromic surveillance of abortions in beef cattle based on the prospective analysis of spatio-temporal variations of calvings	Poisson regression models with over-dispersion	The method by Farrington for baseline and LLR associated with clusters for alarm
	(Bronner et al., 2015c)	Was the French clinical surveillance system of bovine brucellosis influenced by the occurrence and surveillance of other abortive diseases?	Logistic regression model for establishing an association	None
	(Pfeiffer et al., 2021)	Using farmer observations for animal health syndromic surveillance: Participation and performance of an online enhanced passive surveillance system	Generalised linear model, and classification and regression tree (CART)	None
	(Veldhuis et al., 2020)	The Comparison of Three Statistical Models for Syndromic Surveillance in Cattle Using Milk Production Data	Harmonic linear regression for baseline construction. The CUSUM algorithm, Bayesian disease mapping model, and spatiotemporal cluster analysis were compared using the space-time scan statistic.	Deviation from the baseline was used for alarm definition.
	(Faverjon et al., 2019a)	Multivariate syndromic surveillance for cattle diseases: Epidemic simulation and algorithm performance evaluation	The baseline was constructed with Holt-Winters generalized exponential smoothing. Then, multivariate exponentially weighted moving average (MEWMA) and multivariate cumulative sum (MCUSUM) were compared	Deviation from the baseline was used for alarm definition
	(Struchen et al., 2015)	Investigating the potential of reported cattle mortality data in Switzerland for syndromic surveillance	Poisson, negative-binomial regression, and ARMA models	None
	(Torres et al., 2015)	Syndromic surveillance system based on near real-time cattle mortality monitoring	Baseline was constructed, univariate cycling regression model and CUSUM algorithm.	Two types of risk signals were estimated: point risk signals and cumulative risk signals using Control Charts. Data points above the baseline are used for alarm definition
	(Perrin et al., 2010)	Using the National Cattle Register to estimate the excess mortality during an epidemic: Application to an outbreak of Bluetongue serotype 8	Poisson regression model with over-dispersion for prediction	Deviation from the baseline was used for alarm definition
	(Antunes et al., 2017)	Mortality in Danish Swine herds: Spatio-temporal clusters and risk factors	Logistic regression models were used to assess association, and the Bernoulli model for local cluster identifications	None

	(Struchen et al., 2017)	Value of evidence from syndromic surveillance with cumulative evidence from multiple data streams with delayed reporting	Negative binomial (NB) distributions using the Bayesian framework to construct baseline and outbreak signals	The Bayesian likelihood ratio for competing evidence under outbreak
	(Alba-Casals et al., 2015)	Development of new strategies to model bovine fallen stock data from large and small subpopulations for syndromic surveillance use	Auto-regressive integrated moving average (ARIMA) model for mortality patterns and the baseline associated with mortality	Baseline comparisons
	(Fernández-Fontelo et al., 2020)	Enhancing the monitoring of fallen stock at different hierarchical administrative levels: An illustration on dairy cattle from regions with distinct husbandry, demographical and climate traits	Hierarchical time-series and ARIMA for prediction of mortality	Data points above the baseline were used for alarm definition
	(Sala et al., 2020)	Designing a Syndromic Bovine Mortality Surveillance System: Lessons Learned From the 1-Year Test of the French OMAR Alert Tool	General linear model (GLM) (beginning with Poison, Quasi-Poisson GLM and NB). Then EWMA chart, CUSUM chart, Shewhart chart, Holt-Winters, and Historical limits algorithms	Data points above the baseline were used for alarm definition
	(Fernández-Fontelo et al., 2017)	Enhancing syndromic surveillance for fallen dairy cattle: Modelling and detecting mortality peaks at different administrative levels	The classical ARIMA for predicting mortality	Data points above the baseline were used for alarm definition
Slaughterhouse, abattoirs, and meat inspection data stream (12 papers)	(Dupuy et al., 2013b)	Defining syndromes using cattle meat inspection data for syndromic surveillance purposes: A statistical approach with the 2005-2010 data from ten French slaughterhouses	Multiple factor analysis (MFA) combined with K-means and hierarchical ascendant clustering (HAC) were used to establish groups.	None
	(Alton et al., 2010)	Factors associated with whole carcass condemnation rates in provincially-inspected abattoirs in Ontario 2001-2007: Implications for food animal syndromic surveillance	GLM such as Poison and NB regression model links were used in this work	None
	(Vial and Reist, 2014)	Evaluation of Swiss slaughterhouse data for integration in a syndromic surveillance system	ARMA and statistical tests were used to investigate the association	None
	(Dupuy et al., 2015)	Pilot simulation study using meat inspection data for syndromic surveillance: use of whole carcass condemnation of adult cattle to assess the performance of several algorithms for outbreak detection	Baseline outbreak constructed with regression models (Poisson and NB regressions). Then, the Shewhart chart, EWMA, and CUSUM	Data points above the baseline are used for alarm definition
	(Dupuy et al., 2014)	Factors associated with offal, partial and whole carcass condemnation in ten French cattle slaughterhouses	Multivariable multinomial logistic regression analysis	None
	(Alton et al., 2012)	Suitability of bovine portion condemnations at provincially-inspected abattoirs in Ontario Canada for food animal syndromic surveillance	Multilevel Poisson regression modelling was used to evaluate the statistically significant association	None
	(Thomas-Bachli et al., 2014)	Exploring relationships between whole carcass condemnation abattoir data, non-disease factors and disease outbreaks in swine herds in Ontario (2001-2007)	Mixed-effects Poisson model and random-effects NB model were constructed	None

	(Alton et al., 2013)	Comparison of covariate adjustment methods using space-time scan statistics for food animal syndromic surveillance	Space-time scan statistic - (1) animal class adjusted Poisson scan statistic, (2) space-time permutation, (3) multi-level model adjusted Poisson scan statistic, and (4) a weighted normal scan statistic using model residuals	None
	(Alton et al., 2015)	Suitability of sentinel abattoirs for syndromic surveillance using provincially inspected bovine abattoir condemnation data	Multi-level NB regression models were used to construct the rate of condemnation within animal classes	None
	(Vial and Reist, 2015)	Comparison of whole carcass condemnation and partial carcass condemnation data for integration in a national syndromic surveillance system: The Swiss experience	A Poisson model, NB, and zero-inflated NB were used	None
	(Haredasht et al., 2018)	Characterization of the Temporal Trends in the Rate of Cattle Carcass Condemnations in the US and Dynamic Modeling of the Condemnation Reasons in California With a Seasonal Component	Dynamic harmonic regression (DHR) model to predict the rate of carcass condemnations.	None
	(Vial et al., 2015)	A simulation study on the statistical monitoring of condemnation rates from slaughterhouses for syndromic surveillance	Quasi-Poisson regression	Baseline according to Farrington algorithm. Alarm definition by data points above confidence limit and Z score
Laboratory diagnostics data stream (12 papers)	(Dórea et al., 2013a)	Syndromic surveillance using veterinary laboratory data: data pre-processing and algorithm performance evaluation	a Poisson regression model for baseline construction. Compare the effectiveness of (i) Shewhart charts, (ii) CUSUM, and (iii) EWMA for outbreak detections.	Data points above the baseline were used for alarm definition
	(Dórea et al., 2013b)	Exploratory Analysis of Methods for Automated Classification of Laboratory Test Orders into Syndromic Groups in Veterinary Medicine	Natural language processing (NLP), Naïve Bayes classifier, and Decisions trees were used to extract syndromic information from laboratory test requests	None
	(Dórea et al., 2013c)	Retrospective time-series analysis of veterinary laboratory data: Preparing a historical baseline for cluster detection in syndromic surveillance	Fitted Poisson regression model to construct outbreak baselines	Data points above the baseline were used for alarm definition
	(Dórea et al., 2014)	Syndromic surveillance using laboratory test requests: A practical guide informed by experience with two systems	A Poisson regression model was fit to construct the baseline. CUSUM, EWMA, Shewhart charts, and Holt-Winters exponential smoothing were used for outbreak detection performance.	Data points above the baseline were used for alarm definition
	(Warns-Petit et al., 2010)	Unsupervised clustering of wildlife necropsy data for syndromic surveillance	Multiple correspondence analysis (MCA) was used to reduce data to their principal components. Then, HAC and K-Means algorithms were used for clustering.	None

	(Poskin et al., 2016)	Reconstruction of the Schmallenberg virus epidemic in Belgium: Complementary use of disease surveillance approaches	Statistical tests - two-sided Fischer exact tests	None
	(McFadden et al., 2016)	Monitoring an epidemic of Theileria-associated bovine anaemia (Ikeda) in cattle herds in New Zealand	Linear regression model	None
	(Vial et al., 2016)	Methodological challenges to multivariate syndromic surveillance: a case study using Swiss animal health data	Farrington algorithm and two-component model based on Poisson and NB distribution	Data points above the baseline were used for alarm definition
	(O'Sullivan et al., 2012)	Identifying an outbreak of a novel swine disease using test requests for porcine reproductive and respiratory syndrome as a syndromic surveillance tool	Logistic regression using a GLM	None
	(Dórea et al., 2015)	Vetsyn: An R package for veterinary syndromic surveillance	GLM regression for baseline construction, Shewhart, CUSUM, EWMA, Holt-Winters was used for the detection of the outbreak	Data points above the baseline were used for alarm definition
	(Bollig et al., 2020)	Machine learning for syndromic surveillance using veterinary necropsy reports	Compared the following for information extraction and text classification; logistic regression, SVM, CART, Bagging Trees, Random Forest, Gradient tree boosting, and Recurrent Neural Network.	None
	(Burkom et al., 2019)	Equine syndromic surveillance in Colorado using veterinary laboratory testing order data	EARS C3, modified C2, CUSUM, and EWMA charts	Data points above the baseline were used for alarm definition
Clinical data stream (9 papers)	(Amezcua et al., 2010)	Evaluation of a veterinary-based syndromic surveillance system implemented for swine	Mixed effect logistic and linear regression models	None
	(Ruple-Czerniak et al., 2013)	Using Syndromic Surveillance to Estimate Baseline Rates for Healthcare-Associated Infections in Critical Care Units of Small Animal Referral Hospitals	Random effects logistic regression model and Poisson regression	None
	(Ruple-Czerniak et al., 2014)	Syndromic surveillance for evaluating the occurrence of healthcare-associated infections in equine hospitals	Random effects logistic regression model and Poisson regression	None
	(Zurbrigg and Van den Borre, 2013)	Factors associated with good compliance and long-term sustainability in a practitioner-based livestock disease surveillance system	Non-parametric statistical test method - analysis of variance	None
	(Ana et al., 2017)	Syndromic surveillance for West Nile virus using raptors in rehabilitation	A linear regression model and ARMA	None
	(Hale et al., 2019)	A real-time spatio-temporal syndromic surveillance system with application to small companion animals	Spatio-temporal mixed-effects regression model and Bayesian predictive inference using an MCMC algorithm	Bayesian predictive distribution for alarm definition

	(Behaeghel et al., 2015)	Evaluation of a hierarchical ascendant clustering process implemented in a veterinary syndromic surveillance system	HAC algorithm	None
	(Weng et al., 2020)	Evaluation of a novel syndromic surveillance system for the detection of the 2007 melamine-related nephrotoxicosis foodborne outbreak in dogs and cats in the United States	Proportionate diagnostic outcome ratio (PDOR) algorithm	Data points above the baseline were used for alarm definition
	(Hedell et al., 2019)	Surveillance of animal diseases through implementation of a Bayesian spatio-temporal model: A simulation example with neurological syndromes in horses and West Nile Virus	A Bayesian model with a hidden Markov model using Gibbs sampling	The alarm was defined by probability distribution using the Bayesian framework
Online and social media data stream (3 papers)	(Guernier et al., 2016)	Use of big data in the surveillance of veterinary diseases: early detection of tick paralysis in companion animals	Statistical analysis using Spearman's rank correlations and time-series cross-correlations	None
	(Yousefinaghani et al., 2019)	The Assessment of Twitter's Potential for Outbreak Detection: Avian Influenza Case Study	Semi-supervised Naive Bayes and Expectation-Maximization (EM) models	None
	(Robertson and Yee, 2016)	Avian Influenza Risk Surveillance in North America with Online Media	CUSUM technique and a static threshold modelling approach	The baseline was developed with a static 95% CI
Multi data streams (5 papers)	(Faverjon et al., 2016)	Evaluation of a Multivariate Syndromic Surveillance System for West Nile Virus	Generalised linear modelling: fitted regression models based on Poisson and NB distributions. Bayesian framework for combining data from multiple streams	The baseline was constructed in a Frequentist approach—defined alarms with Bayesian hypothesis testing.
	(Andersson et al., 2014)	Using Bayes' Rule to Define the Value of Evidence from Syndromic Surveillance	A general dynamic Poisson model, NB distribution or Poisson-log-normal (PLN) distribution using the GLM framework and Bayesian likelihood ratio framework.	The baseline was constructed in a Frequentist approach. Defined alarm with the likelihood ratios in Bayesian hypothesis testing
	(Amezcua et al., 2013)	Comparison of disease trends in the Ontario swine population using active practitioner-based surveillance and passive laboratory-based surveillance (2007–2009)	NB model	None
	(Cazeau et al., 2019)	Utility of examining fallen stock data to monitor health-related events in equids: Application to an outbreak of West Nile Virus in France in 2015	NB regression model, Shewhart chart, CUSUM, and EWMA	None
	(Tongue et al., 2020)	Improving the Utility of Voluntary Ovine Fallen Stock Collection and Laboratory Diagnostic Submission Data for Animal Health Surveillance Purposes: A Development Cycle	Generalized linear models according to Farrington and colleagues, CUSUM, and NB	Data points above the baseline were used for alarm definition

Furthermore, the Scopus database was searched for existing literature reviews on VSyS and routinely explored data sources using the query in Table 3 to identify previous systematic and literature reviews on the subject. After applying various exclusion and inclusion filters described in section 3.3, the final search returned 4 publications. Then, manual analysis was conducted, which resulted in 1 peer-reviewed publication – Dórea et al. (2011) is the only publication that met the predefined criteria.

However, 2 other publications, Dórea and Vial (2016) and Egates et al. (2015) were identified but not captured by the keyword search. Both papers are referenced in many publications on veterinary syndromic surveillance. Therefore, these 2 new publications and the selected paper were incorporated into this ongoing review work since they focus on previous efforts to classify routinely observed data sources for disease outbreak detection initiatives.

***Table 3**: Search string evolution*

Item	Search Strings	Publication returned
1	"Syndromic Surveillance" OR "Disease Outbreak System" OR "Disease Detection System" AND veterinary OR animal AND "Data source" OR "Data Streams" OR "Data Stream"	17
2	"Syndromic Surveillance" OR "Disease Outbreak System" OR "Disease Detection System" AND veterinary OR animal AND "Data source" OR "Data Streams" OR "Data Stream" AND review	6
3	"Syndromic Surveillance" OR "Disease Outbreak System" OR "Disease Detection System" AND veterinary OR animal AND "Data source" OR "Data Streams" OR "Data Stream" AND review AND PUBYEAR > 2009 AND LIMIT-TO (PUBSTAGE, "final")) AND (LIMIT-TO (LANGUAGE, "English"))	4

***Table 4**: The additional publications selected for the ongoing literature review*

S/N	Authors	Titles	Type
1	(Dórea et al., 2011)	Veterinary syndromic surveillance: current initiatives and potential for development	Literature Review
2	(Dórea and Vial, 2016)	Animal health syndromic surveillance: a systematic literature review of the progress in the last 5 years (2011–2016)	Literature Review
3	(Egates et al., 2015)	Integrating novel data streams to support biosurveillance in commercial livestock production systems in developed countries: challenges and opportunities	Literature Review

3.4 Routinely Explored VSyS Data Sources

The author investigated different data sources from the selected publications in this section. The aim is to identify routinely collected and analysed data streams for VSyS development. Furthermore, each identified data source was critically evaluated to establish the challenges associated with their usage for VSyS development and gaps that still need to be filled.

3.4.1 Animal Production and Farm Activities Data Stream

As part of the daily animal husbandry, data is continuously generated using computer vision techniques, IoT devices and other equipment that measures animals' health parameters and various aspects of productivity (Banakar et al., 2016; Church et al., 2014). These data are generalised as production datasets. Dórea et al. (2011) classified data from this category as herd management datasets. Also, they grouped other routinely explored data sources for VSyS development into the clinical, laboratory, auction markets, and abattoir data sources. However, they argued that scientists regularly explore clinical and laboratory data more than other data sources for VSyS development.

According to Dórea and Vial (2016), animal production data provide a high level of population coverage and the shortest period between a health incident and its possible identification. In the evidence extracted for the ongoing systematic review in Table 2, twenty-four research publications were found to have focused on using animal production datasets for VSyS development. This indicates that the animal production dataset is more explored for VSyS development than other data sources, as publications on the animal production data represent 37% of the total data extracted in Table 2. This observation is contrary to the view held by Dórea et al. (2011).

Marceau et al. (2014) extracted and classified five syndromic indicators to demonstrate a retrospective outbreak of a health event from routinely recorded reproductive datasets on the farm. Each indicator's timeliness and specificity were highly variable and low. From a different perspective, Struchen et al. (2017) argued that data generated on the farm could be exposed to delayed reporting due to the remoteness of farms and the lack of access to the right device to facilitate the direct upload of data to the central processing system even when the data is captured on time. They argued that reports collected on the farm are completed on papers or devices that store data locally, which may require the internet to transfer data into a central database system later. The authors aimed to improve VSyS timeliness and sensitivity, so they

demonstrated their approach using a historical dataset of routinely collected on-the-farm cattle mortality events and incorporated a data analysis technique to account for "delays" in reporting.

Veldhuis et al. (2016) analysed data collected on cattle's reproductive events and milk yield to evaluate the timeliness and sensitivity of a VSyS approach. The study did not result in better sensitivity or timely detection than the traditional surveillance method. As a result, these authors recommended that it might be useful to combine regular surveillance datasets with non-specific animal health data in real-time to improve the timely detection of disease and provide decision-makers with early insights into health events.

Another data source group that falls under the animal production category is animal movement data. Egates et al. (2015) presented a case for the market surveillance data source and its practicality for VSyS development. They argued that since animals are continuously inspected for diseases and certified by veterinarians as they move from farms to sales points, such data could be implemented in the VSyS system for disease outbreak intelligence. The movement of animals from farms to livestock markets was believed to have increased the foot and mouth disease transmission rate in 2001 (Office NA, 2002).

3.4.2 Clinical Data Stream

Direct reporting of health events by practitioners/veterinarians is the primary data contribution to VSyS from clinics and veterinary practices. Usually, veterinarians identify unusual health events or notifiable diseases and report the occurrence by submitting data to disease notification systems (Behaeghel et al., 2015). This method of data contribution is commonly used in the traditional disease surveillance system. However, to collect non-specific datasets in the clinical environments for application in syndromic surveillance, some research communities and private institutions interested in animal health are coordinating a network of clinics to submit daily health visits of animals and test results to a central system for disease surveillance and research purposes (Fernando et al., 2017; SAVSNET, 2019).

However, the major challenge with this data source is sustainability, as many VSyS relying on clinical-based datasets have to provide some form of incentives to veterinary clinics and practices for data transmission (Struchen et al., 2016). Also, many research papers pointed out that most veterinary surveillance platforms provide various incentives to entice veterinarians to participate in data contribution, and the offer ranges from direct reimbursements per a report

submitted to a diagnostic laboratory to free reports on veterinary research and credits for research seminars (SAVSNET, 2019; Sánchez-Vizcaíno et al., 2017). The reason may be because clinical data collection and submission depend on the veterinarian's ability to cooperate and make the extra effort to submit their routine data to a surveillance system.

A major issue with using the clinical data source is the lack of data standards and syndromic classifications between different practice systems recording the animal health data due to a lack of standardised collection framework in the veterinary sector. For example, Singleton et al. (2018) accessed over a million records of animal patient data from SAVSNET. They proposed a text mining approach based on a regular expression technique to identify and classify pharmaceutical agents frequently prescribed to companion animals to determine their efficacy and risk factors. It was noted that different veterinary premises adopted different vocabularies to describe pharmaceutical agents in non-standardised ways. They concluded that their approach's limitation is due to the lack of centrally agreed terms and classification systems for pharmaceutical agents in the veterinary sector, even though human medicine currently benefits from using such standardised terminologies and classification.

Also, Egates et al. (2015) concluded that veterinary systems lack a standardised mechanism for recording animal health events, which normally results in collecting different variables by different veterinary systems or different definitions for the same set of variables. Although many studies still manage to use the data from clinical sources without serious challenges. For example, Hale et al. (2019) recently accessed over a million records of animal patient data from SAVSNET. Similarly, Hedell et al. (2019) obtained the time-series counts of a syndromic dataset of specific diseases submitted by veterinarians to the French network (RESPE) for tracking equine diseases. It may be because the approaches adopted by these studies do not require data to be standardised, and they are not dealing with text classification problems like those encountered by other studies.

3.4.3 Laboratory Diagnostic Data Stream

Laboratory datasets are frequently collected for syndromic surveillance purposes, and this is because they have higher specificity for disease confirmation than clinical and animal production datasets (Egates et al., 2015). Figure 1 explains the relative specificity and timeliness of the laboratory data sources compared to other veterinary data sources. Burkom et al. (2019) also highlighted the laboratory data stream as a useful source of historical and

structured datasets, often available in electronic format to facilitate syndromic classification. However, they underlined the difficulty of using laboratory data sources for solving problems related to data standardisation. Vanderwaal et al. (2017) views this argument the same way. They argued that laboratory datasets usually fail to integrate seamlessly when they are obtained from multiple sources due to disparate data formats and schemas between different systems that generate the data. Similarly, Vial et al. (2016) argued that the lack of commonly adopted data standards in the veterinary sector makes data integration across heterogeneous databases difficult.

Also, the laboratory data stream can be poor in timeliness compared to the clinical-based and animal production datasets (Egates et al., 2015); this is because an animal patient is likely to interface with a surgeon in the clinic or with a practitioner during a visit to a farm where data about a health event may be collected before samples are referred to the laboratory for further investigation. Although the clinic-based and laboratory-based data streams are less captured than animal production datasets, the latter most often relies on automated farm devices for data collection (Church et al., 2014). Also, the population coverage of laboratory submission data can be influenced by practitioner perceptions and experience and the animal's financial and epidemiology state (Egates et al., 2015).

3.4.4 Online and Social Media Data Stream

Three publications were identified from the systematic review under the online media data stream focusing on VSyS implementation. This trend supports the view that the use of online media datasets in veterinary medicine is still in the early stages. Guernier et al. (2016) used Google Correlate to suggest internet search terms relevant to implementing a VSyS strategy in small companion animals, and Google Trends was used to download corresponding search frequency metrics. Their study shows that internet search indicators can monitor health incidents in companion animals, which may help detect outbreaks. As a result, they concluded that online media is useful as complementary data sources for implementing a long-term VSyS strategy.

Yousefinaghani et al. (2019) developed a Twitter-based data analysis framework to automatically monitor disease outbreaks in real-time using tweets from 2017 to 2018, extracting information associated with disease onset and outbreaks in several countries. They validated their approach against the results of a traditional method. They realised that more

than one-third of the real epidemic was reported on Twitter earlier than the official outbreak, confirming the usefulness of online media datasets if the proper technique is applied.

Likewise, Robertson and Yee (2016) investigated the outbreak of avian influenzas using a set of methods that analyses data as it arrived, rather than a purely retrospective analysis of the dataset. The authors' method included an NLP for processing Twitter datasets, and they evaluated a static threshold method against a cumulative-sum dynamic threshold technique. The most significant difficulty highlighted in the last two studies above is the quality of information extracted due to noise. Furthermore, finding useful information within this data stream could be challenging as online media datasets usually contain a high volume of noise, frequently leading to an investigation of many false alarms. However, they concluded that the low cost of developing and maintaining a VSyS that consumes online media datasets could make adopting such a system more attractive.

Also, Yousefinaghani et al. (2019) found that it is difficult to determine whether a tweet is appropriate or not due to the limited contextual information available in short texts. Therefore, they concluded that while their approach helps filter noisy signals from the Twitter data, adding information that further describes non-contextual tweets might increase accuracy.

3.4.5 Slaughterhouse, Abattoirs, and Meat Inspection Data Stream

Many recent studies demonstrated the use of datasets originating from slaughterhouses, abattoirs, and meat inspections for implementing disease surveillance systems with different techniques targeting the improvement of VSyS sensitivity. Haredasht et al. (2018) modelled and predicted the time-series carcass condemnation rate in different time zones. They argued that slaughterhouse condemnation records and slaughter data could support animal and public health surveillance strategies in real-time. The authors' conclusions contradicted the work of both Dórea and Vial (2016) and Egates et al. (2015), who argued that mortality and slaughterhouse data lack specificity and timeliness. As a result, their potential for early detection of diseases remains unclear.

Vial et al. (2015) developed an improved disease outbreak detection system using datasets from post-mortem inspection of slaughtered animals. They discovered that the developed model performed well with large datasets but poorly with small datasets. Therefore, they concluded

that integrating more data streams and simultaneously evaluating more data from other sources could increase VSyS sensitivity and timeliness.

3.4.6 Multi Data Stream

Dórea et al. (2011) highlighted the frequently investigated data sources for VSyS implementation as herd management datasets, clinical sources, laboratory data sources, auction markets, and abattoirs. On the other hand, Egates et al. (2015) reviewed commercial livestock farming and categorised routinely explored data sources as animal production, veterinary clinic, laboratory diagnostic, market surveillance, and slaughter inspection data sources. Similarly, Dórea and Vial (2016) revised the literature review by Dórea et al. (2011) and added the online media data source. Combining these three literature reviews provided the foundation for categorising identified publications into data streams in this research study.

However, in this ongoing study, the author observed that five publications could not fit into any of the groups identified by previous authors because the publications implemented the VSyS approach with multiple data streams. Table 5 shows the list of publications with the identified multiple data streams used for VSyS development. This observation demonstrates the appetite of recent research studies to explore multi-data streams for VSyS implementation because sub-populations are often affected differently by a wide variety of signs and symptoms of a disease. For example, abortion only affects breeding-age female animals, so it would not be sufficient to monitor a single syndromic time series targeted on abortion to identify factors responsible for diarrhoea in an adult cow.

To support the recent interest of researchers in the multi-data stream approach, Veldhuis et al. (2016) argued that combining regular surveillance datasets with non-specific animal health data in real-time might improve rapid syndromic surveillance analysis and provide decision-makers with early insights into disease outbreaks and trends. The authors attempted to demonstrate the multi-data stream approach in their research work by combining the reproduction and milk yield datasets to build a VSyS technique. However, according to the adopted data stream classifications from previous research studies, both datasets belong to the animal production data stream and failed to demonstrate a multi-data stream approach.

***Table 5**: Literature publications with multiple data streams*

S/N	Authors	Title	Data Sources
1	(Faverjon et al., 2016)	Evaluation of a Multivariate Syndromic Surveillance System for West Nile Virus	Clinical, laboratory, and Mortality data from horses and birds
2	(Andersson et al., 2014)	Using Bayes' Rule to Define the Value of Evidence from Syndromic Surveillance	Clinical and Laboratory
3	(Amezcua et al., 2013)	Comparison of disease trends in the Ontario swine population using active practitioner-based surveillance and passive laboratory-based surveillance (2007–2009)	Clinical and Laboratory
4	(Cazeau et al., 2019)	Utility of examining fallen stock data to monitor health-related events in equids: Application to an outbreak of West Nile Virus in France in 2015	Mortality, Clinical, and Laboratory
5	(Tongue et al., 2020)	Improving the Utility of Voluntary Ovine Fallen Stock Collection and Laboratory Diagnostic Submission Data for Animal Health Surveillance Purposes: A Development Cycle	Mortality and Laboratory

3.5 Existing Data Analysis Techniques in VSyS

In this section, the author reviewed the current techniques employed in VSyS and their corresponding data modelling approach for constructing disease outbreak systems and generating alarms to indicate a health event has occurred. Furthermore, to identify how the existing VSyS approach accounts for; (a) associated uncertainties in the data stream or models and (b) account for existing knowledge of the health event under investigation to inform decision-making and communicate the risk of public health concerns. In particular, how the existing methods assess and quantify the generated alarms for decision making. Also, the author established the potential gaps in knowledge that are not yet covered in previous works.

From 65 previously identified publications for this ongoing literature study, 6 applied machine learning algorithms to solve the clustering and classification of disease outbreak problems. 54 publications applied the Frequentist approach in modelling VSyS, in which 25 of the publications constructed outbreak baseline techniques as part of their overall methodology. 5 publications implemented the Bayesian framework approach to solve disease outbreak detection problems, of which only 2 publications based their technique on the Bayesian probability distribution or Bayesian inference approach.

Regardless of the data stream under consideration, VSyS data analysis techniques appear to follow nearly similar steps in most reviewed publications on veterinary disease surveillance.

Also, the author focused on the overall approach of identifying how each publication addressed the associated uncertainties in the data source or the developed surveillance models and handling of previous knowledge about a health event such as the outbreak distribution, the influence of seasons, days of the week or other covariates to inform decision making. Especially when an event occurs very close to the baseline or when an alarm is generated near the threshold. Understanding these techniques will provide the foundation for identifying how existing studies handle risk assessment of disease outbreak signals quantitatively or qualitatively.

3.5.1 Frequentist Statistical Approach

For a standard VSyS approach, the first step focuses on retrospective evaluation of the dataset to uncover any spatial or temporal effects and patterns or identify covariates' influence that needs to be considered in later analysis. This step provides the VSyS model with expected values for each data point at t, as indicated in equation 3.1.

$$Y_t \sim distr\ (\mu_t, \sigma_t^2) \qquad (3.1)$$

where $\mu_t\ and\ \sigma_t^2$ in the equation are the mean and variance that produced a typical background behaviour against which statistical deviation from a baseline can be evaluated. In this step, the Frequentist statistical approach is often used, and the data must be free from any outbreak signals to create a baseline from historical datasets representing a non-outbreak period and behaviour. The commonly used method is the parametric regression model (Vial et al., 2016; Faverjon et al., 2016; Veldhuis et al., 2016).

From the evidence in Table 2, 25 research publications implemented their VSyS techniques by constructing a baseline/threshold for alarms as part of their overall methodology. 22 articles are fundamentally from the Frequentist background, while the other 3 articles implemented the Bayesian methodology. However, they all used one of these baseline methods to construct outbreak thresholds.

1. They applied the entire period of data available to fit a regression model and replaced values above a specific confidence interval (CI), usually 95% CI, with a selected cut-off value. This method was originally proposed by Tsui et al. (2001), and it assumed that observed data points greater than the 95% confidence interval of a model prediction reflect data from an epidemic outbreak.

2. Alternatively, by weighting observations using the inverse of their residuals, this approach reduced the contribution of observations that deviate greatly from their expected values to the model as initially proposed in Farrington's technique (Farrington et al., 1996).

However, the author observed that previous research had to find means of handling temporal covariates such as seasons or days of the week or other known or unknown factors that may substantially affect baseline constructions depending on the objectives of their research work.

Consequently, in the literature examined, the choice of regression technique for dealing with covariates depends on the type of expected data, the extent of the influence of random variables on the model and whether there is a need to account for mixed-effects and fixed effects due to variance changes of individual heterogeneity. For example, Poisson regression or NB models are selected for count data, while linear regression models are applied to continuous data (Struchen et al., 2015; Dupuy et al., 2015).

Veldhuis et al. (2016) proposed the mixed-effect regression technique to handle the presence of random and fixed effects in VSyS modelling. Likewise, other research studies proposed the logistic regression model with either mixed effects or random effect coefficients to model the probability of certain classes or outbreaks of events while accounting for variance changes in individual heterogeneity of each class (Nusinovici et al., 2016; Amezcua et al., 2010; Ruple-Czerniak et al., 2013). The trend in using mixed-effects or random-effects modelling techniques might demonstrate the attempt to tackle challenges faced in mining VSyS data streams due to the stochasticity of surveillance datasets (Faverjon et al., 2019a; Dórea et al., 2019).

In the second step, most publications focused on the Frequentist approach and built the model for VSyS at this stage. In step three, the model is evaluated and tested. The observed value at time t is evaluated against a baseline derived in step 1 to establish whether an alarm should be triggered or not. Researchers frequently monitor data prospectively to detect a possible deviation from the baseline model in real-time or perform predictions to detect any future unexpected deviation from the threshold model, indicating the occurrence of health events or clusters representing possible outbreak signals in time or space-time.

Fernández-Fontelo et al. (2017) analysed the mortality records combined with spatial-temporal datasets using the classical AutoRegressive Integrated Moving Average (ARIMA) technique.

Alba-Casals et al. (2015) also implemented the ARIMA model for investigating mortality patterns and rates; this statistical technique can account for the dependent variable and random variables: in this case, the mortality rate trend and seasonality. Then again, Perrin et al. (2015) applied Scan Statistics to mortality data to identify an abnormal increase. The alarm generation in these studies does not explain how the models discriminate between high and low data points in the outcomes (Veldhuis et al., 2016). Though they presented techniques that relied on deviation from the baseline to generate alarms, the magnitude of the events occurring near the baseline might be difficult to interpret for decision-makers.

Marceau et al. (2014) calculated expected baselines for time-series datasets, and the model parameters were estimated using the maximum likelihood technique. The expected upper confidence limit of each parameter was expressed in relative deviation to avoid frequent false alarms in the model. However, the authors noted that this technique was only effective for detecting deviations that reached a large magnitude. Therefore, they concluded that the approach was not appropriate for detecting small elevations that might occur in the early phase of an epidemic. Veldhuis et al. (2020) compared the CUSUM chart with the Bayesian method; the CUSUM method investigated the cumulative sum of differences between observed and predicted milk. The method could only adjust the mean and variance in the model and did not differentiate between the alarms within the threshold boundaries. The method proposed by Veldhuis et al. (2016) generated a high number of false alarms when the threshold value was lowered and few alarms when the threshold was increased, indicating that the handling of the alarm generation may not provide clear explanations for data points closer or far from the baseline, which may lead to a false interpretation of outbreaks.

Also, many data streams are subjected to different external factors that may not be related to disease outbreaks, such as environmental, welfare, and economic factors. Using the Frequentist approach, these factors may be difficult to account for in the model. The measure of residual or point estimate techniques, such as maximum likelihood or confidence limit, which are popularly used in the Frequentist approach, might find uncertainty estimation difficult since they lack uncertainty measures (Salvatier et al., 2016).

In equation 3.1, the Poisson regression model assumes that the mean of the distribution of observed counts and its variance are equal (Mouatassim and Ezzahid, 2012). Based on this assumption, Poisson's regression analysis techniques have challenges dealing with overdispersion and non-stationarity in the datasets. Therefore, decreasing the possibility of

relying on the predictive distribution of the model for an alarm system. The shortcoming in the Poisson model was identified in Bronner et al. (2015b); they implemented the Poisson regression with over-dispersion and concluded that the technique lacks sensitivity and needs further improvements to handle seasonal variation. Vial et al. (2015) used a quasi-Poisson regression, also known as an improved Farrington algorithm, to detect disease outbreaks during post-mortem inspection of slaughtered animals. They discovered that the algorithm was satisfactory for large datasets but performed poorly for small datasets due to high uncertainty in small datasets. Therefore, they concluded that integrating more data streams and simultaneous evaluation of data sources could increase SyS sensitivity and timeliness.

However, Hedell et al. (2019) argued the need for zero-inflated models to tackle over-dispersion in observations. Likewise, Dórea et al. (2013c) argued the need to use Zero-inflated approaches to address over-dispersion in models. Thomas-Bachli et al. (2014), on the other hand, suggested the use of the NB regression model because the model allows the overdispersion parameter to vary randomly following a beta distribution that can be adjusted. Alton et al. (2015) supported the authors' views, arguing that the NB model allows a beta distribution for overdispersion parameters.

Many publications compared regression modelling techniques with the Statistical Process Control (SPC) charts which have their foundations in quality control and industrial statistical process control; the techniques frequently used in VSyS include the variation of moving average techniques, CUSUM, MCUSUM, and Shewhart charts. The SPC is a graph for studying process changes over time in different research fields. They use a central line for the average population mean and upper and lower lines for the upper and lower control limits, determined by the historical datasets and measured in standard deviations. So, by comparing the current datasets to these lines, scientists can make conclusions about whether the process variation is consistent (in control) or unpredictable (out of control and is affected by any covariates) (Veldhuis et al., 2016; Dupuy et al., 2015; Tongue et al., 2020). The variations of moving averages include simple, linear, cumulative, or weighted forms, which are algorithms for analysing data points by creating a series of averages of different subsets of the full dataset to smooth out random effects and highlight longer-term trends or cycles. They are useful in addressing stochasticity due to trends and seasonality in datasets or models (Sala et al., 2020; Faverjon et al., 2019a).

According to Dupuy et al. (2015), the Shewhart control chart is based on binomial distribution, and it enables the detection of outbreaks through proportions. Also, it is sensitive to random noise in small datasets. However, Sala et al. (2020) argued that the CUSUM and MCUSUM charts tend to pick up small changes in datasets quicker than the Shewhart modelling techniques. Veldhuis et al. (2016) demonstrated a technique that combines the Frequentist statistical approach with the SPC chart. They analysed routinely collected data on cow's reproductive events and milk yields using the linear mixed model to account for fixed and random effects in the data and at the same to construct the baseline model. Then, they constructed the outbreak detection system using a CUSUM chart. The approach was evaluated to determine its sensitivity and timeliness. The authors concluded that the methods investigated did not result in better sensitivity or timely detection due to the system's inability to capture uncertainties in the dataset and the model. The model assumptions were perhaps difficult to incorporate into the model. Furthermore, it is worth noting that the primary assumptions underlying the SPC techniques have been questioned in the context in which it was originally conceived by other studies in the industrial process control field (Woodall, 2000; Velsko and Bates, 2016).

Veldhuis et al. (2020) compared the CUSUM chart with spatio-temporal cluster analysis and the Bayesian disease mapping technique. Generally, the control charts are capable of measuring clustering activities. For example, Faverjon et al. (2019a) combined the cattle mortality dataset with clinical data to implement multivariate event detection methods by comparing the effectiveness of two-directional multivariate control chart algorithms. The techniques are multivariate exponentially weighted moving average (MEWMA) and multivariate cumulative sum (MCUSUM) to model and simultaneously demonstrate the monitoring of multiple epidemics of four different disease indicators. These authors described improved outcomes of the outbreak detection even with small changes occurring in the time series. However, they reported that combining all time-series data into one statistical analysis was a limitation of their technique as it was impossible to quickly identify which time-series contributed to a raised alert unless the raw data were examined.

Dórea et al. (2014) presented a research study on automatically classifying diagnostic laboratory records into syndromes through keyword identification of the information from historical test requests. Their approach was to monitor disease outbreaks and provide early detection of temporal changes using the time-series count of tests submitted to two animal

health diagnostic laboratories. They implemented the temporal aberrations approach by investigating and comparing the effectiveness of CUSUM, EWMA, Shewhart control, and Holt-Winters exponential smoothing algorithm. They concluded that the randomness of each time-series dataset impacted the performance of each of the detection algorithms, and the main cause was the parameters set for each time-series in the model; it was the standard deviations for the control charts, while the confidence interval parameter for the Holt-Winters exponential smoothing algorithm.

Marceau et al. (2014) implemented a univariate time-series modelling with logistic harmonic regression where the dataset was modelled as two variables. The first variable contained time values, and the other was the dependent variable. The timeliness and specificity of each indicator were highly variable and low, respectively. Therefore, the authors argued that the time-series analysis of the five syndromic indicators could be monitored simultaneously to discover the variations and understand the interplay between the predictors in real-time rather than the implemented univariate approach. They concluded that further investigation of the outbreak detection technique was required to improve the VSyS strategies.

Other studies implemented different clustering and classification techniques, including the machine learning approach, to define syndromes and combine data from multiple sources. For example, Dupuy et al. (2013b) used multiple factor analysis (MFA) combined with K-means and hierarchical ascendant clustering (HAC) to group syndromes and define reasons for condemnation extracted from the slaughterhouse data stream. On the other hand, Behaeghel et al. (2015) categorised atypical cases reported by veterinary practitioners using the HAC method on a single data stream. They concluded that the study's outcome was improved; the reason might be that the original data submission also contained confirmed infections. Burkom et al. (2019) implemented a disease detection system using weekly syndromic counts of veterinary laboratory test orders. They applied a rule-set approach for syndromic categorisation and tested it by comparing different statistical analysis methods. Bollig et al. (2020) evaluated the suitability of different machine learning methods for syndromic surveillance, focusing on free-text veterinary necropsy reports from a diagnostic laboratory dataset. They concluded that no single algorithm was superior for describing diseases' temporal and spatial features, especially for discovering epidemiological trends.

The multiple parallel univariate time-series technique is the conventional approach for modelling data across large spatial scales with count data grouped into regions (Fernández-Fontelo et al., 2017). Dórea and Vial (2016) indicated that the technique might not be suitable as disease distribution and administrative borders may not be associated. However, Alba et al. (2015) managed this problem in their study by using hierarchical time-series statistics to account for categorised spatial structure while grouping observations according to geographical zones.

Perrin et al. (2015) proposed a unique technique by dividing France mainland into different geographical scales of regular hexagons. Mortality was monitored individually in each hexagon for VSyS purposes. Hexagon with excessive mortality rate was then compared to the expectation of the Poisson regression model, which was previously calibrated with a historical dataset. Then, a scan statistics was applied to detect clusters of hexagons with high mortality rates. Similarly, the authors recommended that spatial components should be observed with scan statistics. They also suggested that the spatial components should be considered in the modelling even when spatial monitoring is not explicitly required.

Vial et al. (2016) presented a different approach for using the historical laboratory test submission dataset in VSyS and implemented a stochastic modelling technique for a multivariate surveillance system. Most publications reviewed in this current work required training datasets to be free from historical disease outbreak signals to determine the baselines against which future outbreaks can be evaluated. However, Vial et al. (2016) allowed the retention of historical outbreaks and other covariates that may impact the experiment's outcome in their research.

3.5.2 Bayesian Inference Approach

Faverjon et al. (2016) argued in favour of using modelling techniques that can accommodate the stochastic characteristics of datasets, such as the Bayesian approach. They also argued that such a technique could allow researchers to combine data from multiple sources for syndromic analysis. The authors demonstrated this technique and concluded that it offered the opportunity of specifying prior probability and defining the expected impact of distributions in a model as the Bayes likelihood ratio. Bayes likelihood ratio is the relative probability of two hypotheses expressed as a ratio representing a prior belief about a disease outbreak status. The Bayes likelihood ratio is the Bayesian alternative to classical hypothesis testing. Furthermore, when

evidence in favour (or not) of each hypothesis is observed, a posterior belief about the disease status can be evaluated.

Likewise, Andersson et al. (2014) used the Bayes likelihood ratio to calculate the probability of observing an ongoing outbreak over observing no outbreak. However, both authors' approaches above did not take advantage of the Bayesian sampling techniques through the MCMC algorithm to sample from posterior distribution but relied on the Frequentist alarm threshold to detect and interpret outbreaks. Also, no risk assessment methods were included in their techniques.

On the other hand, Hedell et al. (2019) simulated a prior baseline rate of syndromic data sources during a non-outbreak period and predicted the prospective spatial-temporal outbreaks using the disease's historical data with Gibbs sampling. The novelty of their approach is the application of Bayesian methodologies to incorporate existing information as priors into the model, such as syndromic clusters in space and time, syndromic frequency, and the seasonality of the disease. The authors interpreted the disease outbreak signal with a probability distribution, but the performance of the Gibbs sampling in this technique is unclear.

Struchen et al. (2017) incorporated different scenarios involving time delay in data submission under different hypotheses using the Bayes likelihood ratio. They demonstrated that a multivariate approach could yield better performance than a univariate system that currently accounts for most of the techniques incorporated in VSyS through the Frequentist statistical approach. Uncertainties due to seasonal variation in the data under baseline conditions were modelled by fitting NB distributions using dynamic binomial regression based on maximum likelihood. The authors suggested that the Bayes likelihood ratio can perform change point analysis for multiple data streams and account for delayed reporting in the model to detect outbreaks on time. Furthermore, they argued that it could model a small sample size with a short history. However, the authors' work did not explain the implementation and evidence of change point analysis. They suggested using point estimates as input rather than the full posterior distribution to avoid increased complexity and computational time.

The Bayesian disease mapping technique was previously extended to a fully Bayesian framework in Besag et al. (1991) and implemented with MCMC sampling from a probability distribution. In Veldhuis et al. (2020), the Bayesian disease mapping model was constructed on a conditional autoregression algorithm (CAR). The model residuals were weighted by the

square root of the number of herds per district week for the baseline construction to account for uncertainty in residuals from areas with a low cattle herd density. Annual seasonality was considered by including two sine/cosine harmonics as predictors. The model was compared with the CUSUM chart and spatiotemporal cluster analysis, and the authors concluded that the Bayesian approach produced the lowest predictive alert values and performed poorly due to many false alarms. This result identified other aspects of the Bayesian inference approach that need to be considered with great attention during modelling, such as the choice of prior distributions and the variance components to include in the model. Therefore, the authors suggested revising the parameter's settings and improving the model in further research work.

In contrast, Hale et al. (2019) modelled their approach with Bayesian inference using an MCMC algorithm to generate samples from the posterior distribution. The authors combined a spatio-temporal model with the Bayesian inferential framework. The resulting model considered all sources of uncertainty in parameter estimation and prediction and accommodated spatial, temporal, and individual-level covariate estimates. According to the authors, the method used mixed-effects logistic regression and fitted the simulated model samples from the MCMC algorithm. The mixed-effects accounted for spatial and temporal components to produce a binary outcome, but the model was impacted by computation performance as the data grew and the time increased. However, the authors' work did not provide a detailed step-by-step guide on the model development and the MCMC sampling process.

3.6 Quantitative Risk Assessment (QRA) in VSyS

The Scopus electronic citation database was searched for peer-reviewed articles that addressed the review question; "is there any evidence of published research on using QRA techniques to assess VSyS alarms?". Table 6 demonstrates the search string evolution from the initial search string in item 1 to the final search string in item 4, which returned 36 research publications. The various exclusion and inclusion filters described in section 3.2.2 were applied to achieve the final search result. Each publication was re-checked for bias and manually reviewed to ensure that each paper remained focused on the review question, resulting in 8 research articles.

***Table 6**: Search string evolution on quantitative risk assessment in VSyS*

Item	Search Strings	Publication returned
1	"Quantitative Risk Assessment"	3813
2	"Quantitative Risk Assessment" AND surveillance	88
3	"Quantitative Risk Assessment" AND surveillance AND PUBYEAR > 2009 AND (LIMIT-TO (LANGUAGE, "English"))	43
4	"Quantitative Risk Assessment" AND surveillance AND PUBYEAR > 2009 AND (LIMIT-TO (LANGUAGE, "English")) AND (LIMIT-TO (DOCTYPE, "ar") OR LIMIT-TO (DOCTYPE, "re"))	36

Though the QRA technique was found in 7 publications, one common trend among the papers is that the risk model was developed with a proprietary Microsoft Excel spreadsheet macro named @Risk. The publications are listed in Table 7. All the papers focused on assessing surveillance risk through the probability distribution of a disease introduction into a population. For example, Mur et al. (2012) addressed the spatial and temporal variation of surveillance risk quantitatively by implementing the QRA approach in @Risk. However, we found that the effects of variability and uncertainty on the model outcomes were evaluated with sensitivity analysis, and the same trend was observed in the other articles selected for this review.

Kwan et al. (2016) evaluated sensitivity analysis by implementing Spearman's correlation coefficient to rank all the input parameters according to their contributions to the model output variance. The same approach was adopted by Jurado et al. (2019), indicating the method as a standard approach for evaluating the model's sensitivity in the @Risk application. However, the problems and limitations of Spearman's correlation in Armstrong (2019) suggested a more cautious approach regarding its use and the application of alternative methods where appropriate.

Since variability or randomness refers to the inherently stochastic nature of parameters while uncertainty refers to a lack of precise knowledge about a parameter, Vose (2008) *pg.52* argued that the separation of variability from uncertainty when accounting for model parameters during risk assessment is best practice. This author's argument is important because research can reduce uncertainty in future assessments, whereas variability cannot be further reduced. The model presented using the @Risk application simulated both uncertainty and variability together and sampled from the distributions that reflect the uncertainty and variability of the

input variables. This procedure seems to be the only way that the @Risk application can account for variability and uncertainties in the data and the model, and this trend was identified in all the papers reviewed for this section (Mur et al., 2012; Kwan et al., 2016; de Vos et al., 2015; Collineau et al., 2020).

Furthermore, it is worth noting that the @Risk application lacks other effective variants of MCMC simulations, such as the No-U-Turn sampler (NUTS) and Hamiltonian Monte Carlo (HMC) (Salvatier et al., 2016). Additional limitations of using the Excel application for large data analysis include performance bottlenecks, file limits, data size limits, and lack of flexibility in data handling and analysis with advanced modelling techniques. Therefore, this review suggests that QRA techniques in VSyS are still poorly researched, and only a few publications that implemented the quantitative risk assessment method with the @Risk application exist in the literature.

***Table 7**: Publications identified on QRA techniques in VSyS*

S/N	Authors	Title	Technique
1	(Mur et al., 2012)	Quantitative Risk Assessment for the Introduction of African Swine Fever Virus into the European Union by Legal Import of Live Pigs	Monte Carlo simulation in Excel
2	(Kwan et al., 2016)	Quantitative risk assessment of the introduction of rabies into Japan through the illegal landing of dogs from Russian fishing boats in the ports of Hokkaido, Japan (Kwan et al., 2016)	"
3	(Jurado et al., 2019)	Could African swine fever and classical swine fever viruses enter into the United States via swine products carried in air passengers' luggage?	"
4	(Taylor et al., 2019)	A generic framework for spatial quantitative risk assessments of infectious diseases: Lumpy skin disease case study	No description of the technique, and the methodology appears unclear
5	(de Vos et al., 2015)	Risk-based testing of imported animals: A case study for bovine tuberculosis in The Netherlands	Monte Carlo simulation in Microsoft Excel application
6	(Collineau et al., 2020)	A farm-to-fork quantitative risk assessment model for Salmonella Heidelberg resistant to third-generation cephalosporins in broiler chickens in Canada	"
7	(Gierak et al., 2021)	Quantitative risk assessment of the introduction of low pathogenic avian influenza H5 and H7 strains into Poland via legal import of live poultry	"
8	(Dibaba, 2019)	The risk of introduction of swine vesicular disease virus into Kenya via natural sausage casings imported from Italy	"

3.7 Discussion

Dórea and Vial (2016) argued that as electronic records became available to farms, data generated on the farm and individually reported for each animal regarding various aspects of productivity and well-being became the most widely evaluated data source for VSyS development. This trend may be explained since data produced on the farm are created on the fly and are not usually dependent on any particular health events, unlike the clinical and laboratory datasets driven by animal ill health. However, Egates et al. (2015) argued that the animal production data stream could be timely but low in specificity and disease confirmation sensitivity. Such a data stream would be difficult to classify into syndromic indicators for monitoring specific disease outbreaks, which is the foundation of VSyS strategies. Marceau et al. (2014) argued that identifying appropriate health-related data to implement a VSyS is a key challenge to overcome on the farm.

Struchen et al. (2017) argued that data generated on the farm could be exposed to delayed reporting due to the remoteness of farms and the lack of access to the right devices to facilitate the direct upload of data to the central processing system even when the data is captured on time. In addition, the clinical and laboratory data streams may suffer from delayed reporting due to the extra time required to submit data for surveillance purposes, lack of data standards, and poor syndromic classification. Also, clinicians collect health event information from farm visits. This data collection type may suffer delays due to the extra time required to submit the data for surveillance purposes. The need to record animal health data from paper-based to electronic submission usually leads to non-compliance and delays in submitting complete information for surveillance (Dórea and Vial, 2016).

The quality of the information provided by different veterinarians during the same event could be variable irrespective of whether the submission is for voluntary or compulsory surveillance; this is because private businesses majorly drive the veterinary industry. Also, the systems that generate and collect animal health datasets in the private sector are often different and may lack standard communication protocols, which can make the real-time data transmission challenging (Vanderwaal et al., 2017). The strategy is different in human medicine, where a central body is responsible for surveillance, making data contributions from all GP practices, clinics, and hospitals consistent and easily manageable through automatic digital transmission and real-time interoperability of computer systems.

Also, Burkom et al. (2019) argued that the diagnostic data stream still suffers from standardisation problems, primarily when data are sourced from multiple laboratories with different test submission management systems. This is because the electronically transferred data formats and schemas vary significantly between systems due to a lack of centrally agreed standards (SAVSNET, 2019). This argument is viewed the same way by Vanderwaal et al. (2017), who described the lack of an identifiable data framework in veterinary medicine as a challenge making many sources of animal health data not readily usable in statistical models. Similarly, Vial et al. (2016) argued that the lack of commonly adopted data standards in the veterinary sector makes data integration across heterogeneous datasets difficult.

The unwillingness of veterinary clinics and diagnostic laboratories to contribute their data for research and VSyS purposes is a major challenge; this may also be due to a lack of a centrally agreed framework or body in veterinary medicine coordinating such exercise or due to privacy concerns. Furthermore, fear of divulging trade secrets can also be a factor since most laboratories and veterinary practices are private businesses. Therefore, data scarcity is highlighted in this research as a challenge across clinical and laboratory diagnostic data streams.

The major challenge is sustainability in cases where clinicians regularly contribute data for VSyS development since many surveillance platforms relying on animal health datasets have to provide some form of incentive to entice clinicians to participate in data contribution (Struchen et al., 2016). Another challenge is that the datasets used for VSyS modelling are not available simultaneously for real-time outbreak detection, considering the significant variability in the relative timeliness of the animal health data sources, as indicated in Figure 1. For example, the classification of syndromes from low milk production or animal inactivity on the farm has low specificity for disease confirmation but can be timely for modelling outbreaks as they are continuously generated from connected IoT devices. On the other hand, diagnostic laboratory test results have high specificity for disease confirmation but are not always timely (Egates et al., 2015).

Current approaches in VSyS are a variety of statistical modelling techniques that first seek to define baselines/thresholds, which represents the normal behaviour when no disease outbreak is recorded. Abnormal events overlaid on the background noise are evaluated against these baselines to detect outbreaks. In most detection methods, an alarm goes off when the observed

data exceed the population's baseline values. Such VSyS techniques use the threshold/baseline to provide a yes/no qualitative output: “No, there is no outbreak” or “Yes, something unusual is happening in the population”. These alarm-based VSyS techniques seem simple to implement, but the outputs might be difficult to interpret for decision-makers, particularly when events occur very close to the threshold, when there is a slowly increasing outbreak, and when the number of cases reported in each time unit, each week or each day is too small to trigger an alarm.

Most alarm generating techniques produce a high number of false alarms when the threshold value is lowered and few alarms when the threshold is increased; this indicates that the handling of the alarm generation does not provide explanations for data points closer or far from the baseline, which may lead to a false interpretation of alarms (Veldhuis et al., 2016). To inform decision-making, it is also desirable to combine surveillance outcomes with other information, such as associated uncertainties and prior knowledge of the health events under investigation. However, there is no straightforward approach for including these parameters in surveillance modelling when the algorithm is based on an alarm threshold technique.

The most common implementation of VSyS found in the literature is with a single data stream, using the univariate analysis to detect temporal aberration and sometimes modelled with spatial components. Other research highlighted the reduced sensitivity of the univariate approach to disease frequency changes with a higher rate of false alarms (Fischer et al., 2015; Andersson et al., 2014; Robertson et al., 2011). This indicates that no single data source can capture information from all factors responsible for a disease outbreak. Instead, various information can be observed in data extracted from different sources for syndromic surveillance.

Many studies suggested that combining regular surveillance datasets with non-specific animal health data in real-time can improve data availability and timeliness and provide decision-makers with early insights into disease trends (Veldhuis et al., 2016; Dórea and Vial, 2016; Faverjon et al., 2019a). Also, some studies argued that analysing multiple data streams could improve the sensitivity of the outbreak detection algorithms (Struchen et al., 2017; Vanderwaal et al., 2017). However, only a few research studies attempted multivariate analysis to account for covariates and the influence of unexpected variables in their study, and most of those that implemented the multivariate analysis did so using multiple parallels of univariate analysis techniques (Faverjon et al., 2016; Andersson et al., 2014; Tongue et al., 2020).

Furthermore, historical outbreaks create random effects in surveillance datasets (Hedell et al., 2019; Aghaali et al., 2020; Salmon et al., 2015; Manitz and Höhle, 2013; Vial et al., 2016). Many research studies removed historical outbreaks from surveillance datasets to construct thresholds for their techniques. However, Souley Kouato et al. (2018) highlighted that the quantitative estimation of surveillance risk requires uncertainty modelling to determine the effects of random input parameters on risk magnitude. Therefore, some recent studies interested in VSyS modelling argued that retaining the historical outbreak may provide the approach for measuring risk magnitude and the associated uncertainty and help reduce false alarms.

Vial et al. (2016) argued for the use of a stochastic modelling approach to tackle overdispersion and non-stationarity in datasets. According to the authors, stochastic modelling offers more flexibility, allowing for the retention of historical outbreaks in datasets instead of the previously discussed approach where historical outbreaks must be removed from datasets to construct baselines. A method such as the Bayesian technique may provide useful insights into developing a quantitative risk assessment approach. Also, it may allow the analysis of multi-data streams to evaluate the risk of surveillance alarms since more information can be derived from multiple sources for surveillance purposes than relying on a single data stream. Only a few studies demonstrated the ability of the Bayesian framework to combine multiple data streams and retain historical outbreaks in their research work.

3.8 Conclusion

In this chapter, the author identified the following challenges in VSyS development.

1. Data scarcity.
2. Delayed reporting of health incidents.
3. Data has not been available simultaneously from different sources for VSyS modelling.
4. Most VSyS techniques rely on methods that use deviation from baselines to define outbreaks in signal alarms. These systems may be simple to implement but difficult to interpret, particularly when events occur very close to the threshold. In addition, the existing alarm-based VSyS techniques do not differentiate between areas and periods of low and high risk in disease outbreak distribution.

The author also identified existing data sources for VSyS development and categorised them into data streams according to the work of previous studies. However, it was observed that five

publications could not fit into any of the groups defined by previous research works, and this is because they implemented VSyS techniques with multiple data streams. This observation demonstrates the appetite of recent research studies in exploring multi-data streams for VSyS implementation. As a result, a multi-data stream category was recommended in this study.

In real health surveillance practice, all alarms must be assessed to confirm whether they are of public health importance (Lake et al., 2019). To the researcher's knowledge, no research published in animal health surveillance has quantitatively assessed these alarms to inform decision-making. The existing quantitative risk assessment method in the literature emphasised the use of the @Risk Microsoft Excel application for assessing disease introduction into a population. The application is not adaptable for assessing surveillance alarms and has several challenges discussed in section 3.6, making it unsuitable for assessing the VSyS alarms. Also, the application evaluates randomness and uncertainty together in a parameter, which is a limitation. Literature evidence stipulates the evaluation of randomness and uncertainty separately since uncertainty can be improved with more datasets and research, but stochasticity in the dataset may not be improved.

Therefore, the following gaps in knowledge still need to be filled about VSyS and its ability to assess surveillance risk and generate alarms to inform public health decision-making.

GAP 1 – There is no straightforward approach for early quantitative risk assessment of disease outbreak signals with uncertainty measures when the VSyS system is alarm-based.

GAP 2 – There is no straightforward method for evaluating and interpreting risk when the disease outbreak signal is slowly increasing and too weak to trigger an alarm.

GAP 3 – There is a need for an alternative technique to evaluate, interpret, and communicate disease surveillance risk while considering existing knowledge of the disease and the associated uncertainties.

An improved SQRA technique will be required to fill these gaps, and a step-by-step procedure for integrating the SQRA method into VSyS will also be needed. The proposed method is not to replace the existing VSyS alarm system but to supplement it by producing a risk magnitude measurement technique to differentiate between areas and periods of lower and higher risk in disease outbreak distribution. Furthermore, the social media dataset has been used in previous

studies to provide an early real-time disease outbreak system. Therefore, the same approach may accommodate the relative timeliness of data availability in veterinary health surveillance to tackle delayed reporting challenges and implement a system that uses multiple data streams to model VSyS risks. The novelty of the method is the incorporation of social media datasets to tackle the challenges of data scarcity identified in this systematic literature review. Also, implementing an approach that can accommodate the relative timeliness of data availability such that data is introduced into the model whenever they become available as "priori" to update the belief about disease outbreak distribution risk.

Chapter 4 – Data

4.1 Chapter Introduction

SAVSNET was established in 2008 to meet a deficiency in companion animal surveillance. The University of Liverpool solely runs the SAVSNET project to provide syndromic surveillance functions to companion animals and offer a research platform for rapid and actionable interdisciplinary research programmes. There are over 500 veterinary practice premises and commercial diagnostics laboratories within the SAVSNET network, contributing electronic health records, clinician narrative texts, diagnostic laboratory test results, and necropsy reports for surveillance purposes. Although the SAVSNET datasets are not open source, they have been used in some research studies. Table 8 indicates the research articles, their contributions and the data science techniques applied in their studies.

Twitter data and other online media datasets have been sparsely applied to VSyS development. However, studies such as Yousefinaghani et al. (2019) and Robertson and Yee (2016) demonstrated the usefulness of Twitter datasets for developing early disease outbreak systems since the veterinary healthcare data sources suffer several challenges identified in the previous chapter. Combined with the SAVSNET datasets, tweets that suggest an incident of a dog's gastrointestinal infection were collected from 01/01/2018 to 31/03/2020 for the purpose of addressing the shortage of veterinary datasets for syndromic surveillance development. The aim is to validate the use of heterogeneous data sources for modelling an SQRA method and for swiftly assessing the risk magnitudes of syndromic alarms to detect disease outbreaks.

In this chapter, the author discussed and presented the datasets obtained from SAVSNET and Twitter, including the techniques used to acquire them from the sources, the data's attributes, initial processing, and the storage approach. In addition, the author presented some of the difficulties encountered while sourcing the datasets.

Table 8: *Research studies that used the SAVSNET datasets to contribute to knowledge*

S/N	Authors	Major Contributions	Data Science Technique
1	(Hale et al., 2019)	Presented a real-time Spatio-temporal VSyS system to detect outbreaks in close proximity of a veterinary clinic/facility by estimating posterior probability distribution to support decision-making with the ability to measure uncertainty associated with identified clusters.	Bayesian predictive inference within a Spatio-temporal mixed-effects regression model. Used the MCMC algorithm for sampling from a probability distribution.
2	(Sánchez-Vizcaíno et al., 2017)	Presented a technique to explore the association between a range of animal characteristics and socioeconomic factors of pet animals in Great Britain.	Mixed logistic regression models.
3	(Tulloch et al., 2017)	Described the temporal and spatial activity of ticks in companion animals using the Practice data stream	Simple free-text analysis for keyword extraction, regression modelling, and map plotting with QGIS.
4	(Jones et al., 2014)	Used the SAVSNET dataset to establish the feasibility of profiling dogs and cats that are suffering from diarrhoea outbreaks in the UK	Logistic regression models were used to estimate morbidity odds ratios.
5	Garcia-Constantino et al. (2012)	Presented an intuitive way of extracting the contents of a document through text summarization and classification and applied the technique to SAVSNET data.	Hierarchical classifier using the Support Vector Machine (SVM).
6	(Radford et al., 2011)	Proposed a VSyS system using antibacterial prescribing patterns in a small animal dataset hosted by SAVSNET	Statistical analysis and association were used with logistic regression.
7	(Garcia-Constantino et al., 2011)	Developed a technique for text classification from secondary data for text summarization	Used the Term Frequency Inverse Document Frequency (TF-IDF) for the text mining and document retrieval, and the Apriori algorithm for text classification

4.2 Method

Published peer-reviewed research articles that used the SAVSNET datasets or a subset of the data in their work were searched in the Scopus database. The search keyword was simply SAVSNET, and no date range or language restrictions were applied. The search returned 9 publications, of which two articles are survey highlights, and one is a report. The remaining 6 articles were included with a research paper from the literature review section. Therefore, presenting 7 peer-reviewed articles that previously used the SAVSNET datasets in Table 8.

The search for the research datasets began with informal discussions with veterinary scientists and scholars at the company I was employed. The company is renowned in veterinary research and diagnostic laboratory. They provide diagnostic services to all species of animals and are responsible for statutory veterinary syndromic surveillance functions in the Republic of Ireland.

The informal discussions approach was found appropriate for the initial stage of data search efforts since it is a type of qualitative data collection technique used as a general tool to understand people's expertise and experience and uncover their inner perceptions, feelings and attitudes about a subject of interest (Denzin and Lincoln, 2011; Denzin, 1999). Jebreen (2012) also argued that informal discussions or unstructured interviews could be used to effectively collect detailed data about experts' perspectives on a specific topic. For example, Spruit et al. (2014) used the unstructured discussion to create business requirements for the healthcare sector in their study. The authors suggested that unstructured discussions are appropriate when detailed information is required on a subject matter. Therefore, such detailed information can be achieved by clarifying questions and answers casually without a particular structure since they are the best way to develop explorative information.

The informal discussion approach was used to obtain the experts' unbiased opinions on the type of datasets that fit the objectives of this research work and fulfil its overall aim. However, the datasets obtained from the company where I work were homogenous, so it was decided that SAVSNET and Twitter be contacted for heterogeneous datasets. The method of extracting the SAVSNET and Twitter datasets is presented in the following subsections.

4.2.1 Method of Collecting the Veterinary Datasets

The approach undertaken by SAVSNET requires the completion of questionnaires by each participating veterinary practitioner at the end of every consultation. With clinical notes and signalment data, the questionnaire is transmitted in real-time or near real-time to the SAVSNET cloud repository, where the data are stored for further data processing and analysis (SAVSNET, 2019). Sánchez-Vizcaíno et al. (2015) provided the detailed technical procedure adopted by SAVSNET for collecting and processing the veterinary datasets from different sources before storing them securely in the cloud. As a result, the author did not undertake a review of the procedure as it is out of the scope of this ongoing study.

However, accessing data on the SAVSNET platform requires submitting an application to the research committee in charge of the SAVSNET at the University of Liverpool. It is required that applicants provide the details of their research study, including the research aims, objectives of the proposed study, and the proposed analysis method on the datasets. Each application is vetted and reviewed to establish the academic merits and objectives. Furthermore, depending on the requested data size, application to use SAVSNET datasets attracts access fees if successful, but the access fee is waived if a research study demonstrates outstanding academic merits that contribute to one or more SAVSNET objectives. This research study met the requirements for free usage of the SAVSNET datasets for satisfying some of the SAVSNET platform's objectives and for the study's potential academic merits (SAVSNET, 2019).

SAVSNET platform contributed two data streams to this research work, the practice and laboratory diagnostic data streams on gastrointestinal (GI) disease in dogs. The acquired practice data stream was available from 01/03/2018 to 28/02/2020, and the laboratory diagnostic data stream was from 03/01/2017 to 28/02/2020. Each data stream was in a compressed and password-encrypted CSV file to enhance the files' safety, management, quality, and storage. The files were downloaded from a secure link provided by SAVSNET. The ethics approval from the University of Liverpool includes the terms and conditions of using the datasets.

Both data streams downloaded from the SAVSNET platform have been cleaned and pre-processed by the SAVSNET data team and were ready for data analytics and statistical modelling. Therefore, there was no additional cleaning or pre-processing necessary. On the

other hand, the Twitter data stream required data cleaning, pre-processing, complex manipulation, and text tagging, as explained in the following sub-section.

4.2.2 Method of Extracting and Pre-processing the Twitter Datasets

Tweets were extracted from Twitter API and stored in a file in the Google cloud using the Python procedure shown in Figure 6. The comprehensive information on extracting data from Twitter was explained in the research work by Yousefinaghani et al. (2019). However, the difference between their approach and the technique in this study is that they implemented web scraping in a PHP script, but we implemented extraction directly from Twitter API using Python libraries "Pandas" and "Searchtweets". Twitter maintains the Searchtweets library (Twitter, 2020).

The Twitter API provides near real-time access to a collection of tweets matching a specified query, but it requires the developer's account with a monthly subscription to extract bulk tweets. Since the bulk tweet was desirable for this ongoing analysis, a developer's account with a paid subscription was registered. Using Python, the author restricted the extracted tweets to the "English" language. The keywords provided by domain experts were used to identify the occurrences of specific symptoms related to gastrointestinal disease in texts. The keywords were fed into the Twitter API to obtain tweets that show possible elevated GI disease occurrences in dogs in the U.K. The keywords are diarrhoea, vomiting, vomitting, vomit, gastroenteric, gastroenteritis, and gastro. The tweet date was set from 01/01/2018 to 31/03/2020.

The extracted tweets were cleaned up using a Python procedure implemented with the regular expression (RE) libraries to remove irrelevant content. The "bag of words" and "stop words" functions were created and used to further process and clean up the tweets. The snippet of the Python codes used for cleaning and pre-processing the collected tweets is presented in Figure 7. The resulting tweets were manually labelled and checked by domain experts to ensure correct labels and classification. A tweet labelled "relevant" refers to a sick dog demonstrating GI symptoms, while the "irrelevant" label indicates an advert or any other text that did not refer to a sick animal or GI symptoms in a dog.

```
# load credential to connect to Tweeter Search API
premium_search_args = load_credentials("/content/drive/My
Drive/ProjectAA/authentication.yaml",
                                       yaml_key="search_tweets_api_2",
                                       env_overwrite=False)

print (premium_search_args)

# define the query with search keyword
query = "Dog (diarrhoea OR vomiting OR vomitting OR vomit OR gastroenteric
OR gastroenteritis OR gastro)"

# define the rule with start and end dates
rule = gen_rule_payload(query, from_date="2018-01-01", to_date="2020-03-
31", results_per_call=100)

rs = ResultStream(rule_payload=rule,
                  max_results=2500,
                  max_pages=1,
                  **premium_search_args)

tweets = rs.stream()
# create empty list to store the tweet objects extracted from tweet data
stream
list_tweets = list(tweets)

tweet_date = []
tweet_text = []
tweet_lang = []
tweet_quote = []
reply_tweet = []
tweet_user = []

# append each object to the empty list declared above
for tweet in list_tweets:
    tweet_date.append(tweet['created_at'])
    tweet_text.append(tweet['text'])
    tweet_lang.append(tweet['lang'])
    tweet_quote.append(tweet['is_quote_status'])
    reply_tweet.append(tweet['in_reply_to_screen_name'])
    tweet_user.append(tweet['user']['location'])

# convert the lists to Pandas dataframe
df = pd.DataFrame({'Tweet Posted Time (UTC)': tweet_date, 'Tweet Content':
tweet_text,  'Language': tweet_lang,  'Quoted Tweet': tweet_quote, 'Reply
Tweet': reply_tweet, 'Tweet Location': tweet_user,}This )
# save the dataframe to file
df.to_csv("/content/drive/My Drive/ProjectAA/dog_gastro_all.csv")
df.to_csv("/content/drive/My Drive/ProjectAA/cat_gastro_all.csv")
```

***Figure 6**: The Python procedure implemented for extracting raw tweets from Twitter API and saving them in a file*

The first five rows of the tweet datasets before and after applying cleaning and classification procedures are shown in figures 10 and 11. For future research, the labelled tweets could be trained with the Naïve Bayes classification algorithm so that new incoming tweets could be automatically recognised and classified appropriately once cleaned up. However, Twitter cleaning, pre-processing, tagging and classification with the Naïve Bayes algorithm are out of

the scope of this research work. The detailed implementation was already described in the research study by Yousefinaghani et al. (2019).

The extra materials provided with this study contain the full Python codes used for extracting, pre-processing and labelling the tweet dataset. They are included as supplementary material at the link in Appendix.

```
# Select only tweets and discard tweet replies
df = df[df.Tweet_Type != 'Reply']

# Carry out manual labelling of the saved dataset to categorise tweets into
relevant and irrelevant class
# add a new column to hold the labels and copy the contents of Sick column
into Label column
df['Label'] = df['Sick']
# replace labels with numbers for ease of training
repl_label = {"Label":     {"relevant": 1, "irrelevant": 0}}
df.replace(repl_label, inplace=True)

# The tweet dataset needs to be cleaned to get it ready for training
# Define a preprocess function and use pandas apply to apply the function
to each value of the 'Tweet Content' column
# Use tweet preprocessing module
def preprocess_tweet(row):
    text = row['Tweet_Content']
    text = p.clean(text)
    return text
df['Tweet_Content'] = df.apply(preprocess_tweet, axis=1)  # Tweet Content
column has been cleaned to normal text.
# Now, we can apply normal text preprocessing such as
# Lowercasing
# Punctuation Removal
# Replace extra white spaces
# Stopwords removal
# Removing stopwords can potentially help improve the performance as there
are fewer and only meaningful tokens left.
# Thus, it could increase classification accuracy. Even search engines like
Google remove stopwords for fast and
# relevant retrieval of data from the database
def stopword_removal(row):
    text = row['Tweet_Content']
    text = remove_stopwords(text)
    return text
df['Tweet_Content'] = df.apply(stopword_removal, axis=1)

# Remove extra white spaces, punctuation and apply lower casing
df['Tweet_Content'] =
df['Tweet_Content'].str.lower().str.replace('[^\w\s]','
').str.replace('\s\s+', ' ')
```

***Figure 7**: Python code snippet for cleaning and pre-processing the collected tweets*

4.3 Result

The practice data stream contains information about each dog that attended a clinic, the consultation date, species, the main reason for the visit, and binary responses (true or false) to key information extracted from the patient note such as acute vomiting, an anti-vomiting drug prescribed, and antibiotic prescribed. The veterinary practice data file has 2,075,438 rows and 7 columns. The first few rows are presented in Figure 8.

```
print(practice)
```

	consult_date	species	MainReason	QuestionnaireAcuteVomiting	TextMiningAcuteVomiting	AntiVomitingDrugPrescribed	AntibioticPrescribed
0	2018-03-01 00:00:00+00:00	dog	other_unwell	False	False	False	False
1	2018-03-01 00:00:00+00:00	dog	other_unwell	False	False	False	False
2	2018-03-01 00:00:00+00:00	dog	other_healthy	False	False	False	False
3	2018-03-01 00:00:00+00:00	dog	vaccination	False	False	False	False
4	2018-03-01 00:00:00+00:00	dog	post_op	False	False	False	True

***Figure 8**: Practice data stream*

The diagnostic laboratory data has 133,423 rows and 5 columns. The sample of the data stream is illustrated in Figure 9, and it contains the test result date, sample id, species, the name of the assay tested, and a binary result value (negative or positive). The diagnostic data stream contains different assays tested for the presence of GI infections. The assayname column in the data stream contains values such as CampylobacterCulture, Canine Parvovirus PCR, ClostridialEnterotoxin, CanineEntericCoronaVirus, and GiardiaPCR. Further key information about the attributes of both data streams is discussed in chapter 6 during data exploration and visualisation.

```
print(diagnostic)
```

	Unnamed: 0	resultdate	sample_id	SPECIES	ASSAYNAME	RESULTVAL
0	0	09/10/2018	lab_a_1003247967	Canine	CampylobacterCulture	NEGATIVE
1	1	03/01/2017	lab_a_1003640965	Canine	CampylobacterCulture	NEGATIVE
2	2	03/01/2017	lab_a_1003644579	Canine	CampylobacterCulture	NEGATIVE
3	3	03/01/2017	lab_a_1003644607	Canine	CampylobacterCulture	NEGATIVE
4	4	03/01/2017	lab_a_1003644622	Canine	CampylobacterCulture	NEGATIVE

***Figure 9**: Diagnostic data stream*

The processed Twitter dataset describes the daily count of tweets that suggest dog GI infection, and the resulting data file contains 2,623 records and 6 columns, as presented in Figure 11. The

data stream consists of date, tweet contents, tweet type, language, country, tweet location, and the label for tagging whether each tweet suggests a sick animal with GI infection. The label column in Figure 11 shows the conversion of the sick tagged column to the numeric variable for ease of use in statistical modelling and analysis. No retweet contents were included in this data stream to ensure that each record is unique. Furthermore, only the tweet from the United Kingdom was included in the dataset. The attributes of the Twitter data stream are further discussed in chapter 6.

```
print(df.head(5))
```

	Tweet_Posted_Time	Tweet_Content	Tweet_Type	Tweet_Location	Sick
0	31/03/2020 22:25	Oh no dog shite breath is now here in the even...	Tweet	Kensington, London	relevant
1	31/03/2020 20:26	Just cleaned up vomit... twice🤦. My dog vo...	Tweet	On the water	relevant
2	31/03/2020 06:54	Diggy DOG breath is bending over on all fours ...	Tweet	Kensington, London	irrelevant
3	31/03/2020 06:45	Be Careful Checking On Your Dog When He Has Di...	Tweet	Great Britian	irrelevant
4	31/03/2020 05:56	Woken up to the dog having had diarrhoea over ...	Tweet	London	relevant

***Figure 10**: Twitter data stream - before cleaning and pre-processing*

```
print(tweet)
```

	Tweet_Posted_Time	Tweet_Content	Tweet_Type	Tweet_Location	Sick	Label
0	31/03/2020 22:25	oh dog shite breath evening spraying diarrhoea	Tweet	Kensington, London	relevant	1
1	31/03/2020 20:26	just cleaned vomit twice my dog vomited daught...	Tweet	On the water	relevant	1
2	31/03/2020 06:54	diggy dog breath bending fours spraying verbal...	Tweet	Kensington, London	irrelevant	0
3	31/03/2020 06:45	be careful checking on your dog when he has di...	Tweet	Great Britian	irrelevant	0
4	31/03/2020 05:56	woken dog having diarrhoea night pleasant	Tweet	London	relevant	1

***Figure 11**: Twitter data stream - after applying the cleaning and pre-processing procedure*

4.4 Discussion

From Table 8, the closest approach to the method proposed in this ongoing study is the one by Hale et al. (2019), where the authors presented a VSyS system with a spatio-temporal feature to identify disease outbreaks in close proximity to a veterinary clinic. They developed a technique that estimated posterior probability distribution to support decision-making with a measure of uncertainty associated with identified clusters. This work differs from the technique proposed in this ongoing study because they implemented their approach with a Bayesian technique in traditional conditional auto-regressive models (CAR), which can produce spatial-temporal probability distribution.

However, Lee et al. (2014) argued that most CAR priors are globally smooth and have been shown in other studies to be potentially collinear with any covariate that is also globally smooth, leading to poor estimation performance of parameters and exhibiting abrupt step

changes. Hale et al. (2019) have not included the results of the collinearity measure or how this factor was managed in their study. Also, the authors stated that the MCMC computations were a limitation of their technique as it became computational prohibitive with a larger sample. The ability of their technique to identify an outbreak (sensitivity) is dependent on outbreak duration, spatial extent and the rate of incidents presenting at a premise. No detailed step by step guideline was provided to reproduce the authors' approach for using the CAR model and for localising early outbreak detection. The selection and definition of prior parameters for the model are also unclear. Furthermore, the authors' approach was implemented using a single data stream which is the practice data stream.

On the contrary, in this ongoing study, we proposed an approach capable of identifying a subtle increase in the rate of outbreak incidents with no dependence on outbreak duration. The method will allow users to improve the estimation performance of the parameters and hyperparameters compared to using the Bayesian technique in traditional conditional auto-regressive models. Furthermore, we will combine the Twitter data streams with gastrointestinal disease outbreak datasets from SAVSNET, which includes data from laboratory diagnostics and veterinary practices. Therefore, in this research work, the author will implement an SQRA method in VSyS using multiple unrelated data streams to improve the parameter and model estimation performance and provide multiple changepoints analysis with uncertainty measures, providing decision-makers with an estimate of the onset of disease outbreak with credible intervals.

Therefore, the proposed method will provide

a. ability to measure uncertainty in the estimated parameters, hyperparameters, and the model, and the ability to improve their estimation performances.
b. reduction of uncertainties in parameter and model estimation
c. changepoint analysis to discover the onset of disease outbreak with credible intervals
d. updating prior information as more data streams become available
e. ability to incorporate multiple unrelated data streams for outbreak detection

As the author's contribution to knowledge, the detailed step-by-step guide for implementing the method as a risk assessment tool will be produced as part of the outputs of this research work.

However, the author encountered some challenges during the data extraction phase. We had initially thought that Twitter data was free to extract; it turned out that the free method of

extracting twitter datasets through web scraping is illegal and unauthorised (Twitter, 2020). So, the author had to identify a legal means of obtaining the Twitter data. The legal method came with a subscription cost as the researcher had to pay for access to Twitter search API for bulk data extraction. Therefore, the researcher is unaware of a free method to access the Twitter API for research purposes.

The initial plan was to use the datasets from the company I work. However, it was discovered that the datasets were not large enough and were homogenous. On the other hand, the SAVSNET dataset required a rigorous application process, involving the research ethics committee approval at SAVSNET under the University of Liverpool. The entire process for the application was rigorous and challenging. Furthermore, understanding SAVSNET datasets came with challenges as it was very difficult to understand the meaning of the variables in each data stream. The assistance of a senior lecturer from the school of veterinary medicine at the University of Liverpool was required to interpret the variables and advise on the best approach to construct the proposed solution. Also, the size of the datasets was another challenge encountered. The practice dataset was too large to access through a notepad or a CSV viewer. As a result, Panda's data frame procedure was applied to access and visualise the contents of the data files. The files were saved in multi-factor authentication and passworded cloud accounts to secure the data throughout the research investigations.

4.5 Conclusion

In this chapter, the author provided a list of research publications that used the SAVSNET datasets or a subset of the data for a previous research study, including their main contribution and applied data science techniques. Although the list was extracted from the Scopus repository using the systematic approach of searching the academic database, the author acknowledges that the publication list may not be exhaustive as additional articles may not have been discovered using the selected keywords.

Based on the objectives of the selected research articles and their knowledge contribution, it can be concluded that the datasets can be applied in this study since they have been previously used to successfully deliver various VSyS approaches in other research work. Also, they demonstrate typical characteristics of the disease surveillance dataset described by past studies and established in chapter 6 of this thesis.

Furthermore, all the datasets acquired from SAVSNET are useful because they contain datasets of real GI infection outbreaks in dogs in the UK, which can be used to validate the proposed technique in this study. The proposed technique will be developed and tested with the Twitter data stream. Therefore, having access to datasets from real epidemics will assist in validating and comparing the ability of the proposed method to identify regions of high and low risks in the surveillance datasets and whether it corresponds with the same changepoints in the Twitter data stream.

Chapter 5 – Research Methodology

5.1 Chapter Introduction

Harding (1987a) and Harding (1987b) defined research methodology as the rationale, theory, and analysis of a research process. According to these authors, the research method is how evidence is gathered and analysed. Likewise, Abutabenjeh and Jaradat (2018) described the research methodology as a technique and process that begins with broad assumptions and progresses to specific data collection, analysis, assessment, and interpretation methods to produce a research outcome.

Based on the above definitions, the overall research methodology for this ongoing research work is presented in this chapter. The justifications for selecting the overall research methodology are also highlighted, with detailed strategies for developing the proposed SQRA method, including all the steps involved, as outlined in Figure 12. Various data science procedures and software packages for solving the identified research problems were critically evaluated to select the most appropriate method. The author will rely on the datasets presented in the previous chapter to achieve the overall aim of this study since

- the dataset represents real and unspecific animal health data collected from veterinary sites in the UK, consisting of clinical and laboratory data streams
- it consists of the online media data stream – the Twitter dataset
- the collection period of the data streams contains the actual periods of known disease outbreaks to validate the correctness of the proposed method.

The flow diagram of the steps undertaken in the research methodology to obtain the proposed SQRA method is presented in Figure 12, and each stage of the methodology is critically addressed individually in the subsequent chapters.

5.2 Consideration of Data Science Techniques

In this study, the author investigated data science techniques as the first approach toward obtaining the appropriate methods for the SQRA in VSyS. According to Gruson et al. (2019), data science is a human-centred activity that extracts knowledge from complex datasets to generate insights. Vogelius et al. (2020) presented data science as a multidisciplinary field that

uses scientific methods, processes, algorithms, and systems for insights or knowledge extraction from structured and unstructured data.

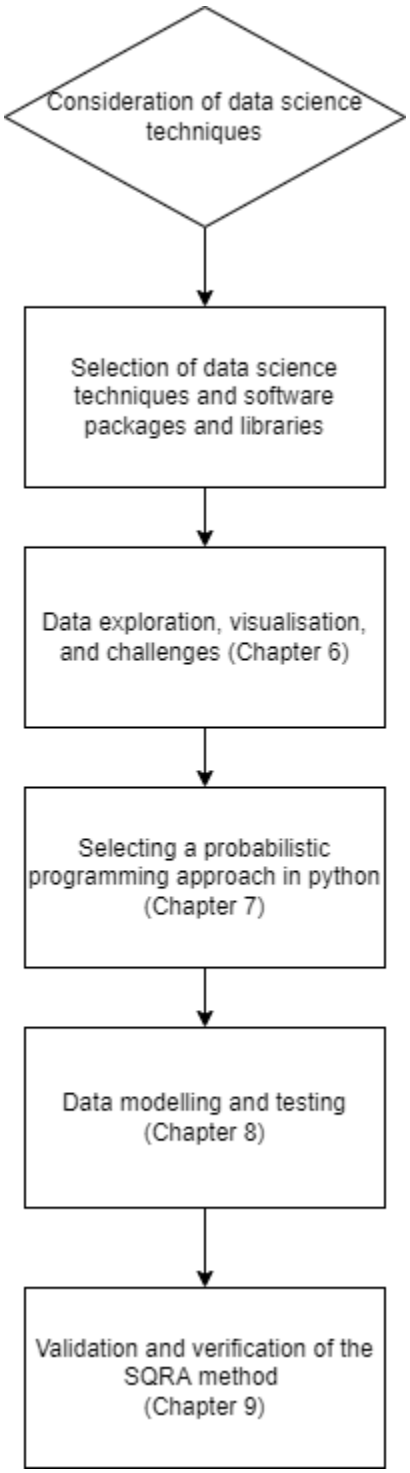

Figure 12: *Flowchart for steps undertaken in the methodology to obtain the SQRA technique*

Likewise, Sanchez-Pinto et al. (2018) described data science as fundamental principles that guide and support the systematic extraction of knowledge and information from datasets.

Furthermore, they described data mining as the actual extraction of knowledge from data using machine learning algorithms or statistical methods that incorporate data science principles. Hand (2007) pg.6 described data mining as a subset of data analysis and a process consisting of many methods and models. In this study, the author considered the concept of data analysis as part of the data science principles, providing the necessary models and methods for obtaining the SQRA technique.

Though there are many data science descriptions due to its infancy stage, the different descriptions suggest combining data science principles, models, and analytical methods to undertake a data science activity. According to Brodie (2019), data science principles include inductive and deductive reasoning, objectivity or lack of bias relative to a given factor, provenance, reproducibility, collaborative and cross-disciplinary methods. Most importantly, data science principles involve understanding the application of scientific principles to data discovery and the underlying evidence-based reasoning for planning and enhancing policy and decision-making. They further explained that a data science activity uses one or more models to represent the parameters that are the critical properties of the phenomenon to be analysed. It often takes multiple models to capture all relevant features in a study. They highlighted the use of analytical methods within a data science pipeline to evaluate specific data features under investigation, such as uncertainties and variability.

The main principle of organising a data science activity is workflow or data science pipeline and its life cycle management (Berman et al., 2018). This is the end-to-end sequence of methods/steps from data discovery to the interpretation and communication of the result in a form that is clear to the domain users. Likewise, Brodie (2019) argued that the state of the art in data science is such that each activity requires a unique pipeline that meets the main objective of the endeavour, as each data science activity is unique. For example, Yousefinaghani et al. (2019) research relied on a data science pipeline that supports the online analysis of social media datasets for VSyS development.

Therefore, in this study, the author considered the concept of data science principles, models, and analytical methods and relied on a data science pipeline to achieve the aim of this research study. The below activities and design of the data science pipeline for the SQRA technique are adapted from the generic workflow presented by Brodie (2019) and are shown in Figure 13 but modified in Figure 14 to meet the objectives of this research work.

1. Primary data sources acquisition, preparation, and storage
2. Selection and acquisition of curated data from data repositories for data analysis
3. Data analysis (Pre-processing, Visualisation, Modelling, and Validation)
4. Results interpretation
5. Result publication or communication

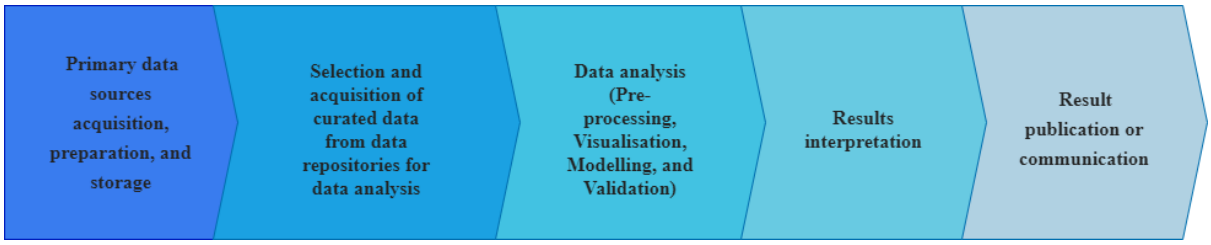

Figure 13: *Generic data science pipeline adapted from Brodie (2019)*

The author identified two common approaches for implementing VSyS techniques in the literature – the Frequentist and Bayesian techniques. Most Frequentist methods rely on deviation from baselines to define syndromic alarms using a point estimate. Also, gaps in knowledge were identified in the literature review chapter where alarms generated by most Frequentist techniques are qualitative risk measures and could not discriminate between areas and regions of high and low risks in data points. However, the literature evidence indicates that quantitative risk estimation requires uncertainty measures (Cardona et al., 2012; Smith et al., 2017; Faverjon, 2017). According to Salvatier et al. (2016), point estimate methods commonly used in the Frequentist techniques, such as maximum likelihood and confidence limit, might find uncertainty estimation difficult since they lack uncertainty measures.

Furthermore, a growing interest among researchers in developing a VSyS that can consume multiple data streams was identified due to scholars realising that more insights can be derived from heterogeneous data sources to account for multiple factors that may be responsible for disease outbreaks. However, the Frequentist technique may find this approach complex due to inconsistent availability of veterinary datasets and data scarcity (Faverjon et al., 2016) since data from different veterinary sources for disease surveillance are difficult to acquire simultaneously. As a result, the author proposed the Bayesian probabilistic programming approach as the main data mining technique for developing the SQRA technique since they can combine different data sources for modelling even when they are not available simultaneously.

Also, they can provide quantitative measures of model parameters with uncertainty measures (Salvatier et al., 2016).

5.2.1 Bayesian Probabilistic Programming

It is very common for disease surveillance activities to produce small observations for syndromic analysis due to data challenges identified in the literature review chapter. Therefore, Faverjon et al. (2016) argued in favour of using a modelling technique that can express both the random characteristics of small datasets and the presence of uncertainties in the data and the model to improve the analysis of VSyS. Likewise, Geweke and Amisano (2010) highlighted the performance of Bayesian inference; they argued that the modelling technique is efficient in providing time-series datasets with predictive distributions that fully and coherently incorporate parameter uncertainties. Some earlier studies also indicated that the Bayesian inference technique could be useful for fitting models where the analytical problem has small observations or where the number of cases reported in each time unit may be too small to trigger an alarm (Struchen et al., 2017; Faverjon et al., 2016; Hedell et al., 2019). However, Briggs et al. (2012) argued that the extent to which an uncertainty analysis can be considered a fit for purpose depends on the decision(s) the model seeks to support.

One of the challenges identified in the literature analysis of VSyS is that the datasets used for modelling may not be available altogether for investigating real-time outbreak detection, considering the significant variability in the relative timeliness of the animal health data sources. Another difficulty may arise when faced with slowly increasing outbreaks. A method that can allow data ingestion whenever data is available for a seamless model update is considered beneficial in this study. Struchen et al. (2017) argued that the Bayesian approach offers a straightforward and transparent way of incorporating such information as priors while updating the model each time new data is available.

Bayesian statistical inference is an approach for updating the beliefs about an outcome after considering new evidence; it is different from traditional statistical inference because it preserves uncertainty in both the data and the model (Pilon, 2015). In Bayesian modelling, probability is interpreted as a measure of belief or confidence in an event occurring. For example, if there is an observation $y_1....y_n$ denoting the count data of daily vomiting incidents in the dog's population, many diseases could have caused vomiting symptoms in the population. In a Frequentist approach, spikes above the baseline generate an alarm to represent possible health incidents. However, alarms near the baseline may not be seen as posing a threat

of an outbreak. In the Bayesian approach, a public health officer might believe an outbreak of disease needs further investigation depending on the distribution of the available datasets. Therefore, Bayesian methodology allows such beliefs to be incorporated into data modelling (Andersson et al., 2014).

In this study, to align Bayesian belief with traditional probability notation, the author assigned a parameter to express the belief that there was an ongoing disease outbreak. This parameter is named the prior probability distribution in Bayesian inference and was expressed in equation 5.1 as

$$f(\theta) \tag{5.1}$$

After introducing new observations, the belief was updated and subsequently interpreted as the probability distribution of the disease outbreak given the new evidence y. The new evidence y is the weekly frequency counts of the GI disease obtained from Twitter and SAVSNET datasets described in Chapters 4 and 6 of this thesis document. When y is introduced into the model as new evidence, the prior belief will be updated to the posterior probability distribution and can be expressed mathematically as

$$f(\theta|y) \tag{5.2}$$

This is the underlying theory of Bayesian methodology. However, it is worth noting that the Bayesian model does not completely discard the prior belief in equation 5.1 after seeing new evidence y, but it re-weights the prior and incorporates the new evidence, which might mean putting more weight, or confidence, on some beliefs than others (Hedell et al., 2019; Pilon, 2015; Ray et al., 2011).

Therefore, by introducing prior uncertainty about events, the model admits that any guess might be potentially wrong. After observing new data, the evidence and other information will update the model's initial belief, making a guess less wrong. In this state, the Bayesian analysis preserves the uncertainty, reflecting the instability of statistical inference of a small data problem. Updating the belief was done via the following equation, known as Bayes' theorem, which was named after its discoverer Thomas Bayes:

Assuming a sample of observations y_1,.....y_n of a random variable is expressed as equation 5.3

$$Y \sim f(y|\theta) \tag{5.3}$$

Where θ is a parameter of the observations' distribution and is considered a random variable. Following the Bayes theorem, the relationship can be re-written, and the model expressed as equation 5.4

$$f(\theta|y) = \frac{f(y|\theta)f(\theta)}{f(y)} = \frac{f(y|\theta)f(\theta)}{\int f(y|\theta)f(\theta)d\theta} \quad (5.4)$$

Where f(θ) is the prior distribution of θ – this is the strength of the belief for θ without considering the evidence, the set of observations y.

The function f(y|θ) is called the likelihood – this is the probability of the evidence generated by a model with parameter θ.

The function f(θ|y) is the posterior distribution, which indicates the refined strength of the belief of θ once the evidence (new observations) have been considered.

Furthermore, f(y) indicates the evidence in the form of observations – it is the probability of the observations as determined by integrating across all possible values of θ.

Note that the f(y) does not depend on θ. Therefore, it can be considered a normalising constant. Also, it is often the case that the integral in equation 5.4 is not easy to compute. However, the relationship can be simplified and expressed as equation 5.5.

$$f(\theta|y) \propto likelihood\ x\ prior \quad (5.5)$$

The MCMC algorithm is a random-walk-based data sampling technique for generating probability distributions. The algorithm has greatly simplified the practical application of Bayesian statistics by overcoming the need for closed-form analytical integration of equation 5.4 (Westera, 2021). As a result, the author relied on an MCMC algorithm to compute the probability distributions of all model parameters in this study.

Furthermore, the author considered the difficulties of implementing the Bayesian approach as it is generally assumed that the MCMC sampler will theoretically converge. However, there is no guarantee that this will happen in a real-case scenario and within reasonable computational time (Yau and Campbell, 2019). For large observations, the Frequentist statistical inference could be more objective. On the other hand, Bayesian techniques could be more intensive due to integration over many parameters in the model. However, the Frequentist inference may be much more unstable for small observations since their estimates have more variance and large confidence intervals (Allenby et al., 2014).

According to Lázaro-Gredilla and ES (2014), some techniques can reduce the computational intensity by using conjugate priors or approximating the posterior distribution using sampling methods or variational inference. Also, Yau and Campbell (2019) outlined the computational frameworks for implementing Bayesian inference in practice, indicating the use of the Bayesian learning technique to analyse high-dimensional big datasets. In this study, the author considered an approach that provides means of diagnosing biases in the model to determine if the model is wrongly specified since it can be argued that handling such biases is important to model validation. If the model takes an extended computational time to fit, it indicates that it is poorly specified. A systematic workflow for diagnosing biases in the model was considered as part of the overall technique in the modified data science pipeline in Figure 14. The three steps constituting the elements of the workflow are;

1. Build the model
2. Identify divergences in the model by performing a diagnosis of the model
3. Re-parametrize the model as needed.

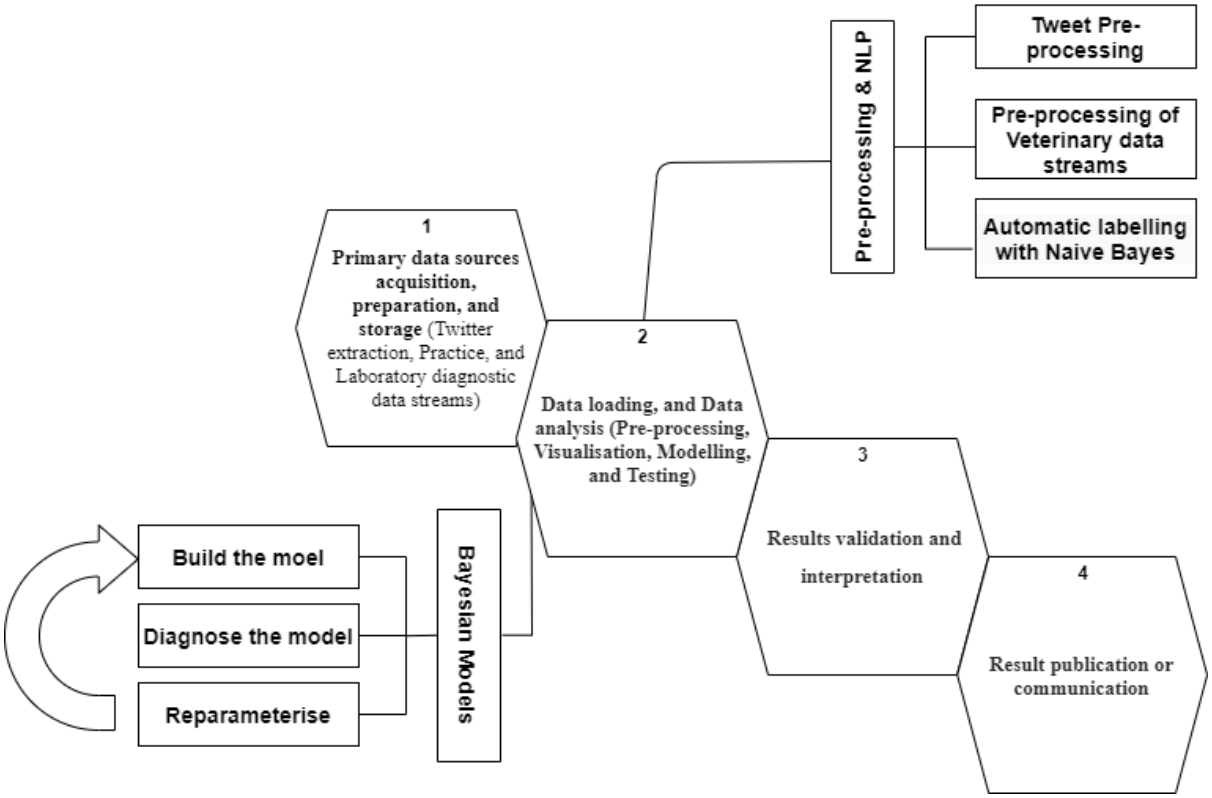

***Figure 14**: Modified data science pipeline for the proposed SQRA method (Adapted from Brodie (2019))*

5.2.2 Changepoint Analysis

Changepoint research has been used to solve time-series problems for several years. The concept was inspired by Page (1963) research on the sequential identification of a shift in the pattern of quality control data collected during the manufacturing process. The concept is now an established method for efficiently understanding critical information based on abrupt changes in data from a variety of disciplines, including environmental data, climatology, oceanography, and biological research studies (Gold et al., 2018; Killick et al., 2010; Reeves et al., 2007; Jarušková, 1997). Changepoint detection can be classified into online and offline methods. Online changepoint analysis can identify abrupt changes near the most recent observation, while offline changepoint analysis can perform a one-off analysis on a historical time-series dataset (Truong et al., 2020).

However, according to Li et al. (2019), the online changepoint detection techniques target data requiring instantaneous analytical responses, but the offline detection approach targets data that requires historical analysis, leading to more accurate detection of events. As a prerequisite for success in this study, the SQRA method must be capable of handling multiple data streams, including historical and real-time datasets.

Changepoint detections are also referred to as anomaly detections, quality control segmentation, structural change detections, breakout detections, and event detections in the literature. According to some studies, it is referred to as changepoint mining because it requires data mining techniques to extract information from the data (Boettcher, 2011). For example, anomaly detection can identify outliers while quality control focuses on the mean and standard deviation's stability, and the changepoint estimation interprets changes to the mean.

Many studies have investigated, classified, and compared changepoint methods. Likewise, several reviews of changepoint methods have been conducted by previous researchers (Tartakovsky and Moustakides, 2010; Li et al., 2019; Reeves et al., 2007). The changepoint methods discussed in previous studies include regression, maximum likelihood estimation, and kernel methods, each of which has a limit on the number of changes it can detect (Truong et al., 2018b). As a result, no additional literature review on changepoint analysis was conducted for this research study. However, many of the detection procedures described in previous reviews are available in Python via the ruptures package and in R via the Strucchange and

Changepoint packages, which are the most comprehensive changepoint detection libraries in both languages.

5.2.3 Implementing Changepoint Analysis

Catastrophic health events may originate from different environmental and biological sources; for example, the Coronavirus 2019 pandemic is believed to have started from a wet market. As a result, risk effects could be mitigated if signs and symptoms were detected from VSyS on time for public health officials to take the necessary actions. However, when using data from multiple veterinary data streams, such as those derived from non-specific health event data sources, identifying and interpreting risk events can be extremely challenging due to the limitations of the current alarm-based VSyS techniques discussed in chapter 3 of this study.

Since a public health risk is an event or an action that may likely cause harm to human or animal health or contribute to disease among humans or animals, such as infections, bioterrorism attacks, and germs (Queensland Health, 2019); as a result, the public health risk is a function of the probability of a catastrophic event occurring. Then, there is the need for a system that could complement the existing alarm-based techniques to interpret areas of high and low-risk regions or a slowly occurring risk. To the best of the author's knowledge, no research publication has proposed a decision support system to complement VSyS and interpret alarms quantitatively while at the same time applying a changepoint analysis.

The author proposed changepoint analysis as part of the method for risk identification using unspecific clinical and non-clinical datasets for interpreting disease outbreaks. Also proposed that the changepoint should be investigated with Bayesian probabilistic programming. This is because the Bayesian probabilistic programming technique can explore changepoints in datasets (Niekum et al., 2015; Tartakovsky and Moustakides, 2010; Salvatier et al., 2016). Furthermore, the author considered the approach of parameterising changepoint within the Bayesian approach to obtain the most appropriate technique for implementing the SQRA method in this study.

5.2.4 Selection of Software Platform for the Proposed SQRA Method

Various software packages and platforms were investigated to identify the most appropriate software libraries for constructing the data science pipeline for the SQRA method. The software packages investigated include Notepad++ 7.9.5, Notepad, TXTCollector 2.0.2,

Microsoft Excel 365, Microsoft Excel combined with @Risk 8.1, PyCharm Professional Anaconda 2019.3.2, R Studio 3.6.2, and Google Colab.

The author applied a subset of the practice data stream to test each software package listed above on a laptop equipped with dual Intel core 7i processors, 16GB RAM, and a Windows 10 x64-based operating system. However, attempts to perform different analysis tasks shown in Table 9 were unsuccessful due to software and computer memory constraints. Table 9 contains a list of all software packages, a description of the tasks, and the identified challenges.

Google Colab with Python was selected for this research work due to the limitations experienced in other software packages investigated. Also, Google Colab is a research platform that allows developers to build machine learning models on powerful hardware such as GPUs and TPUs in the Google cloud. Moreover, it is free to use and provides a serverless Jupyter notebook environment for interactive development (Bisong, 2019). Additional reasons for selecting Python as the programming language and Google Colab as the programming platform include the following:

a) Python and Google Colab are open-source platforms with no cost associated with their use.
b) The author had previously used Python during the research preparation. Also, the author attended many trainings focused on software development in Python and Google Colab for big data analysis. Both Python and Google Colab proved more efficient with few challenges compared to other software packages identified above.
c) The author also used the Python language for previous projects. Therefore, the author is comfortable programming in Python. In addition, Python has well-organized libraries for Bayesian and changepoint analysis. It is portable across multiple computer operating systems and comes with extensive help documentation and a thriving online community. (Patil et al., 2010; Kumar et al., 2019; Truong et al., 2018a).

Tables 10 and 11 summarise the changepoint and Bayesian probabilistic programming libraries available in the Python platform, respectively, along with a brief description of their application, which the author will consider when selecting appropriate packages for the proposed SQRA method. The choice of PyMC3 over Pystan, PyJAGS, Pyro and TensorFlow probabilistic techniques for obtaining the SQRA method is further evaluated in chapter 7.

Table 9: *Software packages considered for data pre-processing and analysis and their task description*

Software Packages	Action	Advantages Observed	Observed Challenges
Notepad	Text file loading and editing	Lines within the data file can be edited and deleted manually	Because notepad does not support batch editing, applying it to the research work was time-consuming as it had to be done manually on individual text files.
TXTCollector 2.0.2	Consolidate all of the folder's text files into a single editable file	Capable of merging files in a folder	Error messages appear because it could not merge all files.
Notepad++ 7.9.5	Text file deletion and merger	Simple to use for deleting lines within each file	This programme cannot delete multiple lines within a file. Its implementation in this study was extremely time-consuming, as it necessitated manually deleting several lines from each file contained within the folder. As a result, implementation was inconvenient.
Microsoft Excel 365	To merge and edit Excel files	It was relatively simple to select and delete data.	Due to memory limitation and row limit of less than 1,048,000, the application could not combine all Excel files for editing.
Microsoft Excel combined with @Risk 8.1	To combine and edit files in Excel for implementing SQRA	Easy to select columns and define a parameter for MCMC implementation.	Memory capacity was an issue, and the application could not separate random and uncertainty parameters during MCMC sampling. @Risk application is proprietary software, and it requires a subscription to use.
R Studio 3.6.2.	Considered for implementing the SQRA technique	Easy to select and manipulate multiple datasets for analysis and to implement the SQRA technique	Memory and performance limitations were experienced while using the application on a computer. It is not easy to implement a cloud version of the application to leverage performance hardware with GPUs and TPUs in the cloud. The author is not proficient in using R statistics for programming.
PyCharm Professional Anaconda 2019.3.2	Considered for implementing the SQRA technique	Easy to manipulate datasets and can be used with Python.	Several performance bottlenecks were experienced due to local PC memory and CPU limitations, leading to a lack of capability to handle large datasets.
Google Colab with Python	It was tested with data pre-processing, data analysis, machine learning modelling, and programming.	Suitable for most data pre-processing and data analysis tasks. Easy to manipulate data and implement the SQRA methods	The platform requires access to reliable and high-speed internet connectivity.

***Table 10**: Changepoint analysis packages in Python*

Changepoint Package/Library	Usage	Summarised Description
ruptures	It obtains offline changepoint using mean changes in datasets.	The ruptures package is designed to perform offline changepoint analysis using algorithms designed in Python language.
changepoint	It performs both single and multiple changepoint by implementing the mean-shift model in time-series data.	The package is developed in R language but provides bindings to some functions to perform changepoint in Python. The package is advanced in R language (Killick and Eckley, 2014).
changefinder	The library is used for online changepoint detection	This library is based on a changefinder algorithm to implement changepoint in time-series datasets.
Bayesian Changepoint Detection	This is a method for getting the probability of a changepoint in time-series data and is used for online and offline detection. Use the temporal changes in the time-series datasets to identify potential changepoints using the predefined user specifications. Available for online and offline changepoints (Salvatier et al., 2016).	The method uses a simple message-passing algorithm directly in Python (Adams and MacKay, 2007). Other approaches use the inferring mechanism of Bayesian probabilistic packages/libraries in Python.

***Table 11**: Probabilistic Programming libraries in Python*

Bayesian Packages / Libraries	Usage	Summarised Description
Pystan	It is used for basic inferential modelling to advanced Bayesian probabilistic programming. Flexible for user-defined parameters and hyper-parameters (Carpenter et al., 2017; Koprivica, 2020).	It is a Python interface to Stan probabilistic platform used for basic to advanced Bayesian inference analysis. It is also capable of variational inference
PyJAGS	It is used for Bayesian inferential modelling and advanced statistical modelling in Python (Plummer, 2003).	PyJAGS is a Python interface to JAGS and provides cross-platform access to the BUGS language. It is a probabilistic programming language for implementing Bayesian inferences.
PyMC3	It is a library for automatic inference based on user-defined parameters, allowing researchers to incorporate assumptions, prior information, and uncertainties.	The PyMC3 is a probabilistic programming framework written in Python that uses Theano to compute gradients via automatic differentiation and compile probabilistic programs on the fly to C for increased speed (Dehning et al., 2020).

Pyro-ppl	Pyro is used to investigate randomly distributed parameters and sample data and perform efficient inference. It implements generic probabilistic inference algorithms such as the No U-turn Sampler, a variant of Hamiltonian Monte Carlo (Bingham et al., 2019).	Pyro is deep probabilistic modelling and focuses on variational inference, backed by PyTorch and developed with Python. It allows users to programme in Python and integrate seamlessly with PyTorch.
TensorFlow Probability	This framework is also called Bayesian Deep Learning, and it uses the TensorFlow Probability framework to investigate randomly distributed parameters and all sorts of statistical methods used to quantify the uncertainty of the prediction. It also focuses on variational inference.	The framework consists of Python, TensorFlow, and TensorFlow probability. It presents the concept of uncertainty in the Artificial Neural Network (ANN) by making the output of the ANN and the weights as random variables (Fanfarillo, 2019).

5.3 Conclusion

The author presented the justification for selecting the appropriate software platforms and packages for developing the proposed SQRA method in this study. It includes the comprehensive data pre-processing and analysis strategies and the justifications for selecting changepoint analysis and Bayesian probabilistic modelling approaches as the core data mining techniques for the proposed SQRA method.

The motivation for using Python is based on the availability of many strong libraries for data pipeline implementations, data analysis, database interaction, and plugins offering interaction with healthcare-related datasets. Also, Python's choice is strongly driven in this work by the author's positive experience as a general-purpose programming and development language. It can be easily installed, and applications can run reliably on all conventional operating systems without changing source codes. Therefore, it is easy to implement a data pipeline using Python to support different scenarios in the cloud. Python has many open and large global communities, with millions of software developers providing online and offline interactions to support users and enthusiasts. Furthermore, Python provides robust support for developing specialised data science pipelines for analysing social media data streams with capabilities for natural language processing, as demonstrated for cleaning the Twitter data stream in this research work.

The following chapter is a continuation of the methodology used to develop the proposed SQRA technique in this study.

Chapter 6 – Data Exploration, Visualisation, and Challenges

6.1 Chapter Introduction

For public health officers to effectively identify which health threats indicate a true outbreak, the ability to "drill down" from generalised data summaries into more specific data analysis is necessary. Therefore, visualisation techniques can enhance the timely investigation of disease outbreaks by generating custom graphs, maps, plots, and spatial-temporal investigation of specific syndromes or data sources.

Although VSyS can automatically detect statistical anomalies in disease surveillance datasets by producing alarms that might need epidemiological investigation. Human understanding is required to interpret these alarms and separate statistically significant events from epidemiologically unimportant ones. To this end, a practical understanding of the datasets through multiple data visualisation techniques is important to aid epidemiologists in deciphering health events in a timely and cost-effective manner. Data visualisation and exploration present the initial capacity for understanding the characteristics and attributes of the datasets under investigation. Also, it assists researchers in selecting appropriate modelling techniques based on the understanding of data characteristics.

In this chapter, the author continues with the research methodology for obtaining the SQRA technique by providing insight into the shapes, structure, and attributes of the datasets described in chapter 4 through visualisation and exploratory techniques. The aim is to understand the underlying data distributions of each data stream to establish the best approach for modelling them. The challenges encountered while visualising the datasets and the solutions applied are also discussed. As mentioned in chapter 4 of this study, the author accessed data from multiple sources, including Twitter, veterinary practices, and diagnostic laboratories, to demonstrate a multi-data streams VSyS approach. Information on the processed data indicates that the Twitter data file has 2,623 rows and 6 columns. The diagnostic laboratory data has 133,423 rows and 5 columns, and the veterinary practice data file has 2,075,438 rows and 7 columns.

6.2 Method

Using the Python Pandas library, each data stream was selected and loaded in Google Colab for investigation. All datasets are time-series data. Therefore, the author decided to sample the practice and laboratory diagnostics data streams by selecting QuestionnaireAcuteVomiting and CampylobacterCulture variables, respectively and calculated the sum of positive cases per week in the diagnostic data stream and the sum of "true instances" in the practice data stream. It was assumed that the attributes of the selected variable from each data stream represent the characteristics of that data source. This approach was undertaken to avoid repeating the same procedure for each variable in the data stream. The author applied the following data visualisation and exploratory analysis techniques to understand the underlying data distributions and characteristics of the selected datasets.

1. Descriptive statistical tools such as summary statistics, frequency distribution plots, histograms, and box plots.
2. Time-series decomposition and analysis such as trend analysis, seasonality, noise components, autocorrelation and partial autocorrelation analysis.

These methods have been used successfully in various research studies and in different domains to understand the underlying structure, shapes, and attributes of time-series datasets (Huybrechts et al., 2014; Hale et al., 2019; Kass-Hout et al., 2012; Tian et al., 2015). For example, Pivovarov et al. (2014) used histograms to explore patients' vital signs measurement gaps. Also, Guerrero et al. (2017) used descriptive statistics and histograms to describe data distribution. Likewise, Polonsky et al. (2019) argued that the most important point to begin a data analysis task is data exploration and visualisation because it rapidly helps to uncover some interesting characteristics of the data under investigation.

As the first step, the author applied descriptive statistical tools to organise and summarise each data stream with the aim of gaining a quick understanding of the dataset and its variables. The concept is useful for uncovering patterns in the datasets and finding their measure of central tendencies, such as finding their averages or mean values and establishing whether a dataset is affected by a large variance. The effect of a large variance in a dataset is overdispersion which often causes difficulty in modelling disease surveillance data accurately (Mouatassim and Ezzahid, 2012).

The author performed autocorrelation measures to inspect the presence and influence of correlation on the datasets. Autocorrelation is the correlation between the dataset and a delayed copy of itself, and it measures repeating patterns or periodic signals in a time-series dataset. It was argued in previous studies that autocorrelation or partial autocorrelation could affect the robustness of changepoint analysis in time-series data if not addressed (Kass-Hout et al., 2012). Therefore, the author applied the first 40 lags of each time-series dataset on itself in figures 25 to 27 to investigate the autocorrelation and partial autocorrelation effects on the data streams.

The time-series line plots in figures 18 to 20 show irregular spaced spikes indicating random effects such as seasonality, trends, and noise in the data streams. Therefore, the author applied time series decomposition tools to each time-series dataset to provide insights into their structures by inspecting the three components — seasonality, trend, and noise. Seasonality describes the periodic signal in the data; a trend describes whether the time series is decreasing, constant, or increasing. Noise describes the residual, which is the unexplained variance and volatility of the time-series datasets (Pettit et al., 2017).

```
# Sample the diagnostic data stream by AssayName - in this case,
# Select diagnostic samples tested for CampylobacterCulture only
diagnostic = diagnostic.loc[diagnostic["ASSAYNAME" ] ==
"CampylobacterCulture" ]
# replace labels with numbers for ease of analysis
repl_label = {"RESULTVAL": {"POSITIVE": 1, "NEGATIVE": 0}y}
diagnostic.replace(repl_label, inplace=True)
diagnostic = diagnostic[["resultdate" , "RESULTVAL" ]]
diagnostic_0 = pd.Series(diagnostic["resultdate" ])
diagnostic_0 = pd.to_datetime(diagnostic_0 )
diagnostic_0.sort_values(ascending=False)
diagnostic_0 = diagnostic_0.dt.strftime("%d-%m-%Y" )
diagnostic = diagnostic.assign(resultdate=diagnostic_0 )
diagnostic = diagnostic.set_index("resultdate" )
diagnostic.index = pd.to_datetime(diagnostic.index)
# rename the RESULTVAL column
diagnostic.rename(columns={"RESULTVAL":"CampylobacterCulture" },
inplace=True)
```

***Figure 15**: The solution applied to format the date column in the procedure*

All the datasets used in this research study required the date column to be formatted in DateTime format and assigned to the dataset index. The aim was to ensure the datasets were ordered over a time interval for ease of modelling. The sample of the Pandas DateTime function applied to arrange and format the date column in the diagnostic data stream is shown in Figure 15. A similar procedure was applied to other data streams in this research work.

6.3 Result

The summary statistics of all the data streams showed no missing observations, and there were no nulls or NAs in the datasets. The Public Health England adopts daily or weekly aggregation of disease surveillance datasets when tracking or analysing infectious disease outbreaks (Noufaily et al., 2019). Also, this approach is commonly adopted by researchers in the literature to avoid the frequently occurring zero counts in the daily datasets (Schmidt and Pereira, 2011). The zero-frequency counts were also observed in the data streams available for this research work. Therefore, the author resampled each data stream into weekly aggregate data frames using the Pandas library.

From the time series plot in Figure 16 and the result of the summary statistics of the Twitter data stream in Table 12, the maximum and minimum data points are 17 and 1 tweet, respectively. It shows many high and low points ranging between 1 and 12 tweets in most weeks, with a very high data point reaching 17 tweets between February and March 2020. The mean value is 6.69 tweets: suggesting an average rate of 6.69 incidents per week from the Twitter data stream. The duration of the Twitter data stream is 118 weeks.

The practice data stream contains outbreak indicators to monitor the GI disease, and the dataset was aggregated into weekly cases where the animal came into the clinic due to acute vomiting. From the time series plot in Figure 17 and the result of the summary statistics of the QuestionnaireAcuteVomiting data stream in Table 12, the maximum and minimum data points are 162 and 26, respectively, and the mean value is 89.34, which indicates an average of 89.34 incidents per week. The duration of the data stream is 105 weeks.

The laboratory diagnostic data stream was sampled with the CampylobacterCulture assay dataset, and it was investigated by calculating the sum of positive cases per week. From Figure 18 and Table 12, the maximum and minimum weekly positive counts of the CampylobacterCulture assay are 267 and 0 incidents, respectively. The mean value is 109.25, indicating an average of 109.25 incidents per week. Also, the duration of the datasets is 165 weeks.

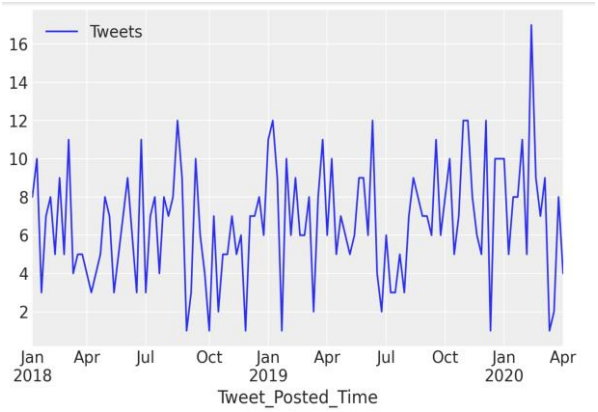

***Figure 16**: Time-series line plot of the Twitter data stream (weekly aggregate)*

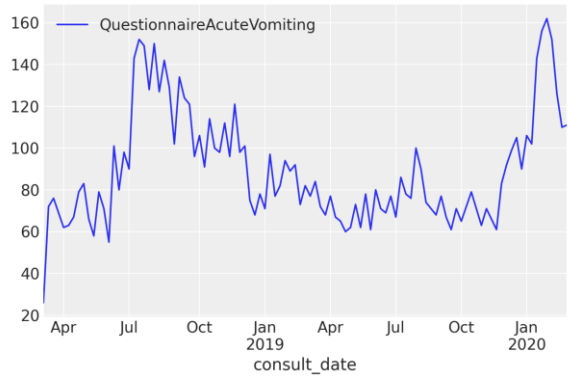

***Figure 17**: Time-series line plot of the QuestionnaireAcuteVomiting (weekly aggregate)*

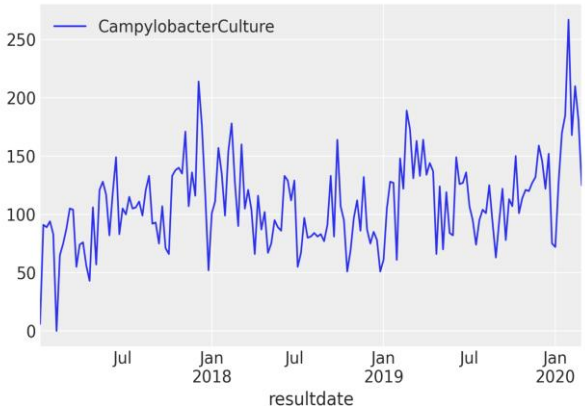

***Figure 18**: Time-series line plot of the CampylobacterCulture assay (weekly aggregate)*

***Table 12**: Summary statistics*

Tweets		QuestionnaireAcuteVomiting		CampylobacterCulture	
Count	118	Count	105	Count	165
mean	6.69	Mean	89.34	Mean	109.25
std	3.01	Std	26.77	Std	37.73
variance	9.01	variance	716.50	variance	1423.61
median	7.0	median	80.00	median	106.00
max	17	max	162	max	267
min	1	min	26	min	0

From the summary statistics in Table 12, the large differences between the mean values of the practice and diagnostics data streams and their variance values indicate that they are substantially overdispersed, which is common in epidemiological or disease surveillance datasets (Schmidt and Pereira, 2011; Zhou et al., 2012).

Using the histogram plots in figures 19 to 21, the author investigated the frequency distributions of each data stream, and the results indicate non-symmetrical unimodal distributions. Furthermore, the measure of central tendencies shows that the frequency distributions for all the data streams are either skewed to the right or left, which means they are not normal distributions. The data distribution is further investigated and verified in chapter 8. As each data stream becomes skewed, the mean cannot explain the central tendencies because the skewed data drag the mean away from the typical central value.

However, the median value remains the most informative as it better explains the central tendencies of the data stream and is not strongly affected by the skewness of the data compared to the mean values (Manikandan, 2011). This trend is explained by the histogram and the summary statistics in Table 12. Furthermore, the standard deviations of the data streams indicate wider frequency distribution in the practice and diagnostic data than the Twitter data, which explains the spread of each data stream.

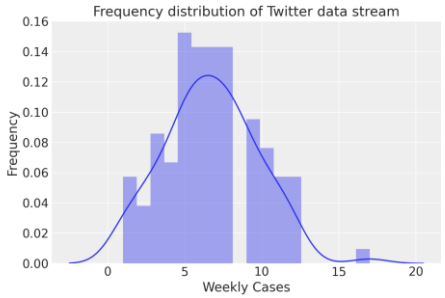

***Figure 19**: Histogram and KDE that describes the frequency distribution of the Twitter data stream*

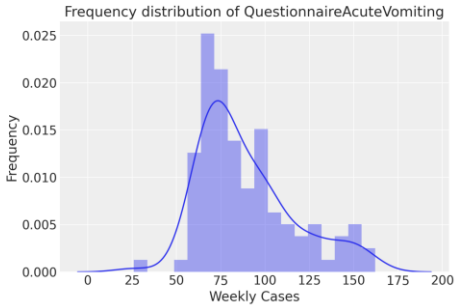

***Figure 20**: Histogram and KDE that describes the frequency distribution of the QuestionnaireAcuteVomiting dataset*

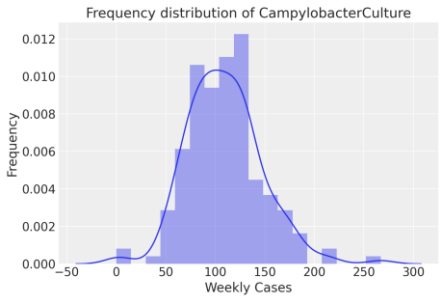

***Figure 21**: Histogram and KDE that describes the frequency distribution of the CampylobacterCulture dataset*

The spread of each dataset was investigated using the box plots as indicated in figures 22 – 24 to establish the presence of outliers. The result shows more extremes and outliers in the upper values of the frequency distribution of the QuestionnaireAcuteVomiting data, followed by the extremes and outliers in both the upper and lower frequency distribution values of the CampylobacterCulture data. The Twitter datasets have a single outlier value in the upper-frequency distribution of the data.

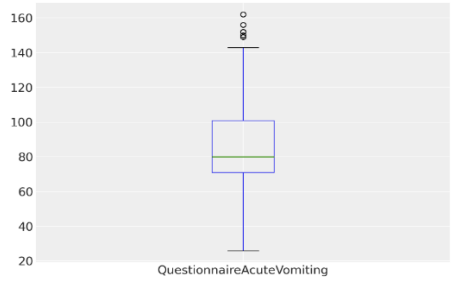

***Figure 22**: Box plot of the distribution of the QuestionnaireAcuteVomiting dataset*

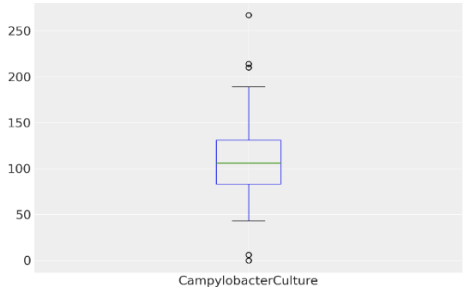

***Figure 23**: Box plot of the distribution of the CampylobacterCulture dataset*

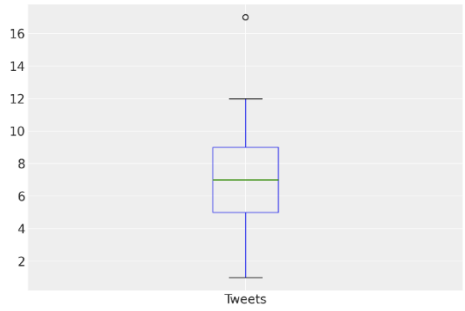

***Figure 24**: Box plot of the distribution of the Twitter dataset*

The autocorrelation and partial autocorrelation plots show that the correlation values are mostly negligible and not above the margins of uncertainty and are represented by the shaded light blue regions in figures 25 to 27. Also, most of the plots indicate values ranging between 0.2 and -0.2. This shows no significant correlations between the time-series datasets and a lagged version of the data over successive time intervals, meaning there are no repeating patterns or periodic signals which might impact changepoint analysis.

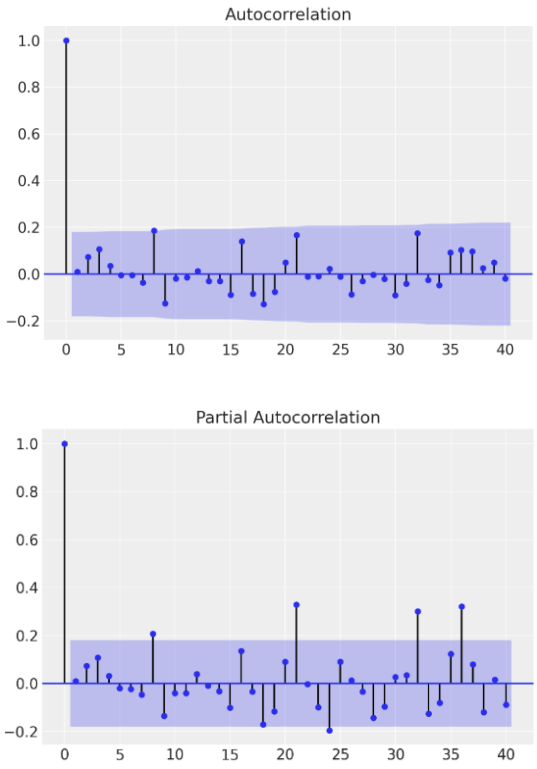

***Figure 25**: Autocorrelation and partial autocorrelation plots of the Twitter dataset*

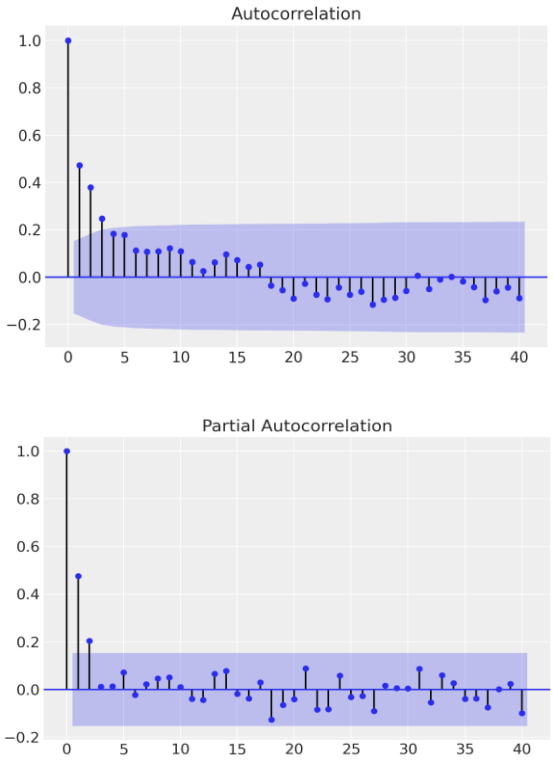

***Figure 26**: Autocorrelation and partial autocorrelation plots of the CampylobacterCulture dataset*

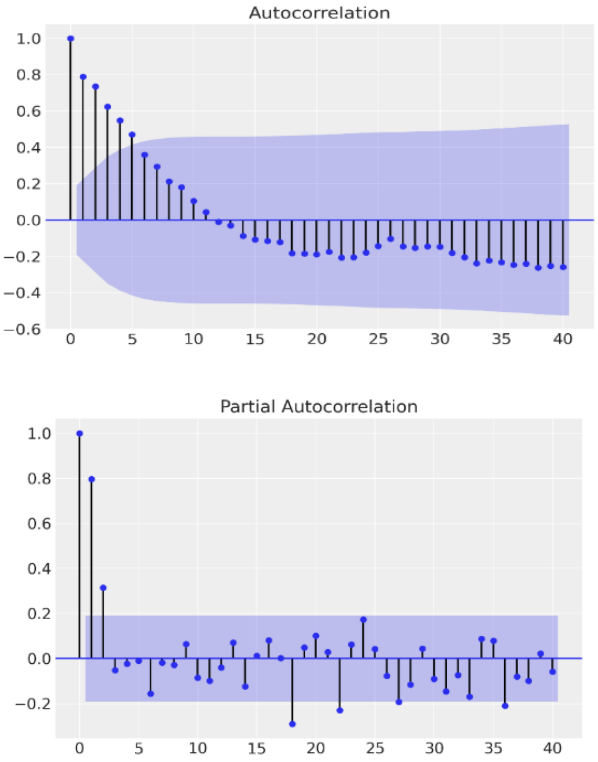

***Figure 27**: Autocorrelation and partial autocorrelation plots of the QuestionnaireAcuteVomiting dataset*

The time-series decomposition shows that the Twitter datasets exhibit an overall progression of an upward trend over time until the peak at 7.25 and then shows the onset of a possible downward trend. The pattern is illustrated in Figure 28, and it suggests an outbreak. The seasonality displays some unclear spikes with high and low points at irregular intervals, indicating no clear seasonal effects. The residual plot shows a random, irregular influence that could not be attributed to the trend or seasonal effects but may be considered noise in the dataset. Figure 29 shows the CampylobacterCulture dataset exhibiting an upward trend that looks like a high spike recovery from a downward trend. The upward trend of the CampylobacterCulture dataset corresponds to the trend in the Twitter dataset, suggesting an interesting phenomenon. An irregular, periodic pattern is shown in the seasonal component of the CampylobacterCulture data. The residual component does not match either the seasonal component or the trend, indicating no influence of periodic patterns in the dataset.

Likewise, in Figure 30, the trend, seasonality, and residual components of the QuestionnaireAcuteVomiting dataset show no relationship or influence on each other.

However, the trend component of the QuestionnaireAcuteVomiting dataset shows a recovery from a downward trend to an upward trend in the dataset, like a sustaining outbreak, which matches the upward sections of the Twitter and CampylobacterCulture trend components.

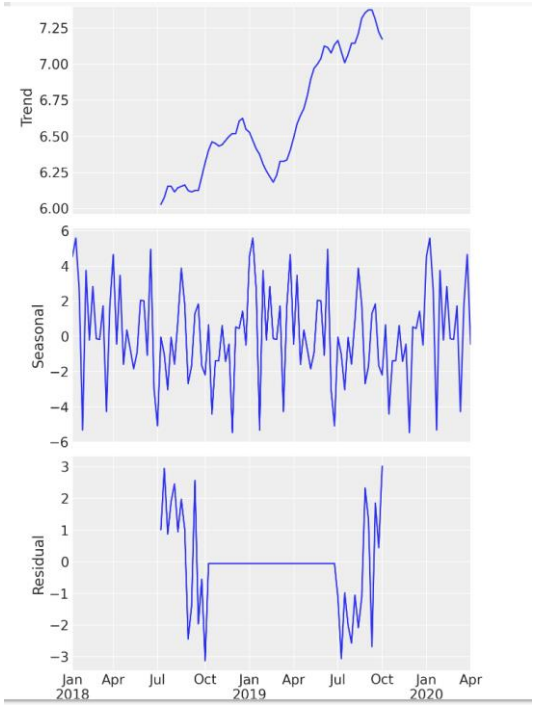

***Figure 28**: Decomposition of the Twitter time-series dataset*

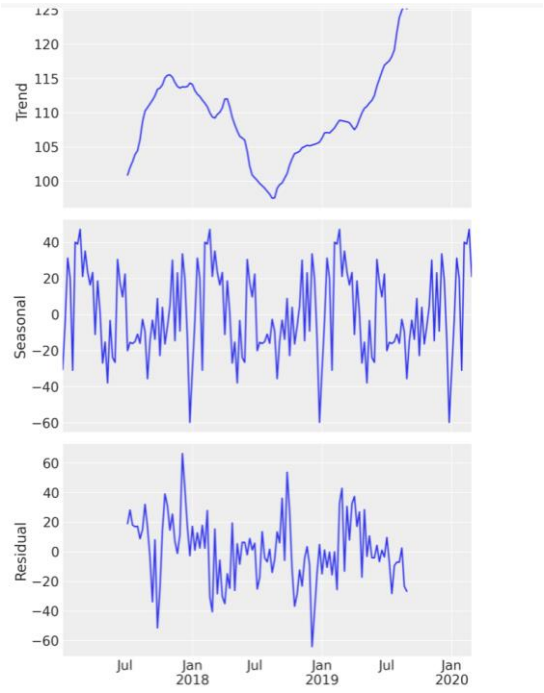

***Figure 29**: Decomposition of the CampylobacterCulture time-series dataset*

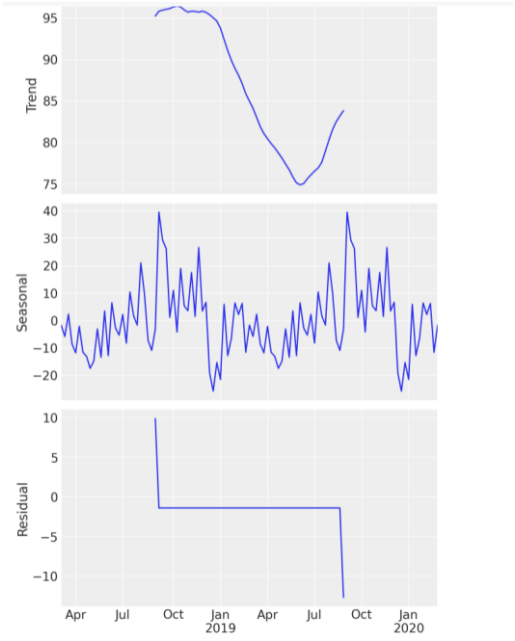

***Figure 30**: Decomposition of the QuestionnaireAcuteVomiting time-series dataset*

6.4 Discussion

From the time-series analysis and data exploration performed, it can be argued that the data streams available for this research work suffer from over-dispersion, and they are non-stationary time-series datasets due to the exhibition of the trend, seasonality, and residual characteristics uncovered during visualisation. Since the datasets are discrete, their underlying distributions might be Poisson or Negative Binomial (Struchen et al., 2015); this will be confirmed in chapter 8 of this thesis with further testing of the data distribution. Furthermore, the features demonstrated in the sampled datasets have been established as common characteristics of the syndromic surveillance datasets. Vial et al. (2016) argued that the use of a stochastic modelling-based approach could offer more flexibility, allowing for the retention of historical outbreaks for overdispersion and non-stationarity. Therefore, it would be reasonable to implement a model that can account for the stochasticity, uncertainties, and prior information to obtain the proposed SQRA method.

A major difficulty encountered during data preparation for exploration was the date format of the time-series datasets. It was assumed that the Pandas' library could automatically recognise date formats in the time-series dataset, but this assumption is wrong. Each data stream was loaded into Pandas' data frame for easy handling after slicing and selecting the required columns for analysis. However, the date column was required in DateTime format, and every

attempt to convert the date column to DateTime format using the Pandas DateTime function was unsuccessful. Most documentation about DateTime conversion in Pandas is based on a conversion from strings, Pandas Series or integers. After several failed attempts, the author investigated the column data type and discovered it was an object type.

Consequently, the date column was converted to Pandas Series and then DateTime format. Further operations on the date column, such as sorting the data, selecting the date part, and assigning the date column to an index column, caused the date to lose its DateTime format. As a result, the author decided to rearrange the Python procedure and re-apply the Pandas DateTime format again at the end of the procedure. The applied solution is provided in Figure 15. It was subsequently used on all time-series datasets selected for this study.

There are several ways of handling the DateTime format type in Python. Python language has a built-in DateTime function that is easily accessible at the Python and C/Cython levels. Numpy DateTime format is available through the Numpy library for handling date data. Also available for date handling is Pandas' data structures, which are DateTime aware and easy to implement. Though Pandas' data structure appears flexible and readable with few performance issues, researchers must not make assumptions about dates when using Pandas.

McKinney (2010) also supports the view that dates must be checked and converted to Datetime format as required and should not be assumed. However, it is unclear why the DateTime handling with Pandas reverted to Pandas' series after the initial conversion in the procedure. It might be because of subsequent operations applied to the date column. The workaround to solve this problem was reapplying the Pandas DateTime function to the date column. However, the available project time for this study is not enough to investigate the issues in detail.

6.5 Conclusion

In this chapter, the author identified various characteristics of the data streams, their structures, and their shapes. All the available data streams for this research work experienced over-dispersion and non-stationarity due to the type of trend, seasonality, and noise characteristics. Therefore, the result of the exploratory analysis and visualisation established the stochastic features of the available datasets. Also, it shows that the data streams are discrete, over-dispersed, non-stationary, and not normally distributed.

Count data is often described by Poisson or NB distribution. However, due to the overdispersion features of each data stream, they failed to conform with the main principle assumption of Poison distributed data, that the mean and variance values must be equal. Previous studies reviewed in this research work in chapter 3 suggested the use of the NB model because they allow the overdispersion parameter to vary randomly following a beta distribution that can be adjusted in the model. Based on these observations and suggestions, the author further investigated the underlying data distributions using the hypothesis testing technique in chapter 8.

Chapter 7 – Selecting a Probabilistic Programming Approach in Python

7.1 Chapter Introduction

The probabilistic programming approach is a paradigm for creating inference procedures by compiling generative probabilistic models using sampling algorithms (Tran et al., 2017). The paradigm could be useful for creating decision support systems in the face of uncertainty since they can measure uncertainties in the data and the resulting models (Fanfarillo, 2019). Also, some research studies argued in favour of applying the probabilistic modelling-based approach to syndromic surveillance datasets because they can offer more flexibility, retain historical outbreaks, handle overdispersion and non-stationarity behaviours, and account for uncertainties in models (Hedell et al., 2019; Aghaali et al., 2020; Salmon et al., 2015; Manitz and Höhle, 2013; Vial et al., 2016).

Probabilities are used to express levels of belief, and probabilistic inferences are used to illustrate logical reasoning in the face of uncertainties (Goodman, 2013). As probabilistic models are used to solve more complex problems, their implementation methods have also become more advanced. However, the model expressiveness may be easily traded off against the computational efficiency of inference generation. Based on this view, Tran et al. (2017) argued that many generic inference engines implement solutions that scale poorly as the problems they tend to solve become complex or the data sizes grow.

Although probabilistic programming is commonly used for restricted classes of statistical models, many of its approaches may lack the flexibility and efficiency required for practical usage with more complex models (Cusumano-Towner et al., 2019). As a result, in this chapter, the author critically investigated various probabilistic programming packages which exist in Python to establish their flexibility and efficiency and subsequently select the appropriate technique for developing the proposed SQRA method.

7.2 Method

The standard probabilistic programming packages available in Python were searched through the Python package repository page. The repository contains all software and libraries developed and shared by the Python community and is available on https://pypi.org. Five packages were found and selected at the time of conducting this research experiment. The packages are Pystan, PyJAGS, Pyro-ppl, TensorFlow Probability, and PyMC3.

The author developed a simple Bayesian probabilistic model that calculates changes in the mean of the time-series dataset and the corresponding time that the changes occurred using each selected package in Python. The model is mathematically illustrated in equation 7.1.

$$\begin{pmatrix} lambda\_1,\ if\ changepoint < t_1 \\ lambda\_2,\ if\ t_1 \leq changepoint < t_2 \\ lambda\_3,\ if\ changepoint \geq t_2 \end{pmatrix} \quad (7.1)$$

Where lambda_1, lambda_2, and lambda_3 are the mean changes, also known as rate, and the corresponding time of the change is the changepoint. The model and its parameters are described in detail in chapter 8, section 8.2.1. The aim of the author in this current chapter is to evaluate how flexibly and efficiently each selected Bayesian package can support the development of the proposed SQRA technique in Python. As a result, the model specified in equation 7.1 was defined with each package and fitted to the Twitter data stream on Google Colab. The model's sampling performance was evaluated using the r_hat value and Arviz diagnostic tool. The results are presented in section 7.3, and further explained in section 7.4 of this chapter.

The r_hat value is the Gelman-Rubin diagnostic which allows the evaluation of the MCMC convergence by providing a numerical summary of the model's performance (Salvatier et al., 2016). Other criteria for investigating each package are the ease of specifying and developing the model in Python language, availability of detailed online documentation and accessible examples on the package use, and integration of the package with the Arviz diagnostic tool. Arviz provides exploratory analysis of Bayesian models so that researchers can estimate model performance and verify convergence. Also, it provides the means for posterior analysis and effective sample size measurement (Kumar et al., 2019). Model convergence and divergence are further discussed in chapter 8.

1. Pystan is a package that provides a Python interface to Stan for Bayesian inference analysis. The inference analysis uses the MCMC method solely based on the No-U-Turn sampler and adaptive form of the Hamiltonian Monte Carlo sampler (Carpenter et al., 2017; Koprivica, 2020). The initial attempt to define a model in Pystan is presented in Figure 31.

```
data {
  real<lower=0> r_e;
  real<lower=0> r_l;

  int<lower=1> T;
  int<lower=0> D[T];
}
transformed data {
  real log_unif;
  log_unif = -log(T);
}
parameters {
  real<lower=0> e;
  real<lower=0> l;
}
transformed parameters {
  vector[T] lp;
  lp = rep_vector(log_unif, T);
  for (s in 1:T)
    for (t in 1:T)
      lp[s] = lp[s] + poisson_lpmf(D[t] | t < s ? e : l);
}
model {
  e ~ exponential(r_e);
  l ~ exponential(r_l);
  target += log_sum_exp(lp);
}
```

***Figure 31**: Pystan model definition*

2. Similar to Pystan, PyJAGS is an interface. It provides Python access to JAGS, a probabilistic programming language written in C++ for implementing Bayesian inferences (Koprivica, 2020). JAGS was written with the following aim to provide a cross-platform engine for the BUGS language to be extensible, enabling users to write their functions, distributions, and samplers. Finally, to provide a platform solely for Bayesian modelling (Plummer, 2003). However, installing PyJAGS on Google Colab required JAGS software to be first installed as a prerequisite. The attempt to write the model in PyJAGS is illustrated in Figure 32.

```
model{
    ## Likelihood
    for(i in 1:N){
      y[i] ~ dpois(lambda[i])
      log(lambda[i]) <- mu[i]
      mu[i] <- inprod(beta[],X[i,])
      }
    ## Priors
    beta ~ dmnorm(mu.beta,tau.beta)  # multivariate Normal prior
}
```

***Figure 32**: PyJAGS model definition*

3. Pyro-ppl is a probabilistic programming language for deep Bayesian modelling built on PyTorch in Python, which can scale to large datasets and high-dimensional models. In addition, Pyro is capable of stochastic variational inference (Bingham et al., 2019). The model specification in the Pyro-ppl package is illustrated in Figure 33; it shows the parameter definition for both mean changes and the corresponding changepoints. The detailed Python code is available at the link indicated for extra resources in Appendix.

```
class Tweet_changepoint(PyroModule):
    def __init__(self, in_features, out_features, b1_mu, b2_mu):
        super().__init__()
        self.linear1 = PyroModule[nn.Linear](in_features, out_features, bias = False)
        self.linear1.weight = PyroSample(dist.Normal(0.5, 0.25).expand([1, 1]).to_event(1))
        self.linear1.bias = PyroSample(dist.Normal(b1_mu, 1.))

        # could possibly have stronger priors for the 2nd regression line, because we wont have as much data
        self.linear2 = PyroModule[nn.Linear](in_features, out_features, bias = False)
        self.linear2.weight = PyroSample(dist.Normal(0., 0.25).expand([1, 1])) #.to_event(1))
        self.linear2.bias = PyroSample(dist.Normal(b2_mu, b2_mu/4))

    def forward(self, x, y=None):
        tau = pyro.sample("tau", dist.Beta(4, 3))
        sigma = pyro.sample("sigma", dist.Uniform(0., 3.))
        # fit lm's to data based on tau
        sep = int(np.ceil(tau.detach().numpy() * len(x)))
        mean1 = self.linear1(x[:sep]).squeeze(-1)
        mean2 = self.linear2(x[sep:]).squeeze(-1)
        mean = torch.cat((mean1, mean2))
        obs = pyro.sample("obs", dist.Normal(mean, sigma), obs=y)
        return mean
```

```
#tensor_data = torch.tensor(reg_data[["daily_confirmed", "days_since_start"]].values, dtype=torch.float)
tensor_data = torch.tensor(tweet[["Tweets"]].values, dtype=torch.float)
#x_data = tensor_data[:, 1].unsqueeze_(1)
x_data = tensor_data[:, 0].unsqueeze_(1)
y_data = np.log(tensor_data[:, 0])

# prior hyper params
# take log of the average of the 1st quartile to get the prior mean for the bias of the 2nd regression line
q1 = np.quantile(y_data, q = 0.25)
bias_1_mean = np.mean(y_data.numpy()[y_data <= q1])
print("Prior mean for Bias 1: ", bias_1_mean)

# take log of the average of the 4th quartile to get the prior mean for the bias of the 2nd regression line
q4 = np.quantile(y_data, q = 0.75)
bias_2_mean = np.mean(y_data.numpy()[y_data >= q4])
print("Prior mean for Bias 2: ", bias_2_mean)
```

```
Prior mean for Bias 1:  1.1455992
Prior mean for Bias 2:  2.3394501
```

```
model = Tweet_changepoint(1, 1,
                      b1_mu = bias_1_mean,
                      b2_mu = bias_2_mean)
# need more than 400 samples/chain if we want to use a flat prior on b_2 and w_2
num_samples =118
# mcmc
nuts_kernel = NUTS(model)
mcmc = MCMC(nuts_kernel,
            num_samples=num_samples,
            warmup_steps = 150,
            num_chains = 2)
mcmc.run(x_data, y_data)
samples = mcmc.get_samples()
```

```
Sample [0]: 100%|██████████| 268/268 [10:20,  2.32s/it, step size=1.11e-03, acc. prob=0.820]
Sample [1]: 100%|██████████| 268/268 [09:28,  2.12s/it, step size=1.67e-03, acc. prob=0.755]
```

***Figure 33**: Pyro-ppl model specification for Twitter data stream*

4. TensorFlow Probability (TFP) is a Python library built on top of TensorFlow; this makes it easy to implement combined deep learning and probabilistic models on modern hardware such as TPU and GPU. TFP provides a wide range of probability distributions and bijectors. In addition, the library is capable of stochastic variational inference. The model specification and parameter definitions in the TFP package are shown in Figure 34, and the full Python code is available at the link indicated in the Appendix for this research's extra materials.

```
[8]  def incident_count_model(incident_rate_fn):
       incident_count = tfd.JointDistributionNamed(dict(
         e=tfd.Exponential(rate=1.),
         l=tfd.Exponential(rate=1.),
         s=tfd.Uniform(0., high=len(tweet)),
         d_t=lambda s, l, e: tfd.Independent(
             tfd.Poisson(rate=incident_rate_fn(np.arange(len(tweet)), s, l, e)),
             reinterpreted_batch_ndims=1)
       ))
       return incident_count

     def incident_rate_switch(ys, s, l, e):
       return tf.where(ys < s, e, l)

     def incident_rate_sigmoid(ys, s, l, e):
       return e + tf.sigmoid(ys - s) * (l - e)

     model_switch = incident_count_model(incident_rate_switch)
     model_sigmoid = incident_count_model(incident_rate_sigmoid)
```

```
num_results = 10000
num_burnin_steps = 3000

@tf.function(autograph=False, experimental_compile=True)
def make_chain(target_log_prob_fn):
  kernel = tfp.mcmc.TransformedTransitionKernel(
      inner_kernel=tfp.mcmc.HamiltonianMonteCarlo(
         target_log_prob_fn=target_log_prob_fn,
         step_size=0.05,
         num_leapfrog_steps=3),
      bijector=[
         # The changepoint is constrained between zero and len(tweet).
         # Hence we supply a bijector that maps the real numbers (in a
         # differentiable way) to the interval (0;len(tweet))
         tfb.Sigmoid(low=0., high=tf.cast(len(tweet), dtype=tf.float32)),
         # Early and late incident rate: The exponential distribution is
         # defined on the positive real numbers
         tfb.Softplus(),
         tfb.Softplus(),
      ])
```

```
    kernel = tfp.mcmc.SimpleStepSizeAdaptation(
        inner_kernel=kernel,
        num_adaptation_steps=int(0.8*num_burnin_steps))

    states = tfp.mcmc.sample_chain(
       num_results=num_results,
       num_burnin_steps=num_burnin_steps,
       current_state=[
           # The three latent variables
           tf.ones([], name='init_changepoint'),
           tf.ones([], name='init_early_incident_rate'),
           tf.ones([], name='init_late_incident_rate'),
       ],
       trace_fn=None,
       kernel=kernel)
    return states

switch_samples = [s.numpy() for s in make_chain(
    lambda *args: target_log_prob_fn(model_switch, *args))]
sigmoid_samples = [s.numpy() for s in make_chain(
    lambda *args: target_log_prob_fn(model_sigmoid, *args))]

changepoint, early_incident_rate, late_incident_rate = zip(
    switch_samples, sigmoid_samples)
```

***Figure 34**: Model specification for fitting Twitter data stream in TensorFlow probability*

5. PyMC3 is an open-source probabilistic programming framework developed in Python, which uses Theano to compute gradients via automatic differentiation and compile probabilistic algorithms to C on the fly for performance. Theano provides many benefits to PyMC3, such as high performance from graph optimisations and compilation to CPU and GPU, while keeping the model definition and code-base pure Python. Also, PyMC3 has the inbuilt capacity to implement a changepoint analysis by declaring it as a parameter. The model specification and parameter definitions in the PyMC3 package are demonstrated in Figure 35. The full Python code is available in the Appendix as part of this research's extra materials.

```
with pm.Model() as twitter_model:

    alpha = pm.Uniform('alpha', lower=0, upper=100)

    lambda_1 = pm.Uniform("lambda_1", lower=0, upper=20)
    lambda_2 = pm.Uniform("lambda_2", lower=0, upper=20)
    lambda_3 = pm.Uniform("lambda_3", lower=0, upper=20)

    tau_1 = pm.Uniform("tau_1", lower=0, upper= number_of_weeks - 1)
    tau_2 = pm.Uniform("tau_2", lower=tau_1, upper= number_of_weeks - 1)

    idx = np.arange(number_of_weeks) # Index

    lamda_0 = pm.math.switch(tau_1 >= idx, lambda_1, lambda_2)
    lambda_ = pm.math.switch(tau_2 >= idx, lamda_0, lambda_3)

    observation = pm.NegativeBinomial("obs", lambda_, alpha, observed=twitter_obs)
```

```
with twitter_model:

 trace = pm.sample(1000, cores=2, tune=2000)
```

```
/usr/local/lib/python3.7/dist-packages/pymc3/sampling.py:468: FutureWarning: In an upcoming release, pm.s
  FutureWarning,
Auto-assigning NUTS sampler...
Initializing NUTS using jitter+adapt_diag...
Multiprocess sampling (2 chains in 2 jobs)
NUTS: [tau_2, tau_1, lambda_3, lambda_2, lambda_1, alpha]
100.00% [6000/6000 07:54<00:00 Sampling 2 chains, 0 divergences]
Sampling 2 chains for 2_000 tune and 1_000 draw iterations (4_000 + 2_000 draws total) took 475 seconds.
```

***Figure 35**: Model specification in PyMC3 model including calls to Arviz libraries*

7.3 Result

The attempts to specify the model and define the parameters with Pystan and PyJAGS packages were unsuccessful. An in-depth understanding of the STAN and JAGS programming languages and their rules is required. However, the available project time for this study is not enough for the author to learn a new programming language in detail to support the two packages.

On investigating the Pyro-ppl package in this study, although the model reports no divergences in the summary statistics, the r_hat values for most sample parameters were very high, indicating a poor performance since r_hat values should be very close to 1 (Salvatier et al., 2016; Carpenter et al., 2017; Gelman et al., 2013). Figure 36 shows the r_hat values of each parameter. Therefore, the investigation of the Pyro-ppl package was discontinued due to poor performance. Divergence is further explained in chapter 8.

```
mcmc.summary()
diag = mcmc.diagnostics()
```

	mean	std	median	5.0%	95.0%	n_eff	r_hat
tau	0.51	0.27	0.37	0.20	0.85	1.04	5.44
sigma	0.22	0.01	0.21	0.20	0.24	15.74	1.14
linear1.weight[0,0]	0.18	0.02	0.18	0.15	0.20	1.71	1.62
linear1.bias	0.60	0.13	0.58	0.39	0.79	1.45	1.86
linear2.weight[0,0]	0.18	0.01	0.18	0.16	0.20	13.24	1.15
linear2.bias	0.49	0.09	0.48	0.35	0.63	16.14	1.02

```
Number of divergences: 0
```

***Figure 36**: Summary statics showing the r_hat values after fitting Pyro-ppl model to the Twitter data stream*

Although TFP demonstrates superior computational speed for sampling the Twitter data stream, indicating high efficiency as the dataset scales in size, the model specification and

parameter definition is more challenging and complex than Pyro-ppl and PyMC3. Figure 37 shows the resulting histogram plots for the posterior distributions of each parameter. The author could not get the Arviz diagnostic tool to work in TFP to investigate the divergences and the performance of the sampling process.

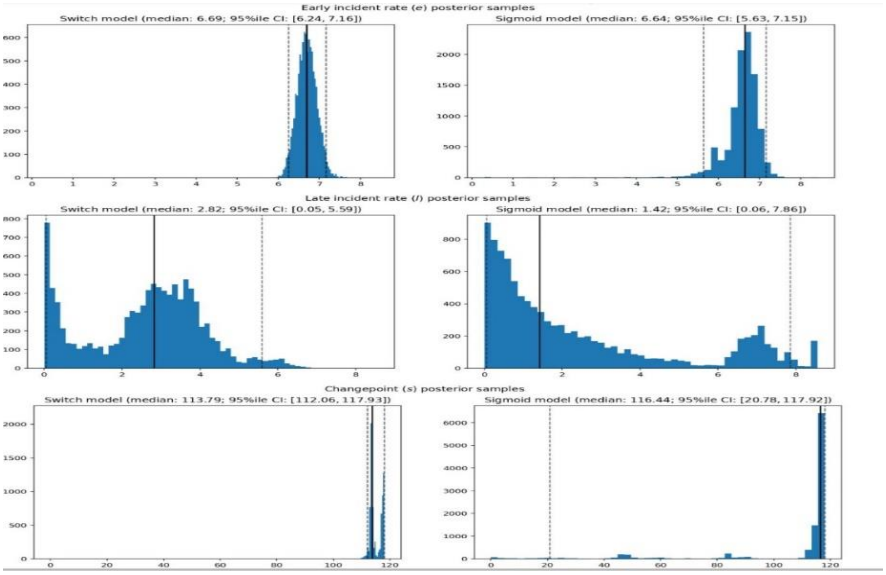

***Figure 37**: Posterior distributions for each parameter modelled in TFP*

Using the Twitter dataset to investigate PyMC3, the author found that the model specification and parameter definitions are relatively easier than other packages examined. Also, many diagnostic parameters could be called directly from the PyMC3 procedure to investigate the model performance. Figure 38 shows that the r_hat values are very close to 1, indicating efficient performance from the samplers.

```
az.summary(trace, round_to=2)
```

```
/usr/local/lib/python3.7/dist-packages/arviz/data/io_pymc3.py:92: FutureWarning: Using `from_pymc3` without the model w
  FutureWarning,
```

	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_mean	ess_sd	ess_bulk	ess_tail	r_hat
alpha	47.16	25.34	14.31	98.00	2.35	1.74	116.34	106.36	128.42	122.60	1.02
lambda_1	6.45	0.36	5.83	7.18	0.04	0.03	95.30	93.83	94.54	130.24	1.03
lambda_2	8.71	2.16	4.90	13.73	0.24	0.17	81.78	79.29	90.87	72.03	1.01
lambda_3	5.17	1.68	2.23	8.10	0.20	0.14	73.05	73.05	70.44	105.01	1.02
tau_1	89.83	21.02	48.08	116.97	2.21	1.62	90.36	84.50	85.67	129.77	1.02
tau_2	108.58	12.91	84.49	116.99	1.09	0.80	141.18	130.04	159.18	197.25	1.01

***Figure 38**: Summary statics showing the r_hat values after fitting PyMC3 model to the Twitter data stream*

7.4 Discussion

Pystan was first released in July 2013; it currently has excellent online documentation with a large community base (Riddell et al., 2020), and it is a matured Bayesian programming package available in Python like PyMC3 and PyJAGS (Koprivica, 2020). Furthermore, Pystan is capable of stochastic variational inference. Also, the author observed that Google Colab does not require installing the package by the user since it is pre-installed and pre-configured on the platform, unlike PyJAGS, which require the installation of JAGS software on Google Colab.

According to previous studies, Pystan supports sampling and optimising inference with posterior analysis. However, using Pystan effectively requires understanding the processes in the Stan programming platform and understanding the Stan language syntax. Attempts to investigate the Twitter data stream with Pystan were frustrating, particularly defining the model with multiple changepoints and feeding the posterior distributions as priors in subsequent datasets. Therefore, the author concluded that the objectives of this study would require advanced knowledge of Stan programming syntax. Achieving such a level of advanced Stan knowledge within the allowed timeframe for this thesis work would be extremely challenging.

Similarly, PyJAGS requires some JAGS syntax knowledge to specify the model effectively. It required the definition of the data variables and the model parameters separately. Then, the definitions of transformed parameters needed to be expressive within the PyJAGS parameter. The model also required the specification of each parameter's distributions and likelihood, similar to the PyMC3 approach. However, the major challenge with using PyJAGS is the lack of detailed documentation and relevant examples to assist in achieving the objectives of the research work. The author observed that PyJAGS does not have as much documentation and advanced examples as PyMC3 and Pystan for advanced Bayesian modelling. Therefore, it was challenging to use the PyJAGS package without a detailed understanding of the processes required to build it. Coupled with the lack of documentation of the package, the time available for the project is not enough to undertake such training.

The Pyro-ppl package was easy to put together because the syntax is pure Python, and the model specification was easy to assemble. It is built on the PyTorch library in Python, and as a result, it has many documentation and online resources to guide its implementation. The author discovered that the outcome of the modelling shows no divergencies, but the r_hat values of each parameter were very high. However, the first version of Pyro-ppl was released

in November 2017, so the package is in its infancy, with a few peer-reviewed studies available on its investigative implementation. Therefore, finding research articles that explain its poor performance on the available Twitter dataset was challenging. The research study by Cusumano-Towner et al. (2019) argued that Pyro-ppl is not as flexible as other probabilistic programming languages in its class due to the lack of expressiveness for custom model implementation.

Similarly, TFP was first released in March 2018, indicating the library is also in its infancy, similar to Pyro-ppl. However, the author observed many online documentation and examples of porting existing models created in other Probabilistic languages to TFP. On investigating the library with the available Twitter dataset, it was discovered that defining the model via the JointDistributionSequential distribution function was the best approach for specifying the model. It is worth noting that the model specification was more challenging to define than other packages investigated. Figure 34 displays the implementation of the JointDistributionSequential for defining prior parameters and the hyper-parameters of the model with their associated distributions. Even more challenging in TFP was the attempt to use the Arviz package to visualise the summary statistics of the posterior distributions compared to the PyMC3 implementation of Arviz for a similar task. Owing to the challenges observed in model specification and the need to define some complex functions to use Arviz with TFP, the author decided to discontinue the investigation of the TFP package.

Like Pyro-ppl and TFP, the author observed that PyMC3 supported the model specification and parameter definition directly in Python code (Salvatier et al., 2016). The PyMC was first released in 2005 as a package in Python (Patil et al., 2010), and in May 2013, the PyMC3 version was released as an upgrade to PyMC. PyMC3 is capable of implementing stochastic variational inferences.

Furthermore, the PyMC3 platform has a large user community and extensive online documentation with examples in various disciplines. In addition, the author observed that the model specifications in PyMC3 are substantially shorter than Pystan, Pyro and TFP, and it was easier to understand since it was written natively in Python. Python is dynamically typed, so the PyMC3 variables get their types from what is assigned to them from Python.

It was observed that the Arviz diagnostic package integrated easily with the PyMC3 library and produced a summary statistic of the model's performance. This allows the visualisation

and investigation of various parameters of the model, including posterior analysis, sample diagnostics, and model comparison. Therefore, model validation and verification were easier to implement in PyMC3 through the Arviz diagnostic tool than in the other investigated Bayesian packages.

7.5 Conclusion

In this chapter, the author investigated Pystan, PyJAGS, Pyro-ppl, TFP and PyMC3 packages to determine the most appropriate probabilistic programming for developing the proposed SQRA method. The flexibility of packages, integration with Arviz visualisation and diagnostic tools, model performance with r_hat, the availability of detailed documentation and online resources, and accessible examples are the key considerations for selecting PyMC3 over other packages.

The author observed that with PyMC3, it was relatively easier to piece together a block of samplers, understand the workflow, and adapt existing examples to fit the objectives of this research work. Furthermore, it was possible to write a procedure that updated priors with posterior distributions as new evidence emerged in the form of data, a feature that fits the main goal of this study. Therefore, based on the effectiveness of PyMC3 over other packages in this chapter, it was selected as the appropriate data modelling technique for developing the proposed SQRA method in this research work.

Chapter 8 – Data Modelling and Testing

8.1 Chapter Introduction

Weak and noisy data often characterise the onset of a disease outbreak. As a result of the noise in the data, early detection of outbreaks is often under uncertainty since noise creates random effects in the data. This has been established as a common feature of disease surveillance datasets in the previous chapters of this study.

Most Frequentist approaches require baseline implementation to be free of the historical outbreak in order to create models of expected behaviour. This often constitutes a real challenge as most disease surveillance datasets may not be free of historical outbreaks and noise, nor will their shape or the risk magnitudes in the datasets be known. However, previous research studies argued that the Bayesian modelling approach for detecting disease outbreaks could offer more flexibility, retain historical outbreaks, and address overdispersion and non-stationarity in the datasets to deal with the noise effects. While such methods have been successfully applied in other research fields, they have only recently received attention in public health surveillance and are sparingly applied to veterinary syndromic surveillance datasets.

Applying Bayesian methodology requires three main parts: the model specification in a mathematical model, the parameter definition, which includes prior distributions of the inference parameters, and the third part is an efficient method for sampling the posterior distributions. The model's likelihood can be derived directly from the model specification either by a mathematical theorem or automated procedure through a sampling algorithm (Gelman et al., 2013). This enables rapid Bayesian fitting of models to data with fully quantifiable uncertainties, as implemented in previous studies such as Hedell et al. (2019) and Hale et al. (2019). It also enables the sampling of parameters from the posterior distribution and evaluating the model outcome through diagnostic tools that measure the model performance.

A common approach is to consider a deterministic iterative model and treat the observed data as a stochastic process. The data is analysed by solving ordinary differential equations or similar equations to address the analysis of the observed data. A formula is then used to compute the likelihood. This technique is considered fast and flexible with small overheads, but using deterministic models in this approach can bias the result. The other approach considers full stochastic models and estimates parameters using the MCMC sampling

algorithm. However, such an approach avoids the biases above but is computationally expensive since, for a given parameter, multiple stochastic trajectories are generated to optimise over all parameter choices (Li et al., 2021). The methodology adopted in this current study is intermediate between the two approaches by defining a deterministic iterative model but using the MCMC algorithm to estimate parameters. The aim is to avoid the biases associated with the deterministic modelling approach without the computational cost of stochastic simulations.

Given the available datasets for this research study and the knowledge of their attributes and characteristics gained in chapters 4 and 6 of this document, the author will specify the proposed model mathematically, define the prior distributions of the model parameters and develop the proposed SQRA model using PyMC3 package in this chapter. The approach will sample the posterior distributions of each unknown parameter using the gradient-based MCMC algorithms to estimate their values.

8.2 Method

Using the Pearson chi-square goodness-of-fit test, the author first performed hypothesis testing to check whether the Poisson distribution adequately describes the Twitter data stream distribution. The theoretical frequencies were compared to the frequencies of the observed data (Jaimes et al., 2005). The calculated chi-square goodness of fit statistic was 55.81, and its p-value was 1.29e-07 which is much smaller than alpha=0.05. Therefore, it was concluded that the Null Hypothesis H0, which says that the Twitter data stream is Poisson distributed, could be rejected at a 95% confidence level. Also, it was observed that the critical chi-square test statistic value to accept H0 at a 95% confidence level was 21.026, which is much smaller than 55.81. The author then concluded that the Poisson distribution could not have described the underlying distribution of the syndromic datasets collected on Twitter; this may be due to the overdispersion of the Twitter datasets.

However, an approach suggested by Rodrıguez (2013) for modelling datasets that suffer from overdispersion is to begin from a Poisson model and add a multiplicative random effect θ to represent unobserved heterogeneity leading to the NB model. Therefore, the author began the modelling procedure for the Twitter data stream with Poisson and then with NB distribution and in the final stage, the resulting outcomes were compared.

To complement the existing VSyS platforms with an SQRA method, which is the primary aim of this study, the parameters and hyper-parameters in the following sections were recommended based on unbiased opinions of the domain experts and the recommendations in the guideline for SyS development by Triple- S Project (2013). Furthermore, the author followed a systematic workflow described in section 5.2.1 of the research methodology chapter, recommending the following steps for model specification in Figure 39.

1. Build the model
2. Identify divergencies in the model through model diagnostics
3. Re-parametrise the model as needed.

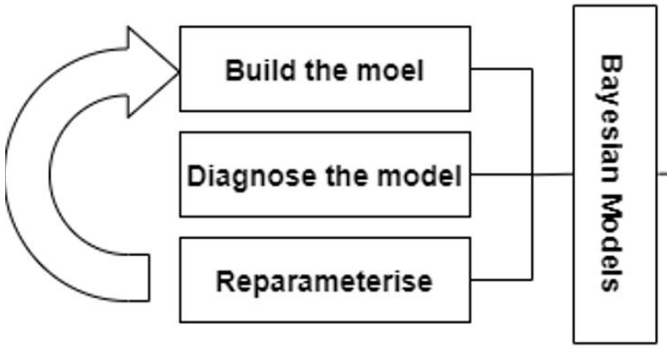

***Figure 39**: Model Specification Workflow in PyMC3*

The author adapted the three primary steps of the Bayesian modelling building block from Gelman et al. (2013) to develop the model workflow above; the three primary steps are

1. Specify a complete probability model that includes all relevant parameters, data, transformations, missing values, and forecasts, and initialise them with prior probability distributions.
2. Calculate the posterior probability distribution of the unknown parameters in the model based on the available dataset.
3. Evaluate the model's quality and suitability by performing model verification.

While the author found each step quite challenging, the second step is the most difficult for non-trivial models and has been a stumbling block for Bayesian methods adoption for decades (Blei, 2014).

8.2.1 Poisson Model Parameters

In section 5.2.1, the author presented the Bayesian inference approach as a technique for providing time-series datasets with predictive distributions that can fully and coherently incorporate parameter uncertainties. From equation 5.1, $f(\theta)$ described the belief about a health event. Whereas from equation 5.2, $f(\theta|y)$ represents the distribution of a health event after new evidence has been introduced into the model. Therefore, equation 5.2 is the posterior probability distribution which indicates the refined strength of the belief after new observations (θ) have been considered. Recall equation 5.4 below;

$$f(\theta|y) = \frac{f(y|\theta)f(\theta)}{f(y)} = \frac{f(y|\theta)f(\theta)}{\int f(y|\theta)f(\theta)d\theta}$$

The denominator f(y), which is the model's evidence or marginal likelihood, cannot be calculated directly and is the expression in the numerator, integrated over all the θ. Therefore, in this study, the Poisson log-likelihood was applied to infer $f(y|\theta)$ using the model likelihood and the MCMC sampling technique (No-U-Turn-Sampler) to compute the denominator f(y) and integrate it over all θ by sampling the posterior distribution.

As a result, the author defined the model parameters in the subsequent steps as follows;

D_t: The number of infection incidents per week, t

r_t: The rate parameters of the Poisson distribution of infection incidents per week t, described by lambda_1, lambda_2, and lambda_3 and expressed mathematically in equation 8.1.

$$\begin{pmatrix} lambda\_1,\ if\ changepoint < t_1 \\ lambda\_2,\ if\ t_1 \leq changepoint < t_2 \\ lambda\_3,\ if\ changepoint \geq t_2 \end{pmatrix} \quad (8.1)$$

Changepoint is the week in which the mean rate parameter changes and is bounded by t_l:t_h as the lower and upper boundaries of the changepoints.

lambda_1: the rate parameter before the first changepoint

lambda_2: the rate parameter after the first changepoint but before the second changepoint

lambda_3: the rate parameter after the second changepoint.

The unknown parameters can be specified as discrete uniform priors for changepoints and exponential distributions for the rate parameters since the probability mass function (pmf) of Poisson log-likelihood is defined by exponential distribution in equation 8.2.

$$f(k \mid \lambda) = \frac{\lambda^k e^{-\lambda}}{k!} \tag{8.2}$$

Rewriting equation 8.2 to obtain

$$f(k \mid \lambda) = \frac{1}{k!} \exp\{k \log \lambda - \lambda\} \tag{8.3}$$

The Poisson rate parameter is λ which assumes λ = Mean = Variance

changepoint_1 ~ Uniform(t_l, t_h) (8.4)

changepoint_2 ~ Uniform(t_l, t_h) (8.5)

lambda_1 ~ exp (1) (8.6)

lambda_2 ~ exp (1) (8.7)

lambda_3 ~ exp (1) (8.8)

The Poisson rate expression is defined in equation 8.9 as

$$D_t \sim \text{Pois}(\lambda = r_t) \tag{8.9}$$

Therefore, the Poisson model can be defined in equation 8.10 by incorporating equation 8.1 into 8.9.

$$D_t \sim \text{Pois}(r_t), \quad r_t = \begin{pmatrix} lambda\_1, & if\ changepoint < t_1 \\ lambda\_2, & if\ t_1 \leq changepoint < t_2 \\ lambda\_3, & if\ changepoint \geq t_2 \end{pmatrix} \tag{8.10}$$

The graphical representation of the Poisson model is illustrated in Figure 40. In Figure 41, equation 8.10 was specified in the PyMC3 package and equations 8.4 to 8.8 were incorporated into the model's specification. A novel deterministic random variable was also introduced as rate r_t, which are the r_0 and r_1 to infer the mean rate parameters of a disease outbreak. The PyMC3 implementation in Figure 41 shows the key parameter definitions in Poisson log-likelihood.

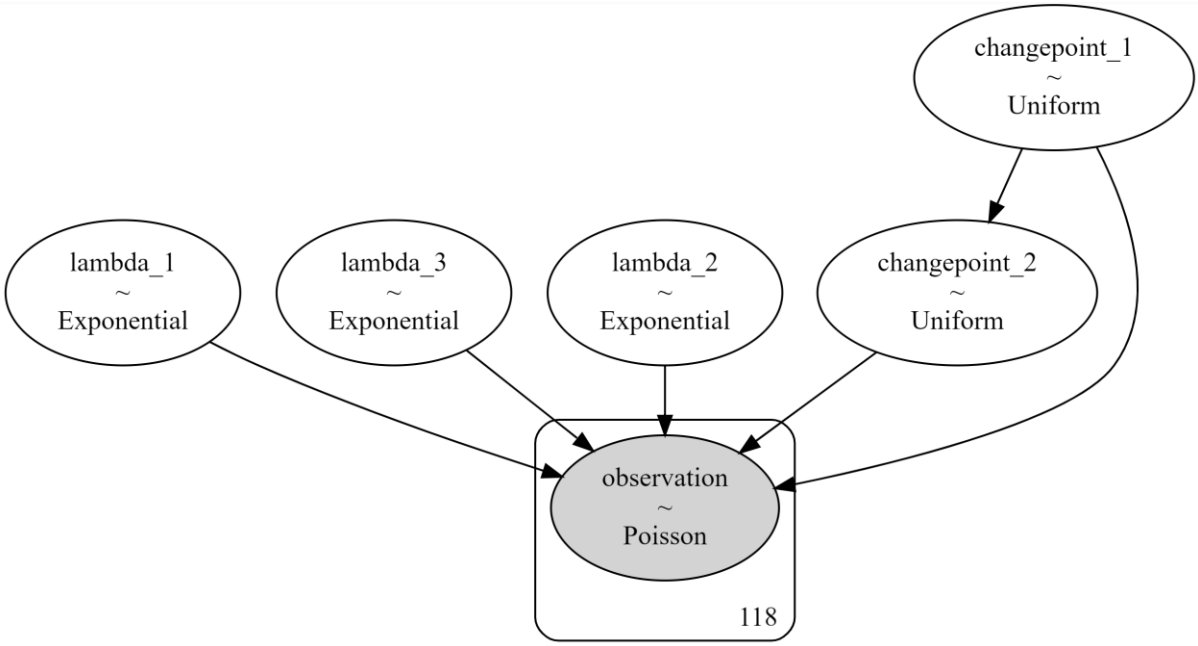

Figure 40: *Graphical Representation of the Model's Specification*

```
with pm.Model() as twitter_model:

    # Hyperparameter for rate
    alpha = 1.0  # the default value of alpha is 1.0 since the rate   can
only be greater than 0, α can be interpreted as the number of prior
observations
    # rate parameters
    lambda_1 = pm.Exponential("lambda_1", alpha)
    lambda_2 = pm.Exponential("lambda_2", alpha)
    lambda_3 = pm.Exponential("lambda_3", alpha)

    changepoint_1 = pm.Uniform("changepoint_1", lower=0, upper=
number_of_weeks - 1)
    changepoint_2 = pm.Uniform("changepoint_2", lower=changepoint_1, upper=
number_of_weeks - 1)

    idx = np.arange(number_of_weeks) # Arrange the Index along time (weeks)

    # The novel form of the deterministic random variable rates of weekly
incidents.
    # It allocates appropriate Poisson rates to weeks before and after
changepoints.
    rate_0 = pm.math.switch(changepoint_1 >= idx, lambda_1, lambda_2)
    rate_1 = pm.math.switch(changepoint_2 >= idx, rate_0, lambda_3)

   # Likelihood
   observation = pm.Poisson("observation", rate_1,observed=twitter_obs)
```

Figure 41*: Poisson Log-Likelihood Specification in PyMC3*

8.2.2 Markov Chain Monte Carlo (MCMC)

The MCMC sampling technique was applied to automatically calculate the posterior probability distribution of each unknown parameter using the available datasets. The expectation was that the MCMC draws of θ were independent by allowing the generation of samples as Markov Chain from a specific posterior distribution.

The technique comprises many sampler types; for example, the Metropolis-Hastings sampler is one of the most basic and adaptable MCMC algorithms. This algorithm generates candidate state transitions from an auxiliary distribution and probabilistically accepts or rejects each candidate based on a random-walk proposal, which is also used in the practical implementation of the Metropolis-Hastings algorithm. While adaptable and simple to use, Metropolis-Hastings's sampling is a random walk sampler that may not be statistically efficient for many models (Salvatier et al., 2016). The first attempt at sampling in this work using the Metropolis-Hastings sampler demonstrated the behaviour described by Salvatier and colleagues, resulting in an inefficient sampling outcome with poor acceptance rates and r_hat values. Furthermore, it was noted that the MCMC sampler in PyMC3 defaulted to Metropolis-Hastings for discrete observations with discrete prior distributions. However, the Hamiltonian Monte Carlo (HMC) can be a powerful tool when sampling continuous variables in this context. By simulating a physical system governed by Hamiltonian dynamics, HMC avoids random walk behaviour while potentially avoiding tricky conditional distributions using the leapfrog algorithms (Betancourt, 2017). The HMC's main disadvantage is the extensive tuning required to make it sample efficient. There are a few parameters that the user must specify:

1. the scaling of the momentum distribution
2. the step size for the leapfrog algorithm
3. and the number of steps that the leapfrog algorithm needs to take

However, when these parameters are incorrectly set, the HMC algorithm may suffer significant efficiency losses. For example, if the process takes too few steps, the simulation becomes a random walk, whereas too many steps result in retracing previously taken paths. On the other hand, Wang and Li (2021) argued that the No U-turn Sampling (NUTS) algorithm, a variant of the HMC, automatically tunes the step size and step number parameters without user intervention. NUTS accomplishes this by constructing a binary tree of leapfrog steps through repeated doubling (Nishio and Arakawa, 2019). When the step trajectory creates an angle

greater than 90 degrees (a u-turn), the doubling stops, and a point is proposed. Thus, NUTS achieves the efficiency of gradient-based MCMC sampling without requiring extensive user intervention to tune Hamiltonian Monte Carlo. NUTS is the default sampling algorithm for continuous variables in PyMC3 (Salvatier et al., 2016). Therefore, the continuous prior distributions were assigned to each unknown parameter in this study, even though they are discrete. The aim is to take advantage of the NUTS sampler's efficiency.

```
with twitter_model:

trace = pm.sample(2500, cores=2, tune=2500)
```

```
Auto-assigning NUTS sampler...
Initializing NUTS using jitter+adapt_diag...
Multiprocess sampling (2 chains in 2 jobs)
NUTS: [changepoint_2, changepoint_1, lambda_3, lambda_2, lambda_1]
100.00% [10000/10000 01:41<00:00 Sampling 2 chains, 1,192 divergences]
Sampling 2 chains for 2_500 tune and 2_500 draw iterations (5_000 + 5_000 draws total) took 102 seconds.
There were 4 divergences after tuning. Increase `target_accept` or reparameterize.
The acceptance probability does not match the target. It is 0.8898751910336967, but should be close to 0.8. Try to increase the number of tuning steps.
There were 1188 divergences after tuning. Increase `target_accept` or reparameterize.
The acceptance probability does not match the target. It is 0.4081324549115282, but should be close to 0.8. Try to increase the number of tuning steps.
The rhat statistic is larger than 1.05 for some parameters. This indicates slight problems during sampling.
```

Figure 42: *The MCMC Sampling with NUTS Sampler*

Figure 42 displays the passing of the MCMC sampling process as a variable and assigning it to trace to draw two chains of 2,500 posterior samples each, using the NUTS sampler. Furthermore, the NUTS sampler auto-tunes the sampling by 2,500 iterations and returns a trace object. The output from the trace is expected to be normally distributed around the true parameters, which is a sign that samples converged towards the target distribution.

8.2.3 Re-parameterising the Model in NB

The Poisson assumption that the mean and observation variance are equal does not always hold in real life. The author proceeded to re-parameterise the model by considering the effects of overdispersion, non-stationarity, and stochasticity in the datasets. Another way to address these effects in the model is to change the distributional assumption to the NB, in which the variance is larger than the mean. Essentially, the NB log-likelihood describes a Poisson random variable whose rate parameter is gamma-distributed. In other words, the NB can be parametrised either in terms of mu or p or in terms of alpha or n. The link between the parameterisations is provided by equation 8.11.

$$mu = \frac{n(1-p)}{p} \qquad (8.11)$$

n = α

mu = μ = r_t
mu: float
Poisson distribution parameter (mu > 0).

α: float
Gamma distribution parameter (α > 0).

p: float
Alternative probability of success in each trial (0 < p < 1).

n: float
Alternative number of target success trials (n > 0)

Therefore, the NB log-likelihood distribution has similar characteristics to the Poisson distribution, except that it has two parameters (μ and α), which can vary its variance independently of the mean. Therefore, the model is expressed as shown in equation 8.12.

$$Y \sim \text{NegativeBinomial}(\mu, \alpha) \quad (8.12)$$

Incorporating equation 8.1 into 8.12 provides the NB implementation of the model with the ability to measure changes in the mean probability distribution as rates and the equivalent changepoints. The resulting NB model is presented in equation 8.13.

$$Y_t \sim \text{NB}(\alpha, r_t), \quad r_t = \begin{pmatrix} lambda\_1, & if\ changepoint < t_1 \\ lambda\_2, & if\ t_1 \leq changepoint < t_2 \\ lambda\_3, & if\ changepoint \geq t_2 \end{pmatrix} \quad (8.13)$$

The dispersion parameter, α in equation 8.13 allows the model to control overdispersion by adjusting variance independently of the mean. The visual representation of the model's specification is demonstrated in Figure 43. The implementation of equation 8.13 in the PyMC3 package is illustrated in Figure 44, indicating the priors and the novel deterministic random variable – rates. The rates calculate the changepoints and corresponding mean of incident occurrences.

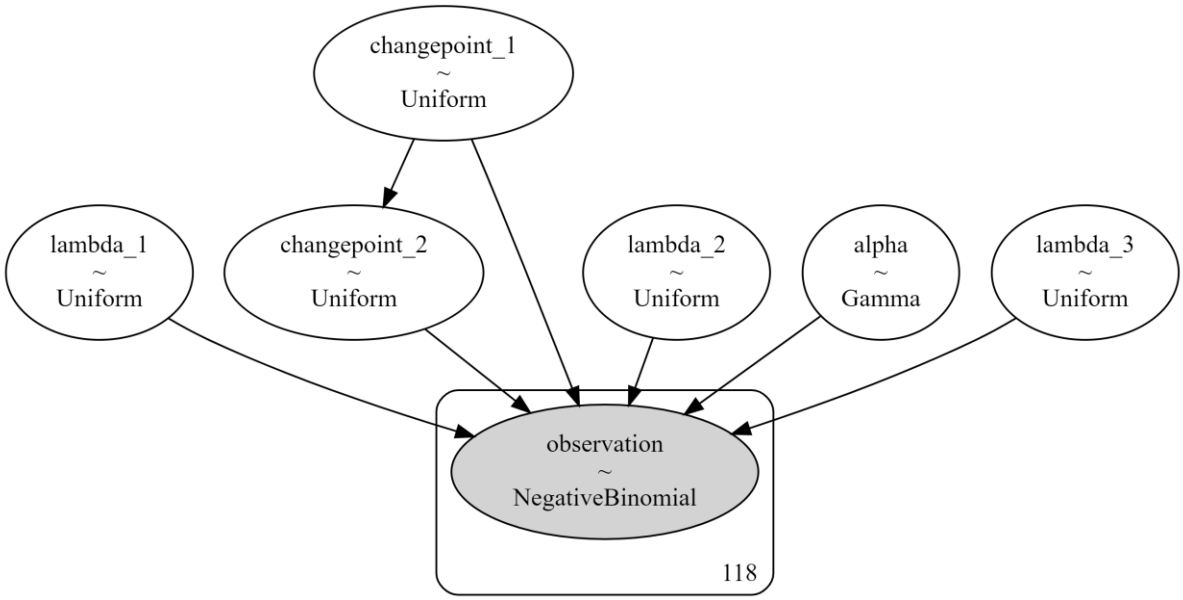

***Figure 43**: Graphical Representation of the Model's Specification*

```
with pm.Model() as twitter_model:

    # The dispersion parameter
    alpha = pm.Gamma('alpha', mu=50, sigma=20)

    # rate parameters
    lambda_1 = pm.Uniform("lambda_1", lower=0, upper=20)
    lambda_2 = pm.Uniform("lambda_2", lower=0, upper=20)
    lambda_3 = pm.Uniform("lambda_3", lower=0, upper=20)

    changepoint_1 = pm.Uniform("changepoint_1", lower=0, upper=
number_of_weeks - 1)
    changepoint_2 = pm.Uniform("changepoint_2", lower=changepoint_1, upper=
number_of_weeks - 1)
    idx = np.arange(number_of_weeks) # Arrange the Index along the weeks

    # The novel form of the deterministic random variable rates of weekly
incidents.
    # It allocates appropriate rates to weeks before and after
changepoints.
    rate_0 = pm.math.switch(changepoint_1 >= idx, lambda_1, lambda_2)
    rate_1 = pm.math.switch(changepoint_2 >= idx, rate_0, lambda_3)

    # Likelihood
    observation = pm.NegativeBinomial("observation", rate_1, alpha,
observed=twitter_obs)
```

***Figure 44**: Model Specification in Negative Binomial Log-Likelihood*

```
with twitter_model:

 trace = pm.sample(2000, cores=2, tune=2000)
```

```
/usr/local/lib/python3.7/dist-packages/pymc3/sampling.py:468: FutureWarning: In an upcoming release, pm.sample
  FutureWarning,
Auto-assigning NUTS sampler...
Initializing NUTS using jitter+adapt_diag...
Multiprocess sampling (2 chains in 2 jobs)
NUTS: [changepoint_2, changepoint_1, lambda_3, lambda_2, lambda_1, alpha]
100.00% [8000/8000 09:56<00:00 Sampling 2 chains, 0 divergences]
Sampling 2 chains for 2_000 tune and 2_000 draw iterations (4_000 + 4_000 draws total) took 598 seconds.
The estimated number of effective samples is smaller than 200 for some parameters.
```

***Figure 45**: The MCMC NUTS Sampler for the Negative Binomial Model*

Finally, the predictive performance of the NB model was evaluated by generating 5,000 samples from the posterior distribution of the model and comparing the observed values to the mean of the posterior observation. The aim was to investigate how closely the estimated posterior distributions of the parameters can generate datasets similar to the observed datasets.

8.3 Result

The Arviz package in Python was applied to access the Poisson model's summary statistics and trace plots to visualise the model's performance. The trace returned the samples from the posterior distribution of each parameter, and the trace object was queried similarly to a dictionary containing a map from variable names to numpy.arrays. The result of the queries is presented in this section.

8.3.1 Poisson Log-likelihood

The acceptance rate is the proportion of proposed values that were not rejected during the sampling process. The Poisson log-likelihood model achieved an acceptance rate of 0.65 and is illustrated in Figure 46.

```
accept.mean()
```

```
0.6549595544926945
```

***Figure 46:** Acceptance Mean Rate of the Poisson Log-Likelihood Model in this Study*

Figure 47 shows the result of the marginal energy distribution of the MCMC NUTS sampler while exploring the samples' space in the Poisson log-likelihood model.

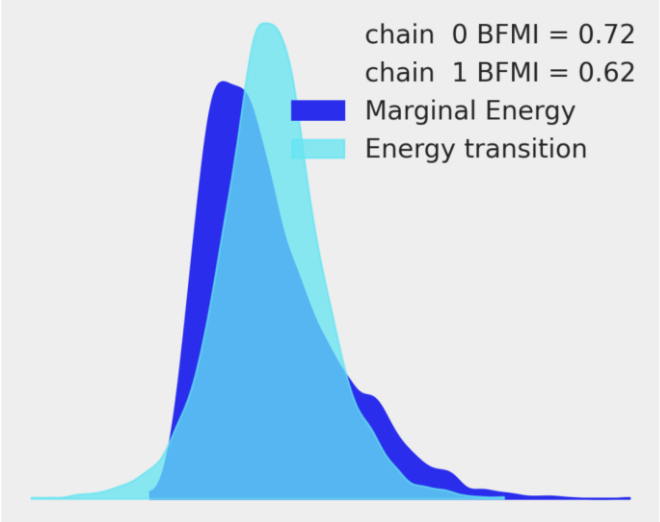

***Figure 47**: Energy Distribution of the Poisson Log-Likelihood Model*

Although not significantly noticeable, the summary statistics of the Poisson model in Figure 48 show that the r_hat values for each parameter are slightly larger than 1.0. The divergence parameter was set to true to investigate where the divergences might have occurred in the model. Figure 49 shows the divergences for the Poisson model, and the black lines underneath the trace plots are the divergences in the model.

	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_mean	ess_sd	ess_bulk	ess_tail	r_hat
lambda_1	6.72	0.27	6.22	7.23	0.02	0.02	146.37	146.37	173.63	135.32	1.02
lambda_2	1.61	1.72	0.00	4.07	0.54	0.39	10.12	10.12	19.81	12.93	1.08
lambda_3	3.84	1.31	1.31	6.49	0.09	0.07	201.63	176.74	213.32	117.94	1.01
changepoint_1	111.87	8.62	109.20	116.99	2.01	1.54	18.35	16.31	46.68	15.99	1.05
changepoint_2	114.21	6.40	113.54	117.00	0.84	0.62	57.84	54.64	78.17	41.17	1.03

***Figure 48**: The Model Summary Statistics*

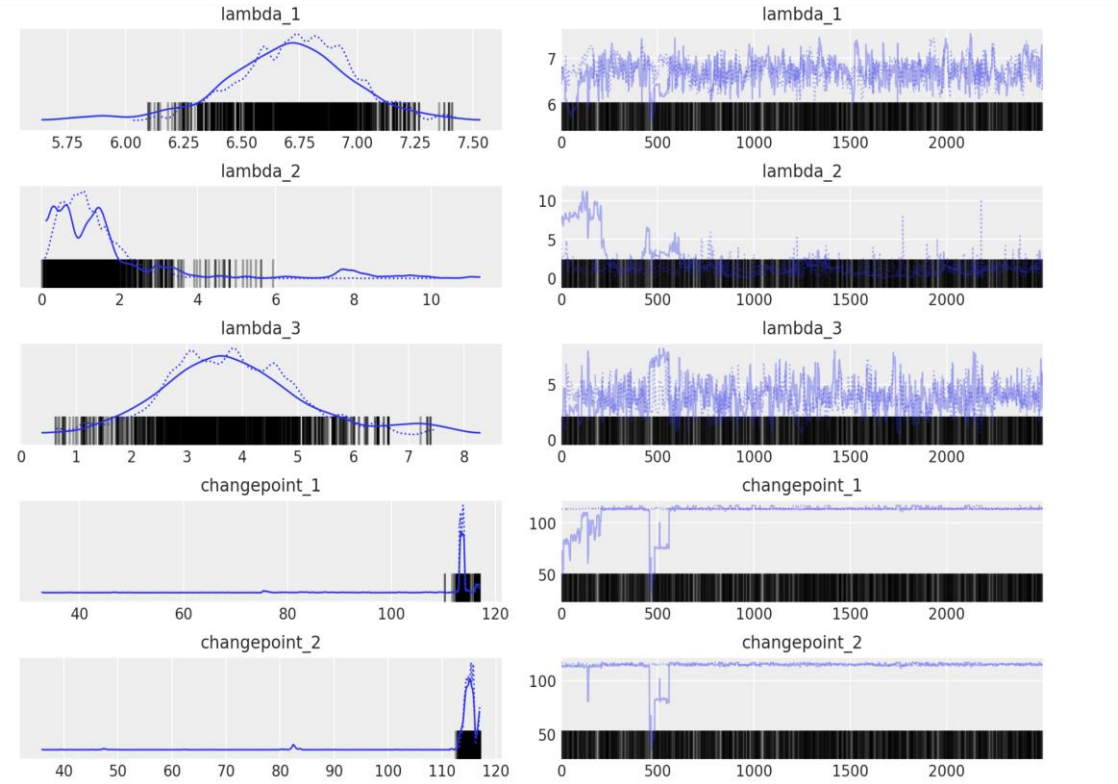

***Figure 49**: Trace Diagnosis to Evaluate the Divergence of Parameters (Poisson Model)*

8.3.2 NB Log-likelihood

For the NB log-likelihood model, it was observed in Figure 45 that the sampler completed the specified iterations without any divergences in the sample distributions. Also, the mean acceptance rate was 0.80, as shown in Figure 50, indicating a better sample acceptance rate.

```
[198] accept.mean()

      0.7986920506421576
```

***Figure 50**: The Mean Acceptance Rate of the Model*

Furthermore, the energy distribution in Figure 51 shows that the MCMC NUTS sampler explored the marginal energy distribution reasonably well. The results of the trace diagnostics in Figure 52 show no divergence, and Figure 53 presents the summary statistics of the NB model. The r_hat values for all parameters were closer to 1.0.

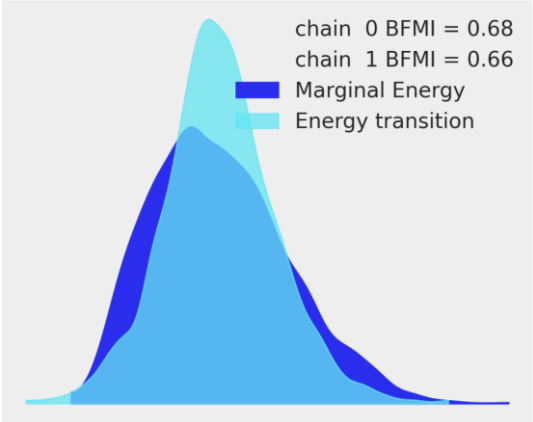

***Figure 51**: Energy Distribution of the Negative Binomial Log-Likelihood Model*

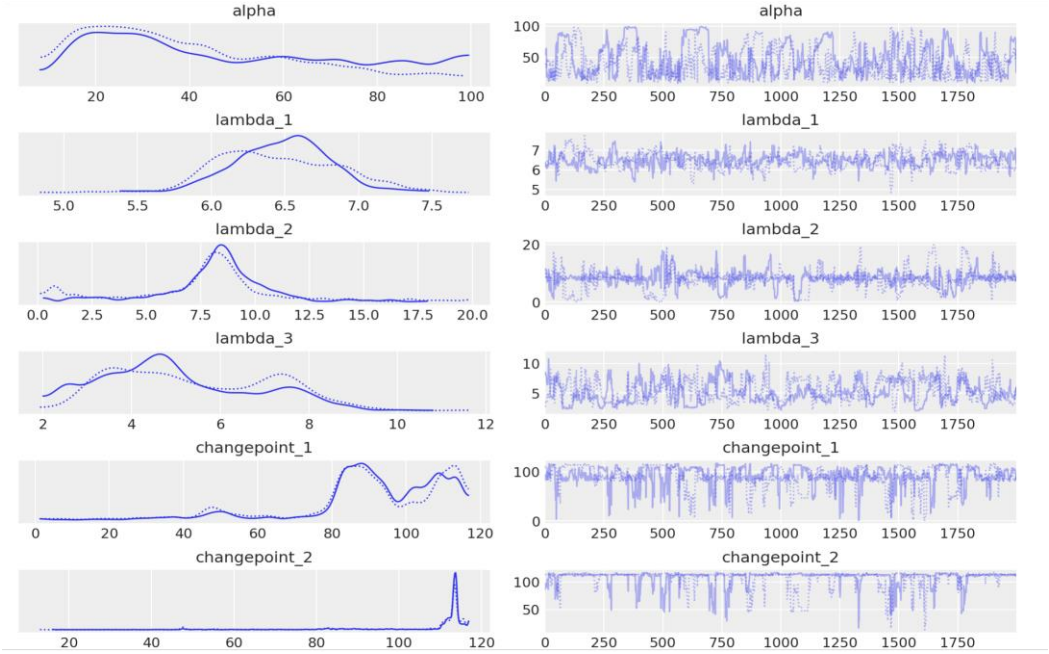

***Figure 52**: Trace Diagnosis of Parameters in the Negative Binomial Model*

	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_mean	ess_sd	ess_bulk	ess_tail	r_hat
alpha	43.96	24.35	10.19	90.75	2.20	1.70	122.06	103.37	138.26	75.81	1.03
lambda_1	6.47	0.37	5.85	7.21	0.03	0.02	140.05	137.30	140.52	247.03	1.03
lambda_2	7.99	3.17	0.16	12.26	0.26	0.19	143.46	143.46	173.80	98.60	1.01
lambda_3	5.24	1.81	2.31	8.27	0.15	0.11	141.93	141.93	148.13	180.44	1.02
changepoint_1	89.83	21.70	46.74	116.99	1.62	1.24	180.20	152.58	143.85	352.98	1.01
changepoint_2	107.05	15.82	78.70	116.99	1.33	0.95	140.71	138.11	167.68	194.60	1.02

***Figure 53**: Summary Statistic of the Negative Binomial Log-Likelihood Model*

Figure 54 presents the result of the predictive performance of the NB model; it shows a very close similarity between the mean of the posterior observation and the observed value.

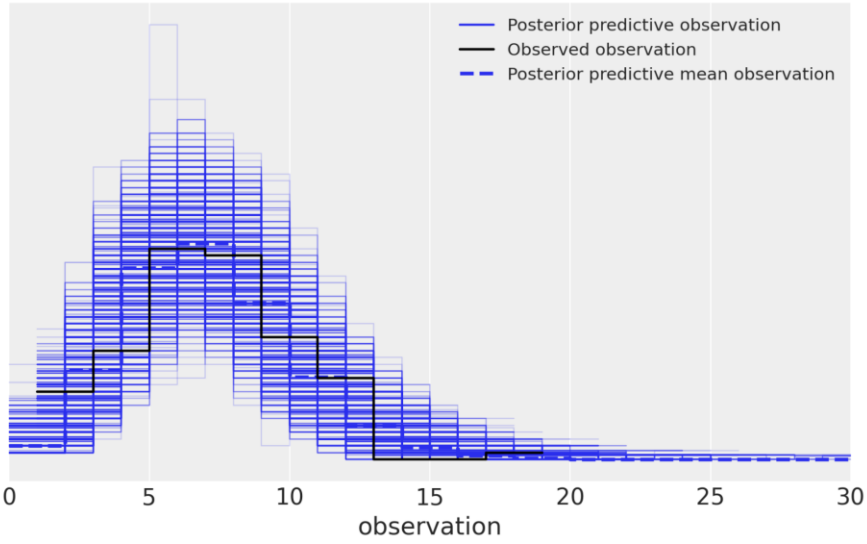

***Figure 54**: Posterior Predictive Performance*

Figure 55 presents the trace plots for all unknown parameters, and each histogram plot presents the estimates of each unknown parameter – the highest density interval (HDI) is the region inside the histogram indicated by the horizontal black lines underneath the trace plots. All estimates within the interval have a higher probability density than those outside the interval. The width of the HDI represents how informative a parameter's estimate might be, which means a large HDI is equivalent to large uncertainty. Therefore, the HDI can be used in the context of uncertainty characterisation of posterior distributions as credible intervals (CI). For PyMC3, the default range of the HDI is 94%, while the points outside the interval represent 6%, indicating 3% on each side of the histogram. The trace plot also provides the mean value of each parameter's distribution.

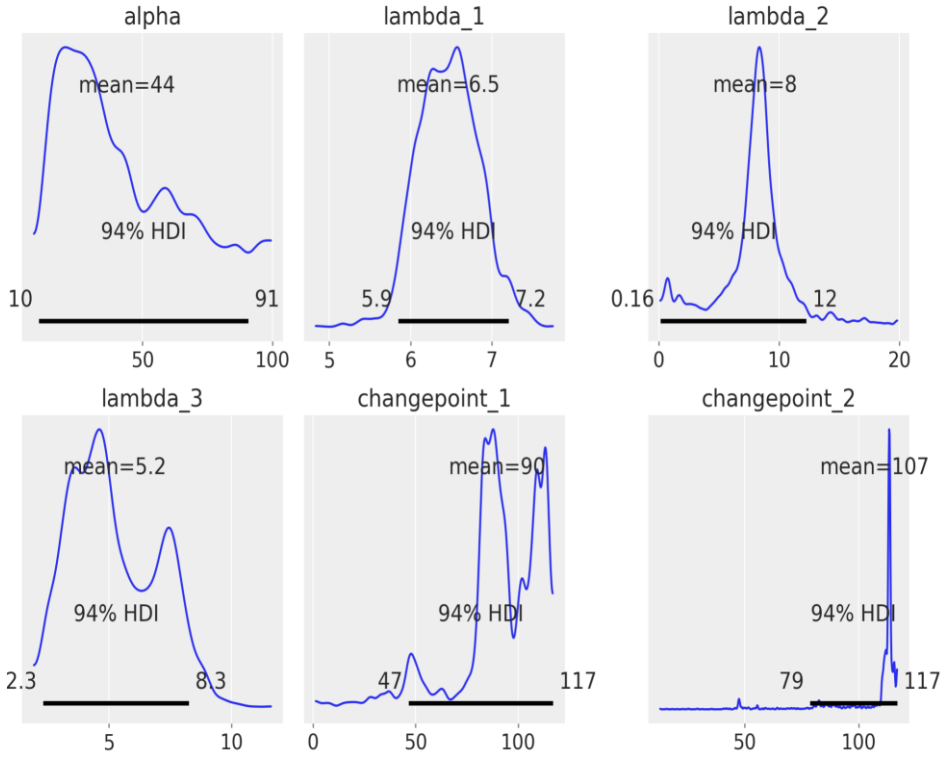

Figure 55: *Trace Plots for the Negative Binomial Log-Likelihood Model*

8.4 Discussion and Conclusion

The acceptance rate is the proportion of proposed values that were not rejected during the sampling process. However, it was reported in the PyMC3 document that the acceptance rate of about 23.4 per cent results is the highest efficiency for Metropolis-Hastings's algorithms. A high acceptance rate, usually about 90 per cent, is a sign that the new samples are being drawn from close points to the existing point. Therefore, the sampler is not exploring the space all that much.

On the other hand, a low acceptance rate could be due to an inappropriate proposal distribution, which can cause the rejection of new samples. PyMC3 package in Python aims to get an acceptance rate between 20 per cent and 50 per cent for Metropolis-Hastings algorithms, about 65 per cent for Hamiltonian Monte Carlo algorithms, and about 85 per cent for NUTS Sampler algorithms (PyMC3 Documentation, 2018). Therefore, the acceptance rate of 0.65 for the Poisson log-likelihood model in this study is well below the expected acceptance rate recommended for the NUTS sampler, which suggests a problem in the model. This might be

due to the characteristics of the available datasets explained in chapter 6, making Poisson not the best modelling approach for the datasets.

In PyMC3, valid inferences from sequences of MCMC samples are based on the assumption that the samples are derived from the true posterior distribution of interest. However, theory guarantees this condition as the number of iterations approaches infinity (Salvatier et al., 2016). Therefore, it is important to determine the minimum number of samples drawn by the MCMC algorithm to ensure a reasonable approximation to the target posterior density. Unfortunately, no universal threshold exists across all problems, so convergence must be assessed independently each time MCMC estimation is performed through a convergence diagnostics method (PyMC3 Documentation, 2018), and the number of samples adjusted accordingly. The convergence was more easily achieved using the NB model than the Poisson model, which can be verified by comparing the results of figures 49 and 52. There were no divergencies in the estimates produced by the NB model.

When the energy transition distribution is narrow relative to the marginal energy distribution, the random walk process in the MCMC sampler will slowly explore the marginal energy distribution, requiring many expensive transitions to survey all relevant energies. This behaviour indicates that the model does not have enough energy to explore the whole parameters' spaces, and the posterior estimation is likely biased. On the other hand, when the two distributions are similar, it means that the sampling algorithm is exploring the marginal energy distribution efficiently, indicating that the closer the similarity between the two energies, the better (Betancourt, 2017). Figures 47 and 51 show that the MCMC sampler in both models explored the samples' spaces reasonably well. In the NB model, the energy distribution ratio was 0.68/0.66, while the Poisson log-likelihood produced 0.72/0.62. Therefore, there is a closer similarity in the energy distributions of the MCMC sampler while sampling in the NB log-likelihood model.

The traces in the Poisson log-likelihood model showed very poor mixing for all the parameters and the chains, as black flatlines could be observed in-between the trace plots in Figure 49, indicating divergencies in the estimates due to the poor performance of the MCMC sampler. Most of the divergences occurred in the posterior highest density regions. Inspecting the summary statistics of the Poisson log-likelihood model in Figure 48, it was observed that the r_hat values of each parameter were reasonably closer to 1.0. Theory suggests that values closer to 1.0 are indicative of high performance. However, the mean effective sample sizes

(ess_mean) were fairly small for most parameters compared to the total number of samples drawn, indicating sampling problems. Two ways to avoid divergences in a model are to increase the tuning samples or the value of the target-accept parameter. However, neither of the approaches produced effective results. Therefore, it was concluded that the Poisson log-likelihood could not effectively support the development of the proposed method due to the complex characteristics of the available data streams discussed in chapter 6.

On the other hand, Figure 50 shows the acceptance rate of 0.80 in the NB model. The result is closer to the prescribed rate in PyMC3 documentation, where 85 per cent was recommended for NUTS sampler algorithms (PyMC3 Documentation, 2018). Furthermore, the energy distribution in Figure 51 shows that the MCMC sampler explored the marginal energy distribution reasonably and had enough energy to discover the whole parameters' spaces and the posterior estimates were derived from true distributions of interest. This may have been responsible for the easy convergence of samples in the model. Furthermore, the r_hat values for all parameters are closer to 1.0, and as a result, it can be concluded that the NB log-likelihood model outperformed the Poisson log-likelihood in developing the SQRA method for VSyS.

The predictive performance of the NB model was evaluated by generating 5,000 samples from its posterior distributions using the estimated parameters. Figure 54 shows a very close similarity between the posterior observations and the observed values, indicating that the model randomly generated posterior observations closely similar to the observed values with little or no deviations.

The summary statistics of the NB model in Figure 53 contain important outputs; the first column contains the parameters that we are interested in, lambda_1, lambda_2, lambda_3, changepoint_1, and changepoint_2. The lambdas represent the rate of occurrence of the disease incidents before the first changepoint, after the first changepoint, but before the second changepoint, and at the second changepoints upward to provide information on risk magnitudes due to the changing trend of the disease outbreak distributions. According to Aven (2012), risk could be defined as a probability distribution. Furthermore, risk can quantify hazards by attributing the probability distribution of being realised to potential harm (National Research Council, 1989).

Since risk is a probability measure, quantitative risk assessment in this study estimates the probability of observing an ongoing disease outbreak, which the lambda parameters described with credible intervals (the HDI). Therefore, the result of the model can be interpreted by the trace plot in Figure 55 as follows. The rate of the GI infection occurrence in dogs was 6.5 incidents per week with a 94% probability that the true value was between 5.9 and 7.2 incidents per week until the changepoint in the 90th week, whereby the rate increased to 8.0 incidents per week with 94% probability that the actual incident rate was between 0.16 and 12 incidents per week. The true changepoint is believed to have occurred between the 47th and 117th week with a 94% probability.

Furthermore, the mean rate of the infection decreased to 5.2 incidents per week, with a 94% probability that the true mean rate was between 2.3 and 8.3 incidents per week. The second changepoint occurred in the 107th week with a 94% probability of anywhere between the 79th and 117th week. The author observed large spans of weeks plausible for changepoints. Also, there were large spans in the mean rate of the disease incident occurrences, indicating that the results were not as informative due to high uncertainties. However, previous studies suggested that uncertainties in the model can be reduced with successive analyses of research datasets. Therefore, the author demonstrated how uncertainty was improved in the NB model and validated the proposed method in the next chapter by incorporating independent observations from heterogeneous data sources.

The technique for developing the SQRA method is divided into eleven stages and is highlighted in Figure 56 below. The detailed step-by-step procedure is subsequently presented and consists of sequential processes applied by the author to accomplish the developed SQRA method in this research work. Therefore, the detailed step-by-step procedure is provided as the main contribution of this research work to public health practice and science.

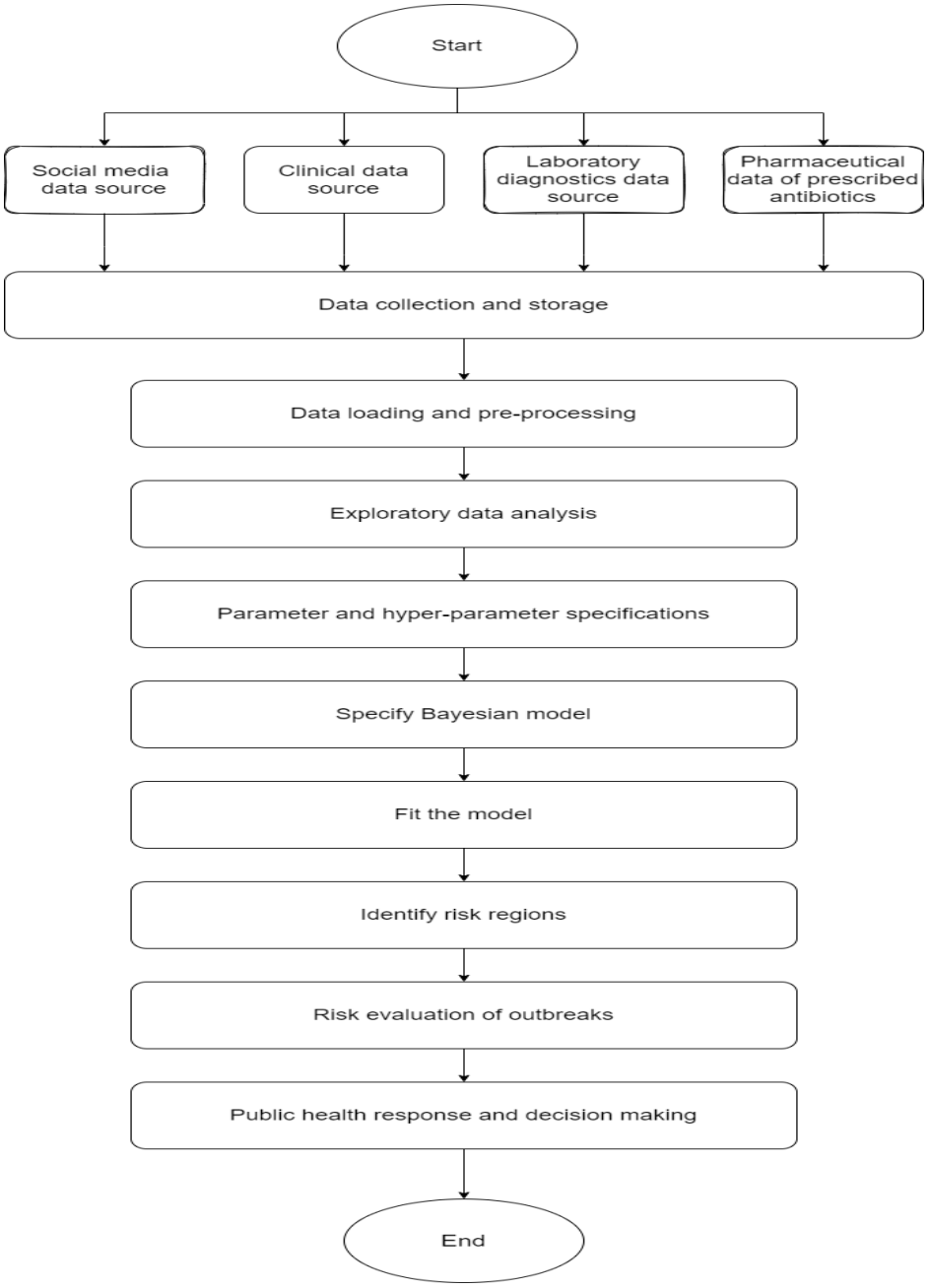

***Figure 56**: Procedure for SQRA Method using Bayesian approach*

The detailed step-by-step procedure for modelling the SQRA method.

1. Identify disease surveillance data sources for which stochastic quantitative risk assessment is required.
2. If social media is one of the sources, identify the optimum means of extracting the data through API with effective keywords.
3. Involve domain experts in defining the keywords.

4. Obtain datasets relating to the infection frequency from the identified data sources and save them in a repository.
5. Load and select the columns required for the proposed SQRA.
6. Investigate data quality and decide on any cleaning and data wrangling exercises required.
7. Examine the summary of the data quality exercise to ensure no observations are missing and no nulls and NAs are present.
8. If missing observations are detected, decide strategy and apply an appropriate technique to fill out the missing observations.
9. Sample each dataset and convert it to time-series data.
10. Investigate the attributes, shapes, and structures of the data by performing descriptive statistics, including frequency distributions, to understand the characteristic of each of the time-series datasets.
11. Investigate the size and spread of the datasets by mean, maximum and minimum values of the observations.
12. Utilise histograms and box plots to visualise the distribution of the datasets to ascertain the underlying distribution of the time series.
13. Perform time-series analysis using autocorrelation and partial autocorrelation to determine repeating patterns or seasonality in the data streams.
14. Perform more time-series analysis with time-series decomposition into seasonality, trend, and residual components to determine the randomness of the observations, time effects, and any potential bias in the data.
15. Furthermore, determine whether the time-series datasets are stationary or non-stationary or under the influence of any seasonal components such as
    a. Seasonality - any periodic pattern
    b. Trend - do the data streams follow a consistent upwards or downward slope
    c. Residual - are there any outlier points or missing values that are not consistent with the rest of the data
16. Determine Bayesian probabilistic programming technique based on flexibility, integration with diagnostic tools such as Arviz, and the ability to specify everything unknown as parameters.
17. Select PyMC3 Bayesian probabilistic programming platform.

18. Test the underlying data distribution and determine which data distribution adequately represents the dataset; use the Pearson chi-square goodness-of-fit test and Poisson log-likelihood to establish the underlying distribution.
19. Define the known and unknown parameters to calculate the change rate of incident occurrences and their corresponding changepoints in each data stream.
20. Determine the appropriate underlying distributions and shapes of each unknown parameter of the model by involving domain experts in disease surveillance.
21. Assign prior values to each unknown parameter based on the appropriate underlying distributions and shapes determined above.
22. Ensure that the priors are defined with continuous data frequency distributions indicating the upper and lower bands depending on the shapes of each dataset instead of discrete data frequency distributions. The reason is to take advantage of the NUT Sampler of the MCMC algorithms instead of the Metropolis-Hastings.
23. Define the Poisson log-likelihood model to use the posterior estimates of previous observations as priors to estimate the posterior of new observations as they are available.
24. Fit the model to a subset of the dataset.
25. Evaluate the model result using Arviz diagnostic tools by plotting traces to visualise the step sizes of the sampler.
26. Check trace acceptance rate and the overall mean acceptance value; low means poor, too high means are not exploring, but about 0.8 means effectively exploring parameter's spaces.
27. Investigate the size of the divergence by accessing the diverging index as an array.
28. Investigate divergence via Arviz plot_pair method to visualise the point at which the parameter space exploration fails to perform.
29. Compare the marginal energy and energy transition plot to verify the sampler's performance.
30. Access the posterior sample distribution plots to investigate performance. Black lines within the distribution plots are divergencies, implying poor performance from the sampler, and the divergences indicate that the model did not converge.
31. Examine the summary statistics, paying attention to r_hat, mean effective sample size – mean_ess and the uncertainty value HDI3% - HDI97%. R_hat values close to 1.0

indicate no problem with the sampling process, but it does not indicate whether or not a divergence has occurred.

32. Ensure the model converges and performs effectively.
33. If the model converges and performs effectively, use the Arviz Model_pred method to sample from the posterior, predict some values, and compare the values with observed values graphically.
34. If the model performance suggests no convergence and is poor, re-parametrise the model and run the iterations again.
35. Re-specify the model in Negative Binomial log-likelihood
36. Define the unknown parameters and hyper-parameters of the model
37. Ensure that the priors are defined with continuous data frequency distributions indicating the upper and lower bands instead of discrete data frequency distributions to take advantage of the NUT sampler.
38. Perform model diagnosis with the Arviz tool as described above from steps 24 - 32.
39. If the model does not converge, re-parametrise the model by changing the distributions of the priors to alternative distributions that closely describe their frequency occurrences.
40. Compare the performance of both models using the outputs from Arviz diagnostic tools.
41. Select the effective model and fit the model to the entire data stream in order of data availability.
42. Perform risk calculation by checking the mean rate of incident occurrences reported by each rate parameter from each data stream and their corresponding changepoints, evaluating the margins of the uncertainties in each parameter.
43. Evaluate the margins of uncertainties to determine whether they are reducing or increasing with each introduced data stream.
44. Determine the regions of risk with a 94% probability that the true (unknown) estimates of the mean rate of disease incident occurrences (risk of outbreaks) and the corresponding changepoints would lie within narrow uncertainty intervals (informative), given the evidence provided by the observed data streams.
45. Based on informative posterior distribution outcomes, decisions can be made to determine whether there is an ongoing disease outbreak risk.

Chapter 9 – Validation and Verification of the SQRA Method

9.1 Chapter Introduction

The most common dataset for evaluating public health risks are measures of disease frequency or the count of infectious diseases observed in a population. In this study, the author agreed with previously published research that the problem with such datasets is overdispersion, non-stationarity, and stochasticity. A common approach emerging from the literature is analysing such count datasets using the Frequentist modelling approach. However, in this work, the author argued in favour of applying the NB modelling approach to surveillance datasets from a Bayesian standpoint.

Fitting VSyS models to social media datasets is still scanty in veterinary public health studies. However, the approach demonstrated in this study fitted the developed SQRA model to the Twitter data stream to demonstrate the usefulness of the social media dataset for analysing disease outbreaks. Despite the high performance of the NB model compared to the Poisson log-likelihood, large uncertainty was found in the model estimates and its parameters. While veterinary disease surveillance is rarely investigated with uncertainty measurement in mind (Smith et al., 2017), other previously published work argued that the risk concept and the notion of uncertainty are related because uncertainty always accompanies risk assessment techniques (Aven, 2012). Therefore, estimating the risk of disease outbreaks with uncertainty measurement may improve risk communication to decision-makers and enhance outbreak detection sensitivity.

Consequently, the proposed and developed SQRA method need to be evaluated in this chapter to examine whether it fulfils the objectives of this research study. To successfully evaluate the proposed method, the author must focus on how the method measures and reduces uncertainties in the model and its parameters and identify risk regions in syndromic surveillance data. According to Vose (2008), the best practice when modelling risk is to separate stochasticity from uncertainty. Future assessments can reduce uncertainty, such as the sequential addition of heterogeneous data sources, whereas stochasticity cannot be further reduced.

As a result, the author will evaluate the developed SQRA method by first fitting the model to the Twitter dataset to infer a possible disease outbreak with associated uncertainties. Subsequently, update the inferences successively using datasets from SAVSNET to reduce

uncertainties in the model and its parameters to validate disease outbreaks. This approach will demonstrate the method's applicability and reliability for risk assessment of disease outbreaks in public health practice. Furthermore, the validation findings and other findings from the research work will be discussed.

9.2 Method

The detailed step-by-step procedure of the proposed SQRA method developed as the main contribution of this research work to public health practice in chapter 8, section 8.4, was applied to fit the NB model to the available datasets sequentially. The Twitter, veterinary practice and laboratory diagnostics datasets were loaded into the Google Colab platform for initial pre-processing. The full descriptions of all datasets were already provided in chapter 4 of this study, with the first five rows shown in figures 8 to 10. The columns that indicate the counts of occurrences of the disease outbreak were selected, pre-processed, and timestamped.

In Figure 8, the relevant columns from the veterinary practice dataset are QuestinairreAcuteVomitting, TextMiningAcutevomitting, AntiVomittingDrugPrescribed, and AntibioticPrescribed. In Figure 9, the relevant columns for the laboratory diagnostics dataset are AssayName and RESULTVAL. Each dataset was aggregated into weekly data since previous studies recommend weekly aggregate as the best practice for modelling epidemiological datasets to avoid zero-inflated frequency (Schmidt and Pereira, 2011).

The model was fitted to the available data streams in the order of their availability in the epidemiological data lifecycle described in chapter 1 and illustrated in Figure 1. It was considered that the Twitter data stream is the most readily available dataset since anyone can access the social media dataset. This was followed by the data generated in the veterinary clinics, hospitals, and farms, triggered by animal visits to veterinary surgery or hospitalisation. The last data in the data cycle was the diagnostic laboratory dataset. The order in which the model was fitted to the data is presented below and illustrated in Figure 57.

1. Twitter data stream
2. Veterinary practice data stream
    a. QuestionaireAcuteVomiting dataset
    b. TextMiningAcuteVomitting
    c. AntibioticDrugPrescribed data
    d. Gastroenteric data

3. Diagnostic Laboratory Dataset
    a. Campylocbactria data

Each data stream was selected from the first week of March 2018. This was to ensure that all data streams began from the same week of the year. Prior values were defined for each unknown parameter as described in chapter 8, section 8.2.3. Our novel deterministic random variable, rate (r_t) was defined in Figure 43 as r_0 and r_1 to infer changes in the mean rate of disease outbreaks and their corresponding changepoints.

Uniform distributions were assigned to define each prior parameter, as indicated in Figure 43. This is because the values of each parameter could be anywhere within a specified range, and this could be better explained with a uniform distribution. For example, the changepoints could have occurred any time during the entire outbreak, and the mean rate of the outbreaks could be any values above or below the true mean rate of the observed data. Since the observed true mean of the Twitter dataset is 6.69, the following priors could be assigned to the mean rates, and this approach is essentially a Bayesian methodology. It allows researchers to assign prior values based on the existing knowledge of the ongoing event.

lambda_1 = pm.Uniform("lambda_1", lower=0, upper=20)

lambda_2 = pm.Uniform("lambda_2", lower=0, upper=20)

lambda_3 = pm.Uniform("lambda_3", lower=0, upper=20)

The corresponding priors of the changepoints are;

changepoint_1 = pm.Uniform("changepoint_1", lower=0, upper= number_of_weeks - 1)

changepoint_2 = pm.Uniform("changepoint_2", lower=changepoint_1, upper= number_of_weeks - 1)

The model was fitted to the Twitter dataset after specifying the priors as above, and PyMC3 triggered the MCMC sampling process using the NUTS technique to sample the posterior probability distribution of each unknown parameter to estimate their expected values.

For successive iterations using other available datasets, prior values were selected for the unknown mean rates by ensuring that a larger range than the true mean of the observed dataset was assigned. Also, the range of the estimated posterior distributions from the previous model was passed as priors to initialise the changepoint parameters in the successive model. The

model specifications and prior definitions for each dataset are presented in Figure 57, indicating the successive iterations for fitting the proposed model to the Twitter and the SAVSNET datasets.

a. Twitter
b. QuestionaireAcuteVomiting dataset
c. TextMiningAcuteVomitting
d. AntibioticDrugPrescribed data
e. Gastroenteric data
f. Campylocbactria dataset

a.

```
with pm.Model() as twitter_model:

    # The dispersion parameter
    alpha = pm.Uniform('alpha', lower=0, upper=100)

    # rate parameters
    lambda_1 = pm.Uniform("lambda_1", lower=0, upper=20)
    lambda_2 = pm.Uniform("lambda_2", lower=0, upper=20)
    lambda_3 = pm.Uniform("lambda_3", lower=0, upper=20)


    changepoint_1 = pm.Uniform("changepoint_1", lower=0, upper= number_of_weeks - 1)
    changepoint_2 = pm.Uniform("changepoint_2", lower=changepoint_1, upper= number_of_weeks - 1)

    idx = np.arange(number_of_weeks) # Arrange the Index along the weeks

    # Our novel form of the deterministic random variable rates of weekly incidents.
    # It allocates appropriate rates to weeks before and after changepoints.
    rate_0 = pm.math.switch(changepoint_1 >= idx, lambda_1, lambda_2)
    rate_1 = pm.math.switch(changepoint_2 >= idx, rate_0, lambda_3)

    observation = pm.NegativeBinomial("observation", rate_1, alpha, observed=twitter_obs)
```

b.

```
with pm.Model() as QuestionnaireAcuteVomiting_model:

    # The dispersion parameter
    alpha = pm.Uniform('alpha', lower=az.hdi(trace[alpha]).min(), upper=az.hdi(trace[alpha]).max())

    # rate parameters
    lambda_1 = pm.Uniform("lambda_1", lower=0, upper=100)
    lambda_2 = pm.Uniform("lambda_2", lower=0, upper=100)
    lambda_3 = pm.Uniform("lambda_3", lower=0, upper=100)


    changepoint_1 = pm.Uniform("changepoint_1", lower=az.hdi(trace[changepoint_1]).min(), upper=az.hdi(trace[changepoint_1]).max())
    changepoint_2 = pm.Uniform("changepoint_2", lower=az.hdi(trace[changepoint_2]).min(), upper=az.hdi(trace[changepoint_2]).max())

    idx = np.arange(number_of_weeks) # Arrange the Index along the weeks

    # Our novel form of the deterministic random variable rates of weekly incidents.
    # It allocates appropriate rates to weeks before and after changepoints.
    rate_0 = pm.math.switch(changepoint_1 >= idx, lambda_1, lambda_2)
    rate_1 = pm.math.switch(changepoint_2 >= idx, rate_0, lambda_3)

    observation = pm.NegativeBinomial("observation", rate_1, alpha, observed=QuestionnaireAcuteVomiting_obs)
```

c.

```
[ ]  with pm.Model() as TextMiningAcuteVomiting_model:

        # The dispersion parameter
        alpha = pm.Uniform('alpha', lower=az.hdi(trace[alpha]).min(), upper=az.hdi(trace[alpha]).max())

        # rate parameters
        lambda_1 = pm.Uniform("lambda_1", lower=0, upper=100)
        lambda_2 = pm.Uniform("lambda_2", lower=0, upper=100)
        lambda_3 = pm.Uniform("lambda_3", lower=0, upper=100)


        changepoint_1 = pm.Uniform("changepoint_1", lower=az.hdi(trace[changepoint_1]).min(), upper=az.hdi(trace[changepoint_1]).max())
        changepoint_2 = pm.Uniform("changepoint_2", lower=az.hdi(trace[changepoint_2]).min(), upper=az.hdi(trace[changepoint_2]).max())

        idx = np.arange(number_of_weeks) # Arrange the Index along the weeks

        # Our novel form of the deterministic random variable rates of weekly incidents.
        # It allocates appropriate rates to weeks before and after changepoints.
        rate_0 = pm.math.switch(changepoint_1 >= idx, lambda_1, lambda_2)
        rate_1 = pm.math.switch(changepoint_2 >= idx, rate_0, lambda_3)

        observation = pm.NegativeBinomial("observation", rate_1, alpha, observed=TextMiningAcuteVomiting_obs)
```

d.

```
[ ]  with pm.Model() as AntiVomitingDrugPrescribed_model:

         # The dispersion parameter
         alpha = pm.Uniform('alpha', lower=az.hdi(trace[alpha]).min(), upper=az.hdi(trace[alpha]).max())

         # rate parameters
         lambda_1 = pm.Uniform("lambda_1", lower=0, upper=1000)
         lambda_2 = pm.Uniform("lambda_2", lower=0, upper=1000)
         lambda_3 = pm.Uniform("lambda_3", lower=0, upper=1000)


         changepoint_1 = pm.Uniform("changepoint_1", lower=az.hdi(trace[changepoint_1]).min(), upper=az.hdi(trace[changepoint_1]).max())
         changepoint_2 = pm.Uniform("changepoint_2", lower=az.hdi(trace[changepoint_2]).min(), upper=az.hdi(trace[changepoint_2]).max())

         idx = np.arange(number_of_weeks) # Arrange the Index along the weeks

         # Our novel form of the deterministic random variable rates of weekly incidents.
         # It allocates appropriate rates to weeks before and after changepoints.
         rate_0 = pm.math.switch(changepoint_1 >= idx, lambda_1, lambda_2)
         rate_1 = pm.math.switch(changepoint_2 >= idx, rate_0, lambda_3)

         observation = pm.NegativeBinomial("observation", rate_1, alpha, observed=AntiVomitingDrugPrescribed_obs)
```

e.

```
with pm.Model() as gastroenteric_model:

    # The dispersion parameter
    alpha = pm.Uniform('alpha', lower=az.hdi(trace[alpha]).min(), upper=az.hdi(trace[alpha]).max())

    # rate parameters
    lambda_1 = pm.Uniform("lambda_1", lower=0, upper=1000)
    lambda_2 = pm.Uniform("lambda_2", lower=0, upper=1000)
    lambda_3 = pm.Uniform("lambda_3", lower=0, upper=1000)


    changepoint_1 = pm.Uniform("changepoint_1", lower=az.hdi(trace[changepoint_1]).min(), upper=az.hdi(trace[changepoint_1]).max())
    changepoint_2 = pm.Uniform("changepoint_2", lower=az.hdi(trace[changepoint_2]).min(), upper=az.hdi(trace[changepoint_2]).max())

    idx = np.arange(number_of_weeks) # Arrange the Index along the weeks

    # Our novel form of the deterministic random variable rates of weekly incidents.
    # It allocates appropriate rates to weeks before and after changepoints.
    rate_0 = pm.math.switch(changepoint_1 >= idx, lambda_1, lambda_2)
    rate_1 = pm.math.switch(changepoint_2 >= idx, rate_0, lambda_3)


    observation = pm.NegativeBinomial("observation", rate_1, alpha, observed=gastroenteric_obs)
```

f.

```
with pm.Model() as campylobacterCulture_model:

    # The dispersion parameter
    alpha = pm.Uniform('alpha', lower=az.hdi(trace[alpha]).min(), upper=az.hdi(trace[alpha]).max())

    # rate parameters
    lambda_1 = pm.Uniform("lambda_1", lower=0, upper=200)
    lambda_2 = pm.Uniform("lambda_2", lower=0, upper=200)
    lambda_3 = pm.Uniform("lambda_3", lower=0, upper=200)


    changepoint_1 = pm.Uniform("changepoint_1", lower=az.hdi(trace[changepoint_1]).min(), upper=az.hdi(trace[changepoint_1]).max())
    changepoint_2 = pm.Uniform("changepoint_2", lower=az.hdi(trace[changepoint_2]).min(), upper=az.hdi(trace[changepoint_2]).max())

    idx = np.arange(number_of_weeks) # Arrange the Index along the weeks

    # Our novel form of the deterministic random variable rates of weekly incidents.
    # It allocates appropriate rates to weeks before and after changepoints.
    rate_0 = pm.math.switch(changepoint_1 >= idx, lambda_1, lambda_2)
    rate_1 = pm.math.switch(changepoint_2 >= idx, rate_0, lambda_3)

    observation = pm.NegativeBinomial("observation", rate_1, alpha, observed=campylobacterCulture_obs)
```

Figure 57: *Model specification showing the successive iteration, prior parameter definitions and the model likelihood*

9.3 Result

9.3.1 The Twitter Data Stream

Figures 58 and 59 illustrate the Twitter model's trace plots and the summary statistics, respectively. They explain the posterior distributions' highest density interval (HDI) for estimating the unknown parameters, such as the rate of incidents, changepoints, and their mean values. From Figure 58, the model produced an expected mean rate of GI outbreak incidents of 6.3 per week with a 94% probability that the true expected value was between 5.5 and 7.1 incidents per week until the changepoint_1 at the 77th week. The mean rate of the GI outbreak incident increased to 7.7 incidents per week with a 94% probability that the true expected value was between 0.37 and 13 incidents per week. The true value of changepoint_1 is believed to have occurred between the 33rd and 108th weeks.

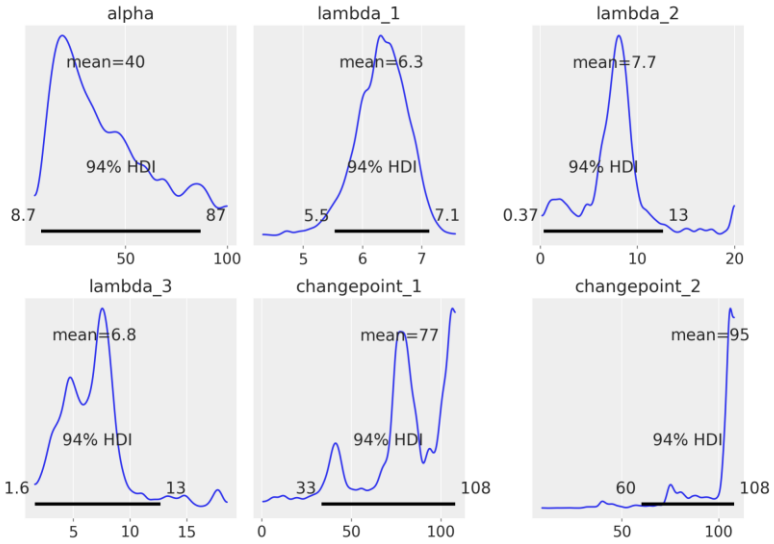

***Figure 58**: Highest Density Intervals of the Posterior Distribution of Parameters for the Twitter Model*

	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_mean	ess_sd	ess_bulk	ess_tail	r_hat
alpha	40.32	23.56	8.65	86.86	1.80	1.35	170.68	152.50	214.59	274.55	1.01
lambda_1	6.32	0.43	5.53	7.13	0.04	0.03	120.33	119.53	119.36	317.05	1.01
lambda_2	7.66	3.41	0.37	12.66	0.35	0.26	93.74	86.26	104.81	121.27	1.03
lambda_3	6.85	3.04	1.61	12.68	0.38	0.27	65.42	65.42	61.95	92.00	1.05
changepoint_1	76.92	24.57	33.37	107.93	2.33	1.82	111.19	91.24	87.37	178.59	1.01
changepoint_2	95.49	17.42	60.02	108.00	1.73	1.31	101.72	89.47	70.94	226.22	1.05

***Figure 59**: Summary Statistic of the Twitter Model*

Then, the mean rate of the outbreak incident decreased to 6.8 with a 94% probability that the true expected value was between 1.6 and 13 incidents per week, and the expected mean rate of the incidents changed in the 95th week with a 94% probability that the true value of the changepoint was between 60th and 108th week. However, there was an overlap between the changepoints (changepoint_1 and changepoint_2) and a large duration in weeks between both changepoints' lower and upper limits, indicating large uncertainties in the weeks plausible for the change to have occurred. Also, overlaps were discovered in the expected mean rate values of the outbreak incidents (lambda_1, lambda_2, and lambda_3). Figure 60 shows the trend of the expected number of outbreak incidents, and the changepoints are illustrated in Figure 61 as orange vertical lines with the credible intervals in orange and green colours, respectively.

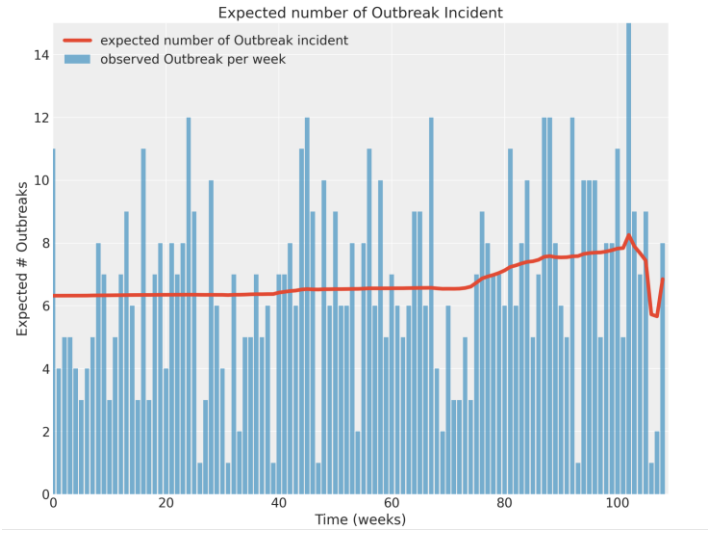

Figure 60: *The Trend of the Outbreak Incidents by the Twitter Model*

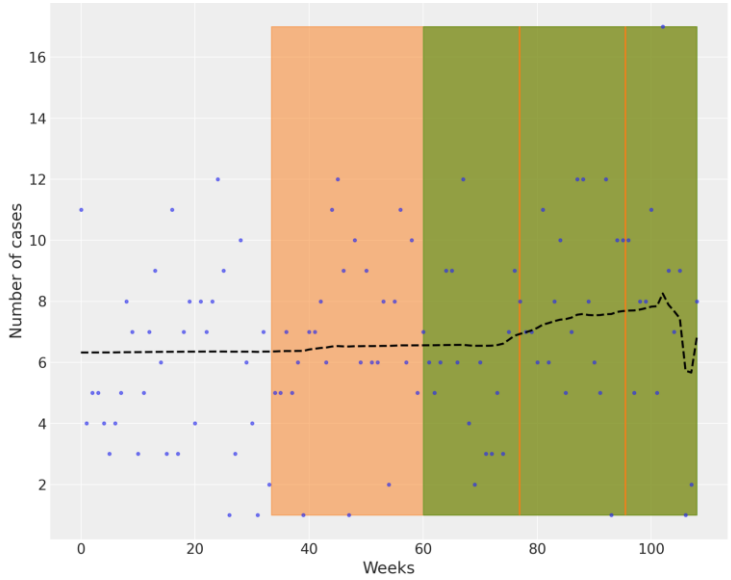

Figure 61*: The Rate of the Outbreak Incidents (Lamda_1, Lambda_2, and Lambda_3) and the Corresponding Changepoints from the Twitter Dataset*

9.3.2 The QuestionaireAcuteVomiting Dataset

Figures 62 and 63 illustrate the QuestionaireAcuteVomiting model's summary statistics and trace plots. The trace plots show an expected mean rate of 95 incidents per week with a 94% probability that the true expected value was between 90 and 100 incidents per week until the changepoint_1 at the 46th week, when the mean rate of the GI outbreak decreased to 75 incidents per week. There is a 94% probability that the credible interval of the true mean rate of incidents was between 69 and 81 per week. The true value of the corresponding

changepoint_1 is believed to have occurred between the 38th and 56th weeks. However, the mean rate of the infection increased to 98 incidents with a 94% probability that the true expected value was between 94 and 100 incidents per week. The change occurred in the 92nd week with a 94% probability that the true value of the changepoint was between the 88th and 96th weeks. Figure 64 shows the trend of the outbreak incidents from the QuestionaireAcuteVomiting dataset, and Figure 65 indicates a clear separation of the changepoints, revealing smaller uncertainties in parameter estimates.

	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_mean	ess_sd	ess_bulk	ess_tail	r_hat
alpha	20.52	3.72	13.55	27.51	0.17	0.12	465.70	458.23	472.03	538.31	1.00
lambda_1	95.36	3.14	90.00	100.00	0.16	0.12	363.21	363.21	336.05	283.68	1.00
lambda_2	74.64	3.92	68.63	80.80	0.34	0.25	136.62	126.65	352.35	239.71	1.01
lambda_3	97.62	5.32	94.18	100.00	0.36	0.27	217.45	198.65	553.68	463.88	1.01
changepoint_1	46.10	7.82	37.96	56.37	0.59	0.46	176.20	146.70	355.32	345.87	1.00
changepoint_2	91.79	2.53	87.92	95.92	0.17	0.12	222.83	206.40	537.45	266.20	1.01

***Figure 62**: Summary Statistics of QuestionaireAcuteVomiting Model*

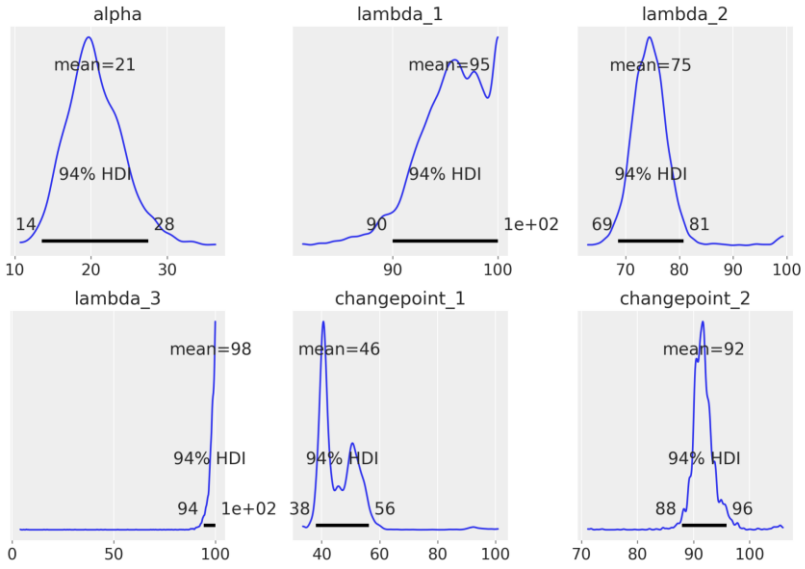

***Figure 63**: Highest Density Intervals of the Posterior Distribution for QuestionaireAcuteVomiting Model*

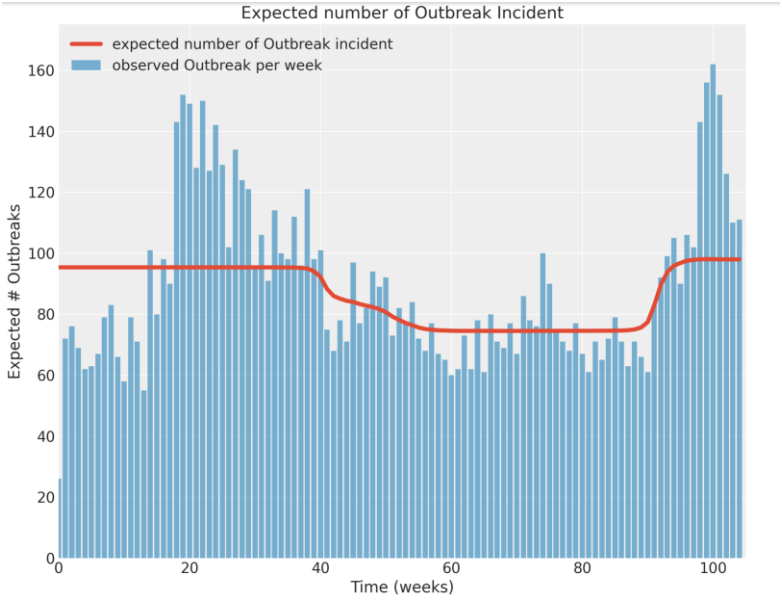

***Figure 64**: The Trend of the Outbreak Incidents by QuestionaireAcuteVomiting Model*

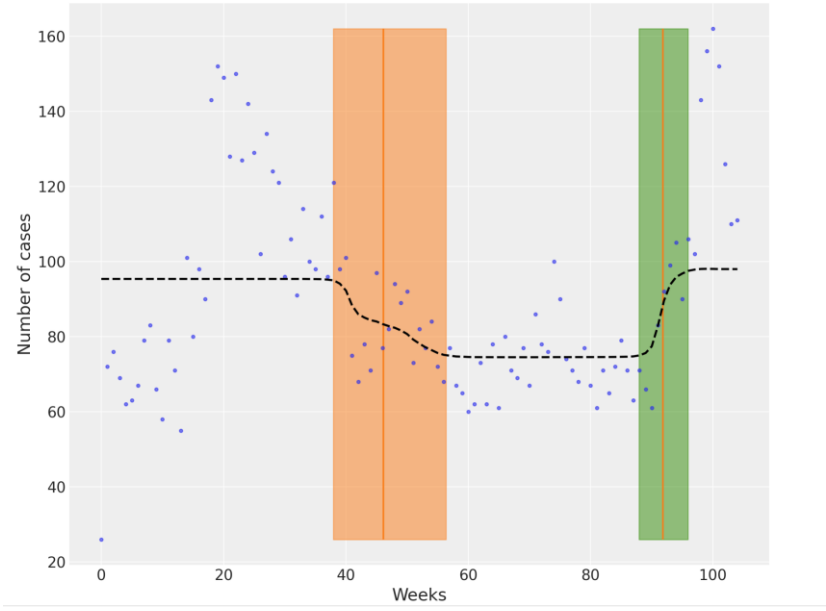

***Figure 65**: The Rate of the Outbreak Incidents (Lamda_1, Lambda_2, and Lambda_3) and the Corresponding Changepoints from QuestionaireAcuteVomiting Dataset*

9.3.3 The TextMiningAcuteVomiting Dataset

Figures 66 and 67 illustrate the model's summary statistics and trace plots. It shows an expected mean rate of 30 incidents per week with a 94% probability, and the true expected value was between 27 and 32 incidents per week. The changepoint_1 occurred in the 41st week when the

mean rate of the GI infection increased to 38 incidents per week with a credible interval between 35 and 41 incidents per week at a 94% probability distribution. The true value of changepoint_1 is believed to have occurred between the 38th and 45th weeks. However, the mean rate of the infection increased to 78 incidents with a 94% probability that the true expected value was between 66 and 92 incidents per week. The changes occurred in the 95th week with a 94% probability that the true value of the changepoint_2 was between the 93rd and 96th weeks.

Figure 68 illustrates the trend of the outbreak incidents. In Figure 69, the smaller range of credible intervals in the changepoint parameters indicates informative estimates that differentiate regions of high and low risks.

	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_mean	ess_sd	ess_bulk	ess_tail	r_hat
alpha	21.98	3.37	16.29	27.51	0.12	0.08	814.97	794.29	759.77	900.48	1.0
lambda_1	29.69	1.41	27.09	32.30	0.05	0.03	944.24	944.24	938.12	1612.46	1.0
lambda_2	37.96	1.48	35.17	40.73	0.05	0.04	760.50	754.21	766.81	1148.73	1.0
lambda_3	77.70	6.92	65.83	91.70	0.30	0.22	543.42	513.12	552.41	540.66	1.0
changepoint_1	40.63	2.65	37.96	44.97	0.09	0.06	890.47	890.47	747.26	679.76	1.0
changepoint_2	94.91	1.14	93.11	95.92	0.04	0.03	812.22	808.56	807.56	1426.34	1.0

***Figure 66**: Summary Statistics of TextMiningAcuteVomiting Model*

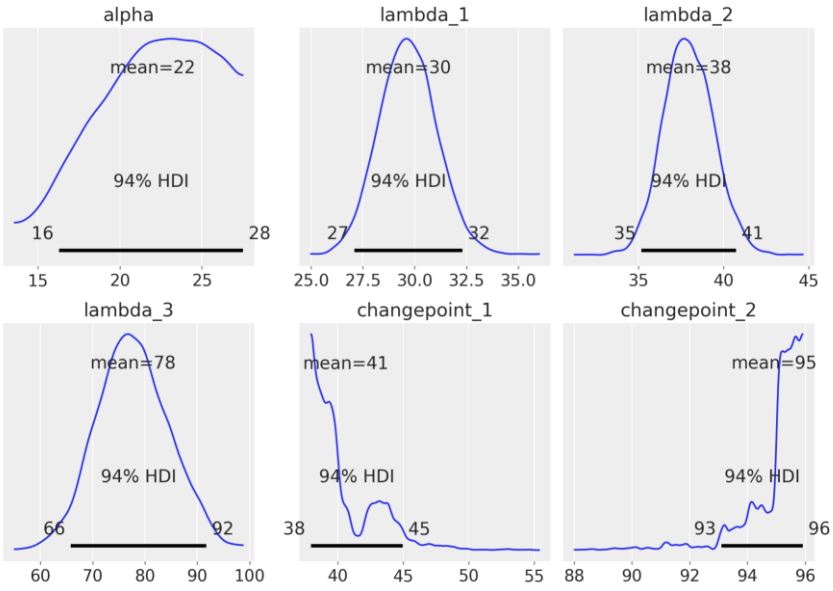

***Figure 67:** Highest Density Intervals of the Posterior Distribution for TextMiningAcuteVomiting Model*

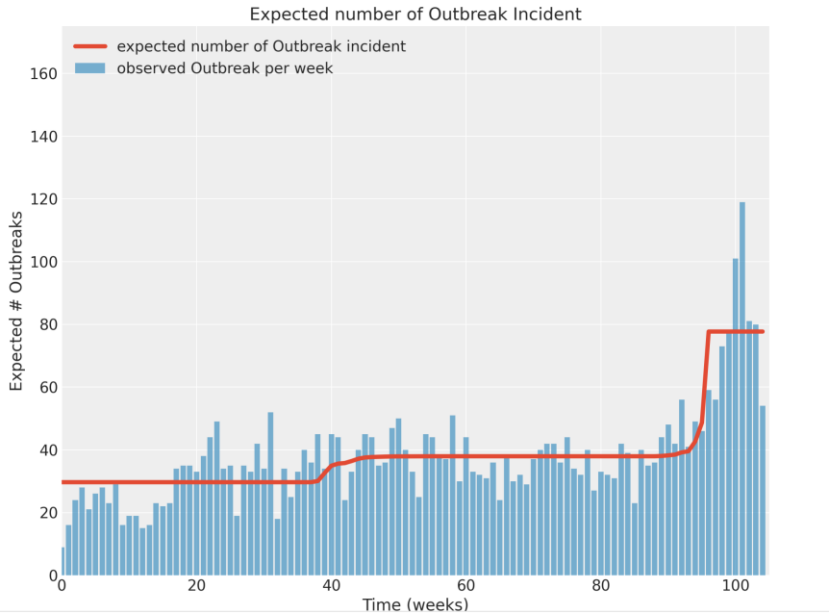

Figure 68: *The Trend of the Outbreak Incidents by TextMiningAcuteVomiting Model*

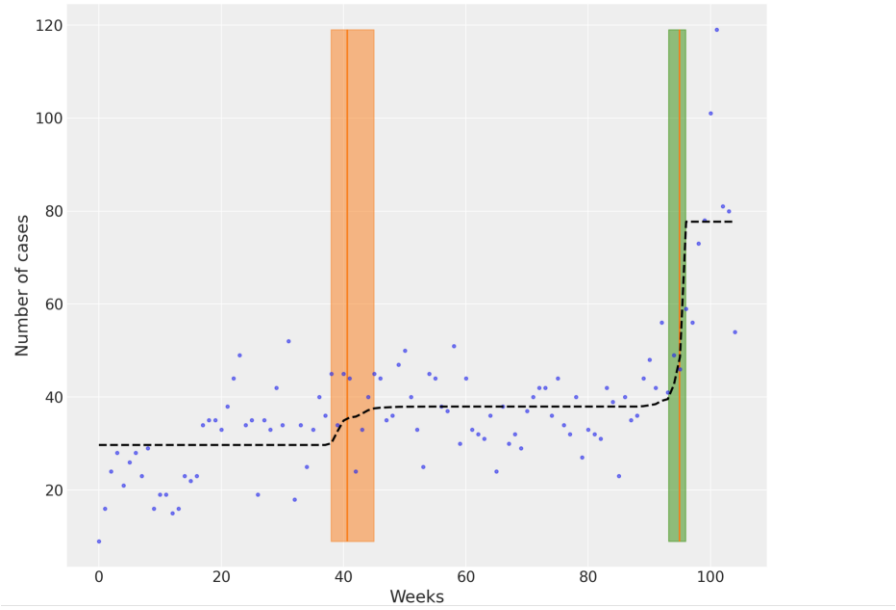

Figure 69: *The Rate of the Outbreak Incidents (Lamda_1, Lambda_2, and Lambda_3) and the Corresponding Changepoints from TextMiningAcuteVomiting Dataset*

9.3.4 The AntiVomitingDrugPrescribed Dataset

The summary statistics and trace plots in figures 70 and 71 show an expected mean rate of 320 incidents per week with a 94% probability that the true expected value was between 301 and 341 incidents per week. The changepoint_1 occurred in the 42nd week when the mean rate of

the GI outbreak increased to 396 incidents per week. There is a credible interval that the true expected value of the mean rate was between 374 and 418 incidents per week with a 94% probability. The true value of changepoint_1 is believed to have occurred between the 38th and 44th weeks. Then, the mean infection rate increased to 835 incidents with a 94% probability that the true expected value was between 734 and 946 incidents per week. The change occurred in the 95th week with a 94% probability that the true value of the changepoint_2 occurred between the 93rd and 96th weeks.

Figure 72 illustrates the trend of the outbreak incidents, and in Figure 73, the parameter estimates and the range of credible intervals in the changepoint are similar to the previous iteration in the TextMiningAcuteVomiting model.

	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_mean	ess_sd	ess_bulk	ess_tail	r_hat
alpha	24.49	2.12	20.76	27.51	0.03	0.02	4432.52	4269.21	3473.83	2530.47	1.0
lambda_1	320.35	10.49	301.29	340.71	0.17	0.12	3842.22	3830.64	3871.48	4069.14	1.0
lambda_2	396.32	11.85	374.20	418.23	0.19	0.13	4039.05	3998.87	4072.42	3028.89	1.0
lambda_3	835.25	57.54	733.68	945.88	1.44	1.05	1605.21	1493.20	1609.57	1037.32	1.0
changepoint_1	41.53	2.09	37.96	44.41	0.03	0.02	4079.71	4079.71	3964.53	3161.48	1.0
changepoint_2	94.79	0.88	93.30	95.92	0.01	0.01	3827.21	3818.93	3546.29	3543.37	1.0

***Figure 70**: Summary Statistics of AntiVomitingDrugPrescribed Model*

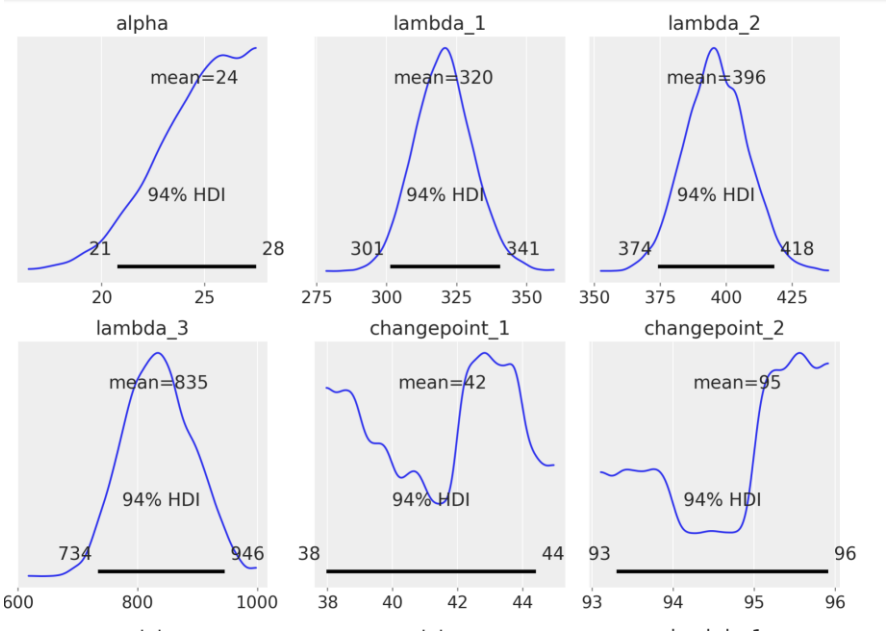

***Figure 71**: Highest Density Intervals of the Posterior Distribution for AntiVomitingDrugPrescribed Model*

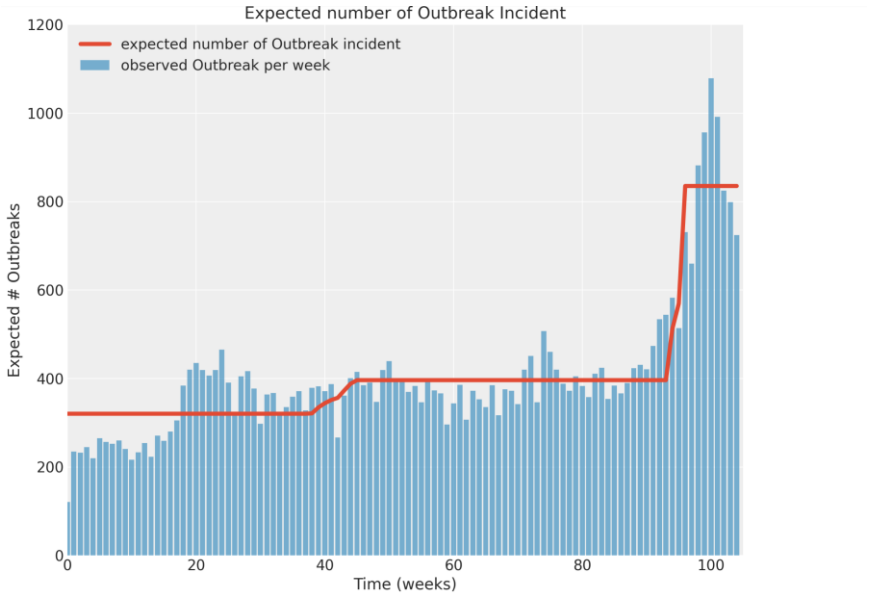

***Figure 72**: The Trend of the Outbreak Incidents by AntiVomitingDrugPrescribed Model*

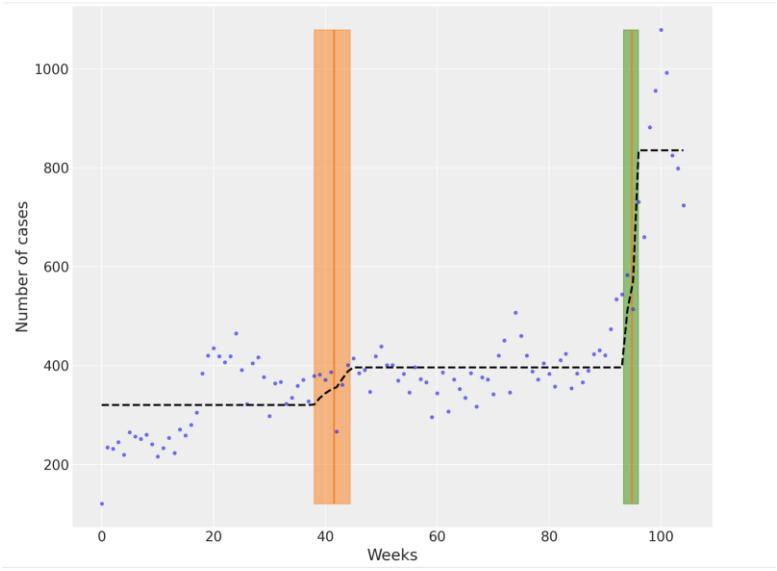

***Figure 73**: The Rate of the Outbreak Incidents (Lamda_1, Lambda_2, and Lambda_3) and the Corresponding Changepoints from AntiVomitingDrugPrescribed Dataset*

9.3.5 The Gastroenteric Dataset

The summary statistics and trace plots in figures 74 and 75 show an expected mean rate of 577 incidents per week with a 94% probability that the true expected value was between 542 and 611 incidents per week until the changepoint_1. The changepoint_1 occurred in the 42nd week

when the mean rate of the GI infection increased to 627 incidents per week. There is a credible interval that the true expected value was between 591 and 659 incidents per week with a 94% probability. The true value of changepoint_1 is believed to have occurred between the 38th and 44th weeks. The mean rate of the infection increased to 963 incidents with a 94% probability that the true expected value was between 908 and 1000 incidents per week. The change occurred in the 95th week with a 94% probability that the true value of the changepoint_2 was between the 94th and 96th weeks. Figure 76 illustrates the trend of the outbreak incidents, and in Figure 77, the parameter estimates and the range of credible intervals in the changepoint parameters remain similar to the two previous iterations.

	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_mean	ess_sd	ess_bulk	ess_tail	r_hat
alpha	24.27	1.86	21.31	27.51	0.03	0.02	4010.34	3982.45	3397.28	2903.46	1.0
lambda_1	577.34	18.59	542.09	611.09	0.28	0.20	4489.79	4458.32	4540.10	4014.62	1.0
lambda_2	626.55	18.22	590.83	659.19	0.31	0.22	3497.85	3443.85	3611.56	2714.24	1.0
lambda_3	963.05	30.37	908.03	999.98	0.46	0.33	4388.67	4285.45	2540.61	1461.84	1.0
changepoint_1	41.54	1.91	38.39	44.41	0.03	0.02	5189.87	5155.73	4838.34	4315.93	1.0
changepoint_2	94.99	0.74	93.56	95.92	0.01	0.01	6393.21	6378.07	5148.95	3947.50	1.0

***Figure 74**: Summary Statistics of Gastroenteric Model*

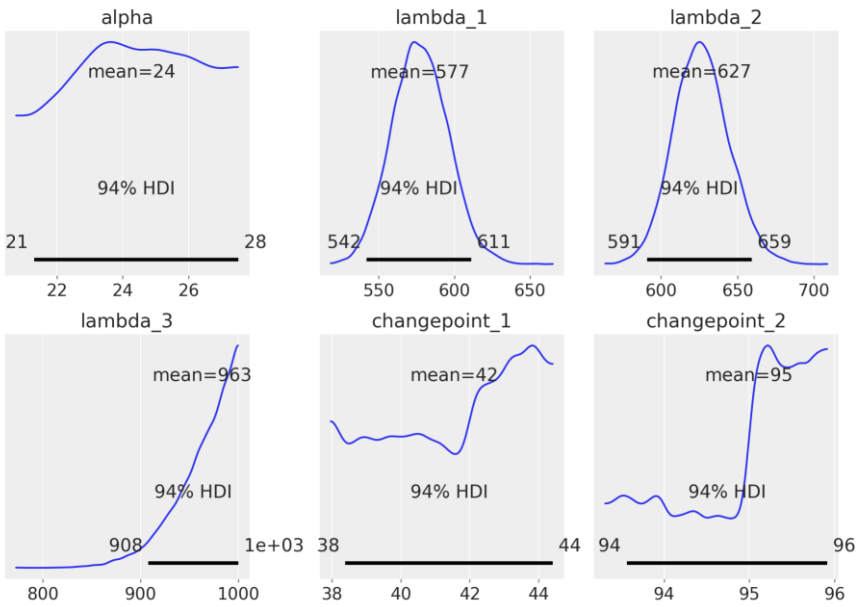

***Figure 75**: Highest Density Intervals of the Posterior Distribution for Gastroenteric Model*

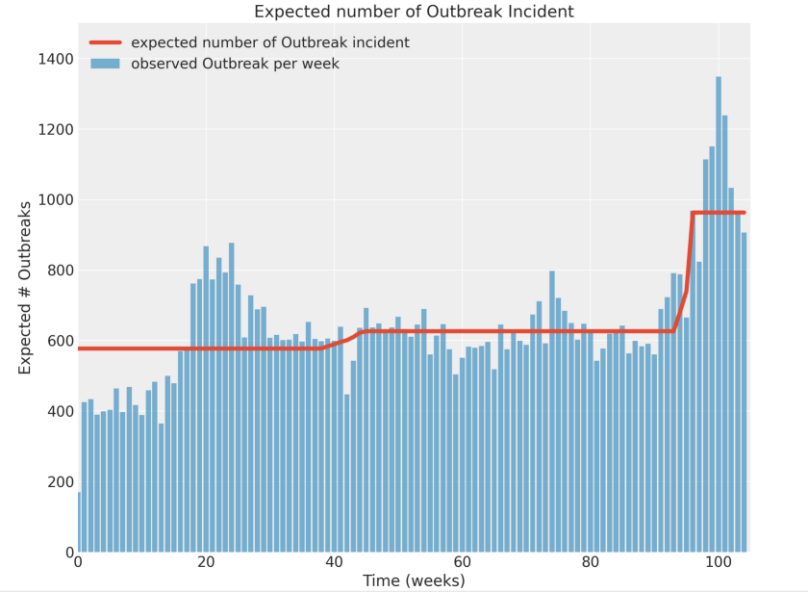

***Figure 76**: The Trend of the Outbreak Incidents by Gastroenteric Model*

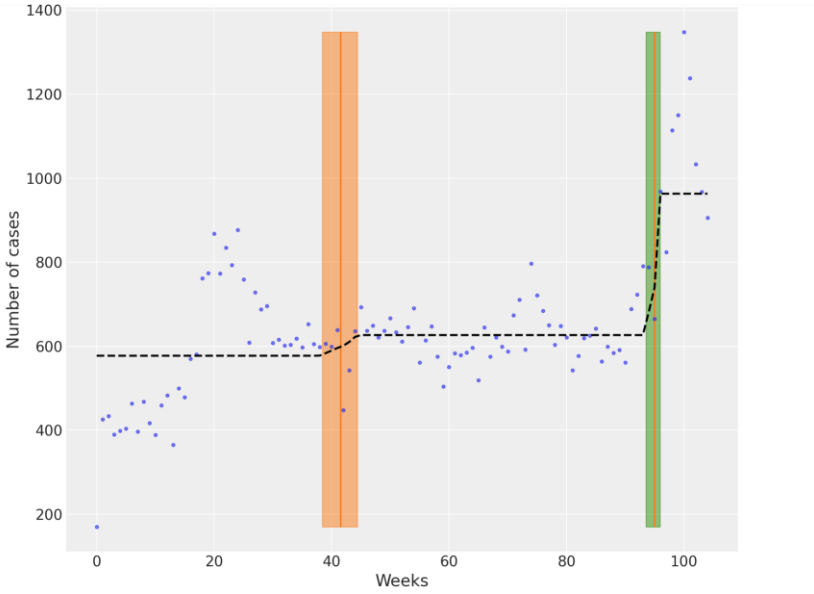

***Figure 77**: The Rate of the Outbreak Incidents (Lamda_1, Lambda_2, and Lambda_3) and the Corresponding Changepoints from Gastroenteric Dataset*

9.3.6 The CampylobacterCulture Dataset

The summary statistics and trace plots in figures 78 and 79 show an expected mean rate of 95 incidents per week with a 94% probability that the true expected value was between 88 and 101 incidents per week until the changepoint_1. The changepoint_1 occurred in the 44th week when the mean rate of the GI outbreak increased to 118 incidents per week. There is a credible interval that the true expected value was between 111 and 126 incidents per week with a 94%

probability distribution. The true value of changepoint_1 is believed to have occurred between the 43rd and 44th weeks. The mean rate of the infection increased to 168 incidents with a 94% probability that the true expected value was between 146 and 194 incidents per week. The change occurred in the 95th week with a 94% probability that the true value of the changepoint_2 was between the 94th and 96th weeks. Figure 80 illustrates the trend of the outbreak incidents, and in Figure 81, the range of the credible intervals in changepoint parameters remains similar to three previous iterations.

	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_mean	ess_sd	ess_bulk	ess_tail	r_hat
alpha	22.47	1.08	21.31	24.47	0.03	0.02	1758.42	1718.42	1728.20	1384.21	1.0
lambda_1	94.54	3.43	88.35	101.18	0.09	0.06	1462.97	1461.17	1464.51	2328.69	1.0
lambda_2	118.07	3.99	110.97	125.95	0.10	0.07	1682.17	1662.76	1711.96	2082.73	1.0
lambda_3	167.88	12.71	146.25	194.44	0.48	0.35	698.09	654.00	702.89	485.52	1.0
changepoint_1	43.81	0.76	42.57	44.41	0.02	0.02	1052.96	1038.81	910.08	1606.13	1.0
changepoint_2	95.33	0.47	94.33	95.92	0.01	0.01	1759.41	1758.75	1694.29	2280.84	1.0

***Figure 78**: Summary Statistics of CampylobacterCulture Model*

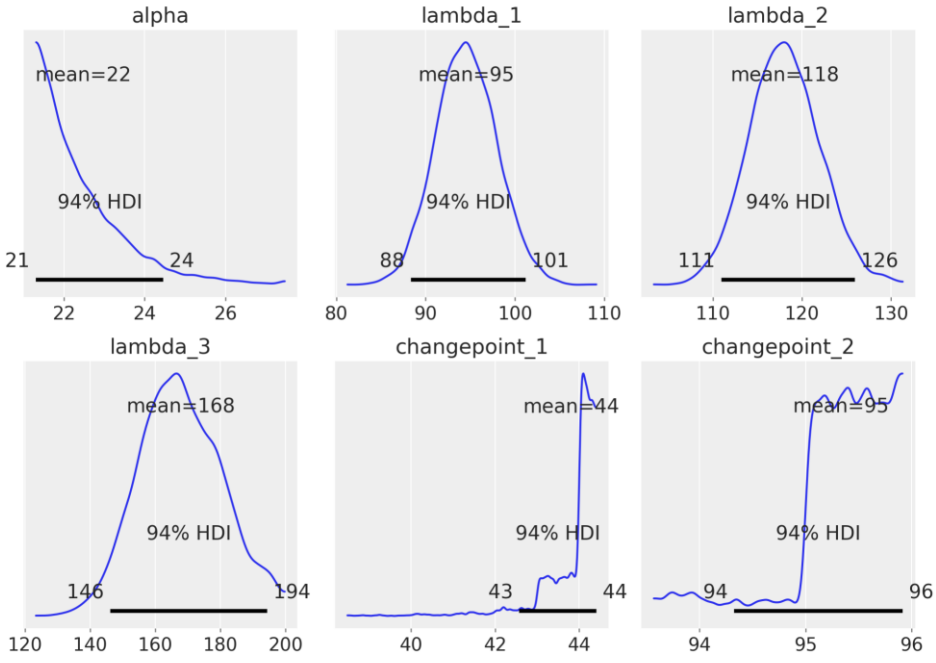

***Figure 79**: Highest Density Intervals of the Posterior Distribution for CampylobacterCulture Model*

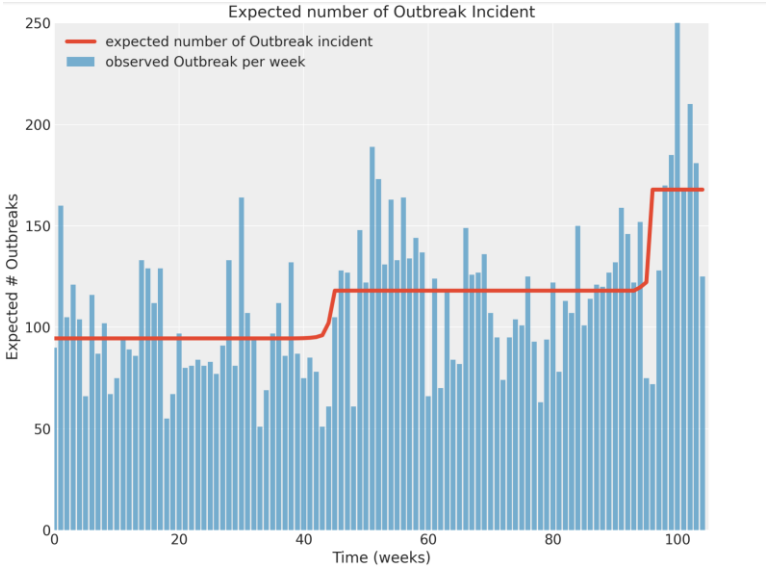

***Figure 80**: The Trend of the Outbreak Incidents by CampylobacterCulture Model*

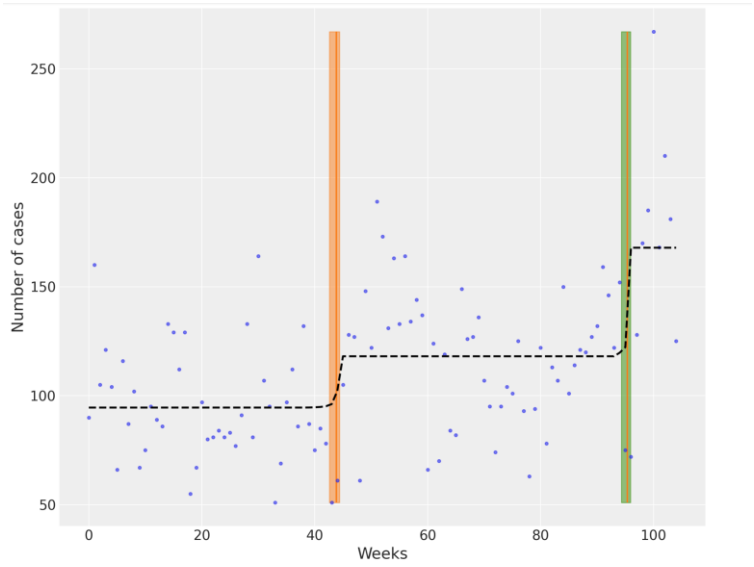

***Figure 81**: The Rate of the Outbreak Incidents (Lamda_1, Lambda_2, and Lambda_3) and the Corresponding Changepoints from CampylobacterCulture Dataset*

9.4 Discussion and Conclusion

As indicated in Table 14, the changepoint_1 in the Twitter model occurred in the 77th week as the mean rate of the GI infection increased, and its true value was between the 33rd and 108th weeks. However, the mean rate of the infection decreased in the 95th week, with a 94% probability that the true value of changepoint_2 was between the 60th and 108th week. It was

observed that the ranges of the changepoint (changepoint_1 and changepoint_2) overlapped with a large duration of weeks between the changepoints' lower and upper limits, indicating a high level of uncertainties in the estimates produced by the Twitter model.

The changepoints are further explained in Figure 61 and are shown as orange vertical lines with the HDI or credible intervals in orange and green colours, respectively. It indicates a lack of clear separation in the credible intervals as they overlap. Furthermore, the black dash line shows the trend of the mean rate of the outbreak incident corresponding to regions of low and high risk of incidents. The large uncertainties show that the plausible range of the changepoints contains uninformative estimates since it could have been anywhere within the credible interval. Although it successfully differentiates regions of high and low risks of disease outbreak incidents, it has little clarity on the exact changepoints. The plausible range could have been anytime between the 33rd and 108th week. The large duration in the range of weeks does not provide clear information to inform decision-making. However, it serves as a starting point to determine the onset of a disease outbreak, as it was proven later that the true outbreak falls in the range of the weeks captured by the Twitter model.

***Table 13**: Model investigated without using posterior distributions as priors*

Order of Data Availability	Dataset	Changepoint_1 (Upper and Lower Limits) in Weeks	Changepoint_2 (Upper and Lower Limits) in Weeks
1	Twitter	83 (29 - 108)	98 (62 -108)
2	QuestionaireAcuteVomiting	47 (38 - 56)	92 (89 - 96)
3	TextMiningAcuteVomiting	16 (15 - 18)	96 (95 - 98)
4	AntiVomitingDrugPrescribed	17 (16 - 18)	94 (91 - 96)
5	Gastroenteric	15 (13 - 16)	95 (93 - 96)
6	CampylobacterCulture	54 (43 - 98)	97 (95 - 104)

***Table 14**: Improvement of changepoint estimate parameters when the model was investigated with posterior distributions as priors*

Order of Data Availability	Dataset	Changepoint_1 (Upper and Lower Limits) in Weeks	Changepoint_2 (Upper and Lower Limits) in Weeks
1	Twitter	77 (33 - 108)	95 (60 -108)
2	QuestionaireAcuteVomiting	46 (38 - 56)	92 (88 - 99)
3	TextMiningAcuteVomiting	41 (38 - 45)	95 (93 - 96)
4	AntiVomitingDrugPrescribed	42 (38 - 44)	95 (93 - 96)
5	Gastroenteric	42 (38 - 44)	95 (94 - 96)
6	CampylobacterCulture	44 (43 - 44)	95 (94 - 96)

Datasets from SAVSNET were introduced into the model according to their availability in the epidemiological data lifecycle presented in Figure 1 to update and improve probabilistic beliefs and demonstrate the improvement of parameter estimates. It was observed that the credible intervals of each changepoint parameter were reduced in range with improved certainties of the estimates for each iteration of the model as new datasets were introduced.

With the introduction of QuestionaireAcuteVomiting datasets, a clear separation was observed between the two changepoints (changepoint_1 and changepoint_2) with narrow ranges of credible intervals. Therefore, unlike the Twitter model, where the changepoints overlap, with large ranges between the upper and lower limits of the expected values of changepoints, the QuestionaireAcuteVomiting model produced smaller uncertainties.

However, the expected mean rate values of the outbreak incidents (lambda_1, lambda_2, and lambda_3) were found to overlap in the credible intervals. Figure 65 indicates a clear separation of the changepoints, revealing smaller uncertainties in parameter estimates. Furthermore, the black dash line shows the trend of the mean rate of the outbreak incident corresponding to regions of low and high risk of incidents. The small uncertainties show that the plausible range of the changepoints contains informative estimates that may differentiate regions of high and low risks of disease outbreak incidents.

For example, changepoint_1 and changepoint_2 are captured in Table 14, showing an improvement in the estimate of both changepoints as more datasets become available. The Twitter dataset presents early and quick insights into what might be going on about the outbreak

by providing an easily accessible dataset to tackle the data scarcity and provide information about a possible outbreak. However, the Twitter dataset is presented with the challenges of lacking informative estimates with high uncertainties. This may be due to the high noise characteristics of the social media datasets.

The QuestionaireAcuteVomiting and other datasets introduced into the model from SAVSNET sources confirmed that both changepoints must have occurred between the 38th and 44th week and between 93rd and 96th week, respectively. This is also supported in Table 14. By observing these estimates, it can be concluded that the Twitter model, although with a wide span of weeks plausible for changepoints to have occurred, the expected estimates reported from other datasets are within the range reported in the Twitter model. Also, from all datasets, the method reported a slightly increasing trend from the 40th week except the QuestionaireAcuteVomiting model, which reported a downward trend in the 40th week.

All data streams appear to have captured a steep upward rise in the number of outbreak incidents in and around the 88th week onwards. Domain experts confirmed that around 90th weeks corresponds to an onset of a real GI infection outbreak in dogs in the United Kingdom. Therefore, the proposed SQRA method can determine regions with low and high risks of a disease outbreak to inform public health practice decision-making.

Furthermore, as part of the validation of our procedure, the model was fitted to all the data streams individually without considering the posterior distributions as priors from previous data in the epidemiological lifecycle. The diagnosis result indicates many divergences in the model for all the data streams considered. It was observed that the estimated mean rate values of both changepoints have wider uncertainties which means that fitting the model to the available data streams in this manner is ineffective, and the outcomes are uninformative to aid the decision-making process. Table 13 shows the parameter estimates.

Chapter 10 – Conclusion

Finally, we recall the research questions guiding this research work and present the answers obtained from different chapters of this thesis. The main research question is "how to develop a procedure in step-by-step detail for analysing heterogeneous data sources and evaluating VSyS alarms to improve public health response and aid decision making?" Due to the scope of the research work, the following sub-questions were formulated to address the primary research question in various chapters of this thesis document. The answers are summarised below.

1. How is risk knowledge and its components perceived in public health surveillance, and which theory or theories can be used to conduct this research?

The author considered knowledge about risk and its components in public health surveillance from the views of the probability distribution, expected values or chance, undesirable events or danger, and uncertainty measures. They also considered public health risk as an event or an action that may likely cause harm to human or animal health or contribute to disease among humans or animals. Factors such as signal, alarm, alert, uncertainty, risk analysis and risk assessment were considered in their application to public health practice.

The author focused on health signal analysis, which is critically essential to inform decision-making, risk assessment, and uncertainty analysis. It is understandable that there cannot be a single best approach to risk management in public health practice. As a result, various theories and conceptual frameworks from different fields were investigated. It was decided that the SDT and Decision theories were the appropriate framework for delivering this research work. The two theories best fit the goal of this research study because they coincide with the objectives of the research work in that the SDT supports the decision-making process, offering an approach to dissecting components of risks and uncertainties and can improve decision-making quality, service quality and public health safety (Cohen et al., 2020; Lynn et al., 2018).

The usefulness of Decision theory in risk management was highlighted, and the implicit dependency of risk management on rules derived from general knowledge and precepts of Decision theory was also explained. As a risk management concept, many research articles indicated the evidence of using SDT or Decision theory individually for research study design and interpretation. However, only Smith et al. (2017) applied both theories to interpret their work on risk assessment techniques for public health disease surveillance from a qualitative perspective. In this present study, the author applied both SDT and Decision theories from

quantitative viewpoints to explain the SQRA method as an approach for differentiating the risk magnitude of disease outbreaks and improving decision-making based on multiple data sources.

2. What are the existing VSyS techniques and data sources? How do they account for uncertainties and previous knowledge of health incidents in a model to evaluate risks? Furthermore, what are the difficulties associated with their use?

The author classified routinely collected VSyS data sources according to the work of Dórea and Vial (2016) and Egates et al. (2015) into animal production and farm activities data stream, clinical, laboratory diagnostic, online and social media, and slaughterhouse, abattoirs, and meat inspection data streams. However, the author discovered five research publications that focused on evaluating multiple data sources for VSyS development which could not be included in the classifications proposed by Dórea and Vial (2016) and Egates et al. (2015).

As a result, in this study, we recommended adding a new VSyS data source category to the existing group, the “multi-data streams”. Also, it was discovered in this work that the animal production and farm activities data stream is the most widely evaluated data source for VSyS development. The major challenges with the use of the identified data sources for VSyS are (i) data scarcity and under-reporting of animal health events since most data sources in the veterinary industry are majorly driven by private businesses (ii) delayed reporting from some of the data sources causes the relative timeliness of data for VSyS implementation to vary greatly, therefore, affecting real-time outbreak detection (iii) lack of data standardisation across systems contributing data since the veterinary industry lack a cohesive framework coordinating the data source contribution efforts.

Furthermore, the author identified the Frequentist statistical approach and the Bayesian inference technique as the common data modelling approaches in VSyS development. Most of the identified research publications applied the Frequentist statistical approach focusing on alarm generation to detect anomalies in the disease surveillance signals. Regardless of the data stream under consideration, VSyS data analysis techniques appear to follow nearly a similar set of steps in most reviewed publications on VSyS.

However, in this study, the author focused on the overall VSyS approach to identify how each publication addressed the associated uncertainties in the data and the models, the handling of the previous knowledge about the health incidents such as outbreak distribution, the influence

of seasons, days of the week or other covariates to inform decision making. Also, the author concentrated on how previous studies interpreted health events that occurred very close to the baseline or were generated near the threshold. It was discovered that nearly all the publications reviewed do not account for uncertainties in alarm generation or incorporate prior information or previous knowledge of the health events in the applied data modelling strategy.

3. Is there any evidence of published research on using QRA techniques to assess VSyS alarms quantitatively?

The author established that the popular approach in the current VSyS technique is first to define baselines/thresholds, which represent the normal behaviour when no disease outbreak is recorded. Abnormal events overlaid on top of the background noise are evaluated against these baselines to detect outbreaks. In most detection methods, an alarm goes off when the observed data exceed the population's baseline values. Most VSyS techniques rely on methods that use deviation from baselines to define outbreaks in signal alarms. These systems may be simple to implement but difficult to interpret, particularly when events occur very close to the threshold or when there is a slowly increasing outbreak or the number of cases reported in each time unit, each week, or each day is too small to trigger an alarm.

The existing alarm-based VSyS techniques do not differentiate between areas and periods of low and high risk in disease outbreak distribution. As a result, most alarm generation techniques produce a high number of false alarms when the threshold value is lowered and few alarms when the threshold is increased; this indicates that the handling of the alarm generation does not provide explanations for data points closer or far from the baseline, which may lead to a false interpretation of alarms.

In practice, all generated alarms must be investigated for the risk of public health concerns. However, most approaches identified in the literature review indicate a manual assessment of the generated alarm, which might require rigorous training, time, and an expert knowledge base. All current QRA techniques in VSyS focus on evaluating disease introduction in a population using the @Risk Microsoft Excel application. To the author's knowledge, no research published in animal health surveillance has quantitatively assessed these alarms to inform decision-making. The existing quantitative assessment method found in the literature does not evaluate the risks of generated VSyS alarms to differentiate regions of high risk from low risk.

4. What techniques in data science can be applied to datasets from social media and veterinary medicine to evaluate risks in VSyS to aid public health decision-making? What is the detailed step-by-step procedure? Furthermore, what are the obstacles to overcome and practical solutions?

In the methodology chapters of this research work, many data science techniques were investigated and tested as part of the strategy to develop the SQRA method with the available datasets. The investigation established that the following data science techniques could be applied to social media and veterinary healthcare datasets to quantify the risk associated with alarms generated during syndromic surveillance analysis.

a) Data exploration and visualisation technique
b) Time-series decomposition and analysis
c) Pearson chi-square goodness-of-fit test technique
d) Bayesian probabilistic technique
e) MCMC sampling technique
f) Changepoint analysis
g) Bayesian evaluation techniques with Arviz

In this study, it was discovered that identifying risk magnitudes of disease outbreaks requires changepoint techniques with probability distributions to identify regions of high and low risks relying on epidemiological datasets. The Bayesian probabilistic technique helps the SQRA method identify and differentiate those regions with low and high risks in disease surveillance signals to inform public health decision-making. The main difference between the alarm-based surveillance system and the proposed method is that this method produces a quantitative measurement of the risk of disease outbreak with the corresponding changepoints and associated uncertainties to inform decision-making.

The author proposed a procedure in step-by-step detail for developing the SQRA method as the contribution of this research work to the body of knowledge. The detailed procedure is described in chapter 8, section 8.4, and the overview of the method is presented below.

a) Identification of data sources suitable for disease surveillance.
b) Collect real-time or historical time-series data from the identified sources.
c) Pre-processing of the data.

d) Investigate the data in order to gain a better understanding of the underlying distributions and data attributes.
e) Identify and define unknown parameters and hyper-parameters with priors using the underlying distributions.
f) Specify the Bayesian probabilistic model to investigate unknown parameters and re-parametrise as needed.
g) Fit the model to the data in order of availability.
h) Estimate unknown parameters and detect variations and changepoints to identify risk regions.
i) Investigate credible intervals and uncertainties in parameter estimates.
j) Risk evaluation of outbreaks.
k) Public health response and decision making.

The major obstacle identified is the data scarcity for veterinary syndromic surveillance. Most of the data owners in veterinary medicine are private businesses that would like to maintain the privacy of their trade secrets. Also, delayed reporting is another major obstacle as most data contributing initiatives have to compensate the collaborating clinics and diagnostic laboratories, and this practice is not sustainable. The lack of compatible information management systems and data standards between the collaborating clinics makes data contribution a painful exercise. For example, SAVSNET spends a lot of effort and resources integrating these datasets for the syndromic surveillance function.

Due to the above challenges, securing datasets for this research work was very challenging. However, the difficulty was overcome by introducing and demonstrating the social media data as the alternative dataset to begin an early syndromic surveillance activity. The approach proposed in this study relies on a probabilistic programming technique for updating outcomes as more datasets are available from independent sources, such as veterinary clinics and diagnostic laboratories. Subsequently, the author proved that the Bayesian probabilistic approach of updating beliefs could help tackle delayed reporting and relative timeliness of data for VSyS implementation.

10.1 Summary

The standard technique for implementing VSyS is the detection of temporal or spatial aberrations in the occurrence of health incidents. Such techniques often employ algorithms that

rely on statistical anomalies of temporal or spatial data to generate alarms when health incidents occur more frequently than expected above a set threshold in the observed population. The algorithms indicate the earliest deviation from baselines to public health officials. According to many previous research studies, the current VSyS technique has a lower sensitivity to changes in disease incident frequencies and a high percentage of false alarms, which may encourage users to disregard potentially valid alarms. All alarms require further investigation to understand the nature and possible cause and establish which alarms are important public health concerns.

In this study, it was discovered that while the outputs of the alarm-based VSyS are straightforward to understand, their interpretations can be challenging to employ in decision-making systems for public health practice, especially when an outbreak is spreading slowly or the number of cases recorded in each time unit is insufficient to generate an alarm. Also, it may be challenging to combine their outputs with other epidemiological facts to influence public health decisions, particularly when it is desirable to incorporate model or parameter uncertainty for estimating surveillance risks quantitatively. The Frequentist approach is the most commonly used method, and as a result, there is no simple way to quantify risk when the algorithm is based on an alarm threshold, which means that alarm-based VSyS may not provide sufficient information to discriminate adequately between lower and higher risks events when modelling disease outbreaks. This is a knowledge gap that was addressed in this thesis work.

It was established that the proposed and developed SQRA method could investigate disease surveillance alarms and evaluate risk magnitudes using the social media, practice, and diagnostic laboratory datasets with the Bayesian probabilistic technique and changepoint analysis. The outputs from the SQRA method demonstrated an efficient approach for identifying disease surveillance risks quantitatively, incorporating existing knowledge of disease outbreaks as prior parameters, and demonstrating uncertainty measurements in the model. Therefore, providing the means of incorporating unknown parameters such as other epidemiological factors as prior parameters and evaluating their posterior distributions to inform decision-making.

To develop the proposed method, the author formulated four sub-research questions to address the main research question due to the scope of the work. The research questions and their answers were discussed under section 10.1 above. At the outset of this research work, the proper data science technique to apply to the available datasets in order to obtain the SQRA

method was unknown. Due to this, the author started by investigating the concept of risk, its theories and components to understand the application of risk measurement to disease surveillance in public health practice, differentiated between risk assessment and analysis and described the key risk terminologies used in this study. Then, the routinely explored data sources for VSyS development were identified, investigated, and the challenges encountered during their usage were critically analysed.

The author identified that the VSyS domain is a poorly investigated area of research due to some challenges, such as data scarcity and under-reporting of animal health incidents. Also, the timeliness of the animal health data for outbreak detection analysis is highly variable as different sources generate data at relatively different times. Therefore, the data for syndromic analysis are not available simultaneously for an effective investigation, which is a challenge to implementing effective veterinary syndromic surveillance systems.

Existing VSyS modelling techniques were investigated to critically analyse their ability to account for associated uncertainties and previous knowledge of ongoing disease activities in the data or the model to interpret alarms and inform decision-making. Then, peer-reviewed analysis of publications on QRA was conducted to highlight research gaps. Most VSyS techniques rely on methods that use deviation from baselines to define outbreaks in signal alarms. Furthermore, the existing alarm-based VSyS techniques do not differentiate between areas and periods of low and high risk in disease outbreak distribution.

Various data science techniques were examined to accomplish the research objectives. The SQRA method was derived, along with the detailed step-by-step procedure presented in chapter 8, section 8.4. After careful consideration and investigation of different data science techniques, the effectiveness and practicability of the techniques as a suitable SQRA method were established through evaluation. The SQRA method was reliable and effective in differentiating and identifying risk regions with corresponding changepoints capable of interpreting VSyS alarms during ongoing disease outbreak incidents. This justifies that the SQRA method developed in PyMC3 can be applied to interpret the VSyS alarms and inform public health decision-making. Particularly when considering the challenges and obstacles identified in this work that make VSyS implementation difficult, such as the data scarcity, delayed reporting, and timeliness of data sources. This SQRA method allows early VSyS implementation starting with readily available datasets such as the Twitter data to quickly infer

a possible outbreak and update inferences when more datasets are available from other sources to quantify risks and validate disease outbreaks.

Finally, the evidence presented in this thesis demonstrates how the Bayesian probabilistic technique implemented in PyMC3 can be used effectively in place of the current Frequentist statistical approach for interpreting VSyS generated alarms. This implies that a technique such as the SQRA method can assist in overcoming some of the challenges identified in this study by allowing quick VSyS implementation with the social media datasets and later incorporating veterinary healthcare datasets whenever they are available to update the outbreak outcomes, which ultimately contribute to science and practice.

10.2 Research Novelty

The novelty of this thesis includes incorporating social media datasets to tackle the challenges of the data scarcity identified in the review. Also, implementing an approach that can accommodate the relative timeliness of data availability such that data is introduced into the model whenever they become available to update the belief about the outbreak distribution risk. With this approach, the author tackled delayed reporting challenges and implemented a system that uses multiple data streams to model VSyS risks. Furthermore, the author introduced a novel deterministic, stochastic variable into the model, r_0 and r_1, that takes advantage of the PyMC3 inbuilt switch function to calculate the changepoints and the corresponding rate of incident occurrences.

10.3 Contribution

In this study, the author proposed the development of an SQRA technique for assessing disease surveillance risk while considering existing knowledge of disease outbreaks and associated uncertainties. The author highlighted gaps in knowledge and existing challenges that make the VSyS implementation and alarm interpretation difficult, particularly when quantifying surveillance risk with associated uncertainties. Few research publications on veterinary syndromic surveillance attempted to look at risk assessment and uncertainty measures under the public health surveillance practice. A thorough search of the relevant literature database yielded no papers that quantitatively evaluated disease surveillance alarm risks or provided a detailed step-by-step guide on how practitioners can implement the recommended technique in this paper. Therefore, the original contribution of this research is a thorough step-by-step

procedure of using Bayesian probabilistic programming and changepoint analysis to quantify and interpret disease surveillance risks using social media and routine healthcare datasets.

10.4 Limitations of the Study

The author recognises the limitations imposed by the time constraints of this study and some limitations imposed by the available datasets. This research could have been conducted differently by employing a comparative design or being more comprehensive. However, this study may need to be compared to other SQRA methods used in other fields to adapt to a comparative design. Additionally, the researcher could have used a longitudinal design to evaluate the research's outcome to demonstrate how changes in the disease surveillance approach could improve public health performance in early disease detection. However, adopting these approaches may extend the duration of the research beyond the time allotted.

Additionally, based on the research objectives and questions, it is possible to conclude that the research design used in this work is appropriate for achieving the objectives of this study. The suggested research designs might have provided additional insights into this study. However, they are not the best fit because they would have created unnecessary distractions and prevented the author from arriving at the thesis's current conclusions for the target sector.

It was extremely challenging to collect surveillance data from multiple veterinary practices and diagnostic laboratories to adequately cover a wider area of interest where disease could occur. Data scarcity is especially problematic in veterinary medicine because private businesses drive the industry. They lack a cohesive framework for data contribution compared to human medicine. Also, data contribution is faced with the problem of delayed reporting making real-time analysis almost impossible. Furthermore, some veterinary practices are concerned about trade secrets and clients' privacy, making data contribution for VSyS very difficult. These issues affected data access to more sources, including the raw healthcare datasets, which might have provided in-depth insights.

10.5 Future Trends

Many research of the size and scope of the current study usually produces conclusions that lead to more research questions. The SQRA procedure developed in this research has been validated and successfully applied to datasets from practice and diagnostics laboratory data streams, providing a framework for addressing several of these questions. However, not all questions

can be answered; the author hopes that the following direction will aid in the future provision of additional answers.

In this study, it has been demonstrated that the mean rate of occurrence and their corresponding changepoints can help determine the risk magnitudes of a disease outbreak from surveillance signals to interpret outbreak alarms and inform public decision-making. However, the researcher manually determined the number of parameters defined for mean rate occurrence and the corresponding changepoints (i.e. Lambda_1, Lambda_2, Lambda_3, Changepoint_1, and Changepoint_2). Instead of manual selection, it would be intuitive to develop a step-by-step procedure that can automatically select the number of parameters for the mean rate of incidents and their corresponding changepoints depending on the shape and structure of the data stream under consideration to assess the risks of the outbreak and interpret alarms.

References

ABAT, C., CHAUDET, H., ROLAIN, J.-M., COLSON, P. & RAOULT, D. (2016). Traditional and syndromic surveillance of infectious diseases and pathogens. *International Journal of Infectious Diseases,* 48**,** 22-28.

ABUTABENJEH, S. & JARADAT, R. (2018). Clarification of research design, research methods, and research methodology: A guide for public administration researchers and practitioners. *Teaching Public Administration,* 36**,** 237-258.

ADAMS, R. P. & MACKAY, D. J. (2007). Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*.

AGHAALI, M., KAVOUSI, A., SHAHSAVANI, A. & HASHEMI NAZARI, S. S. (2020). Performance of Bayesian outbreak detection algorithm in the syndromic surveillance of influenza-like illness in small region. *Transboundary and Emerging Diseases*.

ALBA-CASALS, A., FERNÁNDEZ-FONTELO, A., REVIE, C. W., DÓREA, F. C., SÁNCHEZ, J., ROMERO, L., CÁCERES, G., PÉREZ, A. & PUIG, P. (2015). Development of new strategies to model bovine fallen stock data from large and small subpopulations for syndromic surveillance use. *Épidémiologie et Santé Animale***,** 67-76.

ALBA, A., DÓREA, F. C., ARINERO, L., SANCHEZ, J., CORDÓN, R., PUIG, P. & REVIE, C. W. (2015). Exploring the surveillance potential of mortality data: nine years of bovine fallen stock data collected in Catalonia (Spain). *PloS one,* 10.

ALKHAMIS, M. A., ARRUDA, A. G., VILALTA, C., MORRISON, R. B. & PEREZ, A. M. (2018). Surveillance of porcine reproductive and respiratory syndrome virus in the United States using risk mapping and species distribution modeling. *Preventive veterinary medicine,* 150**,** 135-142.

ALLENBY, G. M., BRADLOW, E. T., GEORGE, E. I., LIECHTY, J. & MCCULLOCH, R. E. (2014). Perspectives on Bayesian methods and big data. *Customer Needs and Solutions,* 1**,** 169-175.

ALTHAUS, C. E. (2005). A disciplinary perspective on the epistemological status of risk. *Risk Analysis: An International Journal,* 25**,** 567-588.

ALTON, G. D., PEARL, D. L., BATEMAN, K. G., MCNAB, B. & BERKE, O. (2013). Comparison of covariate adjustment methods using space-time scan statistics for food animal syndromic surveillance. *BMC veterinary research,* 9**,** 231.

ALTON, G. D., PEARL, D. L., BATEMAN, K. G., MCNAB, W. & BERKE, O. (2012). Suitability of bovine portion condemnations at provincially-inspected abattoirs in Ontario Canada for food animal syndromic surveillance. *BMC veterinary research,* 8**,** 1-13.

ALTON, G. D., PEARL, D. L., BATEMAN, K. G., MCNAB, W. B. & BERKE, O. (2010). Factors associated with whole carcass condemnation rates in provincially-inspected abattoirs in Ontario 2001-2007: implications for food animal syndromic surveillance. *BMC veterinary research,* 6**,** 1-11.

ALTON, G. D., PEARL, D. L., BATEMAN, K. G., MCNAB, W. B. & BERKE, O. (2015). Suitability of sentinel abattoirs for syndromic surveillance using provincially inspected bovine abattoir condemnation data. *BMC veterinary research,* 11**,** 1-9.

AMEZCUA, M. D. R., PEARL, D. L., FRIENDSHIP, R. M. & MCNAB, W. B. (2010). Evaluation of a veterinary-based syndromic surveillance system implemented for swine. *Canadian Journal of Veterinary Research,* 74**,** 241-251.

AMEZCUA, R., PEARL, D. L. & FRIENDSHIP, R. M. (2013). Comparison of disease trends in the Ontario swine population using active practitioner-based surveillance and passive laboratory-based surveillance (2007–2009). *The Canadian Veterinary Journal,* 54**,** 775.

ANA, A., ANDRÉS, M. P., JULIA, P., PEDRO, P., ARNO, W., JULIO, A. & MICHELLE, W. (2017). Syndromic surveillance for West Nile virus using raptors in rehabilitation. *BMC veterinary research,* 13**,** 1-10.

ANDERSSON, M. G., FAVERJON, C., VIAL, F., LEGRAND, L. & LEBLOND, A. (2014). Using Bayes' rule to define the value of evidence from syndromic surveillance. *PloS one,* 9.

ANFARA JR, V. A. & MERTZ, N. T. (2014). *Theoretical frameworks in qualitative research*, Sage publications.

ANTUNES, A. C. L., ERSBØLL, A. K., BIHRMANN, K. & TOFT, N. (2017). Mortality in Danish Swine herds: Spatio-temporal clusters and risk factors. *Preventive veterinary medicine,* 145**,** 41-48.

ARMSTRONG, R. A. (2019). Should Pearson's correlation coefficient be avoided? *Ophthalmic and Physiological Optics,* 39**,** 316-327.

AVEN, T. (2012). The risk concept—historical and recent development trends. *Reliability Engineering & System Safety,* 99**,** 33-44.

AVEN, T. & RENN, O. (2010). *Risk management and governance: concepts, guidelines and applications*, Springer Science & Business Media.

BANAKAR, A., SADEGHI, M. & SHUSHTARI, A. (2016). An intelligent device for diagnosing avian diseases: Newcastle, infectious bronchitis, avian influenza. *Computers and electronics in agriculture,* 127**,** 744-753.

BEHAEGHEL, I., VELDHUIS, A., REN, L., MÉROC, E., KOENEN, F., KERKHOFS, P., VAN DER STEDE, Y., BARNOUIN, J. & DISPAS, M. (2015). Evaluation of a hierarchical ascendant clustering process implemented in a veterinary syndromic surveillance system. *Preventive veterinary medicine,* 120**,** 141-151.

BERMAN, F., RUTENBAR, R., HAILPERN, B., CHRISTENSEN, H., DAVIDSON, S., ESTRIN, D., FRANKLIN, M., MARTONOSI, M., RAGHAVAN, P. & STODDEN, V. (2018). Realizing the potential of data science. *Communications of the ACM,* 61**,** 67-72.

BERZTISS, A. T. (Year) Published. Knowledge and uncertainty. Proceedings. 15th International Workshop on Database and Expert Systems Applications, 2004., 2004. IEEE, 476-480.

BESAG, J., YORK, J. & MOLLIÉ, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics,* 43**,** 1-20.

BETANCOURT, M. (2017). A conceptual introduction to Hamiltonian Monte Carlo. *arXiv preprint arXiv:1701.02434*.

BETTANY-SALTIKOV, J. (2012). How to do a systematic literature review in nursing: a step-by-step guide.

BINGHAM, E., CHEN, J. P., JANKOWIAK, M., OBERMEYER, F., PRADHAN, N., KARALETSOS, T., SINGH, R., SZERLIP, P., HORSFALL, P. & GOODMAN, N. D. (2019). Pyro: Deep universal probabilistic programming. *The Journal of Machine Learning Research,* 20**,** 973-978.

BISONG, E. (2019). Google colaboratory. *Building Machine Learning and Deep Learning Models on Google Cloud Platform.* Springer.

BLEI, D. M. (2014). Build, compute, critique, repeat: Data analysis with latent variable models. *Annual Review of Statistics and Its Application,* 1**,** 203-232.

BODE, F., FERRÉ, T., ZIGELLI, N., EMMERT, M. & NOWAK, W. (2018). Reconnecting stochastic methods with hydrogeological applications: A utilitarian uncertainty analysis and risk assessment approach for the design of optimal monitoring networks. *Water Resources Research,* 54**,** 2270-2287.

BOETTCHER, M. (2011). Contrast and change mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,* 1**,** 215-230.

BOLLIG, N., CLARKE, L., ELSMO, E. & CRAVEN, M. (2020). Machine learning for syndromic surveillance using veterinary necropsy reports. *PloS one,* 15**,** e0228105.

BRIGGS, A. H., WEINSTEIN, M. C., FENWICK, E. A., KARNON, J., SCULPHER, M. J. & PALTIEL, A. D. (2012). Model parameter estimation and uncertainty analysis: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force Working Group–6. *Medical decision making,* 32**,** 722-732.

BRODIE, M. L. (2019). What is Data Science? *Applied data science.* Springer.

BRONNER, A., MORIGNAT, E., FOURNIÉ, G., VERGNE, T., VINARD, J., GAY, E. & CALAVAS, D. (2015a). Syndromic surveillance of abortions in beef cattle based on the prospective analysis of spatio-temporal variations of calvings. *Scientific reports,* 5**,** 1-10.

BRONNER, A., MORIGNAT, E., HÉNAUX, V., MADOUASSE, A., GAY, E. & CALAVAS, D. (2015b). Devising an indicator to detect mid-term abortions in dairy cattle: a first step towards syndromic surveillance of abortive diseases. *PLoS One,* 10**,** e0119012.

BRONNER, A., MORIGNAT, E., TOURATIER, A., GACHE, K., SALA, C. & CALAVAS, D. (2015c). Was the French clinical surveillance system of bovine brucellosis influenced by the occurrence and surveillance of other abortive diseases? *Preventive veterinary medicine,* 118**,** 498-503.

BURKOM, H., ESTBERG, L., AKKINA, J., ELBERT, Y., ZEPEDA, C. & BASZLER, T. (2019). Equine syndromic surveillance in Colorado using veterinary laboratory testing order data. *PloS one,* 14.

BURNHAM, J. F. (2006). Scopus database: a review. *Biomedical digital libraries,* 3**,** 1-8.

BURNS, C. J., WRIGHT, J. M., PIERSON, J. B., BATESON, T. F., BURSTYN, I., GOLDSTEIN, D. A., KLAUNIG, J. E., LUBEN, T. J., MIHLAN, G. & RITTER, L. (2014). Evaluating uncertainty to strengthen epidemiologic data for use in human health risk assessments. *Environmental health perspectives,* 122**,** 1160-1165.

CANFIELD, C. I., FISCHHOFF, B. & DAVIS, A. (2016). Quantifying phishing susceptibility for detection and behavior decisions. *Human factors,* 58**,** 1158-1172.

CARDONA, O. D., VAN AALST, M. K., BIRKMANN, J., FORDHAM, M., MC GREGOR, G., ROSA, P., PULWARTY, R. S., SCHIPPER, E. L. F., SINH, B. T. & DÉCAMPS, H. (2012). Determinants of risk: exposure and vulnerability. *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation: Special Report of the Intergovernmental Panel on Climate Change.* Cambridge University Press.

CARPENTER, B., GELMAN, A., HOFFMAN, M. D., LEE, D., GOODRICH, B., BETANCOURT, M., BRUBAKER, M. A., GUO, J., LI, P. & RIDDELL, A. (2017). Stan: a probabilistic programming language. *Grantee Submission,* 76**,** 1-32.

CARPITELLA, S., CARPITELLA, F., CERTA, A., BENÍTEZ, J. & IZQUIERDO, J. (2018). Managing human factors to reduce organisational risk in industry. *Mathematical and Computational Applications,* 23**,** 67.

CAZEAU, G., LEBLOND, A., SALA, C., FROUSTEY, M., BECK, C., LECOLLINET, S. & TAPPREST, J. (2019). Utility of examining fallen stock data to monitor health-related events in equids: Application to an outbreak of West Nile Virus in France in 2015. *Transboundary and emerging diseases,* 66**,** 1417-1419.

CENTERS FOR DISEASE CONTROL AND PREVENTION. (2020). *National Syndromic Surveillance Program (NSSP); What is Syndromic Surveillance?* [Online]. Available: https://www.cdc.gov/nssp/overview.html [Accessed 22/09/2019.

CHURCH, J. S., HEGADOREN, P., PAETKAU, M., MILLER, C., REGEV-SHOSHANI, G., SCHAEFER, A. & SCHWARTZKOPF-GENSWEIN, K. (2014). Influence of environmental factors on infrared eye temperature measurements in cattle. *Research in veterinary science,* 96**,** 220-226.

CIENFUEGOS, I. (2012). Decision theory and risk management in public organizations: a literature review. *Revista de Gestión Pública,* 1**,** 101-126.

COHEN, A. L., STARNS, J. J. & ROTELLO, C. M. (2020). sdtlu: An R package for the signal detection analysis of eyewitness lineup data. *Behavior Research Methods***,** 1-23.

COLEMAN, M. & MARKS, H. (1999). Qualitative and quantitative risk assessment. *Food Control,* 10**,** 289-297.

COLETTI, A., DE NICOLA, A., DI PIETRO, A., LA PORTA, L., POLLINO, M., ROSATO, V., VICOLI, G. & VILLANI, M. L. (2020). A comprehensive system for semantic spatiotemporal assessment of risk in urban areas. *Journal of Contingencies and Crisis Management,* 28**,** 178-193.

COLLINEAU, L., CHAPMAN, B., BAO, X., SIVAPATHASUNDARAM, B., CARSON, C. A., FAZIL, A., REID-SMITH, R. J. & SMITH, B. A. (2020). A farm-to-fork quantitative risk assessment model for

Salmonella Heidelberg resistant to third-generation cephalosporins in broiler chickens in Canada. *International Journal of Food Microbiology,* 330**,** 108559.

CUSUMANO-TOWNER, M. F., SAAD, F. A., LEW, A. K. & MANSINGHKA, V. K. (Year) Published. Gen: a general-purpose probabilistic programming system with programmable inference. Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation, 2019. 221-236.

DE VOS, C. J., VAN DER GOOT, J. A., VAN ZIJDERVELD, F. G., SWANENBURG, M. & ELBERS, A. R. (2015). Risk-based testing of imported animals: A case study for bovine tuberculosis in The Netherlands. *Preventive veterinary medicine,* 121**,** 8-20.

DEHNING, J., ZIERENBERG, J., SPITZNER, F. P., WIBRAL, M., NETO, J. P., WILCZEK, M. & PRIESEMANN, V. (2020). Inferring change points in the spread of COVID-19 reveals the effectiveness of interventions. *Science,* 369.

DENZIN, N. K. (1999). Interpretive ethnography for the next century. *Journal of Contemporary Ethnography,* 28**,** 510-519.

DENZIN, N. K. & LINCOLN, Y. S. (2011). *The Sage handbook of qualitative research*, sage.

DIBABA, A. B. (2019). The risk of introduction of swine vesicular disease virus into Kenya via natural sausage casings imported from Italy. *Preventive veterinary medicine,* 169**,** 104703.

DONALDSON, L. (2001). *The contingency theory of organizations*, Sage.

DÓREA, F., LINDBERG, A., MCEWEN, B., REVIE, C. & SANCHEZ, J. (2014). Syndromic surveillance using laboratory test requests: A practical guide informed by experience with two systems. *Preventive veterinary medicine,* 116**,** 313-324.

DÓREA, F., WIDGREN, S. & LINDBERG, A. (2015). Vetsyn: an R package for veterinary syndromic surveillance. *Preventive veterinary medicine,* 122**,** 21-32.

DÓREA, F. C., MCEWEN, B. J., MCNAB, W. B., REVIE, C. W. & SANCHEZ, J. (2013a). Syndromic surveillance using veterinary laboratory data: data pre-processing and algorithm performance evaluation. *Journal of the Royal Society Interface,* 10**,** 20130114.

DÓREA, F. C., MUCKLE, C. A., KELTON, D., MCCLURE, J., MCEWEN, B. J., MCNAB, W. B., SANCHEZ, J. & REVIE, C. W. (2013b). Exploratory analysis of methods for automated classification of laboratory test orders into syndromic groups in veterinary medicine. *PLoS One,* 8**,** e57334.

DÓREA, F. C., REVIE, C. W., MCEWEN, B. J., MCNAB, W. B., KELTON, D. & SANCHEZ, J. (2013c). Retrospective time series analysis of veterinary laboratory data: preparing a historical baseline for cluster detection in syndromic surveillance. *Preventive veterinary medicine,* 109**,** 219-227.

DÓREA, F. C., SANCHEZ, J. & REVIE, C. W. (2011). Veterinary syndromic surveillance: current initiatives and potential for development. *Preventive veterinary medicine,* 101**,** 1-17.

DÓREA, F. C. & VIAL, F. (2016). Animal health syndromic surveillance: a systematic literature review of the progress in the last 5 years (2011–2016). *Veterinary Medicine: Research and Reports,* 7**,** 157.

DÓREA, F. C., VIAL, F., HAMMAR, K., LINDBERG, A., LAMBRIX, P., BLOMQVIST, E. & REVIE, C. W. (2019). Drivers for the development of an Animal Health Surveillance Ontology (AHSO). *Preventive veterinary medicine,* 166**,** 39-48.

DUPUY, C., BRONNER, A., WATSON, E., WUYCKHUISE-SJOUKE, L., REIST, M., FOUILLET, A., CALAVAS, D., HENDRIKX, P. & PERRIN, J.-B. (2013a). Inventory of veterinary syndromic surveillance initiatives in Europe (Triple-S project): Current situation and perspectives. *Preventive veterinary medicine,* 111**,** 220-229.

DUPUY, C., DEMONT, P., DUCROT, C., CALAVAS, D. & GAY, E. (2014). Factors associated with offal, partial and whole carcass condemnation in ten French cattle slaughterhouses. *Meat Science,* 97**,** 262-269.

DUPUY, C., MORIGNAT, E., DOREA, F., DUCROT, C., CALAVAS, D. & GAY, E. (2015). Pilot simulation study using meat inspection data for syndromic surveillance: use of whole carcass

condemnation of adult cattle to assess the performance of several algorithms for outbreak detection. *Epidemiology & Infection,* 143**,** 2559-2569.

DUPUY, C., MORIGNAT, E., MAUGEY, X., VINARD, J.-L., HENDRIKX, P., DUCROT, C., CALAVAS, D. & GAY, E. (2013b). Defining syndromes using cattle meat inspection data for syndromic surveillance purposes: a statistical approach with the 2005–2010 data from ten French slaughterhouses. *BMC veterinary research,* 9**,** 88.

EGATES, M. C., LINDSEY, K. H., KEITH, E. & TAMMY, E. (2015). Integrating novel data streams to support biosurveillance in commercial livestock production systems in developed countries: challenges and opportunities. *Frontiers in Public Health,* 3.

ENYA, A., PILLAY, M. & DEMPSEY, S. (2018). A systematic review on high reliability organisational theory as a safety management strategy in construction. *Safety,* 4**,** 6.

EZEKIEL, E., ABOWEI, J. & EZEKIEL, E. (2011). Hazard and risk analysis in culture fisheries. *Research Journal of Applied Sciences, Engineering and Technology,* 3**,** 1108-1117.

FANFARILLO, A. (Year) Published. Quantifying Uncertainty in Source Term Estimation with Tensorflow Probability. 2019 IEEE/ACM HPC for Urgent Decision Making (UrgentHPC), 2019. IEEE, 1-6.

FARRINGTON, C., ANDREWS, N. J., BEALE, A. & CATCHPOLE, M. (1996). A statistical algorithm for the early detection of outbreaks of infectious disease. *Journal of the Royal Statistical Society: Series A (Statistics in Society),* 159**,** 547-563.

FAVERJON, C. (2017). *Risk based surveillance for vector-borne diseases in horses: combining multiple sources of evidence to improve decision making.* Utrecht University.

FAVERJON, C., ANDERSSON, M. G., DECORS, A., TAPPREST, J., TRITZ, P., SANDOZ, A., KUTASI, O., SALA, C. & LEBLOND, A. (2016). Evaluation of a multivariate syndromic surveillance system for west nile virus. *Vector-Borne and Zoonotic Diseases,* 16**,** 382-390.

FAVERJON, C., CARMO, L. P. & BEREZOWSKI, J. (2019a). Multivariate syndromic surveillance for cattle diseases: Epidemic simulation and algorithm performance evaluation. *Preventive veterinary medicine,* 172**,** 104778.

FAVERJON, C., SCHÄRRER, S., HADORN, D. C. & BEREZOWSKI, J. (2019b). Simulation based evaluation of time series for syndromic surveillance of cattle in Switzerland. *Frontiers in veterinary science,* 6**,** 389.

FAVERJON, C., VIAL, F., ANDERSSON, M., LECOLLINET, S. & LEBLOND, A. (2017). Early detection of West Nile virus in France: quantitative assessment of syndromic surveillance system using nervous signs in horses. *Epidemiology & Infection,* 145**,** 1044-1057.

FERNÁNDEZ-FONTELO, A., PUIG, P., CACERES, G., ROMERO, L., REVIE, C., SANCHEZ, J., DOREA, F. C. & ALBA-CASALS, A. (2020). Enhancing the monitoring of fallen stock at different hierarchical administrative levels: an illustration on dairy cattle from regions with distinct husbandry, demographical and climate traits. *BMC veterinary research,* 16**,** 1-13.

FERNÁNDEZ-FONTELO, A., PUIG, P., CACERES, G., ROMERO, L., REVIE, C. W., SANCHEZ, J., DÓREA, F. C. & ALBA, A. (2017). Enhancing syndromic surveillance for fallen dairy cattle: modelling and detecting mortality peaks at different administrative levels. *Epidemiologie et Sante Animale,* 2017**,** 15-26.

FERNANDO, S.-V., PETER-JOHN, M. N., PHIL, H. J., TAREK, M., IAIN, B., SUZANNA, R., SUSAN, D., ROSALIND, M. G., SALLY, E. & ALAN, D. R. (2017). Demographics of dogs, cats, and rabbits attending veterinary practices in Great Britain as recorded in their electronic health records. *BMC Veterinary Research,* 13**,** 1-13.

FIELD, C. B., BARROS, V., STOCKER, T. F. & DAHE, Q. (2012). *Managing the risks of extreme events and disasters to advance climate change adaptation: special report of the intergovernmental panel on climate change*, Cambridge University Press.

FISCHER, A., THRELFALL, A., MEAH, S., COOKSON, R., RUTTER, H. & KELLY, M. (2013). The appraisal of public health interventions: an overview. *Journal of Public Health,* 35**,** 488-494.

FISCHER, A. J. & GHELARDI, G. (2016). The precautionary principle, evidence-based medicine, and decision theory in public health evaluation. *Frontiers in public health,* 4**,** 107.

FISCHER, E. A., ANDERSSON, M. G., FAVERJON, C., LEBLOND, A., GUSSMANN, M., GETHMANN, J. & BØDKER, R. (Year) Published. The joint risk score for vector-borne diseases used for early detection. 14th Conference of the International Society for Veterinary Epidemiology and Economics, 2015. P104.

FONNESBECK, C. J., SHEA, K., CARRAN, S., CASSIO DE MORAES, J., GREGORY, C., GOODSON, J. L. & FERRARI, M. J. (2018). Measles outbreak response decision-making under uncertainty: a retrospective analysis. *Journal of The Royal Society Interface,* 15**,** 20170575.

GAFFNEY, J. E. & ULVILA, J. W. (Year) Published. Evaluation of intrusion detectors: A decision theory approach. Proceedings 2001 IEEE Symposium on Security and Privacy. S&P 2001, 2000. IEEE, 50-61.

GARCIA-CONSTANTINO, M., COENEN, F., NOBLE, P.-J., RADFORD, A., SETZKORN, C. & TIERNEY, A. (Year) Published. An Investigation Concerning the Generation of Text Summarisation Classifiers Using Secondary Data. Machine Learning and Data Mining in Pattern Recognition: 7th International Conference, MLDM 2011, New York, NY, USA, August 30-September 3, 2011Proceedings, 2011. Springer, 387.

GARCIA-CONSTANTINO, M., COENEN, F., NOBLE, P. & RADFORD, A. (Year) Published. Questionnaire free text summarisation using hierarchical classification. International Conference on Innovative Techniques and Applications of Artificial Intelligence, 2012. Springer, 35-48.

GAUBE, S., LERMER, E. & FISCHER, P. (2019). The concept of risk perception in health-related behavior theory and behavior change. *Perceived Safety.* Springer.

GELMAN, A., CARLIN, J. B., STERN, H. S., DUNSON, D. B., VEHTARI, A. & RUBIN, D. B. (2013). *Bayesian data analysis*, CRC press.

GEWEKE, J. & AMISANO, G. (2010). Comparing and evaluating Bayesian predictive distributions of asset returns. *International Journal of Forecasting,* 26**,** 216-230.

GIERAK, A., ŚMIETANKA, K. & DE VOS, C. J. (2021). Quantitative risk assessment of the introduction of low pathogenic avian influenza H5 and H7 strains into Poland via legal import of live poultry. *Preventive Veterinary Medicine,* 189**,** 105289.

GLOBOCNIK, D. (2019). TAKING OR AVOIDING RISK THROUGH SECRET INNOVATION ACTIVITIES—THE RELATIONSHIPS AMONG EMPLOYEES'RISK PROPENSITY, BOOTLEGGING, AND MANAGEMENT SUPPORT. *International Journal of Innovation Management,* 23**,** 1950022.

GOLD, N., FRASCH, M. G., HERRY, C. L., RICHARDSON, B. S. & WANG, X. (2018). A doubly stochastic change point detection algorithm for noisy biological signals. *Frontiers in physiology,* 8**,** 1112.

GOODMAN, N. D. (2013). The principles and practice of probabilistic programming. *ACM SIGPLAN Notices,* 48**,** 399-402.

GROTE, G. (2012). Safety management in different high-risk domains–all the same? *Safety Science,* 50**,** 1983-1992.

GRÖTSCH, V. M., BLOME, C. & SCHLEPER, M. C. (2013). Antecedents of proactive supply chain risk management–a contingency theory perspective. *International Journal of Production Research,* 51**,** 2842-2867.

GRUSON, D., HELLEPUTTE, T., ROUSSEAU, P. & GRUSON, D. (2019). Data science, artificial intelligence, and machine learning: opportunities for laboratory medicine and the value of positive regulation. *Clinical biochemistry,* 69**,** 1-7.

GUERNIER, V., MILINOVICH, G. J., SANTOS, M. A. B., HAWORTH, M., COLEMAN, G. & MAGALHAES, R. J. S. (2016). Use of big data in the surveillance of veterinary diseases: early detection of tick paralysis in companion animals. *Parasites & vectors,* 9**,** 1-10.

GUERRERO, J. I., GARCÍA, A., PERSONAL, E., LUQUE, J. & LEÓN, C. (2017). Heterogeneous data source integration for smart grid ecosystems based on metadata mining. *Expert Systems With Applications,* 79**,** 254-268.

HALE, A. C., SÁNCHEZ-VIZCAÍNO, F., ROWLINGSON, B., RADFORD, A. D., GIORGI, E., O'BRIEN, S. J. & DIGGLE, P. J. (2019). A real-time spatio-temporal syndromic surveillance system with application to small companion animals. *Scientific reports,* 9**,** 1-14.

HAND, D. J. (2007). Principles of data mining. *Drug safety,* 30**,** 621-622.

HARDING, S. (1987a). The method question. *Hypatia,* 2**,** 19-35.

HARDING, S. G. (1987b). *Feminism and methodology: Social science issues*, Indiana University Press.

HAREDASHT, S. A., VIDAL, G., EDMONDSON, A., MOORE, D., SILVA-DEL-RIO, N. & MARTINEZ-LOPEZ, B. (2018). Characterization of the Temporal Trends in the Rate of Cattle Carcass Condemnations in the US and Dynamic Modeling of the Condemnation Reasons in California With a Seasonal Component. *Frontiers in Veterinary Science,* 5.

HEDELL, R., ANDERSSON, M. G., FAVERJON, C., MARCILLAUD-PITEL, C., LEBLOND, A. & MOSTAD, P. (2019). Surveillance of animal diseases through implementation of a Bayesian spatio-temporal model: A simulation example with neurological syndromes in horses and West Nile Virus. *Preventive veterinary medicine,* 162**,** 95-106.

HEFFERNAN, R., MOSTASHARI, F., DAS, D., KARPATI, A., KULLDORFF, M. & WEISS, D. (2004). Syndromic surveillance in public health practice, New York City.

HENNING, K. J. (2004). What is syndromic surveillance? *Morbidity and mortality weekly report***,** 7-11.

HUYBRECHTS, T., MERTENS, K., DE BAERDEMAEKER, J., DE KETELAERE, B. & SAEYS, W. (2014). Early warnings from automatic milk yield monitoring with online synergistic control. *Journal of Dairy Science,* 97**,** 3371-3381.

JAIMES, F., FARBIARZ, J., ALVAREZ, D. & MARTÍNEZ, C. (2005). Comparison between logistic regression and neural networks to predict death in patients with suspected sepsis in the emergency room. *Critical Care,* 9**,** R150-R156.

JARUŠKOVÁ, D. (1997). Some problems with application of change-point detection methods to environmental data. *Environmetrics: The official journal of the International Environmetrics Society,* 8**,** 469-483.

JEBREEN, I. (2012). Using inductive approach as research strategy in requirements engineering. *International Journal of Computer and Information Technology,* 1**,** 162-173.

JIANG, Q., HUANG, X. & TAO, R. (2018). Examining factors influencing internet addiction and adolescent risk behaviors among excessive internet users. *Health communication,* 33**,** 1434-1444.

JONES, K. E., PATEL, N. G., LEVY, M. A., STOREYGARD, A., BALK, D., GITTLEMAN, J. L. & DASZAK, P. (2008). Global trends in emerging infectious diseases. *Nature,* 451**,** 990-993.

JONES, P., DAWSON, S., GASKELL, R., COYNE, K., TIERNEY, A., SETZKORN, C., RADFORD, A. & NOBLE, P.-J. (2014). Surveillance of diarrhoea in small animal practice through the Small Animal Veterinary Surveillance Network (SAVSNET). *The Veterinary Journal,* 201**,** 412-418.

JURADO, C., PATERNOSTER, G., MARTÍNEZ-LÓPEZ, B., BURTON, K. & MUR, L. (2019). Could African swine fever and classical swine fever viruses enter into the United States via swine products carried in air passengers' luggage? *Transboundary and emerging diseases,* 66**,** 166-180.

KARIM, K. E. & SIEGEL, P. H. (1998). A signal detection theory approach to analyzing the efficiency and effectiveness of auditing to detect management fraud. *Managerial Auditing Journal*.

KASS-HOUT, T. A., XU, Z., MCMURRAY, P., PARK, S., BUCKERIDGE, D. L., BROWNSTEIN, J. S., FINELLI, L. & GROSECLOSE, S. L. (2012). Application of change point analysis to daily influenza-like illness emergency department visits. *Journal of the American Medical Informatics Association,* 19**,** 1075-1081.

KHAN, O. & BURNES, B. (2007). Risk and supply chain management: creating a research agenda. *The international journal of logistics management*.

KILLIANOVA, T. (2013). Risky Behavior. *In:* GELLMAN, M. D. & TURNER, J. R. (eds.) *Encyclopedia of Behavioral Medicine.* New York, NY: Springer New York.

KILLICK, R. & ECKLEY, I. (2014). changepoint: An R package for changepoint analysis. *Journal of statistical software,* 58**,** 1-19.

KILLICK, R., ECKLEY, I. A., EWANS, K. & JONATHAN, P. (2010). Detection of changes in variance of oceanographic time-series using changepoint analysis. *Ocean Engineering,* 37**,** 1120-1126.
KITCHEL, T. & BALL, A. L. (2014). Quantitative theoretical and conceptual framework use in agricultural education research. *Journal of agricultural education,* 55**,** 186-199.
KIVITS, J., RICCI, L. & MINARY, L. (2019). Interdisciplinary research in public health: the 'why'and the 'how'. *Journal of Epidemiology and Community Health,* 73**,** 1061.
KOPRIVICA, M. (2020). Comparison of software packages for performing Bayesian inference. *Neural Network World,* 30**,** 283.
KUMAR, R., CARROLL, C., HARTIKAINEN, A. & MARTÍN, O. A. (2019). ArviZ a unified library for exploratory analysis of Bayesian models in Python.
KWAN, N. C., OGAWA, H., YAMADA, A. & SUGIURA, K. (2016). Quantitative risk assessment of the introduction of rabies into Japan through the illegal landing of dogs from Russian fishing boats in the ports of Hokkaido, Japan. *Preventive Veterinary Medicine,* 128**,** 112-123.
LAKE, I. R., COLON-GONZALEZ, F. J., BARKER, G. C., MORBEY, R. A., SMITH, G. E. & ELLIOT, A. J. (2019). Machine learning to refine decision making within a syndromic surveillance service. *BMC Public Health,* 19**,** 1-12.
LARSSON, P., DEKKER, S. W. & TINGVALL, C. (2010). The need for a systems theory approach to road safety. *Safety science,* 48**,** 1167-1174.
LAYTON, R., SMITH, J., MACDONALD, P., LETCHUMANAN, R., KEESE, P. & LEMA, M. (2015). Building better environmental risk assessments. *Frontiers in bioengineering and biotechnology,* 3**,** 110.
LÁZARO-GREDILLA, M. & ES, U. M. (Year) Published. Doubly stochastic variational Bayes for non-conjugate inference. International Conference on Machine Learning, 2014. Citeseer.
LEE, D., RUSHWORTH, A. & SAHU, S. K. (2014). A Bayesian localized conditional autoregressive model for estimating the health effects of air pollution. *Biometrics,* 70**,** 419-429.
LEVESON, N. G. (2002). A new approach to system safety engineering. *Manuscript in preparation, draft can be viewed at http://sunnyday. mit. edu/book2. pdf*.
LEVESON, N. G. & STEPHANOPOULOS, G. (2013). A system-theoretic, control-inspired view and approach to process safety.
LI, Y., CHEN, J. & FENG, L. (2012). Dealing with uncertainty: A survey of theories and practices. *IEEE Transactions on Knowledge and Data Engineering,* 25**,** 2463-2482.
LI, Y., LIN, G., LAU, T. & ZENG, R. (2019). A review of changepoint detection models. *arXiv preprint arXiv:1908.07136*.
LI, Y. I., TURK, G., ROHRBACH, P. B., PIETZONKA, P., KAPPLER, J., SINGH, R., DOLEZAL, J., EKEH, T., KIKUCHI, L. & PETERSON, J. D. (2021). Efficient Bayesian inference of fully stochastic epidemiological models with applications to COVID-19. *Royal Society Open Science,* 8**,** 211065.
LINVILLE, P. W., FISCHER, G. W. & FISCHHOFF, B. (1993). AIDS risk perceptions and decision biases.
LUSTED, L. B. (1971). Decision-making studies in patient management. *New England Journal of Medicine,* 284**,** 416-424.
LYNN, S. K. & BARRETT, L. F. (2014). "Utilizing" signal detection theory. *Psychological science,* 25**,** 1663-1673.
LYNN, S. K., BUI, E., HOEPPNER, S. S., O'DAY, E. B., PALITZ, S. A., BARRETT, L. F. & SIMON, N. M. (2018). Associations between feelings of social anxiety and emotion perception. *Journal of Behavior Therapy and Experimental Psychiatry,* 59**,** 40-47.
MADOUASSE, A., MARCEAU, A., LEHÉBEL, A., BROUWER-MIDDELESCH, H., VAN SCHAIK, G., VAN DER STEDE, Y. & FOURICHON, C. (2013). Evaluation of a continuous indicator for syndromic surveillance through simulation. application to vector borne disease emergence detection in cattle using milk yield. *PloS one,* 8**,** e73726.
MADOUASSE, A., MARCEAU, A., LEHÉBEL, A., BROUWER-MIDDELESCH, H., VAN SCHAIK, G., VAN DER STEDE, Y. & FOURICHON, C. (2014). Use of monthly collected milk yields for the detection of

the emergence of the 2007 French BTV epizootic. *Preventive veterinary medicine,* 113**,** 484-491.

MANIKANDAN, S. (2011). Measures of central tendency: Median and mode. *Journal of pharmacology and pharmacotherapeutics,* 2**,** 214.

MANITZ, J. & HÖHLE, M. (2013). Bayesian outbreak detection algorithm for monitoring reported cases of campylobacteriosis in Germany. *Biometrical Journal,* 55**,** 509-526.

MARCEAU, A., MADOUASSE, A., LEHÉBEL, A., VAN SCHAIK, G., VELDHUIS, A., VAN DER STEDE, Y. & FOURICHON, C. (2014). Can routinely recorded reproductive events be used as indicators of disease emergence in dairy cattle? An evaluation of 5 indicators during the emergence of bluetongue virus in France in 2007 and 2008. *Journal of dairy science,* 97**,** 6135-6150.

MARTIN-MORENO, J. M., HARRIS, M., JAKUBOWSKI, E. & KLUGE, H. (2016). Defining and assessing public health functions: a global analysis. *Annual review of public health,* 37**,** 335-355.

MCFADDEN, A., VINK, D., PULFORD, D., LAWRENCE, K., GIAS, E., HEATH, A., MCFADDEN, C. & BINGHAM, P. (2016). Monitoring an epidemic of Theileria-associated bovine anaemia (Ikeda) in cattle herds in New Zealand. *Preventive veterinary medicine,* 125**,** 31-37.

MCFALL, R. M. & TREAT, T. A. (1999). Quantifying the information value of clinical assessments with signal detection theory. *Annual review of psychology,* 50**,** 215-241.

MCKINNEY, W. (Year) Published. Data structures for statistical computing in python. Proceedings of the 9th Python in Science Conference, 2010. Austin, TX, 51-56.

MEESTER, W. J., STEIGINGA, S. & ROSS, C. A. (2017). A brief history of Scopus: The world's largest abstract and citation database of scientific literature. *Research analytics: Boosting university productivity and competitiveness through Scientometrics***,** 31.

MIKES, A. & KAPLAN, R. S. (Year) Published. Towards a contingency theory of enterprise risk management. 2014. AAA.

MILLER, J. G. (1972). Living systems: The organization. *Behavioral Science,* 17**,** 1-182.

MORBEY, R. A., ELLIOT, A. J., CHARLETT, A., VERLANDER, N. Q., ANDREWS, N. & SMITH, G. E. (2015). The application of a novel 'rising activity, multi-level mixed effects, indicator emphasis'(RAMMIE) method for syndromic surveillance in England. *Bioinformatics,* 31**,** 3660-3665.

MOUATASSIM, Y. & EZZAHID, E. H. (2012). Poisson regression and Zero-inflated Poisson regression: application to private health insurance data. *European actuarial journal,* 2**,** 187-204.

MOULE, P. & HEK, G. (2011). *Making sense of research: An introduction for health and social care practitioners*, Sage.

MUBAMBA, C., RAMSAY, G., ABOLNIK, C., DAUTU, G. & GUMMOW, B. (2018). Is syndromic data from rural poultry farmers a viable poultry disease reporting tool and means of identifying likely farmer responses to poultry disease incursion? *Preventive veterinary medicine,* 153**,** 84-93.

MUR, L., MARTÍNEZ-LÓPEZ, B., MARTÍNEZ-AVILÉS, M., COSTARD, S., WIELAND, B., PFEIFFER, D. U. & SÁNCHEZ-VIZCAÍNO, J. M. (2012). Quantitative risk assessment for the introduction of African swine fever virus into the European Union by legal import of live pigs. *Transboundary and emerging diseases,* 59**,** 134-144.

MUSA, I., PARK, H. W., MUNKHDALAI, L. & RYU, K. H. (2018). Global Research on Syndromic Surveillance from 1993 to 2017: Bibliometric Analysis and Visualization. *Sustainability,* 10**,** 3414.

NATIONAL RESEARCH COUNCIL (1989). *Improving risk communication*, National Academies.

NATIONAL RESEARCH COUNCIL (2009). *Science and decisions: advancing risk assessment*, National Academies Press.

NIEKUM, S., OSENTOSKI, S., ATKESON, C. G. & BARTO, A. G. (Year) Published. Online bayesian changepoint detection for articulated motion models. 2015 IEEE International Conference on Robotics and Automation (ICRA), 2015. IEEE, 1468-1475.

NISHIO, M. & ARAKAWA, A. (2019). Performance of Hamiltonian Monte Carlo and No-U-Turn Sampler for estimating genetic parameters and breeding values. *Genetics Selection Evolution,* 51**,** 1-12.

NOUFAILY, A., MORBEY, R. A., COLÓN-GONZÁLEZ, F. J., ELLIOT, A. J., SMITH, G. E., LAKE, I. R. & MCCARTHY, N. (2019). Comparison of statistical algorithms for daily syndromic surveillance aberration detection. *Bioinformatics,* 35**,** 3110-3118.

NUSINOVICI, S., MADOUASSE, A. & FOURICHON, C. (2016). Quantification of the increase in the frequency of early calving associated with late exposure to bluetongue virus serotype 8 in dairy cows: implications for syndromic surveillance. *Veterinary research,* 47**,** 18.

O'SULLIVAN, T. L., FRIENDSHIP, R. M., PEARL, D. L., MCEWEN, B. & DEWEY, C. E. (2012). Identifying an outbreak of a novel swine disease using test requests for porcine reproductive and respiratory syndrome as a syndromic surveillance tool. *BMC veterinary research,* 8**,** 1-11.

OFFICE NA. (2002). *The 2001 Outbreak of Foot and Mouth Disease. Technical report.* [Online]. Available: http://www.nao.org.uk/wp-content/uploads/2002/06/0102939.pdf [Accessed 13/09/2019.

OPATOWSKI, L., OPATOWSKI, M., VONG, S. & TEMIME, L. (2020). A One-Health Quantitative Model to Assess the Risk of Antibiotic Resistance Acquisition in Asian Populations: Impact of Exposure Through Food, Water, Livestock and Humans. *Risk Analysis*.

OSANLOO, A. & GRANT, C. (2016). Understanding, selecting, and integrating a theoretical framework in dissertation research: Creating the blueprint for your "house". *Administrative issues journal: connecting education, practice, and research,* 4**,** 7.

OYAS, H., HOLMSTROM, L., KEMUNTO, N. P., MUTURI, M., MWATONDO, A., OSORO, E., BITEK, A., BETT, B., GITHINJI, J. W. & THUMBI, S. M. (2018). Enhanced surveillance for Rift Valley Fever in livestock during El Niño rains and threat of RVF outbreak, Kenya, 2015-2016. *PLoS neglected tropical diseases,* 12**,** e0006353.

PAGE, E. (1963). Controlling the standard deviation by CUSUMS and warning lines. *Technometrics,* 5**,** 307-315.

PATIL, A., HUARD, D. & FONNESBECK, C. J. (2010). PyMC: Bayesian stochastic modelling in Python. *Journal of statistical software,* 35**,** 1.

PERRIN, J.-B., DUCROT, C., VINARD, J.-L., MORIGNAT, E., GAUFFIER, A., CALAVAS, D. & HENDRIKX, P. (2010). Using the National Cattle Register to estimate the excess mortality during an epidemic: Application to an outbreak of Bluetongue serotype 8. *Epidemics,* 2**,** 207-214.

PERRIN, J.-B., DURAND, B., GAY, E., DUCROT, C., HENDRIKX, P., CALAVAS, D. & HÉNAUX, V. (2015). Simulation-based evaluation of the performances of an algorithm for detecting abnormal disease-related features in cattle mortality records. *PloS one,* 10.

PETTIT, C. J., TANTON, R. & HUNTER, J. (2017). An online platform for conducting spatial-statistical analyses of national census data across Australia. *Computers, Environment and Urban Systems,* 63**,** 68-79.

PFEIFFER, C., STEVENSON, M., FIRESTONE, S., LARSEN, J. & CAMPBELL, A. (2021). Using farmer observations for animal health syndromic surveillance: Participation and performance of an online enhanced passive surveillance system. *Preventive Veterinary Medicine,* 188**,** 105262.

PILON, C. D. (2015). Probabilistic programming and Bayesian methods for hackers. Addison-Wesley Professional.

PIVOVAROV, R., ALBERS, D. J., SEPULVEDA, J. L. & ELHADAD, N. (2014). Identifying and mitigating biases in EHR laboratory tests. *Journal of Biomedical Informatics,* 51**,** 24-34.

PLUMMER, M. (Year) Published. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. Proceedings of the 3rd international workshop on distributed statistical computing, 2003. Vienna, Austria., 1-10.

POLONSKY, J. A., BAIDJOE, A., KAMVAR, Z. N., CORI, A., DURSKI, K., EDMUNDS, W. J., EGGO, R. M., FUNK, S., KAISER, L. & KEATING, P. (2019). Outbreak analytics: a developing data science for

informing the response to emerging pathogens. *Philosophical Transactions of the Royal Society B,* 374**,** 20180276.

POSKIN, A., THÉRON, L., HANON, J.-B., SAEGERMAN, C., VERVAEKE, M., VAN DER STEDE, Y., CAY, B. & DE REGGE, N. (2016). Reconstruction of the Schmallenberg virus epidemic in Belgium: Complementary use of disease surveillance approaches. *Veterinary Microbiology,* 183**,** 50-61.

PYMC3 DOCUMENTATION. (2018). *PyMC3 Inference* [Online]. Available: https://docs.pymc.io/api/inference.html [Accessed 22/01/2010.

QUEENSLAND HEALTH. (2019). *Public health risks* [Online]. Available: https://www.health.qld.gov.au/public-health/industry-environment/environment-land-water/public-health-risks#:~:text=A%20public%20health%20risk%20is,harmful%20substances%20in%20the%20environment [Accessed 27/11/2019.

RADFORD, A., NOBLE, P., COYNE, K., GASKELL, R., JONES, P., BRYAN, J., SETZKORN, C., TIERNEY, A. & DAWSON, S. (2011). Antibacterial prescribing patterns in small animal veterinary practice identified via SAVSNET: the small animal veterinary surveillance network. *Veterinary Record***,** vetrecd5062.

RAY, J., MARZOUK, Y. & NAJM, H. (2011). A Bayesian approach for estimating bioterror attacks from patient data. *Statistics in Medicine,* 30**,** 101-126.

REEVES, J., CHEN, J., WANG, X. L., LUND, R. & LU, Q. Q. (2007). A review and comparison of changepoint detection techniques for climate data. *Journal of applied meteorology and climatology,* 46**,** 900-915.

RENN, O. (1992). Concepts of risk: a classification.

RHODES, T. (1997). Risk theory in epidemic times: sex, drugs and the social organisation of 'risk behaviour'. *Sociology of Health & Illness,* 19**,** 208-227.

RIDDELL, A., HARTIKAINEN, A. & CARTER, M. (2020). *PyStan (3.0.0)* [Online]. Available: https://pypi.org/project/pystan [Accessed 20/10/2020.

ROBERTSON, C., SAWFORD, K., GUNAWARDANA, W. S., NELSON, T. A., NATHOO, F. & STEPHEN, C. (2011). A hidden Markov model for analysis of frontline veterinary data for emerging zoonotic disease surveillance. *PLoS One,* 6.

ROBERTSON, C. & YEE, L. (2016). Avian influenza risk surveillance in North America with online media. *PloS one,* 11.

RODRıGUEZ, G. (2013). Models for count data with overdispersion. *Addendum to the WWS,* 509.

RODRÍGUEZ, M. & DÍAZ, I. (2016). A systematic and integral hazards analysis technique applied to the process industry. *Journal of Loss Prevention in the Process Industries,* 43**,** 721-729.

RUDDELL, B. L., ZASLAVSKY, I., VALENTINE, D., BERAN, B., PIASECKI, M., FU, Q. & KUMAR, P. (2014). Sustainable long term scientific data publication: Lessons learned from a prototype Observatory Information System for the Illinois River Basin. *Environmental Modelling and Software,* 54**,** 73-87.

RUPLE-CZERNIAK, A., ACETO, H., BENDER, J., PARADIS, M., SHAW, S., VAN METRE, D., WEESE, J., WILSON, D., WILSON, J. & MORLEY, P. (2013). Using syndromic surveillance to estimate baseline rates for healthcare-associated infections in critical care units of small animal referral hospitals. *Journal of veterinary internal medicine,* 27**,** 1392-1399.

RUPLE-CZERNIAK, A., ACETO, H., BENDER, J. B., PARADIS, M., SHAW, S., VAN METRE, D., WEESE, J., WILSON, D., WILSON, J. & MORLEY, P. (2014). Syndromic surveillance for evaluating the occurrence of healthcare-associated infections in equine hospitals. *Equine veterinary journal,* 46**,** 435-440.

SALA, C., VINARD, J.-L., PANDOLFI, F., LAMBERT, Y., CALAVAS, D., DUPUY, C., GARIN, E. & TOURATIER, A. (2020). Designing a Syndromic Bovine Mortality Surveillance System: Lessons Learned From the 1-Year Test of the French OMAR Alert Tool. *Frontiers in veterinary science,* 6**,** 453.

SALATHÉ, M. (2016). Digital pharmacovigilance and disease surveillance: combining traditional and big-data systems for better public health. *The Journal of infectious diseases,* 214**,** S399-S403.
SALMON, M., SCHUMACHER, D., STARK, K. & HÖHLE, M. (2015). Bayesian outbreak detection in the presence of reporting delays. *Biometrical Journal,* 57**,** 1051-1067.
SALVATIER, J., WIECKI, T. V. & FONNESBECK, C. (2016). Probabilistic programming in Python using PyMC3. *PeerJ Computer Science,* 2**,** e55.
SANCHEZ-PINTO, L. N., LUO, Y. & CHURPEK, M. M. (2018). Big data and data science in critical care. *Chest,* 154**,** 1239-1248.
SÁNCHEZ-VIZCAÍNO, F., JONES, P. H., MENACERE, T., HEAYNS, B., WARDEH, M., NEWMAN, J., RADFORD, A. D., DAWSON, S., GASKELL, R. & NOBLE, P. J. (2015). Small animal disease surveillance. *Veterinary Record,* 177**,** 591-594.
SÁNCHEZ-VIZCAÍNO, F., NOBLE, P.-J. M., JONES, P. H., MENACERE, T., BUCHAN, I., REYNOLDS, S., DAWSON, S., GASKELL, R. M., EVERITT, S. & RADFORD, A. D. (2017). Demographics of dogs, cats, and rabbits attending veterinary practices in Great Britain as recorded in their electronic health records. *BMC veterinary research,* 13**,** 218.
SAVSNET. (2019). *Small Animal Veterinary Surveillance Network* [Online]. Available: https://www.liverpool.ac.uk/savsnet/using-savsnet-data-for-research/ [Accessed 22/09/2019.
SCHMIDT, A. M. & PEREIRA, J. B. M. (2011). Modelling time series of counts in epidemiology. *International Statistical Review,* 79**,** 48-69.
SINGLETON, D. A., SÁNCHEZ-VIZCAÍNO, F., ARSEVSKA, E., DAWSON, S., JONES, P. H., NOBLE, P. J. M., PINCHBECK, G. L., WILLIAMS, N. J. & RADFORD, A. D. (2018). New approaches to pharmacosurveillance for monitoring prescription frequency, diversity, and co-prescription in a large sentinel network of companion animal veterinary practices in the United Kingdom, 2014–2016. *Preventive Veterinary Medicine,* 159**,** 153-161.
SLOVIC, P., FISCHHOFF, B. & LICHTENSTEIN, S. (1977). Behavioral decision theory. *Annual review of psychology,* 28**,** 1-39.
SMITH, G. E., ELLIOT, A. J., IBBOTSON, S., MORBEY, R., EDEGHERE, O., HAWKER, J., CATCHPOLE, M., ENDERICKS, T., FISHER, P. & MCCLOSKEY, B. (2017). Novel public health risk assessment process developed to support syndromic surveillance for the 2012 Olympic and Paralympic Games. *Journal of Public Health,* 39**,** e111-e117.
SMITH, T. A. (2010). The Challenge of Safety Management in the 21st Century.
SOULEY KOUATO, B., DE CLERCQ, K., ABATIH, E., DAL POZZO, F., KING, D. P., THYS, E., MARICHATOU, H. & SAEGERMAN, C. (2018). Review of epidemiological risk models for foot-and-mouth disease: Implications for prevention strategies with a focus on Africa. *PloS one,* 13**,** e0208296.
SPRUIT, M., VROON, R. & BATENBURG, R. (2014). Towards healthcare business intelligence in long-term care: an explorative case study in the Netherlands. *Computers in Human Behavior,* 30**,** 698-707.
STRUCHEN, R., HADORN, D., WOHLFENDER, F., BALMER, S., SÜPTITZ, S., ZINSSTAG, J. & VIAL, F. (2016). Experiences with a voluntary surveillance system for early detection of equine diseases in Switzerland. *Epidemiology & Infection,* 144**,** 1830-1836.
STRUCHEN, R., REIST, M., ZINSSTAG, J. & VIAL, F. (2015). Investigating the potential of reported cattle mortality data in Switzerland for syndromic surveillance. *Preventive veterinary medicine,* 121**,** 1-7.
STRUCHEN, R., VIAL, F. & ANDERSSON, M. G. (2017). Value of evidence from syndromic surveillance with cumulative evidence from multiple data streams with delayed reporting. *Scientific reports,* 7**,** 1-12.
TARTAKOVSKY, A. G. & MOUSTAKIDES, G. V. (2010). State-of-the-art in Bayesian changepoint detection. *Sequential Analysis,* 29**,** 125-145.

TAYLOR, R. A., BERRIMAN, A. D., GALE, P., KELLY, L. A. & SNARY, E. L. (2019). A generic framework for spatial quantitative risk assessments of infectious diseases: Lumpy skin disease case study. *Transboundary and emerging diseases,* 66**,** 131-143.

THOMAS-BACHLI, A. L., PEARL, D. L., FRIENDSHIP, R. M. & BERKE, O. (2014). Exploring relationships between whole carcass condemnation abattoir data, non-disease factors and disease outbreaks in swine herds in Ontario (2001–2007). *BMC research notes,* 7**,** 185.

TIAN, H.-Y., YU, P.-B., LUIS, A. D., BI, P., CAZELLES, B., LAINE, M., HUANG, S.-Q., MA, C.-F., ZHOU, S. & WEI, J. (2015). Changes in rodent abundance and weather conditions potentially drive hemorrhagic fever with renal syndrome outbreaks in Xi'an, China, 2005–2012. *PLoS Negl Trop Dis,* 9**,** e0003530.

TONGUE, S. C., EZE, J. I., CORREIA-GOMES, C., BRÜLISAUER, F. & GUNN, G. J. (2020). Improving the utility of voluntary ovine fallen stock collection and laboratory diagnostic submission data for animal health surveillance purposes: a development cycle. *Frontiers in veterinary science,* 6**,** 487.

TORRES, G., CIARAVINO, V., ASCASO, S., FLORES, V., ROMERO, L. & SIMÓN, F. (2015). Syndromic surveillance system based on near real-time cattle mortality monitoring. *Preventive veterinary medicine,* 119**,** 216-221.

TRAN, D., HOFFMAN, M. D., SAUROUS, R. A., BREVDO, E., MURPHY, K. & BLEI, D. M. (2017). Deep probabilistic programming. *arXiv preprint arXiv:1701.03757*.

TRIMPOP, R. M. (1994). *The psychology of risk taking behavior*, Elsevier.

TRIPLE- S PROJECT (2013). Guidelines for Designing and Implementing a Syndromic Surveillance System. *Saint-Maurice, France: Triple S*.

TRUCCO, P., CAGNO, E., RUGGERI, F. & GRANDE, O. (2008). A Bayesian Belief Network modelling of organisational factors in risk analysis: A case study in maritime transportation. *Reliability Engineering & System Safety,* 93**,** 845-856.

TRUONG, C., OUDRE, L. & VAYATIS, N. (2018a). ruptures: change point detection in Python. *arXiv preprint arXiv:1801.00826*.

TRUONG, C., OUDRE, L. & VAYATIS, N. (2018b). Selective review of offline change point detection methods. *arXiv preprint arXiv:1801.00718*.

TRUONG, C., OUDRE, L. & VAYATIS, N. (2020). Selective review of offline change point detection methods. *Signal Processing,* 167**,** 107299.

TSUI, F.-C., WAGNER, M. M., DATO, V. & CHANG, C. (Year) Published. Value of ICD-9 coded chief complaints for detection of epidemics. Proceedings of the AMIA Symposium, 2001. American Medical Informatics Association, 711.

TULLOCH, J., MCGINLEY, L., SÁNCHEZ-VIZCAÍNO, F., MEDLOCK, J. & RADFORD, A. (2017). The passive surveillance of ticks using companion animal electronic health records. *Epidemiology & Infection,* 145**,** 2020-2029.

TWITTER. (2020). *Twitter API - Real-time access to the global conversation, right at your fingertips.* [Online]. Available: https://developer.twitter.com/en/docs/twitter-api/tools-and-libraries [Accessed 06/05/2020 2020].

VANDERWAAL, K., MORRISON, R. B., NEUHAUSER, C., VILALTA, C. & PEREZ, A. M. (2017). Translating Big Data into Smart Data for Veterinary Epidemiology. *Frontiers in veterinary science,* 4**,** 110.

VELDHUIS, A., BROUWER-MIDDELESCH, H., MARCEAU, A., MADOUASSE, A., VAN DER STEDE, Y., FOURICHON, C., WELBY, S., WEVER, P. & VAN SCHAIK, G. (2016). Application of syndromic surveillance on routinely collected cattle reproduction and milk production data for the early detection of outbreaks of Bluetongue and Schmallenberg viruses. *Preventive veterinary medicine,* 124**,** 15-24.

VELDHUIS, A., SWART, W. A., BROUWER-MIDDELESCH, H., STEGEMAN, J. A., MARS, M. H. & VAN SCHAIK, G. (2020). The comparison of three statistical models for syndromic surveillance in cattle using milk production data. *Frontiers in veterinary science,* 7**,** 67.

VELSKO, S. & BATES, T. (2016). A conceptual architecture for national biosurveillance: moving beyond situational awareness to enable digital detection of emerging threats. *Health security,* 14**,** 189-201.

VIAL, F. & REIST, M. (2014). Evaluation of Swiss slaughterhouse data for integration in a syndromic surveillance system. *BMC veterinary research,* 10**,** 1-12.

VIAL, F. & REIST, M. (2015). Comparison of whole carcass condemnation and partial carcass condemnation data for integration in a national syndromic surveillance system: The Swiss experience. *Meat science,* 101**,** 48-55.

VIAL, F., THOMMEN, S. & HELD, L. (2015). A simulation study on the statistical monitoring of condemnation rates from slaughterhouses for syndromic surveillance: an evaluation based on Swiss data. *Epidemiology & Infection,* 143**,** 3423-3433.

VIAL, F., WEI, W. & HELD, L. (2016). Methodological challenges to multivariate syndromic surveillance: a case study using Swiss animal health data. *BMC veterinary research,* 12**,** 288.

VOGELIUS, I. R., PETERSEN, J. & BENTZEN, S. M. (2020). Harnessing data science to advance radiation oncology. *Molecular oncology,* 14**,** 1514-1528.

VOSE, D. (2008). *Risk analysis: a quantitative guide*, John Wiley & Sons, pg 52.

WAGNER, M. M., TSUI, F.-C., ESPINO, J. U., DATO, V. M., SITTIG, D. F., CARUANA, R. A., MCGINNIS, L. F., DEERFIELD, D. W., DRUZDZEL, M. J. & FRIDSMA, D. B. (2001). The emerging science of very early detection of disease outbreaks. *Journal of public health management and practice,* 7**,** 51-59.

WANG, Y. & LI, P. (Year) Published. Algorithm and Hardware Co-Design for FPGA Acceleration of Hamiltonian Monte Carlo Based No-U-Turn Sampler. The 32nd IEEE International Conference on Application-specific Systems, Architectures and Processors, 2021.

WARNS-PETIT, E., MORIGNAT, E., ARTOIS, M. & CALAVAS, D. (2010). Unsupervised clustering of wildlife necropsy data for syndromic surveillance. *BMC Veterinary Research,* 6**,** 1-11.

WEIL, R. S., SCHWARZKOPF, D. S., BAHRAMI, B., FLEMING, S. M., JACKSON, B. M., GOCH, T. J., SAYGIN, A. P., MILLER, L. E., PAPPA, K. & PAVISIC, I. (2018). Assessing cognitive dysfunction in Parkinson's disease: An online tool to detect visuo-perceptual deficits. *Movement Disorders,* 33**,** 544-553.

WENG, H.-Y., GAONA, M. A. & KASS, P. H. (2020). Evaluation of a novel syndromic surveillance system for the detection of the 2007 melamine-related nephrotoxicosis foodborne outbreak in dogs and cats in the United States. *PeerJ,* 8**,** e9093.

WESTERA, W. (2021). Comparing Bayesian Statistics and Frequentist Statistics in Serious Games Research. *International Journal of Serious Games,* 8**,** 27-44.

WHO. (2020). *Public Health Surveillance* [Online]. Available: https://www.who.int/immunization/monitoring_surveillance/burden/vpd/en/ [Accessed 08/09/2020.

WILSON, R. & CROUCH, E. (1987). Risk assessment and comparisons: an introduction. *Science,* 236**,** 267-270.

WINSLOW, C.-E. (1920). THE UNTILLED FIELDS OF PUBLIC HEALTH. *Science,* 51**,** 23-33.

WOODALL, W. H. (2000). Controversies and contradictions in statistical process control. *Journal of Quality Technology,* 32**,** 341-350.

WOODS, M. (2009). A contingency theory perspective on the risk management control system within Birmingham City Council. *Management Accounting Research,* 20**,** 69-81.

WORLD HEALTH ORGANIZATION (2012). Rapid risk assessment of acute public health events. World Health Organization.

YAU, C. & CAMPBELL, K. (2019). Bayesian statistical learning for big data biology. *Biophysical reviews,* 11**,** 95-102.

YOUSEFINAGHANI, S., DARA, R., POLJAK, Z., BERNARDO, T. M. & SHARIF, S. (2019). the Assessment of twitter's potential for outbreak Detection: Avian Influenza Case Study. *Scientific reports,* 9**,** 1-17.

ZHANG, H. & MALONEY, L. T. (2012). Ubiquitous log odds: a common representation of probability and frequency distortion in perception, action, and cognition. *Frontiers in neuroscience,* 6**,** 1.

ZHOU, M., LI, L., DUNSON, D. & CARIN, L. (Year) Published. Lognormal and gamma mixed negative binomial regression. Proceedings of the... International Conference on Machine Learning. International Conference on Machine Learning, 2012. NIH Public Access, 1343.

ZHU, J. & LIU, W. (2020). A tale of two databases: The use of Web of Science and Scopus in academic papers. *Scientometrics***,** 1-15.

ZIEMANN, A., FOUILLET, A., BRAND, H. & KRAFFT, T. (2016). Success factors of European syndromic surveillance systems: a worked example of applying qualitative comparative analysis. *PloS one,* 11.

ZURBRIGG, K. J. & VAN DEN BORRE, N. M. (2013). Factors associated with good compliance and long-term sustainability in a practitioner-based livestock disease surveillance system. *The Canadian Veterinary Journal,* 54**,** 243.

Appendix

The web links to extra materials used in this research work are below

https://uelac-my.sharepoint.com/:f:/r/personal/u1437194_uel_ac_uk/Documents/Thesis_extras?csf=1&web=1&e=gwntBv

https://drive.google.com/drive/folders/1ALHLD1IP0za4SPjsVDY_-R4CWgQ5wfV9?usp=sharing

University of East London

Dear Victor

Application ID: ETH2122-0019

Original application ID: ETH1920-0217

Project title: Analysis of Heterogeneous Data Sources for Veterinary Syndromic Surveillance to Improve Public Health Response and Aid Decision Making

Lead researcher: Mr Victor Adejola

Your application to Ethics and Integrity Sub-Committee was considered on the 29th of September 2021.

The decision is: **Approved**

The Committee's response is based on the protocol described in the application form and supporting documentation.

Your project has received ethical approval for 4 years from the approval date.

If you have any questions regarding this application please contact your supervisor or the secretary for the Ethics and Integrity Sub-Committee.

Approval has been given for the submitted application only and the research must be conducted accordingly.

Should you wish to make any changes in connection with this research project you must complete 'An application for approval of an amendment to an existing application'.

The approval of the proposed research applies to the following research site.

Research site: University of East London

Principal Investigator / Local Collaborator: Mr Victor Adejola

Approval is given on the understanding that the UEL Code of Practice for Research and the Code of Practice for Research Ethics is adhered to.

Any adverse events or reactions that occur in connection with this research project should be reported using the University's form for Reporting an Adverse/Serious Adverse Event/Reaction.

The University will periodically audit a random sample of approved applications for ethical approval, to ensure that the research projects are conducted in compliance with the consent given by the Research Ethics Committee and to the highest standards of rigour and integrity.

Please note, it is your responsibility to retain this letter for your records.

With the Committee's best wishes for the success of the project

Yours sincerely

Fernanda Silva

Administrative Officer for Research Governance

University of East London

Dear Victor

Application ID: ETH1920-0217

Original application ID: ETH1920-0140

Project title: A Framework to Support the Integration and Analysis of Heterogeneous Data Sources in Veterinary Syndromic Surveillance

Lead researcher: Mr Victor Adejola

Your application to Arts and Creative Industries School Research Ethics Committee was considered on the 7th of May 2020.

The decision is: **Approved**

- *In view of the COVID-19 pandemic, the University Research Ethics Sub-Committee (URES) has taken the decision that all postgraduate research student and staff research projects that include face-to-face participant interactions, should cease to use this method of data collection, for example, in person participant interviews or focus groups. Researchers must consider if they can adapt their research project to conduct participant interactions remotely. The University supports Microsoft Teams for remote work. New research projects and continuing research projects must not recruit participants using face-to-face interactions and all data collection should occur remotely. These regulations should be followed on your research until national restrictions regarding Covid-19 are lifted. For further information please visit the Public Health website page https://www.gov.uk/government/organisations/public-health-england*

The Committee's response is based on the protocol described in the application form and supporting documentation.

Your project has received ethical approval for 2 years from the approval date.

If you have any questions regarding this application please contact your supervisor or the secretary for the Arts and Creative Industries School Research Ethics Committee.

Approval has been given for the submitted application only and the research must be conducted accordingly.

Should you wish to make any changes in connection with this research project you must complete 'An application for approval of an amendment to an existing application'.

The approval of the proposed research applies to the following research site.

Research site: University of East London

Principal Investigator / Local Collaborator: Mr Victor Adejola

Approval is given on the understanding that the UEL Code of Practice for Research and the Code of Practice for Research Ethics is adhered to.

Any adverse events or reactions that occur in connection with this research project should be reported using the University's form for Reporting an Adverse/Serious Adverse Event/Reaction.

The University will periodically audit a random sample of approved applications for ethical approval, to ensure that the research projects are conducted in compliance with the consent given by the Research Ethics Committee and to the highest standards of rigour and integrity.

Please note, it is your responsibility to retain this letter for your records.

With the Committee's best wishes for the success of the project

Yours sincerely

Fernanda Silva

University of East London

Administrative Officer for Research Governance

University of East London

Dear Victor

Application ID: ETH1920-0140

Original application ID: ETH1920-0073

Project title: A Framework to Support the Integration and Analysis of Heterogeneous Data Sources in Veterinary Syndromic Surveillance

Lead researcher: Mr Victor Adejola

Your application to Arts and Creative Industries School Research Ethics Committee was considered on the 6th of February 2020.

The decision is: **Approved**

The Committee's response is based on the protocol described in the application form and supporting documentation.

Your project has received ethical approval for 2 years from the approval date.

If you have any questions regarding this application please contact your supervisor or the secretary for the Arts and Creative Industries School Research Ethics Committee.

Approval has been given for the submitted application only and the research must be conducted accordingly.

Should you wish to make any changes in connection with this research project you must complete 'An application for approval of an amendment to an existing application'.

The approval of the proposed research applies to the following research site.

Research site: University of East London

Principal Investigator / Local Collaborator: Mr Victor Adejola

Approval is given on the understanding that the UEL Code of Practice for Research and the Code of Practice for Research Ethics is adhered to.

Any adverse events or reactions that occur in connection with this research project should be reported using the University's form for Reporting an Adverse/Serious Adverse Event/Reaction.

The University will periodically audit a random sample of approved applications for ethical approval, to ensure that the research projects are conducted in compliance with the consent given by the Research Ethics Committee and to the highest standards of rigour and integrity.

Please note, it is your responsibility to retain this letter for your records.

With the Committee's best wishes for the success of the project

Yours sincerely

Fernanda Silva

Administrative Officer for Research Governance

University of East London

Dear Victor

Application ID: ETH1920-0073

Project title: A METADATA FRAMEWORK FOR DIMENSIONALITY REDUCTION OF BIG DATA

Lead researcher: Mr Victor Adejola

Your application to University Research Ethics Sub-Committee was considered on the 14th of January 2020.

The decision is: **Approved**

The Committee's response is based on the protocol described in the application form and supporting documentation.

Your project has received ethical approval for 2 years from the approval date.

If you have any questions regarding this application please contact your supervisor or the secretary for the University Research Ethics Sub-Committee.

Approval has been given for the submitted application only and the research must be conducted accordingly.

Should you wish to make any changes in connection with this research project you must complete 'An application for approval of an amendment to an existing application'.

The approval of the proposed research applies to the following research site.

Research site: University of East London

Principal Investigator / Local Collaborator: Mr Victor Adejola

Approval is given on the understanding that the UEL Code of Practice for Research and the Code of Practice for Research Ethics is adhered to.

Any adverse events or reactions that occur in connection with this research project should be reported using the University's form for Reporting an Adverse/Serious Adverse Event/Reaction.

The University will periodically audit a random sample of approved applications for ethical approval, to ensure that the research projects are conducted in compliance with the consent given by the Research Ethics Committee and to the highest standards of rigour and integrity.

Please note, it is your responsibility to retain this letter for your records.

With the Committee's best wishes for the success of the project

Yours sincerely

Fernanda Silva

University of East London

Administrative Officer for Research Governance