Retinopathy of Prematurity and blood oxygen saturation: confirmation of the relationship with high fidelity data

by

K. Emmanuel Shiron Fernando

A Thesis Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of

Health Science in Health informatics

In

The Faculty of Health Sciences Program

University of Ontario Institute of Technology

August, 2017

Certificate of Approval

Abstract

Limitations in the area of knowledge discovery amongst physiological data streams are a challenge that emanate from the existing but still overlooked requirement for harnessing and using valuable medical information embedded in physiological data streams that are acquired from bedside patient monitoring devices. This thesis presents a standardised, medical informatics approach, illustrated through a single case study, taken to identify a solution for challenge.

The combination of the CRoss Industry Standard Process for Temporal Data Mining (CRISP-TDM) approach that supports temporal data mining of physiological data streams and the CRISP-DM_0 aspect of the methodology that performs null hypothesis driven, confirmatory data mining, known as the CRISP-TDM_0, has been proposed by a previous research work. This research used this CRISP-TDM_0 as the standardised approach to managing, reporting and performing retrospective clinical research in the case study.

This research was designed to confirm the relationship between high fidelity blood oxygen saturation data and Retinopathy of Prematurity (ROP). The temporally abstracted (TA) behaviors of blood oxygenation saturation levels (SpO_2) of nine premature neonates were analysed using the Artemis platform, an information architecture that complies with the Big Data concept [10], [22]. The results were then correlated with the clinically detected ROP.

An initial clustering analysis of the modelling phase was performed based on the visualisation of patterns resulted from hourly SpO_2 TA profiles. This approach enabled the emergence of three clusters: cluster 1 with subjects 3 and 8 both having episodes each hour of some over and under oxygenation; cluster 3 comprising subjects 1, 2, 4, 5, 6, and 10 with predominant over oxygenation and cluster 2 comprising subject 9 having a degree of over and under oxygenation that fell in between those of the above two clusters. A sub-clustering based on the outcome for ROP identified subjects 3 and 8 (cluster 1) as both having stage 3 ROP; subject 9 (cluster 2) as having stage 1 ROP; and subjects 1, 2, 4, 5, 6, and 10 (cluster 3) as having stage 0 ROP. Another sub-clustering based on gestational age (GA) that was performed within the initial sub-clustering demonstrated an association as subjects of cluster 1 belonging to ≤ 26 weeks GA, subject 9 of cluster 2 as 26 weeks GA and those of cluster 3 with ≥ 27 weeks GA. An additional sub-

clustering based on known risk factors for ROP; sex, birth weight, occurrence of respiratory distress syndrome and neonatal sepsis did not correlate as a discriminator for ROP.

The CRISP-TDM_0 methodology that operated within Artemis could be demonstrated as the standardised approach as it allowed us to follow its process mining steps in this research.

Acknowledgements

Thank you Professor Carolyn McGregor, you gave me excellent guidance over the years of my studies. Your patience and the belief you had in me was the greatest strength right through the entire Master's journey. You consistently endorsed this thesis to be my own work, but steered me in the accurate direction whenever you thought I needed it.

Dr. Andrew James, thank you for your untiring, unconditional support and guidance. Your encouraging thoughts and words right through to the end of my studies are immensely appreciated.

Professor Paul Yielder, Professor Jennifer Percival and Kathleen Smith, unwittingly all of you have made a profound impact on my life through your teaching for which I am truly grateful.

Last but not least, dear Rishi, this past few years has been an incredible experience and I would like to thank you for giving me the opportunity to be associated with Professor Carolyn McGregor and Dr. Andrew James, without you that would not have been possible.

Publications Related to this Thesis

Correlation of Retinopathy of Prematurity and Blood Oxygen Saturation in Neonates using Temporal Data Mining: A Pilot Study; K.E.S. Fernando, C. McGregor AM, *Senior Member, IEEE* and Andrew G. James. A platform presentation at The 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC'16) at Disney's Contemporary Resort at Walt Disney World® Resort, Lake Buena Vista (Orlando), Florida USA.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

Knowledge discovery of patterns in physiological data streams that are pathophysiological behaviours previously unknown for medical conditions is emerging as a new trend for the delivery of neonatal critical care. However, there are limitations to this process of knowledge discovery: a broader general challenge that has an impact on neonatal care and the lack of a standardised informatics methodology operating within a computational analytic environment is contributing to this limitation. This thesis demonstrates how a single case study designed from the domain of neonatal intensive care is employed to find an elucidation to this broader general challenge.

Patient monitoring devices are capable of measuring various physiological parameters, i.e. electrocardiography (ECG), heart rate (HR), blood pressure (BP), blood oxygen saturation level (SpO_2), pulse rate (PR), respiration rate (RR), capnography, which is the process of monitoring the concentration or partial pressure of carbon dioxide in the exhaled respiratory gases, and body temperature. Physiological data streams that are acquired from these patient monitoring devices exhibit early indicators of potentially life threatening conditions such as nosocomial infections (NI) and these points of interest in the data stream may precede the existing clinical practice detection of the disease [1]. Therefore, studying the evolution of physiological data parameters over time is more important than relying on instant parameter values because the potential for new medical knowledge discovery embedded in the patterns of streams is greater than those of conventional spot recordings [1]. However, manually prepared NICU medical notations in the past summarised continuously streamed physiological data only by intermittent readings and were not conducive to recording abnormal behaviours among the multiple streams that occur more frequently. For instance, transient falls in BP, HR and RR variability and fluctuations in SpO_2 levels, which may be of critical importance for clinical management of the premature neonate, can go undetected through conventional spot recordings [2].

The increasing volumes of patient data generated by Neonatal Intensive Care Unit (NICU) patient monitoring devises are as dynamic as the environment within which they are produced. Present day electronic records display physiological data streams captured directly from the bedside medical devices at a high frequency and these data can be visualised in a temporal, graphical format. The data streams so produced are in the calibre of waveform data sampled at

over 500 readings a second and the human brain cannot process and analyse the data as rapidly and in real-time as a computational analytic environment would perform [3]. Similarly, the insight that can be gleaned from studying multiple streams of physiological data a premature neonate would generate at a given time is important because these streams are not isolated from each other. Even though NICU patient monitoring devices generate high frequency, high volume, highly dimensional real-time physiological data streams and these diagnostic data components are virtual representations of certain physiological stream behaviours peculiar to a given clinical event that may manifest before typical clinical signs appear, the requirement for harnessing and using the valuable medical information embedded in them still seems to remain overlooked. If an approach is taken to discover the knowledge embedded in these combined physiological data streams rather than in conventional intermittent spot recordings, that knowledge can be used to support prompt clinical decision making and timely medical intervention which in return would provide enormous medical, social and economic benefits.

In order to address the limitation in the area of knowledge discovery in these physiological data streams which is the broader medical problem under discussion in this thesis, the necessity arises of a data mining (DM) method that perform data storage and multi-dimensional data analysis because traditional statistical methods are not well suited to evaluating the probability of coincidental patterns in high-dimensional data sets. This prompts the need for a dynamic data mining (DM) method that can meet the needs of high-frequency, high-volume, real-time, multi-dimensional time series data. Accordingly, recent research has supported the need for sophisticated DM systems that combine TAs frequently as a preprocessing step to DM. This approach of integrated Temporal Data Mining (TDM) of multi-dimensional time series data can cross correlate data from multiple streams for a combined TA model as a complex indicator of condition onset.

The studies that dealt with positively correlating ROP with a higher blood oxygen saturation level, variability, fluctuation and duration of exposure etc., [5], [6], [7], [8] had not used a standardized approach to capture and analyze real-time blood SpO_2 data, which results in incomplete information about the study being reported leading to difficulties in repeatability as a result. Researchers have tried to solve these issues and found their approaches showing certain limitations with regards to sensitivity and specificity of the methodology they used even though

most of data mining methods and temporal abstractions (TAs) they used were capable of demonstrating conditions exhibiting certain physiological stream behaviors. Therefore this study approach has been designed to address that issue by using patient characteristics to gain better individual patient understanding through retrospective data mining and using TAs to improve sensitivity and specificity by creating subgroups of physiological behaviors.

Even though the results of Cirelli et al. [13]'s study on analysing neonatal blood SpO_2 every second in real-time using the Artemis platform [2], [3], [4], provided a better representation of immature retina's exposure to potentially damaging oxygen levels, a standardised approach to support the research continuum in this area is currently unavailable. This limitation in the area of knowledge discovery of physiological data streams which is the broader medical problem can be viewed as emanating from the lack of a standardised informatics methodology to use this information to its optimum potential.

1.1 CRoss Industry Standard Process for Data Mining, CRISP-TDM$_0$ methodology

By way of a Neonatal Intensive Care Unit (NICU) case study, Catley et al. [3] investigated the current Cross Industry Standard Process for Data Mining (CRISP-DM) approach and the comparisons they made to the areas of more established applications of artificial intelligence in medicine revealed that CRISP-DM is insufficient to meet the needs of integrated Temporal Data Mining (TDM) systems. As a result, Catley et al. [3] proposed an extended CRISP-DM approach when modeling clinical systems that apply Data Mining (DM) and Temporal Abstractions (TA). This resultant CRISP-TDM methodology through TDM could aid reporting results of TDM-based clinical investigations on multi-dimensional time-series data.

In order to conduct both exploratory and confirmatory data mining of physiological data streams in neonatal critical care, Heath and McGregor's work involved construction of data mining tools and methodologies that proposed the CRISP-DM$_0$ aspect of CRISP-DM methodology [9]. This further enables the use of exploratory data mining as a tool to find unknown patterns or relationships to create a hypothesis. An alternative null hypothesis could then be created and tested during iterative, explanatory, or confirmatory, data mining. This thesis is built on the previous research work that proposed the CRoss Industry Standard Process for Data Mining, CRISP-TDM$_0$ methodology that merges CRISP-TDM and CRISP-DM$_0$ aspects. The following subsections of this chapter detail how the CRISP-TDM$_0$ was utilised as the standardised

informatics methodology to manage report, perform and document this retrospective, population based, cross sectional, analytical study.

The CRISP-TDM_0 methodology is integrated within the knowledge discovery component of Artemis platform, a computational analytic environment devised for modeling Intelligent Data Analysis (IDA) [3]. These features can assist clinicians to isolate consistent temporal behaviours in physiological data streams across multiple patient data sets representing common behaviour at the onset of disease [3], [4]. Within the perspective of this research, this is performed by correlating the temporally abstracted behaviours of hourly blood SpO_2 data to the onset of ROP. These aspects are discussed in detail in chapter 4, Research Construction and Methodology of this thesis.

1.2 The Artemis Platform Architecture

Artemis is an established real-time analytic system that captures, transmits, processes, stores in real-time and conducts analysis of high frequency physiological data streams acquired from bedside monitoring devices. By means of temporal data mining of time series data it is capable of identifying relationships between physiological data streams and specific clinical conditions for multiple patients in support of retrospective clinical research studies [4].

The data acquisition component of the Artemis platform enables the provision of real-time synchronous medical device data and asynchronous clinical data and information extracted from a Clinical Information Management System (CIMS) which are then forwarded to an online Analysis component that operates in real-time. For this real-time component, Artemis employs IBM's InfoSphere Streams, a novel streaming middleware system that processes data in real-time and then enables data storage within the Data Persistency component. Stream processing was initially supported by the Stream Processing Application Declarative Engine (SPADE) together with the SPADE language, which was the initial programming language for IBM's InfoSphere Streams version 1.0 middleware. This has now been upgraded to the stream processing language (SPL) which was used in this research as part of InfoSphere Streams version 3.1. For the Knowledge Extraction component, Artemis uses temporal data mining that supports the discovery of condition onset behaviours in physiological data streams and associated clinical data. New knowledge, once tested through clinical research, is transferred for use within the

Online Analysis component through the Redeployment component, which translates the knowledge to a SPADE representation [3], [4].

The Artemis platform can enable the instantiation of clinical rules and can support multidimensional analysis. Further, this platform can be provided not only through a local installation but also through cloud computing, a phenomenon which allows service of critical care through a Health Analytics-As-A-Service model that is of interest for small remote hospitals where infrastructure for information technology is comparatively limited. By way of the cloud computing approach, raw physiological data streams and related clinical data can be transmitted securely over the Internet, with de-identified patient data, for processing at the cloud computing site. All these features complement screening susceptible neonates early for better intervention which in turn would bring about enormous economic and social benefits without a negative impact on patient care [4].

This research utilises the Artemis framework's Service based multidimensional Temporal Data Mining ($STDM^n_0$) capacity to perform process mining that enables knowledge discovery of a new condition onset pathophysiological and physiological data stream behaviour: high fidelity SpO_2 data streams, a surrogate marker for the oxygen level of the retinal vasculature at any given second to the onset of Retinopathy of Prematurity (ROP), which is an eye pathology that has detrimental consequences for the child's vision. ROP has a predilection for premature neonates where elevated blood oxygenation levels and their fluctuations have been deemed as pronounced associations amongst several other risk factors [5], [6], [7], [8]. Accordingly, a more detailed analysis of the blood oxygen saturation level paradigms using the Artemis platform is performed and those values and pattern visualisations are compared to the severity of ROP occurring before 35 completed weeks postmenstrual age (PMA). This research contributes to the larger Artemis Project currently in progress in the Health Informatics Research Laboratory at University of Ontario Institute of Technology (UOIT).

1.3 Big Data in Neonatal Intensive Care

The ever increasing volume in medical data termed 'Big Data' has resulted in major challenges faced by intensive care units (ICUs)because this is a pool of data sets, so massive and multifaceted, it rules out the opportunity of processing by conventional data processing methods. Nevertheless, the data generated by these patient-monitoring devices are a Big Data challenge

with near real-time restrictions for processing medical algorithms designed to predict certain neonatal conditions. The near real-time implementation of algorithms intended for early prediction of neonatal clinical conditions is associated with a demand in computational capability because these algorithms are designed to process huge quantities of physiological data and need to be executed for each individual patient. The limited scope of the memory of the devices used in ICUs can store data approximately for 72 hours, is a challenge for contemporary Big Data analytics. These data cannot be deleted as soon as a patient is discharged, rather need to be stored for a considerable time so that medical professionals can extract only relevant clinical information from vast amounts of data to reduce treatment costs. Nevertheless, these data can be preserved for future study purposes that can improve patient care. Both paper and most of available electronic charting mechanisms allow storage of limited number of values per hour of a particular physiological parameter and it is highly likely that the medical staff follow a qualitative than a quantitative approach to judge how stable a patient is [67]. Hence, Big Data has become the only solution for secure and efficient storage as well as retrieval of medical data [68]. Artemis leverages Streams to deliver a Big Data based solution for neonatal critical care [10], [11], [12]. A detailed account on this will be given in Chapter 2.

1.4 Research Motivation

A motivation of this research is to apply recent computational knowledge and information technology (IT) to neonatal physiological monitoring to support real-time data mining of the vast amounts of unused data. Determining whether the Artemis framework can support data exploration and clustering based on patient characteristics that in turn will lead to identify trends and patterns in patient specific physiological data streams to improve real-time clinical management and clinical decision support is a motivation as well. Another motivation of this research is being able to use data mining as it offers a newer approach to the analysis of physiological data streams for clinical research on stored physiological data streams to deduce new findings for condition onset prediction indicators based on patient characteristics.

1.5 Retinopathy of Prematurity (ROP)

1.5.1 The scope of this clinical problem

This study is important because ROP is not only a common blinding disease in children in the developed world, despite current treatment, but also is becoming increasingly prevalent in the

developing world [14], [15]. Recent advancements in neonatal care have increased the survival of preterm neonates, and this consequently has resulted in an increase in the incidence of ROP. There is a potential for enormous economic and social benefits of screening programmes and early intervention, if carried out properly [13], [17].

1.5.2 Risk factors and pathogenesis of Retinopathy of Prematurity

The identified risk factors for ROP are the early neonatal course and relative hyperoxaemia in very low birth weight infants, especially when alternating with periods of hypoxaemia [6] [14]. Elevated glucose [15], hypercapnia, ventilator therapy, and blood transfusions [16] are some of the other known risk factors.

ROP is a vasoproliferative retinopathy occurring due to injury to the differentiating primitive capillary meshwork of the premature retina [16]. ROP progresses in two phases [15]: the first phase is characterised by delayed retinal vascular growth after birth and partial regression of existing vessels during the interval from birth until postmenstrual age 30-32 weeks; the second phase is characterised by hypoxia-induced pathological retinal neovascularization which begins between 32-34 weeks postmenstrual age. Vessels surviving this insult form arteriovenous shunts and in 90% of cases, cells inside these shunts divide and produce normal capillary mesh works leading to a complete regression of this abnormal process, however, in 10%, cells proliferate and grow on the surface of retina and vitreous resulting in excessive vasoproliferation leading to traction on the retina and detachment leading to a spectrum of long term morbidities ranging from myopia to blindness [16].

1.5.3 Retinopathy of Prematurity: the state of understanding

Evidence-informed practices for oxygen administration form the framework for guidelines to maintain the SpO_2 level at 85-92% to reduce the risk for ROP and other complications while increasing the benefits of oxygen [13]. Another study has revealed that for the ongoing observation and management of preterm infants, a SpO_2 target range of 85-93% seems to be the most appropriate, with alarm limits set within 1-2% of this target [16]. However despite an improved understanding of ROP and ongoing investigations of supplemental therapeutic oxygen including recent clinical trials—Surfactant, Positive Airway Pressure, Pulse Oximetry Randomized Trial, SUPPORT [7] and Benefits of Oxygen Saturation Targeting, BOOST [17]—the best oxygen profiles to reduce ROP risk while optimizing preterm infant health and

development remain unknown [7], [18]. Therefore, all the above information together with the knowledge gathered from surveys carried out on NICU guidelines for oxygen administration and monitoring continue to reveal the existence of wide variances in the target range used for blood oxygen saturation levels [6]. Nevertheless, targeting oxygen saturation levels has been a constant challenge despite extensive investigations carried out on preterm neonates [7], and these challenges can be appreciated within the context of previous related work carried out on this area. A study by Hagadorn et al. [8] has shown that SpO_2 had remained in the target range only 50% of the time; and the fluctuations that occur are such that, a timely intervention cannot be performed since occurring episodes of deviations have reverted back to normal before a healthcare provider discriminates the potential significance of the event. Furthermore, the results of a most recent study by Cirelli et al. [13] have shown that the second by second monitoring of SpO_2 levels, in contrast to the conventional hourly monitoring, reveals significant differences in the pattern of oxygen variation. The above issues therefore suggest that the monitoring of SpO_2 more frequently is a definite requirement for improved care of these immature infants.

1.6 Aims and objectives of the research

It was assumed that the close interrelationship of the steps of CRISP-TDM_0 will enable one step to flow to the next by providing a structure to plan out the remainder of the research based on the changed pathway in the event that changes were required and this will in turn help to manage the medical data streams research.

Likewise, it was assumed that adhering to the steps proposed by CRISP-TDM_0 and reporting the information as required by the phases as research progressed will lead to demonstrate the CRISP-TDM_0 as standardised approach for reporting the research on medical data streams. It was further assumed that following the phases of CRISP-TDM_0 would lead to greater detail provided about the study rather than traditional approaches of performing the research. The design of this research was based on the inference that the SpO_2 measured in high fidelity is a better representation of immature retina's exposure to oxygen. Hence adhering to the results of this analysis should enable a risk score for future development of ROP.

1.6.1 Primary objective

To demonstrate the capability of the standardised informatics methodology for managing, reporting and performing retrospective clinical research using physiological data streams from the domain of neonatal intensive care.

1.6.2 Secondary objectives

a) To demonstrate the instantiation of this design within the Artemis platform

b) To develop a risk assessment score for the onset of ROP based on the existing ROP stages (stage 0-5).

c) To develop a real-time algorithm for the monitoring of blood oxygen saturation levels to assist with oxygen administration and minimise the risk for the development of ROP.

1.7 Research questions

1. Can the extensions of CRISP-TDM methodology enable the complete management, reporting and completion of a clinical research study involving physiological streaming data that incorporates null hypothesis testing?
2. Is there a correlation between the blood oxygen saturation levels of immature newborn infants, born at less than 32 weeks gestational age, during the first two weeks of life and the severity of retinopathy of prematurity occurring before 35 completed weeks postmenstrual age?
3. Is it possible to calculate the probability at postnatal ages three days, seven days, ten and fourteen days for the severity of retinopathy of prematurity occurring before 35 completed weeks postmenstrual age?

A discussion on whether those objectives could be achieved and answers to the above research questions will be available at the conclusion section of this thesis.

1.8 Scope of the thesis

The scope of this thesis is embedded within the extensions to CRISP-DM proposed by Catley et al. [3] to form CRISP-TDM together with the extensions by McGregor and Heath to create CRISP-DM_0 and the integration of these extensions within the CRISP-TDM_0 model as it stands within the context of this research.

CRISP-TDM extensions are provided in detail in chapter 4 relate to additions in: phase 1 – Business Understanding; phase 2 – Data Understanding; phase 4 – Data Modelling; phase 5 – Evaluation and phase 6 –Deployment to the CRISP DM model. These extensions were suggested as per analysis by Catley, et al. [3] that revealed those phases of the CRISP 1.0 model are limited in their ability to describe certain issues that has an impact on the knowledge discovery aspect of the research. These issues will be discussed in detail in chapter 2 of this thesis.

The scope of this thesis includes designing a method to capture neonatal blood oxygen saturation levels in high fidelity, to conduct a more detailed analysis of these saturation level paradigms using the Artemis platform and comparing those values to the clinically detected ROP stages. To propose a design for a real-time algorithm for monitoring blood oxygenation levels to assist with a safer means of oxygen administration for the premature neonate is within the scope of this study. Determining how these algorithms could be designed based on the principles of stream based computing as the paradigm within which they were coded fits within the scope of this thesis as well. In keeping with this, the question of whether or not blood SpO_2 level analysis proposed in this work is considered a suitable problem for streams computing approaches to solve is also within the scope of this thesis. However, the technical coding and development of the algorithm that would run in real-time as well as the visualisation of the algorithms results are not within the scope of this thesis research.

The presentation and visualisation of the algorithms results will be continued by the Artemis team in future research. The output of results within the range of this thesis includes text-based output, concordance and sensitivity and specificity reports in comparison to expert clinical annotations.

1.9 Contribution of this research thesis to Health Informatics

The CRISP-TDM_0 methodology is the health informatics contribution. This contribution can be viewed in relation to the need for development of a new research tool that is of utmost importance within the background of a second by second analysis of blood oxygen saturation levels, a surrogate marker for the oxygen level of the immature retinal vasculature, at any given second is superior to an analysis of hourly data. Hence this research used CRISP-TDM_0 methodology incorporated within the multidimensional approach of the Artemis platform that

used time-series data and temporal data mining to conduct a more detailed analysis of the blood oxygen saturation level paradigms and compare those values to the remaining ROP stages.

1.10 Thesis overview

The remainder of the thesis is structured as follows. Chapter 2 presents the informatics literature review, focusing on current frameworks for enabling data mining and temporal abstractions and in particular entities included within these temporal abstractions that further define temporal condition onset of illness based on patient characteristics such as gestational age and SpO_2. In chapter 3 the clinical aspect of ROP is introduced through a clinical literature review. The extent of the problem in a global perspective, pathology, risk factors for ROP and various associations of arterial oxygenation with the onset ROP are discussed in detail. Chapter 4: Research Construction and Methodology fully describes the existing multi-agent framework, including integration of relevant aspects of the extended data mining model, Cross Industry Standard Process for Data Mining to support temporal data mining CRISP-TDM and the version that facilitate null hypothesis testing on real-time series physiological data streams CRISP-DM_0. This combined CRISP-DM_0 and CRISP-TDM model known as CRISP-TDM_0 has been proposed as the standardised approach for managing, reporting and performing retrospective clinical research using medical data streams. Chapter 5 deals with Informatics Data Analysis and clinical correlation of Retinopathy of Prematurity. This section presents the clustering analysis of the modelling phase within the context of this case study using visualisations of the TA based process mining results of temporally abstracted hourly scores of final nine study participants. Chapter 6 discusses the overall study under three aspects: reflecting the success of applying CRISP-TDM_0 to this research; focusing on the effects of risk factors that occurred as adverse clinical events for ROP and reflecting them on the study subjects and finally, exploring the possibilities in the new areas of research based on the knowledge gathered through this study. Chapter 7 concludes the thesis by summarising it and linking the intended objective goals to the achievements made through this research experiment. Further, it discusses the limitations of the study and future advancements.

Chapter 2: Informatics Literature Review

This section of the thesis presents the literature from an informatics perspective, highlighting the importance of knowledge discovery through Temporal Data Mining in multidimensional physiological data streams within the background of existing limitations to the knowledge discovery in neonatal critical care. This review further demonstrates, using several research studies as examples, how the integration of the CRISP-TDM methodology within the Artemis platform that performs real-time data analysis of physiological data streams, is used for retrospective clinical research and prospective clinical decision support. Depending on the requirement for a more rigorous validation of hypothesis in experimentations in healthcare than any other sector, the necessity to construct data mining tools and methodologies to conduct both exploratory and confirmatory data mining is discussed together with the extended aspect of CRISP-DM that support null hypothesis which is known as CRISP-DM_0. This section also demonstrates how a background is set for the demonstration of CRISP-TDM_0, which is the merged CRISP- TDM and CRISP-DM_0 aspects as the standardised informatics methodology that is being used in this research study.

2.1 Streams deployment in healthcare

2.1.1 Advantage of knowledge discovery in physiological data streams

Within the NICU individual premature neonates undergo rapid transitions in specific patient characteristics that manifest as physical growth, gaining weight, development and changes in HR, RR and BP etc. These changes occur normally due to the advancing postnatal age but could be a manifestation of another pathophysiological condition.

Critically ill babies often give rise to abnormal variations in their vital parameters more frequently and these patterns of varying vital parameters within physiological data streams may reflect significant information for determining neonate's survival and long-term morbidity i.e. frequent transient falls in blood pressure and blood oxygen saturation in neonatal critical care may warn a neonatologist an impending disease [3]. Interestingly, some of these variations in the vital parameters occur a considerable time before the actual onset of physical signs [12], [30]. Supporting that, recent medical research literature has reported that physiological data streams exhibits early indicators of potentially life threatening nosocomial infections (NI) and that these points of interest in the data precede the existing clinical practice detection [1], [30]. Hence,

being able to capture these data patterns of interest before a detrimental outcome of the disease situation occurs would lead a neonatologist to make clinical decisions promptly. This will reduce the cost of neonatal critical care which usually amount to much more than that of adult care due to the associated comparatively longer hospital stay. This can be supported by the following reference i.e. over 40% of the total of over 3 million patients those were admitted to critical care units in Canada belonged to an age group of over 63 years. This data that belong to 2003-04 further mention that neonatal patients comprised only 15% of the total admissions. However, the further analysis of data reveals that the average hospital stay for a neonatal patient is almost three times as long as that of other units. What reveals through this data is crucial provided the associated higher cost in critical care in intensive care units and NICUs that account for comparatively higher percentage of inpatient direct expenses [19], [20].

Most of the NICU medical records in the past years were largely manually prepared paper notes where nursing staff often manually summarised continuously streamed physiological data to single readings at 30 or 60 minute intervals [3]. These hand notations of data were not conducive to recording abnormal behaviours among the multiple streams that frequently occur minute by minute or second by second. Hence, there would have been numerous instances where these variations have gone undetected and unrecorded through those conventional intermittent spot recordings in the past clinical practice. These abnormal data stream behaviours may be of critical importance for neonatal survival [2] hence, [30] cannot be left undetected. Modern day NICU patient physiological monitoring equipment from electrocardiograms to devices that measure temperature and blood pressure to blood oxygen sensors and much more are designed to capture the signals pertaining to the varying vital parameters and to generate vast array of high frequency, high volume, multidimensional data streams in real-time. Nevertheless these devices provide a mechanism to output these data streams at higher speeds as well. These electronic records display physiological data streams captured directly from the bedside medical monitoring devices at a high frequency enabling the visualisation of data in a temporal, graphical format. For example, the Philips Component Monitoring System accepts multiple sensor modules each can produce multiple data streams both in numeric and wave-form. Hence, in numeric form, one reading every 1024 milliseconds (ms) and in wave-form, four data values in the wave data stream arrive every 32ms, and fast wave, with 16 values for every 32ms [3].

However, the cognitive capacity of a neonatologist or in fact any human does not allow processing data of that calibre.

2.1.2 Limitation to the knowledge discovery in healthcare

Even though these physiological data streams may exhibit early indicators of potentially life threatening conditions such as neonatal sepsis and these points of interest in the data streams precede the point of diagnosis through existing clinical practice [1], the requirement for harnessing and using the valuable medical information embedded in them still seems to remain somewhat overlooked. Zhang emphasized the importance of real-time data mining as 'our expanded system of real-time data collection and algorithm development demonstrated that patient-specific learning in real-time is a feasible approach to developing alarm algorithms for monitoring purposes in the ICU' [21]. Nonetheless, there is potential for increasing trends in knowledge discovery in clinical stream data. Currently there is a growing body of research that employs data mining and temporal abstractions (TA) to demonstrate that a given condition exhibits certain physiological stream behaviours [22], [23].

This limitation of knowledge discovery in these physiological data streams has become a broader medical challenge that can be viewed as emanating from the lack of a standardised informatics methodology to manage report, perform and document retrospective informatics research and a real-time computational analytic platform to support retrospective informatics research and prospective clinical decision making based on those results of retrospective informatics research.

While some NICUs have now replaced this manual data capture with electronic feeds to electronic health records, it is not at the same frequency that is being generated which Artemis can now capture and utilise. The data so produced can be visualised in a temporal, graphical format as well. But the challenge is that the human brain cannot process and or analyse the data as rapidly as Artemis. Harnessing these data and analysing them in real-time would deliver benefits to neonatal critical care unlike any other industry. Therefore, using patient characteristics to gain better understanding of individual patients through retrospective data mining and making use of the TAs to improve sensitivity and specificity by creating subgroups of physiological behaviours can deliver a greater potential within the purview of neonatal care [24]. This has prompted the need for an environment to detect and capture real time streaming data and analyse them to identify stream behaviours through a TA-based dynamic DM method

that can meet the needs of high-frequency, high-volume, real-time, multidimensional time series data where data elements each representing a dimension that can vary in value and characterise an item of interest.

2.1.3 Real-time data analysis for the knowledge discovery in physiological data streams

The importance of applying real-time data analysis for the knowledge discovery can be further demonstrated through the following examples.

Chen et al. [25] designed a study to identify risk factors associated with ROP and to explore the interrelationships between prominent risk factors. The central idea of this study was to determine if exposure to one risk factor alters the impact of the second or third risk factor. This study concluded that neonatal sepsis, higher oxygen exposure and a low GA are not only independently associated with a significantly increased risk for the development of ROP, but also interact beyond additive and even multiplicative patterns. This important conclusion was based on detailed analysis of oxygen exposure at 28 days that was assessed through high-frequency ventilation data collected in real-time. This approach provided a more accurate reflection of oxygen exposure data than intermittent spot recordings. However, in that study they did not use a structured approach to write up the medical devices used, data collection frequency and experimental design for example to ensure that a complete set of relevant information was reported.

Holmström et al. [26] aimed to evaluate possible neonatal risk factors for ROP in a population-based cohort of preterm, very low birth weight, infants. The study identified 15 possible risk factors, including continuous positive airway pressure (CPAP) readings, which were collected in real-time. A completed univariate analysis showed an association between ROP and respiratory distress syndrome (RDS), bronchopulmonary dysplasia (BPD), septicaemia, intraventricular haemorrhage (IVH), and the use of ventilator as well as continuous positive airway pressure. Targeting oxygen saturation levels has been constantly challenging despite extensive investigations of the vulnerable preterm population [27]. Hagadorn et al. [8] have shown that SpO_2 remains in the target range only 50% of the time. The fluctuations in SpO_2 occur in such a manner that a timely intervention cannot be performed because the episodes of deviation have reverted back to the target range before a health care professional determines the potential

significance of the event [8]. As a result the provision of complex techniques to assess trends and patterns in all the data rather than the traditional approach of down sampling is needed.

2.2 Data Mining through the CRISP-TDM

In terms of delivering medical care, it is important to study the evolution of one or more parameters of physiological data over time than instant parameter values. Due to the play of synergism, a separately monitored stream, such as blood pressure or heart rate, falls within normal parameters, but the combination of several streams with some specific value ranges can turn out to be a predictor of impending illness. Because Streams is performing analytics on moving data instead of just looking for out-of-bound values, it not only has the potential to save lives, but it also helps to drive down the cost of healthcare [12]. Because physiological data are multidimensional in nature [4], for the purpose of studying this evolution of multiple parameters of physiological data over time, Data mining (DM) is employed. Furthermore, DM is important in multi-dimensional analysis because traditional statistical methods are not well suited to evaluating the probability of coincidental patterns in high-dimensional data sets.

2.2.1CRISP-DM

The CRoss Industry Standard Process for Data Mining (CRISP-DM) methodology was developed in 1996 with the goal of being industry, tool and application neutral. Repeated references to the methodology by analysts have established it as the de facto standard for data mining [4]. This provides an overarching structured approach to carrying out projects that involve the discovery of new knowledge from trends and patterns in data and breaks down data mining projects into six different phases: business understanding, data understanding, data preparation, modelling, evaluation and deployment [4], [31].

Each phase of a project, the tasks involved and the relationships between them provide together an overview of the data mining lifecycle shown in Figure 2.1.

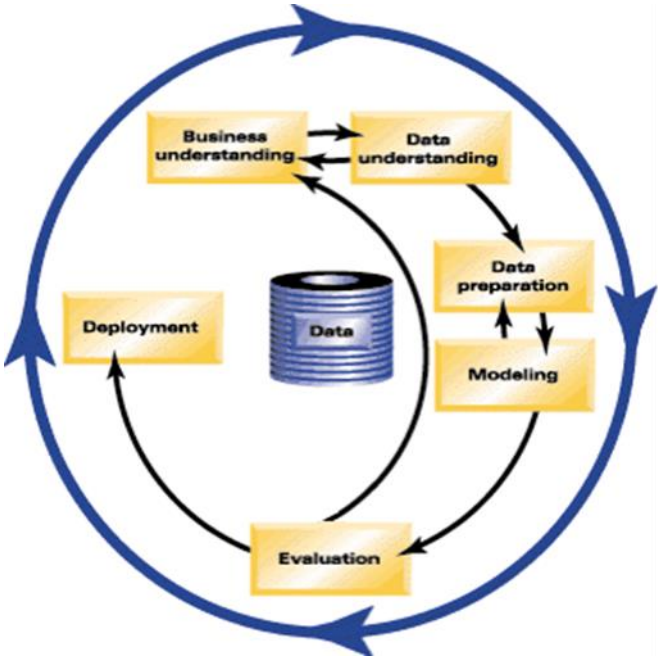

Figure 2.1: CRISP-Data Mining process model [last accessed May 2, 2017] ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/14.2/en/CRISP_DM.pdf]

2.2.2 Temporal Data Mining

When it comes to neonatal care, the multiple streams of physiological data that a premature infant generates are not isolated from each other. Therefore the need arises for TA approaches that cross correlate data from multiple streams for a combined TA model as a complex indicator of condition onset. This in turn has led to growing recognition of the need for specialised techniques for mining temporal data using TDM systems [3].

2.2.3 CRISP-TDM

An investigation based on a NICU case study on the CRoss Industry Standard Process for Data Mining (CRISP-DM) approach for modeling Intelligent Data Analysis (IDA) was carried out by Catley et al. [2], [4]. Their aim was to study CRISP-DM approach for modeling IDA that perform temporal data mining of time-series data with a view to assess the applicability of CRISP-DM to provide a methodology to model the intricacies of multidimensional time series data.

Since the needs of time-series data drive an evolution in DM, and the methodology for CRISP-DM does not include provision for integrated Temporal Data Mining (TDM) of multi-dimensional time series data, Catley et al. [2], [4] concluded that the methodology for CRISP

DM does not include provision to compare the systems from either a clinical or IDA-based perspective. As the number of such integrated TDM systems i.e. Temporal Data Mining (TDM) of multidimensional time-series data which is increasingly forming the data source for sophisticated DM systems that combine temporal abstraction (TA) frequently as a pre-processing step to DM continues to grow, this limitation would become even more significant [2], [4]. Since this consequently had an impact on the knowledge discovery in streaming data, Catley et al. [2], [4], [22] proposed extensions to the CRISP-DM methodology to support temporal data mining for reporting the results of TDM-based clinical investigations on multidimensional time-series data.

The extensions presented in Figure 2.2 below addressed the limitations of CRISP 1.0 model that had an impact on the knowledge discovery aspect of the research.

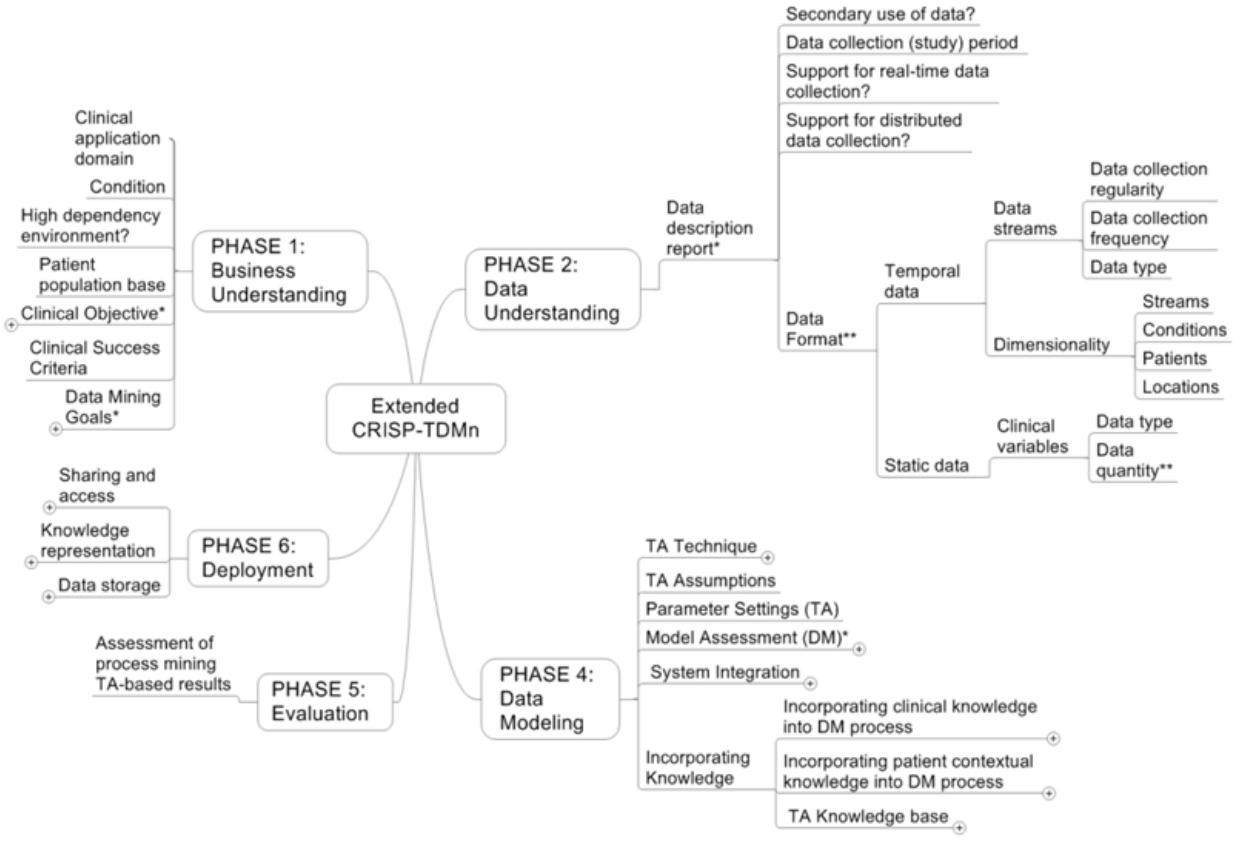

Figure 2.2: Illustration of the details on the extensions added to the phases of CRISP-DM Model to make CRISP- TDM [22]

These extensions are further detailed in the Table 2.1.

Phases to which extensions were applied	extended to	Remarks
Business understanding (Phase 1)	incorporate the clinical objectives	These extensions support clinical investigation and exploratory data mining on clinically relevant, population-based information
Data understanding (Phase 2)	incorporate temporal aspects of multidimensional data of the clinical investigation	This is performed through the clinical study
Data modelling (Phase 4)	apply relevant details in IDA knowledge management and IDA system integration; provides support for integrating techniques, such as DM and TA, in a closed or open loop workflow (Figure 2.3).	In contrast CRISP 1.0 (CRISP-DM) is devised to apply several DM techniques to reach at one which offers the best results.
Data evaluation (Phase 5)	enable the assessing of process mining of TA based results	The study was evaluated by assessing the process mining of TA based results when correlated with the other clinical data.
Deployment (Phase 6)	incorporate the provision of a methodology for describing system storage. In this step the current study temporal abstractions are integrated within the Artemis platform. These data together with those of future research will be employed as a new tool for real-time patient monitoring that aid clinical decision support.	Hence it is of utmost importance that a methodology which is relating to TA based systems requires provision for system storage where both the raw data and TAs are archived for future use.

Table 2.1 Details of the extensions added to the phases of CRISP-DM Model to make CRISP- TDM [22]

2.2.4 Designing of clinical algorithm through the extensions of CRISP-TDM

Designing a clinical algorithm through clinical Decision Support system integration can be discussed in relation to the IDA system integration branch of the extended CRISP-TDM methodology [22]. It is a two-step process, as illustrated in figure 2.3.

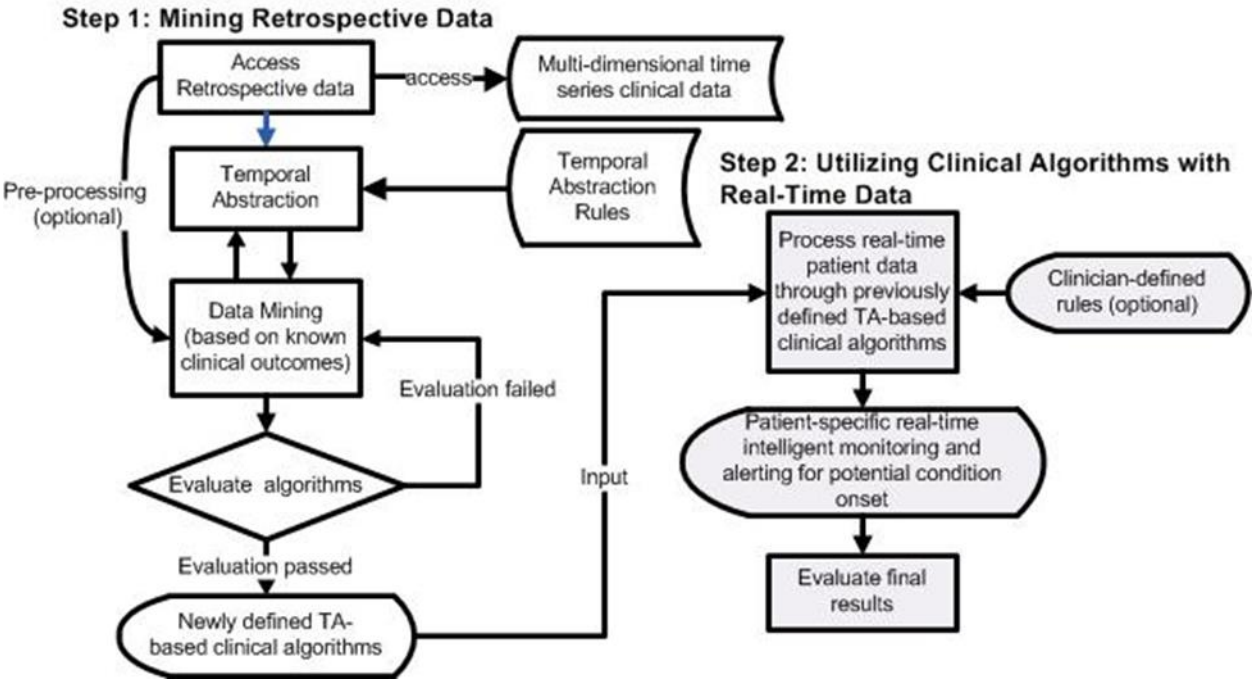

Figure 2.3: Organization of clinical algorithm formation through Clinical Decision Support system integration [22]

In Step 1, clinical algorithms are developed based on retrospective clinical data streams with known outcomes. TA is performed on the data, transforming the data streams into a time-stamped, interval-based representation by extracting the most relevant data features, such as states, trends and temporal relationships. The temporally abstracted data is then processed using predictive DM approaches with the goal of identifying early indicators for complications of interest. This represents a tightly coupled closed loop system, with an optional circular approach to TA and DM workflow modeling. The results are evaluated based on clinician-defined requirements and the results of current gold standards for clinical outcome prediction. DM parameters are continually tuned in an effort to meet or exceed both the clinician requirements and current gold standards. If this process is successful, the end result will be a newly defined clinical algorithm for the early prediction of disease based on retrospective clinical data [22].

Step 2 involves utilising the clinical algorithms developed in Step 1 to provide patient context-specific intelligent monitoring and alerting on real-time patient data streams. Co-mining, which enables the integration of DM results with expert knowledge, is also possible, in that additional input may be received in the form of clinician-defined rules [22].

In terms of workflow ordering, the data is processed using TA techniques defined in Step 1, and feed through a TDM system using the clinical algorithms developed in Step 1. In the event that the TDM process indicates the potential early onset of a condition of interest, the intelligent patient monitoring system indicates this knowledge and the results are evaluated [22].

2.2.5 Demonstration of CRISP-TDM phases within a hypothetical NICU case study

In keeping with the growing recognition for the need of specialized techniques for mining temporal data using TDM systems, Catley et al. [22] has presented an example designed to demonstrate the CRISP-TDM methodology within a hypothetical NICU case study of population-based analysis to provide patient-centric outcomes. This illustrates the framework's support for multiple streams of multiple patients with multiple potential conditions.

2.3 The importance of integrating a scientific method to support null hypothesis testing

Any data mining carried out on medical data sets will be of restricted value to clinicians and the medical field if it has not been produced using a rigorous scientific-method driven approach [28]. This approach involves a cycle of observation, followed by hypothesis generation and experimentation [29].

The experiments in the healthcare space require much more rigorous validation of a hypothesis than any other sector because of the higher risk potential to cause patient harm associated with erratic clinical decision making based on incomplete, inaccurate or erroneous clinical research conclusions. As a result, an overarching research process or investigative method on medical data must formulate and validate a null hypothesis [28], [29] by driving the research through the Data Preparation, Modelling and Evaluation phases of the CRISP-DM process model. This necessitates the construction of data mining tools and methodologies to conduct both 'exploratory' and 'confirmatory' data mining. Motivated by that, Heath and McGregor proposed the extension of CRISP-DM to support null hypothesis [9] and this extended model, known as CRISP-DM_0, enabled the use of exploratory data mining as a tool to find unknown patterns or relationships to create hypotheses. An alternative null hypothesis could then be created and tested during iterative explanatory data mining within CRISP-DM_0. Once a clear statement of the null hypothesis has been identified, factors/attributes within the dataset that are not related to the null hypothesis must be removed thereby leaving only factors/attributes of relevance to the null hypothesis.

As already described in Chapter 1, this thesis is built on the previous research work that proposed the CRoss Industry Standard Process for Data Mining, CRISP-TDM_0 methodology that merges CRISP-TDM and CRISP-DM_0 aspects as the standardised informatics methodology to manage report, perform and document this retrospective, population based, cross sectional, analytical study. The CRISP-TDM_0 methodology is integrated within the knowledge discovery component of Artemis platform.

2.4 Big Data Analytics in Intensive Care Units: challenges and applicability

The Artemis is one such platform that complies with the Big Data concept [67], [68]. High frequency physiological data streams are an untapped data resource that have the potential to significantly improve clinical decision support through online health analytics platforms that leverage high-speed physiological data together with other electronic health record data [10], [11]. The authors of 'Harness the Power of Big Data' define the Big data using four Vs; namely volume, variety, velocity and veracity [12]. Some analysts include other V-based descriptions, such as variability and visibility as well.

a) Volume

This characteristic can be considered as giving the most obvious portrayal of the Big Data concept. The peculiarity of all voluminous statistics is that their volumes grow bigger and go out of date the moment they are quoted. For example aggregations that were initially measured in petabytes (PB) are now measured in zettabytes (ZB) which is a trillion gigabytes (GB) that is equal to a billion terabytes (TB). The rapidity with which this volume characteristic propagates is such that any future estimate of voluminous statistic in a particular statistical literature goes out of date the moment a draft of the current estimate is finalized. However, the opportunities that are associated with this amount of huge data are impressive, just as impressive are the challenges associated with managing this amount of data [12].

b) Variety

This characteristic complements how a decision-making is processed by capturing both structured and unstructured data components. Ultimately the decision is made through a process of making sense out of unstructured data and combining them with structured data [12]. The medical data belong to two broader categories: structured and unstructured. Structured data are HR, RR, the presence or absence of infection and clinical information regarding the developing

retina etc. Then, within the context of this research, a unique and/or peculiar pattern of variation in the blood levels of SpO_2 is unstructured data and this phenomenon can be demonstrated through an example that relates to the tonal variation in the voice of a disappointed customer. The unique pattern in the tone will give an insight to the identity of the person in relation to the particular situation he/she is facing due to the unmet demands. When this concept is applied to the physiological data streams, for example not just the variability but the particular pattern of variability in the levels of SpO_2 that is characteristic to a particular situation, it could give us an individualised idea of the patient's clinical condition. Just as we can combine the known details e.g. name, residential address billing information etc. of the above customer, which mimics structured data and the insight that can be gleaned from the peculiar tonal variation to get an overall identity of the person, the structured and unstructured physiological data can be combined to create a very personalised data model. Currently, the Artemis platform is being equipped with the capacity to deal with unstructured data.

c. Velocity

This is the rate at which data arrives at an enterprise, is processed and well understood. Whether it relates to the process of understanding of the health of a patient, the condition of a loan portfolio or operations and functionalities of a traffic system, being able to more swiftly understand and respond to data signals is advantageous [12]. The IBM Big Data platform is equipped with IBM InfoSphere Streams (Streams) as a component. This is a real-time streaming data analytics engine that provides fast, flexible, and scalable processing of continuous streams of time-sequenced data packets. Accordingly Streams enables advanced analysis across diverse data types with very high messaging data rates and very low latency. Furthermore, the IBM Big Data platform can seamlessly move the analytic artifacts that are being harvested at rest and apply that insight to the streaming data thereby making the analytic model adaptive and smarter as it applies learned intelligence to the data [12].

d. Veracity

Veracity is the Big Data characteristic that refers to the data quality and this deals with data issues that occur within the confines and in relation to the huge amount of data. The Big Data platform is equipped with tools that handle Big Data's veracity and they are capable of discarding noise and transform the data into trustworthy insights [12].

2.4.1 Big Data challenge

Medical data, defined as all data that relate to the health of a patient belong to two categories; continuous and non-continuous. This classification is based on the recording continuity of data. Continuous data is described as data recordings continued without intermission or with a recurring regularly after minute interruptions i.e. pulse rate or temperature. Some of the examples for non-continuous data are clinical data, medical images or laboratory data.

The patient monitoring devices usually measure different physiological parameters at different frequencies. Some physiological parameters need higher frequency sampling than others in order to achieve a high accuracy i.e. temperature measurement requires low sampling rates as compared to heart rate measurements because the temperature changes are comparatively slow. The continuous data is the source of the problem addressed by the real-time big data analytics systems. At times the measurements obtained by the NICU patient physiological monitors are lost and the need to retrieve previous data would become impossible if physicians do not unequivocally program these devises to save measurements. Nevertheless, a monitoring device can only store limited data in the internal memory. Hence this massive data generated per unit time demands a high computing and storage capacity to be supported by the real-time big data analytic systems. Furthermore, the data generated by these patient monitoring devises are a Big Data problem with near real-time restrictions for processing medical algorithms designed to predict certain neonatal ailments.

The near real-time implementation of algorithms intended for early prediction of neonatal clinical conditions is associated with a demand in computational capacity because these algorithms are designed to process huge quantities of physiological data and need to be executed for each individual patient. Nevertheless, a high secondary storage facility is a prerequisite to sustain all historical physiological data of all patients. Extraction of knowledge in the form of behavior patterns from historical data will allow predicting certain neonatal pathologies and let neonatologists and health informatics specialists later use this knowledge in the development of new predictive algorithms. Moreover, historical data is mandatory to validate algorithms [67], [68].

2.4.2 Existing Solutions

A clinical decision support system (CDSS) is a system that incorporates health information technology within the context of health care and is devised to assist health professionals with tasks pertaining to clinical decision making. Most CDSS systems comply with non-continuous data i.e. clinical data, diagnostic test data and relevant information that aid detection of diseases. The Artemis platform is a real-time, big data analytics system and may be the only CDS system that has been proposed to be compliant with continuous data streams [67], [68].

2.5 Real-time analytics for retrospective clinical research and prospective clinical decision support

2.5.1 The Artemis Project

This is a framework to capture, process, and store in real-time, multiple time-series data at higher frequencies for multiple patients [4], [30].

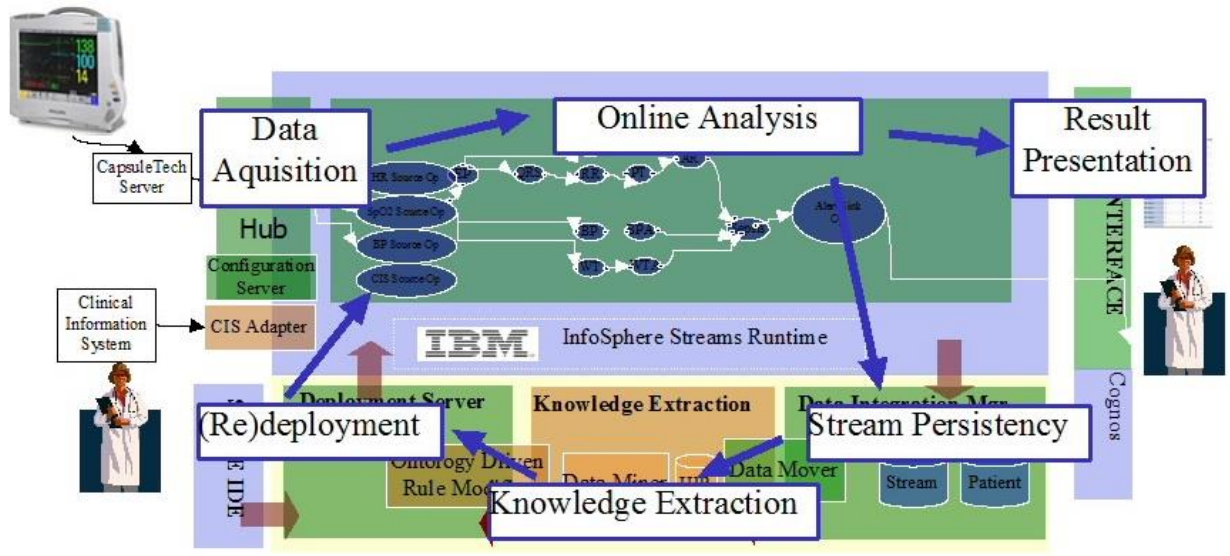

Figure 2.4: Organization of the Artemis platform architecture [30]

Research investment by the Canada Research Chair program together with an IBM First-of-a-Kind award has resulted in the creation of the Artemis platform. It was deployed first in the NICU at The Hospital for Sick Children, Toronto. As depicted in the above figure 2.4, the platform comprises five components. The data acquisition and online analysis components of the platform are located at The Hospital for Sick Children, Toronto while the data persistency, knowledge extraction, and redeployment components reside in the Health Informatics

Laboratory at the University of Ontario Institute of Technology (UOIT), Oshawa, Ontario, Canada. The Data Acquisition component enables the provision of real-time synchronous medical device data i.e. physiological data streams, from multiple physical monitoring devices and asynchronous Clinical Information Management System (CIMS) data i.e. clinical information and information from the laboratory information system.

A set of hardware and software elements, provided by Capsule Tech Inc., provide the interface between the wide range of patient monitoring devices and the Artemis platform. A DataCaptor terminal unit gets the input from these patient physiological monitors and converts the RS-232 output to an Internet Protocol (IP) stream through Ethernet. Then these data are forwarded to a Capsule DataCaptor Interface Server that can support up to 500 monitoring machines connected at a given time. This server filters the received data to extract data streams necessary for a particular study, converts the data streams to the Artemis format and sends them to the Medical Data Hub which creates concurrent data streams for the streaming system. Additionally, the clinical information system (CIS) adapter gets patient information from CIMS. Then data from all these sources are streamed to the Artemis clinical rules. These clinical rules are medical algorithms that can either be defined by a neonatologist or proposed through knowledge extracted from medical data streams, laboratory investigation results and direct patient observations.

The data are then subjected for analysis within the Online Analysis component that processes and stores the raw data and derived data from multiple infants at the rate they are generated. This function is carried out through IBM's InfoSphere Streams a novel streaming middleware system that processes data in real-time applying the appropriate algorithm. The Online Analysis component delivers scalable processing of multiple streams. The programming of the stream computing system is done with the Stream Processing Application Declarative Engine (SPADE) language which is a high-level declarative language for high performance distributed stream processing systems. SPADE language, which is the programming language for IBM's Info Sphere Streams middleware permits flexible composition of parallel and distributed dataflow graphs.

The original data and generated analytics are then stored in the data persistency component. The Data Integration Manager (DIM) has a set of SPADE operators and these operators interact with

an open database system in order to store data in appropriate databases as and when the DIM receives data. The data so stored in the database of data persistency component, and are formatted for data mining, are being periodically moved to the repository in the knowledge extraction component.

The knowledge extraction component of Artemis operates a data stream-data mining framework that aids multidimensional, real-time data analysis. This is carried out through the utilization of the newly proposed temporal data mining approach that supports discovery of condition onset behaviours in physiological data streams and associated clinical data. The operation of this component results in clinical rules and data explicitly personalised to enable physicians and neonatologists to perform clinical research in a range of neonatal clinical conditions.

New knowledge, once tested through rigorous clinical research techniques, is transferred to use within the Online Analysis. This is performed through the Redeployment component which translates the newly acquired knowledge into a SPADE representation and feeds new clinically validated algorithms to the online analysis component [4], [30], [32], [67], [68].

2.5.2 Artemis Cloud

The organisation of Artemis platform is such that the positioning of Data Acquisition and Online Analysis components are in the same site as well as the medical devices that provide data. This represents a negative price tag model for or remote, rural, small hospitals. Artemis Cloud is a cloud computing platform and by enabling the function of computing resources outside the hospital, reduces technical costs, staff and required infrastructure to bring about appositive economic impact. This allows patients to access better care without the necessity to refer them to an urban health centre.

Artemis Cloud is built on the support of software through the Software-as-a-Service and the facilitation of storage of persistent data through the Data-as-a-Service model. In the Artemis Cloud, the interface with Artemis is performed through the web services that allow interaction of hospitals with the system and this allows hospitals to gain access to persistent data stored in the cloud. Currently the Artemis Cloud has been deployed on secured servers located at UOIT. The data received by the Cloud from hospitals are in the HL7 format which is the international standard for transfer of clinical and administrative data between software applications [67], [68].

2.6 Clinical Research Use of Artemis

2.6.1 Neonatal apnoea and spells monitoring in neonatal intensive care

In most clinical settings that deal with neonatal care watching for apnoeic events or spells and recording them is usually carried out by a nurse and this is one such example for manual documentation of significant medical events. This recording is usually done as and when a spell occurs with simple device alarms based on one stream of data and when it is perceived to be of significance. However, these manual recordings of episodes of apnoea are limited in its degree of detail and a recent workshop on apnoea of prematurity by the National Institutes of Child Health and Human Development too concluded its significance by highlighting the fact that the available technology is not routinely used to document real-time events associated with apnoea [32], [33].

Catley et al. [32] have presented a framework to support multi-dimensional analysis of real-time physiological data streams and clinical data focusing on the detection of episodes of central apnoea. This framework has been instantiated within the Artemis platform. To accurately identify and classify preterm spells Thommandram et al. [33] developed precise, robust computational algorithms for the Artemis platform which enables capture and temporal analysis in real-time of multi-stream, high-fidelity physiological data of multi-patient origin at the speed generated by the bedside medical devices. For this venture, heart rate and blood oxygen saturation levels were sampled at a rate of 1 Hz. The impedance respiratory wave data were sampled at a rate of 62.5 Hz and electrocardiographic data were sampled at a rate of 1000 Hz. This was performed to provide a means of effective clinical decision support.

Medical literature defines five main types of neonatal spell: central, mixed, and obstructive apnea; isolated bradycardia; and isolated desaturation. Pugh et al. [34] completed a detailed review of the literature as the initial step in performing preclinical verification of algorithms. They aimed to determine the unique temporal pattern of change in heart rate, breathing pattern, and blood oxygen saturation associated with each of the type of spell. They assessed the accuracy of the algorithms before clinical testing. This preclinical verification study against manual annotations was performed through evaluating the functional behaviour of each algorithm against 24 hours of clinically annotated physiological data. Since this demonstrated

high levels of correlation with each of the algorithms in the system, Pugh et al. concluded that computational algorithms based on physiological rules could be used to detect neonatal spells.

2.6.2 Retrospective analysis for a novel Premature Infant Pain Profile using Artemis

Proper alleviation of pain in premature neonates is of utmost importance that poorly managed pain has been shown to result in an increased susceptibility for late onset neonatal sepsis (LONS) and intraventricular haemorrhage together with increased incidence in mortality [35]. What is currently practiced is the assignment of a PIPP score, a numeric value calculated at different time intervals, as ordered by neonatologist. Therefore, continuously monitoring the neonatal physiological parameters and focusing on the changes that occur in them in response to nociceptive stimuli together with automating the documentation of pain scores in contrast to the previous ad hoc documentation is a more reliable way of assessing neonatal pain. Motivated by that background Naik et al. [35] designed a novel Premature Infant Pain Profile called novel Artemis Premature Infant Pain Profile (APIPP). This focused on four physiological variables heart rate (HR), mean arterial pressure (MAP), respiratory rate (RR), and blood oxygen saturation level (SpO_2). The researchers intended creating rules to deploy the novel scale into a model that can predict real-time nociceptive events and deploying this novel APIPP within the Artemis platform using GA to identify nociceptive stimuli in real-time [35].

As an extension to the above research, Naik et al. [36] used retrospective data from Artemis database to compare the performance between the Premature Infant Pain Profile (PIPP) used in the NICU of The Hospital for Sick Children, Toronto and the novel APIPP through Artemis platform. The absolute and the relative values of the multiple physiological data streams; HR, MAP, RR, and SpO_2 were used to correlate differing GAs to validate a reliable APIPP score. Through the retrospective equation of the novel APIPP with the PIPP to evaluate the effectiveness and validity with regards to increased specificity and sensitivity of the novel scale being proposed, it should be possible to better identify nociceptive stimuli and therefore improve the use of drugs and non-pharmacologic interventions for pain management. The addition of physiological factors that has not been used before in pain scales holds the potential to reflect a more accurate quantification of pain in premature infants. The novel scale built upon the PIPP scale by using additional physiological factors will then be deployed prospectively within the Artemis platform to predict nociceptive events before any potential pain provoking procedures.

Also, a more ambitious algorithm will be deployed in Artemis, which should then enable not only detection but also quantification of pain in neonates as mild, moderate, or severe. This algorithm is programmed in IBM Info sphere Streams using the SPL language and applies retrospectively on the Artemis patient data to output a pain score per second which will be aggregated over an hour [35], [36].

2.6.3 Using high fidelity SpO_2 data streams to provide a better representation of immature retina's exposure to oxygen

As already described in chapter 1, Cirelli et al. [13] used Artemis, to ascertain if the collection of second by second data provides a better representation of retinal exposure to oxygen than an infrequent, intermittent spot reading. This study revealed that the second by second monitoring of SpO_2 levels showing significant differences in the pattern of variation in contrast to conventional hourly monitoring [13]. They also showed that Artemis is capable of producing more accurate representation of the immature retina's exposure to oxygen from the higher frequency data recordings, as these include all the episodic events in the activity of the hour, which provides a better understanding of oxygen fluctuation ranges that affect the physiological status of the neonate. Based on the fact that this second by second analysis of blood oxygen saturation as a surrogate marker for the oxygen level of the retinal vasculature at any given second is far superior, the research presented in this thesis employs the Artemis platform together with CRISP-TDM_0 methodology and this will be discussed in detail in the following chapters.

2.7 Conclusions and implications on this research

At the beginning of this chapter an insight into the streams deployment in healthcare is given by highlighting the advantage of real-time data analysis for the knowledge discovery in physiological data streams and existing limitations to the knowledge discovery in healthcare. Since this thesis is built on previous research work that proposed the CRISP-TDM_0 methodology that merges CRISP-TDM and CRISP-DM_0 aspects and this CRISP-TDM_0 is used in this research as the standardised informatics methodology, this sets the background for the formulation of first research question; 'Can the extensions of CRISP-TDM methodology enable the complete management, reporting and completion of a clinical research study involving physiological

streaming data that incorporates null hypothesis testing?' The answer to this research question is given in chapter 7.

Further, an account on the Artemis platform and some of the other research projects based on Artemis i.e. apnoea monitoring in neonatal intensive care, developing computational algorithms for the Artemis platform to help managing preterm spells and devising a novel premature infant pain profile using Artemis is given. Through these examples an insight in to the Artemis framework is given highlighting its capability to capture, process, and store in real-time, multiple time-series data at higher frequencies for multiple patients and to perform retrospective clinical research and prospective clinical decision support [4], [30]. However, what related mostly to this thesis was the previous study on high fidelity SpO_2 data streams that provided a better representation of immature retina's exposure to oxygen. Even though this study by Cirelli et al. [13] did analyse neonatal blood SpO_2 every second in real time using the Artemis platform [2], [3], [4], and that provided a better representation of immature retina's exposure to potentially damaging oxygen levels, a standardised approach to support the research continuum in this area is currently unavailable. Depending on that background, this section of the thesis highlights the importance of utilising the CRISP-TDM_0 methodology as the standardised informatics methodology to manage report, perform and document this study. This sets the background to discuss in chapter 4 how this research was conducted through the phases of CRISP data mining process model. Further this sets the scene for formulating second and third research questions; 'is there a correlation between the blood oxygen saturation levels of immature newborn infants, born at less than 32 weeks gestational age, during the first two weeks of life and the severity of ROP occurring before 35 completed weeks postmenstrual age?', and 'is it possible to calculate the probability at postnatal ages three days, seven days, ten and fourteen days for the severity of ROP occurring before 35 completed weeks postmenstrual age?' The answers to these research questions will be given at the conclusion chapter of this thesis.

Chapter 3: Clinical Literature Review

This section of the thesis presents the clinical aspect of retinopathy of prematurity (ROP). Together with the socioeconomic consequences of ROP, the effects of blood oxygen as a contributor to the causation of ROP and various studies that were conducted to elucidate the effects of blood oxygen are discussed in detail. The clinical staging of ROP and screening methods of ROP are discussed in detail as well.

3.1 Clinical issues associated with prematurity and low birth weight

Preterm birth is associated with greater difficulty with transition from childhood to adolescence and adulthood [37]. Premature survivors are at increased risk for impaired neurodevelopmental outcome that would manifest as cognitive abnormalities, fine or gross motor delay, cerebral palsy, vision and hearing losses [38]. Persistence of multiple abnormal neurologic signs in the first 12 to 18 months is ominous and emergence of vision impairment, seizures, feeding issues, delay of head growth would lead to neurodevelopmental and/or psychiatric disabilities [38]. These could manifest in adolescents and young adults as higher susceptibility to cerebral palsy, intellectual disability, cognitive impairment, learning disability, and executive dysfunction. Specific psychological and behavioral problems including attention deficit hyperactivity syndrome, general anxiety, and depression are other consequences as well [38]. In addition they are more amenable to social and emotional difficulties that demand financial support and consume societal resources [37]. Other neonatal complications of prematurity and low birth weight such as bronchopulmonary dysplasia, necrotising enterocolitis, retinopathy of prematurity, intraventricular haemorrhage, poor growth, presence of congenital anomalies are also associated with an increased risk of poor outcome [38] that among those affected, fewer complete higher education, find and keep meaningful employment, and live independently from their parents [37]. Among the numerous sequelae of prematurity and low birth weight, retinopathy of prematurity has the most adverse repercussions [37], [38].

3.2 Retinopathy of prematurity - extent of the problem in a global perspective

Retinopathy of prematurity (ROP) is the most common cause of visual impairment and blindness in children and it could persist through adulthood to become a lifelong sequel [39]. However, it is one of the most common causes of preventable blindness in preterm neonates. Unfortunately, ROP is emerging as a third epidemic in middle-income countries due to the increasing survival

of preterm neonates, insufficient monitoring of oxygen saturation and lack of ROP screening guidelines in most neonatal units [40], [41]. In the developing world too, increased survival of low birth weight newborns have led to an increase in ROP. When the statistical data of the developed world is considered, in Brooklyn, NY (2010), all stages of ROP occurred in 81% of preterm newborns born less than 1000 grams birth weight, with 5.7% progressing to stage 3 plus requiring therapy [40], [41].

3.3 Pathology

3.3.1 What is Retinopathy of Prematurity?

ROP is a neovascular disease in preterm newborns characterised by an early vaso-obliterative phase leading to ocular hypoxia and aberrant regulation of ocular vascular growth factors and a later vasoproliferative phase of pathologic neovascularisation leading to cicatricle (scar) formation, retinal detachment and blindness [15].

3.3.2 Pathogenesis: the evolution of ROP in two phases

The first phase begins with a delayed retinal vascular growth after birth and partial regression of existing vessels followed by a second phase of hypoxaemia-induced pathological vessel growth and proliferation [15].

In the human retinal blood vessel development is initiated during the fourth month of gestation. Retinal vascularisation which originates at the center of optic disc, progresses radially outwards towards the ora serrata, which is the serrated junction between the retina and the ciliary body and reaches the retinal periphery just before birth [15]. As a result, infants born prematurely have incompletely vascularised retinae with a peripheral avascular zone. Hence, the gestational age at birth determines the area of the avascular zone [15].

Normally in utero the blood is only about 70% saturated compared to 100% in full term infants in room air. The normal PaO_2 in utero is 30 mm Hg but a normal infant breathing room air will have a PaO_2 of 60-100 mm Hg. As the infant matures the non-vascularised retina becomes increasingly metabolically active and in the absence of an adequate vascular system, this leads to tissue hypoxaemia [15].

The first phase of ROP occurs from birth to postmenstrual age 30-32 weeks. In premature infants vascular growth that would normally occur in utero slows or ceases and is accompanied by

regression of developed retinal vessels. The relative hyperoxaemia of the extra-uterine environment as well as supplemental oxygen given to premature infants are thought to be responsible for this process [15]. The second phase of ROP is characterised by hypoxia induced retinal neovascularisation which begins between 32-34 weeks postmenstrual age. New vessels form at the junction between the vascularised and avascular zone of the retina [15]. Over time, this pathological growth of vessels produces a fibrous scar extending from the retina to the vitreous gel and lens. Retraction of this scar tissue can separate the retina from the retinal pigment epithelium, resulting in a retinal detachment and likely blindness [15].

3.4 Risk factors for ROP

Several risk factors for ROP have been studied over the past 50 years. Among them, general immaturity and prolonged oxygen therapy have been consistently related to disease onset [42]. First observed in the nursery and then supported by animal studies, ROP is said to be associated with excessive oxygen use [15], [40]. In addition, very low birth weight (VLBW) and GA make important independent contributions to the overall incidence of severe ROP [43]. Chen et al. [25], aimed at identifying and exploring the interrelationships between prominent risk factors for ROP, and found that oxygen associated ROP risk was more prominent among infants of 23-25 weeks gestational age while infection associated ROP risk was higher among infants born at 28-29 weeks. The study further suggested that neonatal sepsis, oxygen exposure and low GA are not only independently associated with a significantly increased risk of ROP but also interact beyond additive and even multiplicative patterns [25]. Recent reports suggest that elevated glucose may also play a role [40]. Nevertheless, asphyxia and apnoea are found to be high risk factors for ROP as well [42], [44]. It is further understood that the progression of ROP is multifactorial and may be associated with other risk factors such as multiple gestation, intracranial haemorrhage, anaemia, prolonged mechanical ventilation and multiple transfusions [42].

Even though premature infants have lower vitamin E levels in serum the clinical studies on the role of vitamin E in the causation of ROP gave conflicting results hence the efficacy of vitamin E prophylaxis remains inconclusive [25], [40].

3.5 Oxygen and Retinopathy of Prematurity

3.5.1 Historical background

In 1604 the Polish alchemist, Michal Sedziwoj, (Sendivogious) discovered oxygen. He did not publish his findings and the discovery did not receive much fanfare. Therefore, oxygen had to be 'rediscovered' as it were by Karl Wilhelm Scheele in 1772, followed by Priestly in 1774. In the late 1700s Antoine Lavoisier renamed this compound oxygène. In 1780, another Frenchman, Francois Chaussier used oxygen for the revival of 'near-dead' infants but it was not until the 1930s that oxygen began to be used liberally in neonates. Doctor J.H. Hess described an oxygen unit for premature and very young infants in the 1930s. This was followed by another publication by C.C. Chapple in 1938 on an incubator for infants. In this unit, 100% oxygen gas was delivered at four liters per minute, which gave a FiO_2 of approximately 0.46 in the isolate [41]. Oxygen rapidly came into use in the care of preterm infants in the 1940s. Blood oxygen levels could not be measured and oxygen was given on the basis of clinical assessment. High concentrations of oxygen were commonly prescribed for many days in pursuit of increased survival [45].

3.5.2 Oxygen as a risk factor for ROP

The adverse consequences of this high oxygen delivery only became clear several years later. Between 1942 and 1945 Dr. Terry collected 117 examples of a novel type of blindness in infants born premature characterised by a thick fibrotic membrane in the retrolental space, the etiology of which was not yet elucidated [41]. Retrolental fibroplasia (RLF), now known as ROP, was first described in 1942 and a link with oxygen therapy was suggested in 1951 [45]. It was not until 1951 that Dr. Campbell first suspected a role for supplemental oxygen in the etiology of this new blindness entitled RLF. Trials showed that prescribing high concentrations of oxygen for many weeks increased the risk of RLF when compared with a more restricted approach where the prescribed concentration was lower and the duration of treatment was limited [45].

3.5.3 The possible optimal blood oxygen saturation level for premature retina

Supplemental oxygen used to treat respiratory distress syndrome of prematurity and bronchopulmonary dysplasia has long been associated with ROP. Following Terry's description in 1942 of RLF [46] to confirm the role of supplemental oxygen in the etiology of this new disease, in 1952, Dr. Patz performed an NIH-funded randomized trial subjecting infants to either

high, 100% oxygen, or restricted oxygen, supplemental oxygen only when showing evidence of desaturation. Of those exposed to high oxygen, 61% developed RFL vs.16% in the restricted oxygen group suggesting that oxygen was playing a role in the disease causation [41], [47].

The chief purpose of one of the largest studies, 'the cooperative study' [45] involving a cohort of 586 premature neonates was to determine by means of a controlled clinical trial whether the incidence of RLF was positively associated with the length of exposure of the premature infant to oxygen and whether restricting oxygen to that amount considered necessary to prevent anoxia would influence the survival rate. This study was designed to deliver widely different amounts of oxygen to the two groups of premature infants while all other factors associated with the care of the infants were kept as similar as possible. They found that the incidence in ROP in the routine group was higher and statistically significant more than that in the curtailed-oxygen group ($p < 0.01$) while the mortality rates remained without differing significantly between the groups [48]. After the results of this study were published in 1956, restricted oxygen therapy was widely adopted and in the period that followed, neonatal mortality from respiratory distress syndrome increased. It has since been estimated that as many as 16 additional deaths may have resulted from oxygen restriction for each case of blindness from ROP that was prevented [45]. This newly found awareness of the role of prolonged supplemental oxygen use in the aetiology of RLF led practitioners to restrict oxygen exposure and unfortunately face an increased incidence of cerebral palsy.

Dennery in her publication discusses a multicenter study of 1080 infants weighing < 1800 gm, where a median duration of supplemental oxygen exposure of less than two days was associated with a 17.4% incidence of cerebral palsy and an 8.7% incidence of ROP whereas a median supplemental oxygen exposure of greater than 10 days was associated with a 5.8% incidence of spastic diplegia but a 21.7 % incidence of ROP. This early study is one such example that illustrates the delicate balance between too little and too much supplemental oxygen in premature infants and suggests that oxygen is a necessary evil that must be used in moderation in the care of premature neonates [41].

3.5.4 Duration of administering supplemental oxygen and occurrence of ROP

During 1976-77, a second national cooperative study was designed by Kinsey et al. [49] to study the relation between PaO_2 and RLF prospectively. Blood gas studies were performed on 589 of

these infants, 66 of whom had a diagnosis of RLF; in 27 of these 66, some grade of mostly non blinding cicatricial disease developed. The frequency of RLF was highest among infants of lowest birth weight. RLF is associated with concentration of oxygen administered in the lightest birth weight group, but the strongest association, aside from birth weight, was with time in oxygen. The occurrence of RLF was found to be unrelated to PaO_2, as determined by the limited information available from intermittent sampling. Publications on several earlier studies, namely those by Patz et al. [47] and Lanman et al. [50] have shown that premature infants treated with oxygen regimens of longer duration and/or higher concentrations of inspired oxygen have significantly more ROP than control.

3.5.5 Oxygen use in neonatal acute and convalescent neonatal care and the onset of ROP

a. Oxygen use in acute neonatal care

In a survey of 144 NICUs, Anderson et al. found the rate of retinal ablation surgery was increased among infants cared for in NICUs that used higher maximum target levels of oxygen saturation, as compared with infants in NICUs that used lower target levels. Results of this survey show a 'gradient of risk' towards less retinal ablation surgery when the maximum SpO_2 is < 98% in the first two weeks following birth ($p < 0.05$) [51]. Sola et al. concluded that adequate SpO_2 monitoring equipment (Masimo Signal Extraction Technology®) accompanied by implementation of a strict clinical practice of oxygen administration and monitoring which avoids high SpO_2 and minimise wide fluctuations from the time of birth and during the first few weeks of life was associated with a significant decrease in severe ROP. Furthermore, they found that these settings are accompanied by no increase in mortality and/or developmental anomalies and a decrease in BPD [52].

The SUPPORT trial used a two by two factorial design to compare blood oxygen saturation targets of 85-89% vs. 91-95% among 1316 infants greater than 24 completed weeks GA and less than 28 weeks GA. The primary outcome measure was a composite of severe ROP, defined as the occurrence of threshold retinopathy, the need for surgery or both. When the results were analysed, they found that the combined outcome of severe retinopathy or death was not different between the two groups and that death prior to discharge occurred more frequently in the lower saturation group. In contrast, severe retinopathy among survivors was reduced significantly in the lower saturation group. The incidence of bronchopulmonary dysplasia as well was reduced in

the lower saturation group. Subsequent analyses demonstrated that across both randomised arms saturations in excess of 96% were strongly associated with increased risk of ROP [7]. These observational studies together suggest that lower oxygen saturation targets might improve visual outcomes, yet the feasibility of maintaining infants in these targets and the safety of them have to be rigorously tested.

b. Oxygen use in convalescent care

The STOP-ROP trial [53] was carried out to determine the efficacy and safety of supplemental therapeutic oxygen in neonates with diagnosed pre threshold ROP. The probability of progression to threshold ROP and the need for peripheral retinal ablation were considered to measure the efficacy and safety of supplemental therapeutic oxygen. The study randomised premature neonates with confirmed pre-threshold ROP in at least one eye and a median pulse oximetry saturation of < 94% to a conventional oxygen arm with pulse oximetry targeted at 89-94% saturation or a supplemental arm with pulse oximetry saturation targeted at 96-99%. Participants were assessed through weekly eye examinations performed by examiners masked to treatment assignments until each study eye had reached either an adverse end point; defined as reaching threshold criteria for laser or cryotherapy in at least one study eye or a favourable endpoint defined as regression of ROP into zone III for at least two consecutive weeks or occurrence of full retinal vascularisation. A reassessment of eyes, pulmonary status, growth, and interim illnesses was carried out at three months.

Even though this study shows a small favourable effect in preventing the supplemented group (96-99%) from reaching threshold (41% vs 48%), it was not statistically significant (CI = 0.52-1.01). Analysis of the results further reveals that supplemental oxygen did not significantly reduce the number of infants requiring peripheral ablative surgery. Nevertheless, supplemental oxygen increased the risk of adverse pulmonary events including pneumonia and/or exacerbations of chronic lung disease and the need for oxygen, diuretics, and hospitalisation at three months postmenstrual age [45], [54].

In order to determine whether retaining a higher oxygen saturation level in extremely preterm infants with a long-term dependence on supplemental oxygen magnifies growth and neurodevelopmental outcomes measured at a corrected age of 12 months (primary aim), and to determine whether a higher oxygen-saturation levels had favourable or adversarial physical or

psychosocial effects on infants or parents (the secondary aims), a trial on Benefits of Oxygen Saturation Targeting (BOOST) was launched [17].

This was a multicenter, double-blind, randomised, controlled trial involving 358 premature neonates born at a GA < 30 weeks and who continued to receive supplemental oxygen at 32 weeks postmenstrual age (PMA). A random assignment was done to a target functional oxygen-saturation range of either 91-94% (standard-saturation group) or 95-98% (high-saturation group) as measured through pulse oximeters with devised algorithms to assess functional oxygen saturation. The researchers employed a central telephone randomisation method to camouflage the treatment-group assignments. Likewise, a dynamic balancing method was used to stratify the randomization. Neonates, after undergoing randomisation, were assigned an explicit study oximeter, which after the calculation of oxygen saturation level through the normal means, was modified to display a value 2% higher than the actual saturation in neonates in the standard-saturation group and 2% lower than the actual saturation in neonates in the high-saturation group. The care givers and parents who remained unaware of the actual ranges being targeted, were then advised to maintain the neonate's oxygen saturation range at 93-96 %. This double-blind targeting of the assigned saturation range of 93-96% was maintained for the duration of the infant's oxygen therapy and compliance with this was assessed by downloading of each infant's oxygen saturation data two times per week.

Results when analysed with regards to the primary outcome i.e. bringing about a positive impact on growth or development in preterm neonates, were not in favour of a beneficial effect imparted by targeting a functional oxygen-saturation range of 95-98 % compared to 91-94%. Additionally a long-term dependence on supplemental oxygen was noted. Likewise results with regards to the analysis of secondary outcome showed no significant differences between the groups in the incidence in ROP of any stage or in the frequency of the need for surgical treatments.

The results of both the STOP-ROP trial and BOOST trial suggest that the possibility for ophthalmic intervention may be reduced when a higher oxygen-saturation range is targeted in extremely preterm infants with more severe ROP. Since the neonates in the BOOST study were randomly assigned to treatment through different oxygen target ranges at 32 weeks PMA, before threshold ROP usually develops, the observed effect of different target oxygen-saturation ranges

on ROP becomes more interesting. However, the differences between the treatment groups were not significant at the $P<0.05$ level and call for validation in larger studies [17], [45], [54].

The progression of pre-threshold ROP to threshold ROP may be occurring through the effect of different array of physiological parameters that were not taken in to consideration in the STOP-ROP trial. Even though the results of the STOP-ROP trial seems to have weakly validated the research finding that the progression to threshold ROP occurs due to accumulation of angiogenic factors in the avascular, hypoxaemic retina, one reason for the STOP-ROP trial's failure to achieve its objective may be the data simply have not supported the introduction of the novel use of oxygen [69]. This is one such example where the importance of studying the evolution of one or more parameters of physiological data over time than relying on instant parameter values can be highlighted.

Because the potential for new medical knowledge discovery embedded in the patterns of streams is greater than those of conventional spot recordings and premature infants generate multiple streams of physiological data at a given time and these streams are not isolated from each other, recent research has supported the need for TA approaches that cross correlate data from multiple streams for a combined TA model as a complex indicator of condition onset [22].

The BOOST trial has employed pulse oximetry to assess the target oxygen-saturation ranges and devised algorithms to assess functional oxygen saturation. The compliance with this was assessed by downloading of each infant's oxygen saturation data two times per week. Even though the STOP-ROP and BOOST trials have described a technically complex and labor-intensive endeavors, both these trials have not used a standardised approach to capture and analyse real-time blood SpO_2 data, which results in incomplete information about the study being reported leading to difficulties in repeatability as a result. Hence applying TA approaches and employing a standardised approach to capture and analyse real-time blood SpO_2 data could have probably given a better out come by achieving their objective goals.

3.5.6 Varying low or high oxygen saturation targets during the two phases of Retinopathy of Prematurity

Chen et al. analysed the above studies and other non-randomised studies published prior to 2009 in a systematic review and found that the ideal oxygen saturation target might vary by the period

of use. According to them a low oxygen saturation level, 90-96% prior to 32 weeks PMA, reduced the risk of ROP while high oxygen saturation level 94-99% after 32 weeks PMA, remained favourable for a better outcome [54], [55], [56].

The beneficial effect of low and high oxygen saturation on severe ROP can be explained by the two successive phases in the pathogenesis of ROP. The first vaso-obliterative phase of ROP is triggered by hyperoxaemia between birth and 30-32 weeks PMA. Supplemental oxygen suppresses vascular endothelial growth factor (VEGF), which results in the cessation of normal vessel growth and regression of existing vessels. The second proliferative phase begins around 32-34 weeks gestation and is associated with an increased VEGF expression in the retina caused by relative hypoxaemia which results in pathologic neovascularization. Thus, supplemental oxygen might be used therapeutically at appropriate time points to down regulate VEGF expression and to limit the neovascular complications of ROP [54], [55], [56]. Chen et al. speculated that low oxygen saturation in the first phase combined with high oxygen in the second phase of ROP pathogenesis might achieve greater protection than administering low oxygen alone. However, this should be carried out only with a sufficiently powered randomised controlled trial on optimal oxygen delivery in the early and late stages of ROP and this need to be accompanied by a long-term visual, pulmonary, and neurodevelopment follow-up [55].

3.5.7 Graded oxygen saturation SpO_2 targets as a preventive measure of Retinopathy of Prematurity

Retinal metabolism and role of growth factors are different during vaso-obliterative and vasoproliferative phases of ROP. Prior to 2003, Arora et al. [56] targeted oxygen saturation at 83-89% range in preterm infants born < 29 weeks GA. Since 2003 they implemented graded oxygen saturation targets in the following way. In infants born < 29 weeks GA, saturation target was 83-89% until 33 weeks PMA. The saturation target was 90-94% between 33-36 weeks PMA and more than 94% for over 36 weeks PMA while breathing oxygen. They collected maternal and neonatal data on infants born < 29 weeks as two cohorts: Group 1, before graded oxygen saturation targets and Group 2, after the application of graded oxygen saturation targets. The two cohorts were matched for birth weight.

The analysis of results showed that ROP and laser surgery rates significantly decreased between the two cohorts. There was no significant difference in mortality rate between the two groups

(8.4 vs. 7.1%, p = 0.32). Depending on this significant decrease in ROP and laser surgery rates and the trend towards lower mortality after implementing graded oxygen saturation targets, it was concluded that use of graded oxygen saturation targets based on PMA, rather than low or high saturation targets during the two different phases of ROP may be a better approach. However, they have suggested the need for further studies to test this hypothesis [56].

This study had not measured the SpO_2 data streams as well and highlights how the limitation in the area of knowledge discovery in SpO_2 streams has become a medical problem that emanates from the existing but still overlooked requirement for harnessing and using valuable medical information embedded in these SpO_2 streams. In the domain of neonatal care, this study too is an example that highlights the importance on focusing on the evolution of one or more parameters of physiological data over time than relying on instant parameter values because the potential for new medical knowledge discovery embedded in the patterns of streams is greater than those of conventional spot recordings [22]. Within the context of the above study, analyzing the temporally abstracted behaviors of high fidelity blood SpO_2 streams and confirming their relationship to the occurrence of ROP would have given a better in sight to the effect of graded SpO_2 saturations to the onset of ROP.

3.5.8 Role of hypoxaemia as a causative factor of ROP

Too little oxygen also can have deleterious effects on the premature retina. Hypoxaemic damage in the developing retina has been characterised by retinal ganglion cell (RGC) death, swelling of Müller cells, changes in the retinal pigment epithelium and increased vascular leakage. Inflammation has been suggested as a major factor involved in the pathogenesis of hypoxaemia-induced retinopathy. It was suggested that hypoxaemia may initiate inflammation by direct activation of amoeboid microglial cells (AMCs), hence, may play a pivotal role in the pathophysiological mechanism of hypoxaemic damage to the RGCs in the neonatal retina through increased production of pro inflammatory cytokines. Increased production of TNF-α and IL-1β by the activated AMCs was accompanied by an up regulated expression of TNF-R1 and IL-R1 on the RGCs suggesting that binding of the cytokines to their respective receptors would be one of the major factors involved in RGC death [57].

3.5.9 Arterial oxygen fluctuation and Retinopathy of Prematurity in Very Low Birth Weight Infants

York et al. suggested that hyperoxaemic and hypoxaemic episodes, duration of oxygen therapy, oxygen fluctuation etc, are not exclusive of one another. Longer periods on oxygen therapy would provide more opportunities for damaging fluctuations, supporting the correlation between duration of therapy and ROP. Given the multifactorial nature of ROP, the dissection of arterial oxygen fluctuation from other strongly associated variables is not a trivial exercise [43].

York et al. launched a retrospective study of 231 infants weighing < 1500 g at birth to determine the influence of arterial oxygen fluctuation on the development of threshold ROP. Fluctuation in partial pressure of dissolved arterial oxygen (PaO_2) was expressed as coefficient of variation (CoV) for each infant. They investigated the relationship between CoV at three intervals and the risk of developing threshold ROP. The odds ratio (OR) of developing threshold ROP versus pre-threshold ROP was associated with an increase in the CoV during the oxygen therapy. Hence they concluded that VLBW infants experiencing fluctuation of arterial oxygen tension are at higher risk of threshold ROP [43].

Targeting oxygen saturation levels has been a constant challenge as it has been found that SpO_2 remained in the target range only 50% of the time and the fluctuations that occur are such that a timely intervention cannot be performed since occurring episodes of deviations have reverted back to normal before a health care professional discriminates the potential significance of the event [7], [8]. Cirelli J. et al. [13] studied neonatal blood SpO_2 every second and analysed them in real-time using the Artemis platform. The data of their study provided a better representation of immature retina's exposure to potentially damaging oxygen levels. The possibility of capturing the potentially unnoticeable deviations of blood oxygen levels is more when compared to conventional methods. Hence incorporation of a medical informatics based standardised approach such as CRISP-TDM_0 approach within Artemis, to correlate temporally abstracted behaviours of hourly blood SpO_2 data to the onset of ROP would have increased the credibility of the research study of York et al.

3.6 Classification of Retinopathy of Prematurity

Current assessment of ROP involves clinical grading by expert examination of the location and pathological stage of retinal changes. The International Classification of ROP originally

published in 1984, was expanded in 1987, and revisited in 2005 due to advances in knowledge of the disease and improved technology. This classification defines ROP by its location which depends on the zones involved; extent denoted by the number of clock hours involved and severity by staging. Further, it defines plus and pre-plus disease. Three zones are centered on the optic nerve and Zone I is defined as a diameter twice the distance between the fovea and the center of the optic nerve. Clinically this is approximately the area of the retina seen through a 28D lens when the view is centered on the optic nerve. Zone II is a circle that extends from the nasal ora serrata toward the temporal ora serrata. Zone III is a crescent encompassing the temporal area of the retina to the temporal ora serrata that is not included in zone II [58]. The extent of the disease is defined in clock hours; for example, three o'clock indicates the nasal aspect of the right eye and the temporal aspect of the left eye. Care is made to note consecutive and total clock hours involved when monitoring disease progression. ROP is described by individual clock hours however, staging for the entire eye is based on the most severe presentation of ROP [58].

Stages of retinopathy of prematurity

Stage 1 Demarcation line separates avascular from vascularized retina
Stage 2 Ridge arising in region of demarcation line
Stage 3 Extra retinal fibro-vascular proliferation or neovascularization
Stage 4 Partial retinal detachment
Stage 5 Total retinal detachment

The five stages of ROP are illustrated in figures 3.1 and 3.2.

Some authors mention stage 0 ROP which is not defined within the International Classification of Retinopathy of Prematurity. It is defined as the presence of an area of retina that has yet to acquire vasculature. This entity could be easy to miss at first as a 'normal' finding in all premature infants [60], [61].

Pre-plus disease denotes a state of more vascular tortuosity than normal but insufficient for diagnosis of Plus disease in which vascular dilatation and tortuosity can be seen to have involved at least two quadrants of the eye [58], [59], [60].

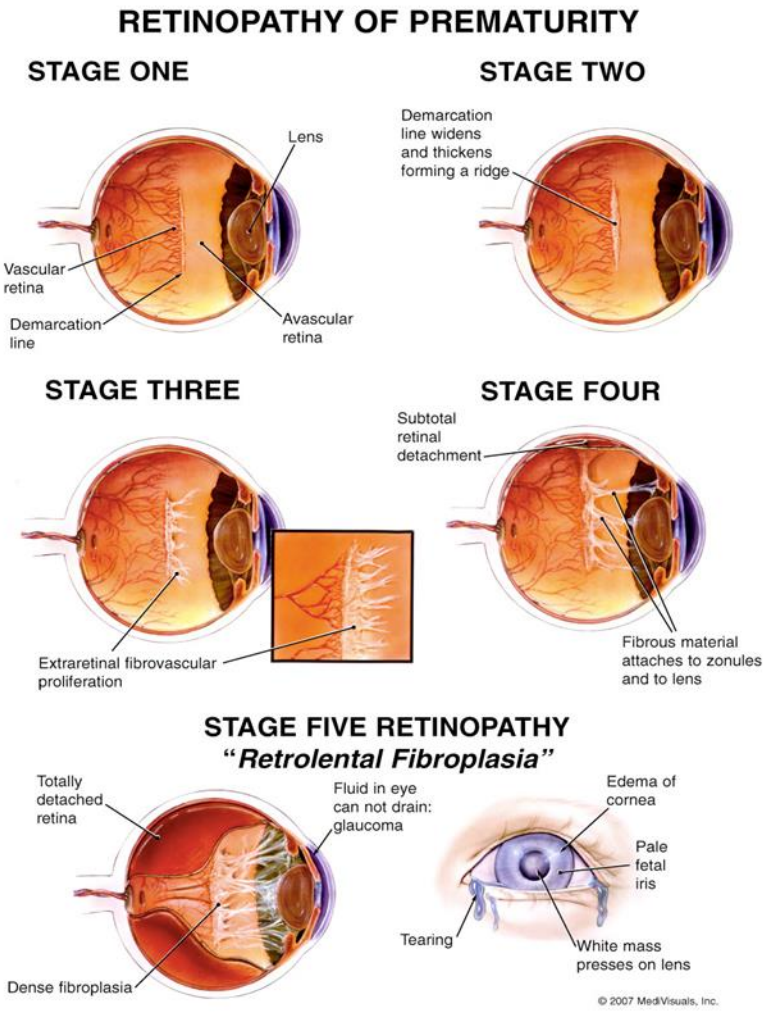

Figure 3.1: Schematic illustration of abnormalities associated with the stages of ROP (last accessed May 2, 2017)

https://www.google.ca/search?q=rop&source=lnms&tbm=isch&sa=X&sqi=2&ved=0ahUKEwidx8SSnsjTAhVJ9WMKHXNrC3sQ_AUIBigB&biw=1517&bih=708#imgrc=zL7bHjrH1-P7hM:)

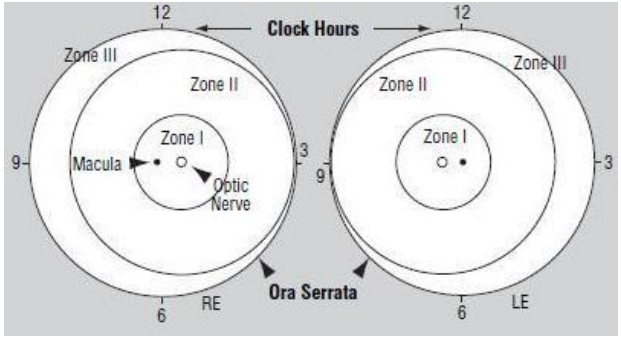

Figure 3.2: Coronal view of retinae depicting the zone borders and clock hours used to describe the location and extent of ROP [59]

3.7 Diagnosis of Retinopathy of Prematurity

3.7.1 Whom to screen

In 1998, the Canadian Paediatric Society (CPS) recommended screening infants born at 30 weeks GA or less, or with a birth weight of 1500 g or less. The birth weight criterion was included because of uncertainty in determining GA. The same criteria were recommended by the American Association of Paediatrics (AAP) in 2006. The 2008 United Kingdom guideline development group reviewed 23 articles involving 10,481 screened babies and found only one infant requiring treatment who was more than 30 weeks GA and weighing more than 1250 g at birth and recommended screening all infants with GA up to 30 6/7 weeks or with birth weights of less than 1251 g. However, more recent publications have described the Canadian experience, and documented that only one infant requiring therapy for ROP fell outside the GA of 30 6/7 weeks or less and 1250 g or less. This provided additional support for using the criteria of GA of 30 6/7 weeks or less or birth weight of 1250 g or less [59], [60], [61].

Screening of infants with a birth weights between 1251 g and 2000 g would be appropriate if the neonatologist believes the baby to be at high risk, as suggested by the presence of severe and unstable respiratory disease, hypotension requiring inotropes and prolonged ventilator or oxygen therapy, because of the severity and complexity of the neonatal clinical course [59], [60], [61].

3.7.2 When to screen

Data from two large clinical trials, the Multicenter Trial of Cryotherapy (CRYO-ROP) study and the Light ROP (LIGHT-ROP) study have been taken in to consideration as a means of providing evidence-informed criteria to define the appropriate ages and retinal ophthalmoscopic signs that determine when to commence and conclude acute phase ROP screening. The conclusions have been made after examining eyes sequentially in 4099 infants with BW < 1251 g (CRYO-ROP study) and in 361 infants with BW< 1251 g and GA < 31 weeks (LIGHT-ROP study). The recommendations have been developed to ensure that eyes that had a high probability of requiring treatment would be identified in a timely manner, while at the same time minimizing the number of examinations for infants at low risk. Timing of the initial examination is based on both PMA and chronological age (CA) [61].

	Age in weeks at initial examination	
Gestational age at birth (weeks)	Postmenstrual age (weeks)	Chronological age (weeks)
22	31	9
23	31	8
24	31	7
25	31	6
26	31	5
27	31	4
28	32	4
29	33	4
30	34	4
31	35	4
32	36	4

Table 3.1: Timing of the Initial eye examination designed to detect at least 99% of serious retinopathy of maturity [61]

According to the studies, the initial eye examination should be conducted by 31 weeks post menstrual age or 4 weeks chronologic age, whichever is the later [61].

3.7.3 Scheduling follow-up examinations

One week or less follow-up	One to two week follow-up	Two week follow-up	Two to three week follow-up
Stage 1 or 2 ROP in zone I	Immature vascularisation (stage 0) in zone I	Stage 1 ROP in zone II	Stage 1 or 2 ROP in zone III
Stage 3 ROP in zone II	Stage 2 ROP in zone II	Regressing ROP in zone II	Regressing ROP in zone III
	Regressing ROP in zone I		

Table 3.2: American Academy of Pediatrics' suggested follow-up examination schedule [61]

3.7.4 Duration of Retinopathy of Prematurity screening

ROP screening can be discontinued when any of the three signs: infant's attainment of 45 weeks postmenstrual age without the development of pre-threshold ROP or worse; progression of retinal vascularisation into zone III without previous zone II ROP; and full vascularisation is present, indicating that the risk of visual loss from ROP is minimal or passed [61].

3.8 Conclusions and implications on this research

The forgoing literature review has further highlighted that research has found that ROP is positively correlated with a higher blood oxygen saturation level, variability, fluctuation and duration of exposure etc. Some of the potentially important studies carried out in the area have given inconclusive results suggesting that there is potential for new medical knowledge discovery embedded in the patterns of the blood SpO_2 streams that would yield valuable results than those in conventional intermittent spot recordings. However, these studies have not used a standardised approach to capture and analyse real-time blood SpO_2 data, which has resulted in incomplete information about the study being reported leading to difficulties in repeatability as a result. Researchers have tried to solve these issues and found their approaches showing certain limitations with regards to sensitivity and specificity of the methodology they used even though most of data mining methods and TAs they used were capable of demonstrating conditions exhibiting certain physiological stream behaviors. Cirelli J. et al. [13] studied neonatal blood SpO_2 every second and analysed them in real-time using the Artemis platform. Even though this data provided a better representation of immature retina's exposure to potentially damaging oxygen levels, a standardised approach to support the research continuum in this area is currently unavailable.

Therefore this study approach has been designed to address that issue by using patient characteristics to gain better individual patient understanding through retrospective data mining and using TAs to improve sensitivity and specificity by creating subgroups of physiological behaviors. The CRISP-TDM_0 [22] approach has been proposed as the standardised methodology to manage, report and perform this retrospective clinical research. To demonstrate the CRISP-TDM_0 approach within Artemis, the temporally abstracted behaviors of hourly blood SpO_2 data have been correlated to the onset of ROP.

Chapter 4: Research Construction and Methodology

The standardised methodology of this research is built on the previous research work that proposed the CRoss Industry Standard Process for Data Mining, CRISP-TDM_0 methodology. Since this research utilized CRISP-TDM_0 as the standardised informatics methodology to manage report, perform and document this study, this section of the thesis through the same standardised CRISP-TDM_0 methodology, concentrates on setting the back ground to find answers to the following research questions;

1. Can the extensions of CRISP-TDM methodology enable the complete management, reporting and completion of a clinical research study involving physiological streaming data that incorporates null hypothesis testing?
2. Is there a correlation between the blood oxygen saturation levels of immature newborn infants, born at less than 32 weeks gestational age, during the first two weeks of life and the severity of ROP occurring before 35 completed weeks postmenstrual age?
3. Is it possible to calculate the probability at postnatal ages three days, seven days, ten and fourteen days for the severity of ROP occurring before 35 completed weeks postmenstrual age?

Setting the background to find answers to the above questions is performed by capturing data that identify relationships between physiological data streams, blood SpO_2 and specific clinical conditions , ROP and by subjecting time-series data to temporal data mining through concurrent multi-patient and multi-stream temporal analysis in real-time. The CRISP-$TDM_{0,}$ the standardised informatics methodology in this research is integrated within the knowledge discovery component of Artemis platform that serves as a multidimensional environment that enables knowledge discovery of new behaviours in physiological streams prior to the onset of a clinical event through a mechanism that watches for these in real-time.

4.1 The research construction

The construction of this research can be viewed encompassing three key components namely:

a) Acquisition of data

The patient monitoring devices in Neonatal Intensive Care Units are increasingly generating vast volumes of physiological data and these data stream so produced are as dynamic as the environment within which they are produced.

b) Temporal data mining

This has been viewed as originating from the intersection of three different disciplines namely statistics, computer science and database management, which include the areas of artificial intelligence and machine learning [62].

c) Multidimensional mechanism to watch for the new behaviours in physiological data streams in real-time

This mechanism is the Cross Industry Standard Process for Data Mining approach for modeling Intelligent Data Analysis (IDA) that performs temporal data mining CRISP-TDM_0, which is integrated in to the Artemis platform [2], [3], [4]. In this study, the CRISP-TDM_0 which is the combined methodology of CRISP-DM_0 and CRISP-TDM is used as the standardised approach to perform retrospective clinical research using medical data streams. This is further detailed in the following sections of this chapter. Whether performing knowledge discovery using statistics, data mining, or a technique based on machine learning, the processes of business understanding, data understanding and data modelling are commonly required for all. In keeping with this requirement, this research employed the functionality of these three phases for managing and reporting the results generated through knowledge discovery. This approach has great potential to ensure consistency and completeness in reporting the results of our knowledge discovery experimental activity. The CRISP-TDM_0, as the standardised approach in this study enabled the use of exploratory data mining as a tool to find unknown patterns or relationships to create hypotheses. The speculation was that an alternative null hypothesis can then be created and tested during iterative explanatory (confirmatory) data mining activity. Once a clear statement of the null hypothesis is identified, factors and attributes within the dataset that are not related to the null hypothesis can be removed leaving only the factors and attributes of concern to the null hypothesis. This can be achieved by driving the Data Preparation, Modelling and Evaluation phases through the CRISP-TDM_0.

4.2 Instantiation of CRISP-TDM_0 within the Artemis platform

This research presents an instantiation of the CRISP-TDM_0 framework's capacity of process mining and knowledge representation which is detailed as follows. Further, the flexibility and multidimensionality of this framework within the NICU where multiple streams of physiological data being captured from multiple patients, being watched for the onset of multiple potential

conditions that can occur concurrently with patients located in multiple NICU locations is demonstrated.

4.3 Framework for Process Mining and Process Representation in neonatal care

This research utilized CRISP-TDM_0 framework's Service Based Multi-dimensional Temporal Data Mining capacity to perform process mining that enables knowledge discovery of a relationship between a new condition onset pathophysiology and physiological data stream that is generated within the monitoring devices of the SickKids' neonatal intensive care unit. These are ROP and blood oxygenation levels measured as SpO_2 streams respectively.

4.4 Demonstration of the CRISP-TDM_0 Framework within the research study at NICU

As detailed in chapter 3, many studies have shown that ROP is positively correlated with higher blood oxygen saturation level, variability, fluctuation and duration of exposure [6], [14], [15], [16]. Hence, the pattern of variability of arterial SpO_2 level was used as a marker to indicate the possible risk of ROP in the absence of a confirmed 'clinical' observation.

4.4.1 The CRISP-TDM_0 in knowledge discovery in critical care

Using the six phases of the data mining process model, as detailed in chapter 2, this section demonstrates the experimental process of this research. This experimental process, through the analyses of the obtained results, correlates the SpO_2 variability to the onset of ROP within the neonatal cohort. It further correlates the relationship of GA to the causation of ROP within this study.

Phase 1: Application of Business Understanding phase to the study

The objective of this phase is to determine and document the overall context of this clinical research study. During this phase ethics approval would normally be sought for the research and the required application and other supplementary documentations were submitted. Accordingly, Research Ethics Board approval was granted from The Hospital for Sick Children, Toronto and the University of Ontario, Institute of Technology. The following subsections detail the content required for this phase based on CRISP-TDM_0.

a) Clinical Application Domain:

A retrospective analytical/observational study of a cohort of premature neonates who were enrolled in an REB approved Artemis study ((SickKids REB#100013567 and UOIT REB#09-002).) being conducted at in the NICU of The Hospital for Sick Children, Toronto.

b) Condition:

This is 'Retinopathy of prematurity'. As the next step, a study data subset for the data mining was established. This was performed by defining the new study, 'Retinopathy of prematurity and blood oxygen saturation: confirmation of the relationship with high fidelity data'. Then the clinical condition of interest, Retinopathy of Prematurity was defined as 'a vasoproliferative disorder of the eye affecting premature neonates which progresses in two phases: the first phase that begins with delayed retinal vascular growth after birth and partial regression of existing vessels, followed by the second phase of hypoxia-induced pathological vessel growth [15].

c) Availability of a high dependency environment:

The study was carried out on new born infants admitted to the NICU of The Hospital for Sick Children, Toronto, which is a designated high dependency environment.

d) Patient population base:

This is a cohort of nine premature neonates ≤ 32 weeks gestational age. These subjects were selected on primary inclusion criteria; gestational age at birth ≤ 32 weeks; admitted to the NICU at ≤ 3 days of birth; minimum length of stay in NICU ≥ 14 days. A total of 101 neonates satisfied the primary inclusion criteria. Then secondary inclusion criteria; availability of Artemis data for ≥ 10 days, and placement in an Artemis bed space within 3 days of admission to NICU were applied to the selected 101 neonates and 70 exclusions were made as shown in table 4.1 below.

For the remaining 31 patients, the cumulative data availability period of 19 patients was less than the stipulated total duration to make use in the study. Those 19 patients were excluded as well and our further data analysis was carried out only on the remaining 12 patients. However, at the time of performing temporal abstractions on these 12 subjects, we found that complete data recordings for another two patients were lacking and in one subject the available data recording was not properly calibrated due to technical reasons. After excluding these three subjects the ultimate patient cohort comprised only of 9 patients (figure 4.1).

e) Clinical Objectives:

The objectives were designed to determine whether a new pathophysiological event for the condition ROP could be determined by analysing the designated physiological data stream, SpO_2, prior to the clinical diagnosis of ROP or evident earlier than the current clinical guideline measures, and in each instance, the condition or event will be quantified. These objectives are as follows.

Primary objective:

To demonstrate the capability of standardised informatics methodology to managing, reporting and performing retrospective clinical research using physiological data streams from the domain of neonatal intensive care.

Secondary objectives:

a) To demonstrate the instantiation of this research design within the Artemis platform.

The CRISP-TDM_0 approach within Artemis led to greater detail provided about the study rather than traditional approaches to managing, reporting and performing retrospective clinical research results using medical data streams.

b) To develop a risk assessment score for the onset of ROP based on the existing ROP stages (stage 0-5).

In the case of this study, what performed was just an assessment of the correlation between SpO_2 and other clinical markers with the onset of ROP.

c) To develop a real-time algorithm for the monitoring of blood oxygen saturation levels to assist with oxygen administration and minimise the risk for the development of ROP.

This should involve transition of the algorithm to run in real-time within the online part of Artemis and this work is currently in progress.

A detailed account as to whether these objectives were achieved will be given in chapter 7 of this thesis.

f) Data Mining Temporal Abstraction (TA) goals:

To perform a retrospective data mining of the Temporal Abstractions of arterial SpO_2 in target, above target and below target ranges.

Phase 2: Application of Data Understanding phase to the study

During this phase, rigorous analysis of data was carried out. Through the use of $STDM^n_0$, the data were gathered using $STDM^n_0$ processing agents and contained within the data management layer of the architecture. The following subsections detail the content required for this phase based on CRISP-TDM.

a) Primary or secondary use of data:

Data for this research was collected utilising the Artemis platform and the data for a cohort of hundreds of neonatal infants were primarily collected to support another ethics approved research study entitled: Real-time, multi-dimensional temporal analysis of complex high volume physiological data streams for the identification of condition onset predictors for nosocomial infection in the neonatal intensive care unit (SickKids REB#100013567 and UOIT REB#09-002). As a result it is a secondary use of data.

b) Data collection period:

The assigned data belonged to a period from August 1, 2009 through June 30, 2014 inclusive.

c) Support for real-time data collection:

While this research reports a retrospective analysis, the Artemis platform enables real-time prospective studies on data. The Artemis platform enabled real-time data collection through bedside monitoring devices and streams of raw SpO_2 were captured by the Knowledge Extraction component of the Artemis platform and contained within the data management layer of the architecture.

d) Support for distributed time data collection:

This study was developed in a single site; however, Artemis Cloud platform supports multi-centre studies.

e) Data collection regularity:

Subjects were analysed regularly for the duration of their stay in the NICU. Blood oxygen saturation physiological data is routinely acquired continuously from neonates in this NICU. However, not all bed spaces in the NICU were connected to Artemis within the data collection period. As a result regular data collection could only occur when the infants were located in a bed space where data was being sent to Artemis.

f) Data collection frequency:

For data analysis, blood oxygen saturation level data (SpO2) data were sampled by $STDM^n_0$ processing agents at a frequency of second by second when it could be collected.

g) Streams:

Second by second SpO_2 values were collected by the Massimo SET pulse oximeters or Philips MP70 bedside medical devices.

h) Conditions:

Additional clinical information were obtained retrospectively from the NICU's Clinical Information Management System (CIMS) and the Hospital's Electronic Patient Chart (EPC) and these data included date of birth, gestational age (GA), post menstrual age (PMA), birth weight (BW), severity of respiratory failure, details of any respiratory support given, haemoglobin status and details on other clinical events such as infections and blood transfusions etc.

i) Subjects:

The selection of subjects for the study was performed according to the following inclusion criteria:

Primary inclusion criteria:

Gestational age at birth ≤ 32 weeks; admitted to the NICU at ≤ 3 days of birth; minimum length of stay in NICU ≥ 14 days. Then all neonates enrolled in the Artemis study were screened for their eligibility. A number of eligible neonates were excluded because of their missing critical data elements (e.g. outcome for ROP, etc.). A total of 101 neonates satisfied the primary inclusion criteria.

Secondary inclusion criteria:

We then applied the following secondary inclusion criteria to the selected 101 neonates: availability of Artemis data for ≥ 10 days, and placement in an Artemis Bed space within 3 days of admission to NICU. Accordingly we made the 70 following exclusions as shown in table 4.1.

For the remaining 31 patients, the cumulative data availability period of 19 patients was less than the stipulated total duration to make use in the study. Further data analysis was carried out only on the remaining 12 patients after excluding those 19 patients. However, when performing TAs

on these 12 subjects, complete data recordings for another 2 patients were found to be lacking and as a consequent these 2 subjects were excluded. Then the remaining subjects were given a number from Subject 1 to subject 10. However, we discovered that the available data recording for subject number 7 was not properly calibrated due to technical reasons and this subject too was excluded from the study. Therefore the study patient cohort comprised only of 9 patients as demonstrated in the figure 4.1.

Reasons for exclusion	No Artemis data	Insufficient days of Artemis data	Delayed placement in Artemis bed space
Number excluded	26	27	17

Table 4.1: Details on the subject exclusions carried out in this study

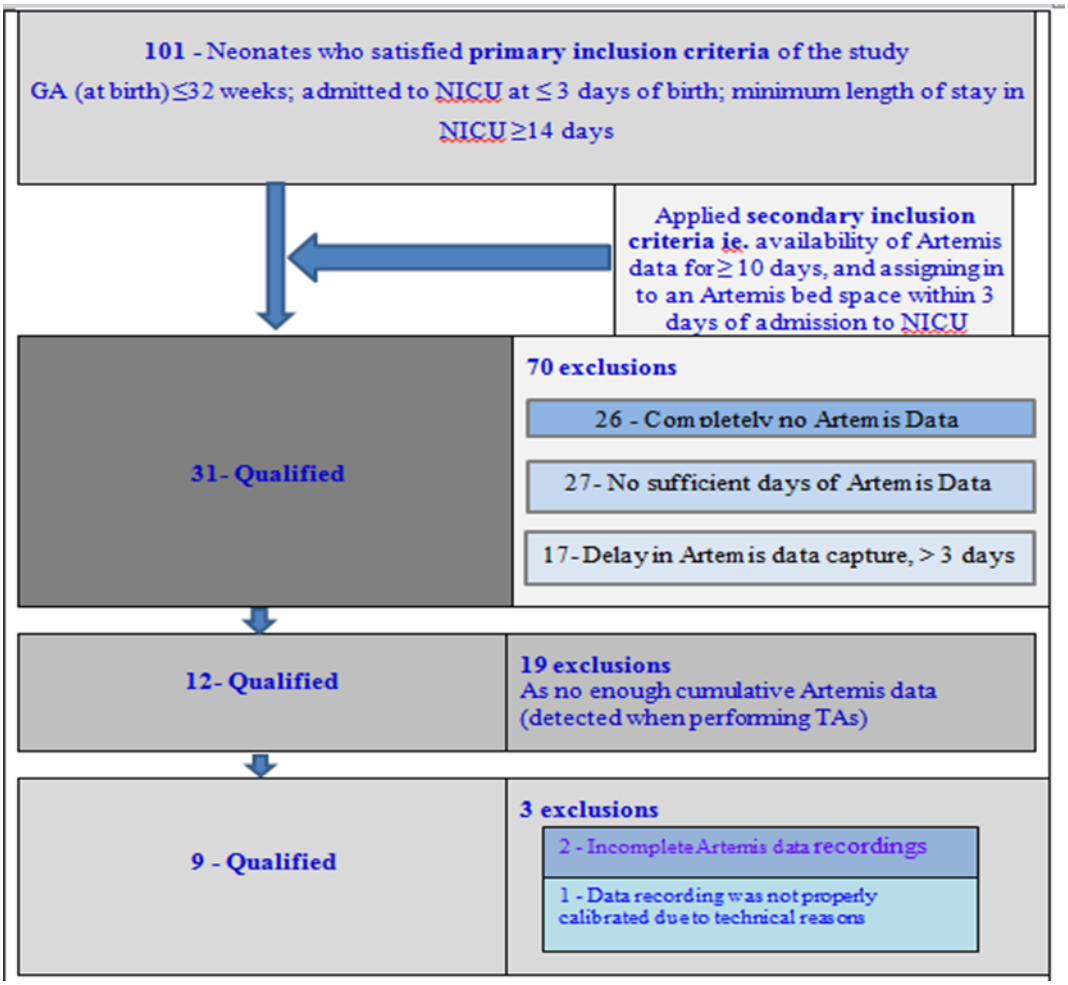

Figure 4.1: Diagram of the steps involved in the subject exclusion process of this study

Phase 3: Application of Data Preparation phase to the study

During this Data Preparation the SpO_2 data streams for the patients who qualified for this research study were selected for Data Modelling in the next phase.

Data loading: Data for this research were collected utilising the Knowledge Extraction component of Artemis platform to extract raw, second by second SpO_2 values from the Massimo SET pulse oximeters or Philips MP70 bedside medical devices. All SpO_2 from the infant's stay were selected so that identification both of the moment a neonate was detected to have developed ROP and the point of experiencing a transition in the stage of clinical severity could be supported. The $STDM^n_0$ processing agents were employed for this purpose and the gathered data were contained within the data management layer of the architecture. SpO_2 levels were loaded into $STDM^n_0$ Patient Physiological tables and clinically assessed ROP stages and other clinical data were loaded into $STDM^n_0$ Patient Clinical source tables.

Further, we utilised previously captured data contained within the NICU's Clinical Information Management System (CIMS) and the Hospital's Electronic Patient Chart (EPC). These data included the date of birth, gestational age (GA), post menstrual age, birth weight (BW), and ROP severity reached by each neonate, documented in stages. The type of therapy given, documented as none, laser vaso-obliteration, or intraocular drug therapy was recorded. If the neonate had experienced respiratory failure, its severity was also noted as well as the details of any respiratory support given (none, oxygen only, CPAP, CMV, HFV etc.). If the neonate had contracted any infections or undergone specific clinical events pertaining to prematurity, such details together with neonate's haemoglobin status were noted.

Phase 4: Application of Data Modelling phase to the study

The modeling component was performed within $STDM^n_0$ by the Processing Agents and this prepared data for data mining. The data modelling component was devised to incorporate clinical knowledge as well as patient contextual knowledge into the data mining process while enabling various data mining algorithms which accommodate temporal data to propose and analyse correlations. This enabled knowledge discovery for this research study objectives and is detailed further below.

a) The seven level breakdowns of extended CRISP-TDM$_0$ Phases into sub tasks.

Table 4a and 4b illustrate the integration of a scientific method to the research within the modelling phase of the CRISP-TDM process creating CRISP-TDM$_0$. This provides support for exploratory data mining through data mining rule-set generation and through the functionalities of Data Preparation, Modelling and Evaluation phases.

The rules generated from this layer were reviewed by a neonatologist (AJ) and health informatics specialist (CM): 'Significant Rule Sets' were selected for further analysis as shown in Table 4a.

Phases	Designing the ROP – SpO_2 model; insert blood SpO_2 parameters to the design; simulation of the ROP – SpO_2 model (optional); acquire physiological data streams; analyse results
Generic tasks	Synchronised the onset of ROP with the physiological variable, blood SpO_2
Specialised tasks	Tested the correlation between onset of ROP with the physiological variable, blood SpO_2
Data mining rule-set generation	When designing the ROP model designated limits of target ranges in blood SpO_2 were defined. The defined target range for SpO_2 was 87-93%; $SpO_2 > 93$ % was considered above target range; and $SpO_2 < 87$ % was considered below target range. The artifact was categorised when the device detected that the leads were off.
Select significant rule sets	Hourly TAs were created to down sample the SpO_2 data streams from second by second readings to an hourly summary. Readings were grouped and tallied based on the rule-defined ranges described below. The associated clinical knowledge based TA rule consists of three, distinct disjunctive clauses: 1) If $87\% < SpO_2 < 93\%$ (within target range) at any time during an hour, hourly TAs are performed and documented as total seconds within the target range. 2) If $SpO_2 > 93$ % (above target) any time during an hour, hourly TAs are performed and documented as total seconds above the target range. 3) If $SpO_2 < 87$ % (below target) any time during an hour, hourly TAs are performed and documented as total seconds of stay below target range.

Table 4.2a: First five subtasks carried out in this study

This research study formulated the null hypothesis as 'there is no correlation between ROP stages and detected blood SpO_2 levels'. The next steps were carrying out clinical investigative support for null hypotheses driven confirmatory data mining and running a statistical process to test the null hypothesis to find whether it would yield a sound outcome on the population cohort.

As depicted in Table 4.2b, the formulated null hypothesis ideally would have been carried forward onto the confirmatory data mining process on a confirmed final data set to generate a statistically sound outcome. However, the ultimate number of participants in the study was limited to 9. Therefore running a statistical process to test null hypothesis was not carried out.

Formulate null-hypothesis	There is no correlation between ROP stages and detected blood SpO_2 levels
Run statistical process to test null hypothesis	Model ROP output and physiological variable, blood SpO_2

Table 4.2b: Last two sub tasks left for a possible future study on ROP

Tables 4.2a and 4.2b collectively show the seven level breakdowns of extended CRISP-DM_0 Phases into sub tasks.

b) Temporal Abstractions:

To achieve the Data Mining TA goals, Temporal Agents within $STDM^n_0$ were utilised to perform quantitative analysis of the temporal abstractions performed on Arterial SpO_2. Then we performed population-based retrospective supervised data mining on the temporally abstracted behaviours of SpO_2 for all the neonates in the final 9 patient cohort.

As detailed in figure 4.2, temporally abstracted behaviours of SpO_2 in all three ranges (within, above and below target range) were plotted against the time for each neonate from the moment of commencing Artemis data capture in an Artemis bed space. TAs were performed on blood oxygenation data as long as data would become available and later correlated with clinical findings to identify both the moment a neonate was detected to have developed ROP and the point neonate experienced a transition in the stage of clinical severity.

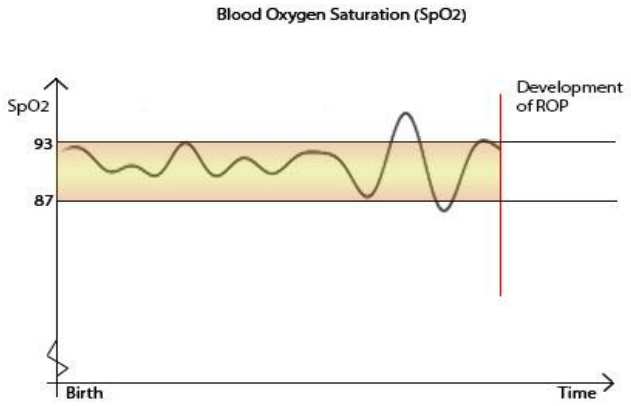

Figure 4.2: Illustration on how blood SpO_2 can fluctuate through the defined ranges; target, above target and below target

The following figure 4.3 shows SpO_2 hourly temporal abstraction profiles in one of our study participants with number of seconds in the hour over target in orange (warning), within target in green (good), below target in blue (skin tone colour for low oxygen levels) and purple for artifact.

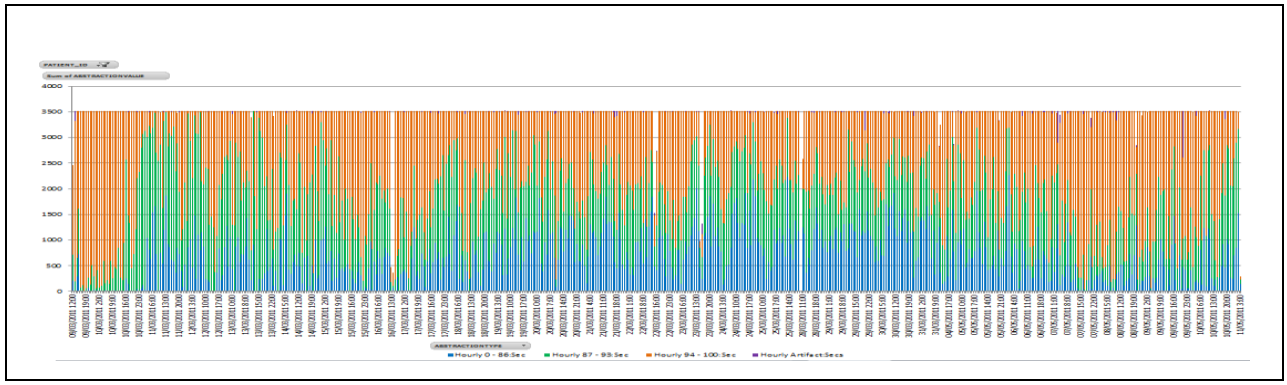

Figure 4.3: TA based process mining results displayed in a stack bar chart

For the data profile parameter analysis, summaries were created on an hourly basis for the already defined SpO_2 87-93% target range and the ranges 0-86% and 94-100% defined below and above target respectively. Artifact was categorised when the device detected that the leads were off.

Temporally abstracted behaviours of SpO_2 in all three defined ranges were plotted against the time for each neonate and correlated with clinically assessed ROP stage to identify both the

moment a neonate was detected to have developed ROP and the point neonate experienced a transition in the stage of clinical severity.

c) Model Assessment:

This denotes an assessment of a model selected as the determinant of specificity for a research. Assessment of a model selected as the determinant of specificity for a research is based on how efficacious the model is when compared to the objectives of the study as proposed in Phase 1.

d) System Integration:

This process is executed in relation to the IDA system integration branch of the extended CRISP-TDM methodology and the integration of current study temporal abstractions within the Artemis platform is performed. This translation is performed with a view to employ it in real-time patient monitoring in the event the current and future research results would allow the use of this approach in clinical decision support. This leads to the designing of clinical algorithm through Clinical Decision Support system integration as already illustrated in figures 2.2 and 2.3.

e) TA Parameter setting:

Usually in this step, the optimal value for the threshold would be determined based on iterative analysis. However, this research study considered the assigned levels based on the gold standard as detailed in chapter 3.

Phase 5: Application of Evaluation phase to the study

In this is the step evaluation of the study/clinical investigation is carried out by assessing the process mining of TA based results when correlated with the other clinical data. Accordingly in this step assessment of the population-based retrospective supervised data mining we performed on the temporally abstracted behaviours of SpO_2 was completed.

Phase 6: Application of Deployment phase to the study

This is the process of transition of the algorithm to run in real-time within the online component of Artemis. This work is currently in progress.

Designing of clinical algorithms through Clinical Decision Support system integration can be executed in relation to the IDA system integration branch of the extended CRISP-TDM methodology [22]. It is a two-step process, as discussed in chapter 2 and illustrated in figure 2.3.

In this step the current study temporal abstractions are integrated within the Artemis platform. These data together with those of future research will be employed as a new tool for real-time patient monitoring that aid clinical decision support.

The next chapter, chapter 5 will detail the aspects of informatics data analysis and its clinical correlation with ROP.

Chapter 5: Informatics Data Analysis and Clinical Correlation with ROP

5.1 Result presentation

This section of the thesis presents the results of hourly TA summaries of SpO_2 to support the research to determine whether there were possible associations between SpO_2 physiological data streams and the onset of ROP. This was performed by subjecting time-series data to temporal data mining. In keeping with the requirement for a multidimensional environment that enables knowledge discovery of new behaviours in physiological streams prior to the onset of a clinical event and a multidimensional mechanism to watch for these in real-time, this research used the Artemis platform and integrated CRISP-TDM_0 methodology as the standardised approach for the propagation of this research study.

The following figures; 5.1 to 5.9 are the visualisations of the hourly TA summaries from the modelling phase within the context of this case study. These figures present visualisations of the nine remaining patients as per the exclusions as detailed in chapter 4, inclusion criteria of the Data Understanding phase.

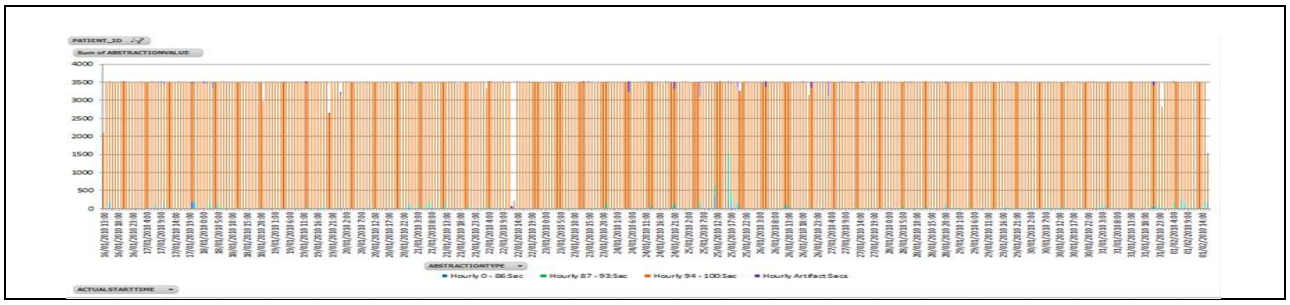

Figure 5.1: TA-based Process mining results for subject 1

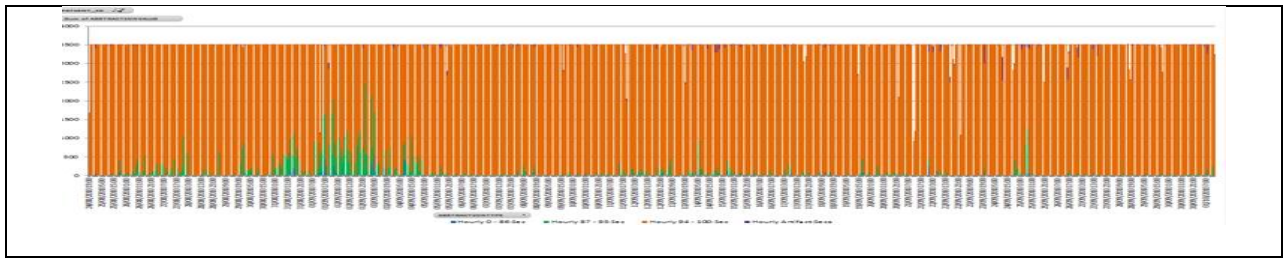

Figure 5.2: TA-based Process mining results for subject 2

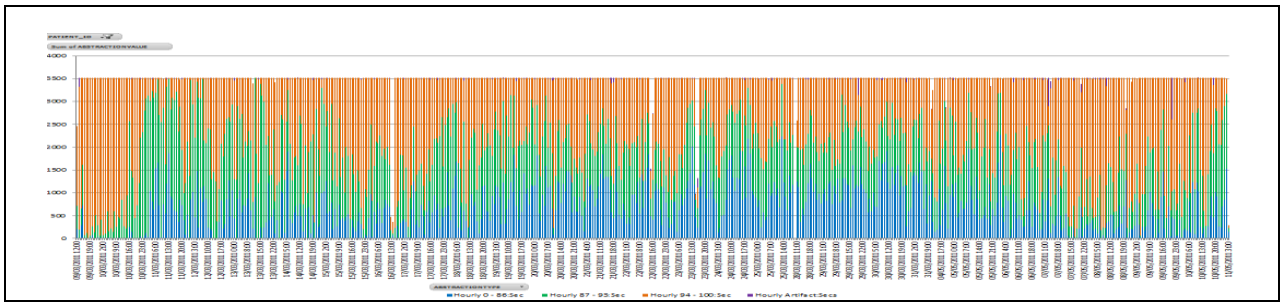

Figure 5.3: TA-based Process mining results for subject 3

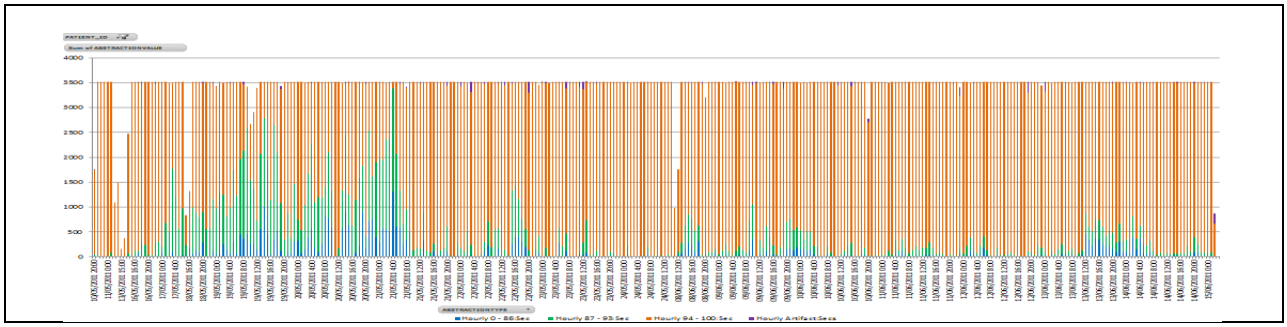

Figure 5.4: TA-based Process miningresults for subject 4

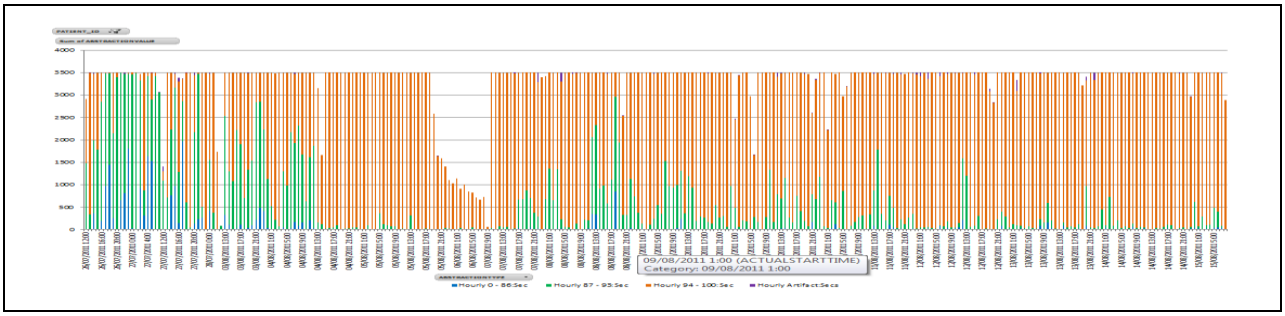

Figure 5.5: TA-based Process mining results for subject 5

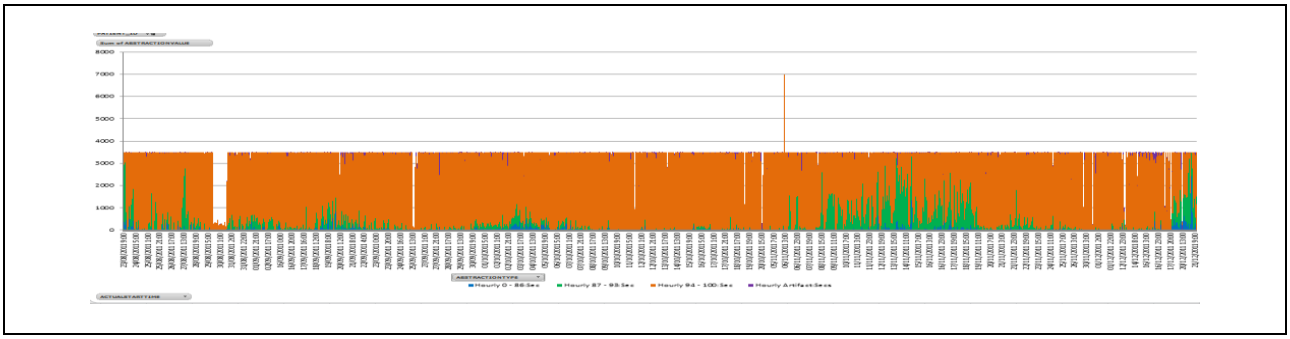

Figure 5.6: TA-based Process mining results for subject 6

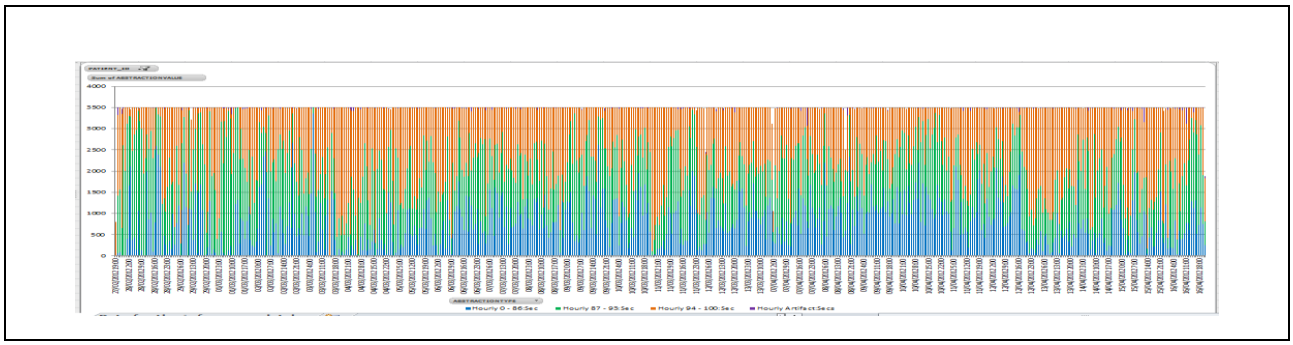

Figure 5.7: TA-based Process mining results for subject 8

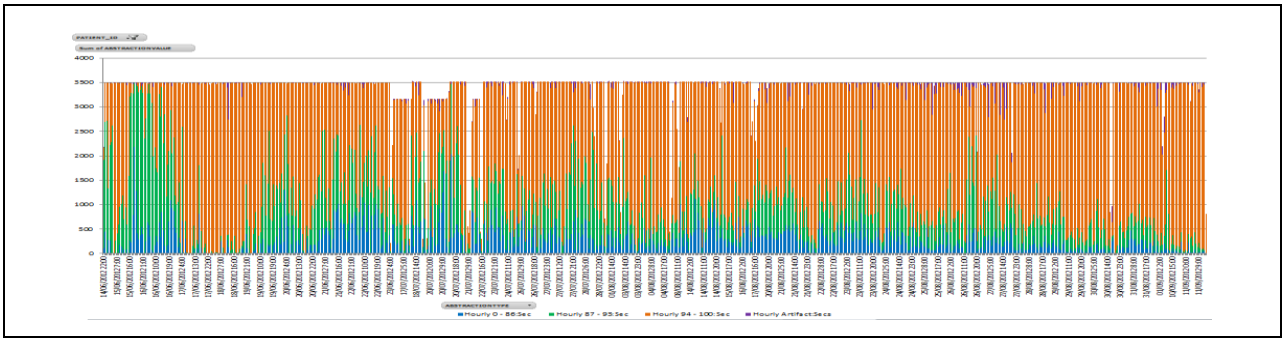

Figure 5.8: TA-based Process mining results for subject 9

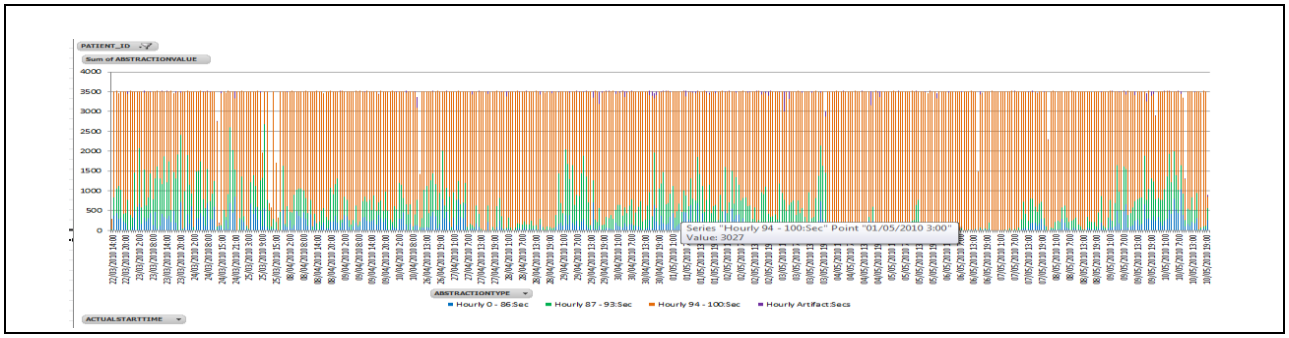

Figure 5.9: TA-based Process mining results for subject 10

As already detailed in relation to figure 4.3 in the preceding chapter, the above visualisations are the actual recordings of temporally abstracted behaviours of SpO_2 levels of within target, above target and below target ranges along with artifact plotted against the time for each neonate from the moment of commencing data capture in an Artemis bed space. Capturing and extracting those data was continued as long as the data was available to identify both the moment a neonate was detected to have developed ROP and the point of experiencing a transition in the stage of clinical severity.

As this research required analysis of SpO_2 as the only physiological data stream, the temporal abstractions are based on univariate analysis of SpO_2 as combined with GA.

5.2 Initial clustering performed based on visualisation of nine study participants

Initial clustering was performed based on the visualisation of overall distribution of the ROP hourly scores within the categories of below, within target, above and artifact ranges. Three clusters emerged namely: cluster 1 with subjects 3 and 8 having episodes each hour of over and under oxygenation; cluster 2 with subject 9 having a significant degree of over and under oxygenation that is pronounced more than that of cluster 1 but to a lesser extent than that is seen in cluster 3; and cluster 3 comprising subjects1, 2, 4, 5, 6 and 10 with predominant over oxygenation. The visual inspection was further supported by mathematical review of percentage of time under, in target and over oxygenation as shown in table 5.2.

5.3 Clustering performed based on ROP stages within the initial clustering

After performing the initial clustering, the already detected clinical ROP status of each infant was taken into consideration to assess whether there was a correlation between the infants in each cluster. Subjects from cluster 1 [subjects 3 and 8] developed stage 3 ROP; the subject from cluster 2 [subject 9] developed stage 1 ROP and it was a less severe form than that observed for subjects from cluster 1; cluster 3 [subjects; 1, 2, 4, 5, 6 and 10] did not develop ROP.

5.4 Clustering performed based on gestational age (GA) within the initial clusters

An additional clustering was performed based on GA within the initial clusters and that demonstrated a possible association: cluster 1 with subjects of $\leq$ 26 weeks GA, cluster 2 with 26 weeks GA and cluster 3 with $\geq$ 27 weeks GA.

5.5 Subsequent clustering performed based on Respiratory Distress Syndrome (RDS)

An attempt was made to fit these study participants into the above three clusters depending on whether they acquired RDS or not and this showed the following possible associations. RDS occurred in all subjects with a comparatively lower GA of $\leq$ 26 weeks at birth (clusters 1 and 2) and in each case surfactant had been administered. All of them acquired ROP ranging from stage 1 to stage 3. Among those with a GA of $\geq$ 27 weeks (cluster 3), there were two instances with no onset of RDS. Out of the remaining four cases that experienced RDS, only one participant received surfactant therapy. The participants in this cluster 3 showed no onset of ROP.

S	GA	BW	Sex	RDS	SRT	Jaundice	PTX	Sepsis	Blood Tx	Over oxygenation	Under oxygenation	ROP
1	31	810	M	No	No	Yes	Yes	No	No	98.92%	0.175%	0
2	29	1210	M	Yes	No	Yes	Yes	Yes	Yes	97.86%	0.155%	0
3	26	810	F	Yes	Yes	Yes	Yes	Yes	Yes	51.74%	16.98%	3
4	28	1080	M	Yes	No	Yes	Yes	No	No	88.895%	3.005%	0
5	31	1400	F	No	No	No	No	No	No	78.34%	2.885%	0
6	27	746	F	Yes	Yes	Yes	Yes	Yes	No	89.456%	1.938%	0
8	25	780	M	Yes	Yes	Yes	Yes	Yes	Yes	40.66%	22.07%	3
9	26	1010	M	Yes	Yes	Yes	Yes	No	Yes	74.12%	7.25%	1
10	27	955	M	Yes	No	Yes	Yes	No	Yes	78.44%	6.44%	0

Table 5.1: Clustering of subjects based on some of the known risk factors for ROP (colour coded for the convenient differentiation of clusters)

As shown in the research by Termote et al. [63] there is a possible association among RDS, surfactant replacement therapy (SRT) and ROP onset. They concluded that SRT increases the risk of developing ROP but is not associated with more severe forms of ROP. However, the findings in the study reported here must be validated through multivariate regression analysis to come to the same conclusion and these steps will be carried out in a future extension of this study on a larger patient cohort.

5.6 Sub-clustering performed based on sex and birth weight

A second sub-clustering was performed based on sex that did not correlate with the initial clustering. A third sub-clustering was attempted on birth weight (BW) and it too did not correlate with the initial clustering (Table 5.2).

	Average % Under oxygenation	Average % Target oxygenation	Average % Over oxygenation	GA	Sex	BW
S1	0.175%	1.18%	98.92%	31	M	810g
S2	0.155%	1.639%	97.86%	29	M	1200g
S3	16.98%	30.87%	51.74%	26	F	810g
S4	3.005%	7.81%	88.895%	28	M	1080g
S5	2.885%	18.53%	78.34%	31	F	1400g
S6	1.938%	8.164%	89.456%	27	F	746g
S8	22.07%	36.76%	40.66%	25	M	780g
S9	7.25%	17.59%	74.12%	26	M	1010g
S10	6.44%	14.68%	78.44%	27	M	955g

Table 5.2: Clustering based on the maximum reached ROP stage and GA (colour coded for convenient differentiation of clusters)

5.7 Sub-clustering performed based on known risk factors for ROP

Another sub-clustering was attempted based on other general clinical events that could have acted as risk factors, based on the chapter 3 for the onset of ROP (Table 5.3). Some of the risk factors considered in this study are the appearance of jaundice, whether phototherapy given, occurrence of sepsis and whether blood transfusions were given. These aspects will be discussed in detail in the following chapter; 'Discussion' which focuses on the effects of risk factors that occurred as adverse clinical events for ROP and reflects them on this study subjects.

The validation of sepsis and other risk factors; jaundice, administering phototherapy, occurrence of sepsis and blood transfusions as strong associations to the onset of ROP within the context of

this study could not be performed due to the low number of participants that prevented a statistical analysis.

	Occurrence for 6 subjects who did not develop ROP	%	Occurrence for 3 subjects who developed ROP	%
GA ≤ 26 weeks	0	0	3	100
Male	4	66.66	2	66.66
Female	2	33.33	1	33.33
RDS	4	66.66	3	100
SRT	1	16.66	3	100
Appearance of jaundice	5	83.33	3	100
Phototherapy given	5	83.33	3	100
Sepsis	2	33.33	2	66.66
Blood transfusions given	2	33.33	3	100

Table 5.3: Some of the subject criteria and some major risk factors for ROP; clustered to show the association with ROP

Chapter 6: Discussion

This research study demonstrates the applicability of the CRISP-TDM_0 framework as a structured approach to reveal previously unknown trends and patterns in physiological data streams. This was carried out through its dynamic DM approach for retrospective data and through comparison of multiple abstractions regardless of the actual time of data collect and diagnosis.

Although previously it was sufficient for static data mining (DM) methods to be applied to static data, the rapidly advancing demand for clinical stream data to be stored and analysed has impelled the need for dynamic data mining methods that can meet the needs of high-volume, high velocity, high-frequency, time-series data (aka Big Data). Furthermore, these data are multi-dimensional in nature where data elements each represent a dimension that can vary in value and characterise an item of interest.

This discussion is based three aspects: 1) reflecting the success of applying CRISP -TDM_0 within the case study context used in this research; 2) focusing on the effects of risk factors that occurred as adverse clinical events for ROP and reflecting them on our study subjects; and finally 3) exploring the possibilities in the new areas of research based on the knowledge gathered through this study. This aspect will be discussed in the conclusion chapter.

6.1 Reflection on the success of CRISP-TDM_0 methodology as a means of managing, reporting and performing this study

We were able to follow the process of CRISP-TDM_0 as detailed in chapter 4. This led to greater detail provided about the study rather than traditional approaches to performing and reporting the research results. Specifically the additional details are within the phases of CRISP-TDM_0 methodology which are incorporated in to the Artemis platform.

During this study, the phases of the CRISP data mining process model could be followed and results could be obtained as shown in the preceding chapter 5. By means of informatics data analysis and correlating these results to the outcome ROP, the CRISP-TDM_0 can be designated as the ideal methodology for managing and reporting results from the knowledge discovery experimental activity that has a greater potential to ensure consistency and completeness. In terms of delivering medical care, it is important to study the evolution of one or more parameters

of physiological data over time than an instant parameter values. The multiple streams of physiological data that a premature infant generates are not isolated from each other. Therefore, we need TA approaches that cross correlate data from multiple streams for a combined TA model as a complex indicator of condition onset. Data mining is important in multidimensional analysis because traditional statistical methods are not well suited to evaluating the probability of coincidental patterns in high-dimensional data sets [3]. The flexibility and multidimensionality of the Artemis framework within the NICU allows multiple streams of physiological data to be captured from multiple patients, to be monitored for the onset of multiple potential conditions that can occur concurrently with patients located in multiple NICU locations. In this study, blood oxygenation streams in the form of SpO_2 were captured from multiple patients and were assessed in the context of one condition, from a dataset from one location.

As already discussed in Chapter 4, a scientific method driven approach based on the cycle of observation - hypothesis formation - experimentation, when applied to generate results of data mining in clinical data streams would yield more important outcomes that favour a reasonable clinical practice. Likewise forming a null hypothesis and getting it validated through the same process of experimentation is important [9], [28], [29]. This concept is suggestive of constructing data mining tools and methodologies to conduct both 'exploratory' and 'confirmatory' data mining. The CRISP-TDM_0 as the standardised approach in this study enabled the use of exploratory data mining as a tool to find unknown patterns or relationships to create hypotheses. Keeping in with this requirement, this research employed the functionality of business understanding, data understanding and data modelling phases for managing and reporting the results generated through knowledge discovery. In this study we could achieve those steps.

Likewise, testing of an alternative null hypothesis too is important. According to the speculation of this research an alternative null hypothesis was created and a clear statement of the null hypothesis was identified as 'There is no correlation between ROP stages and detected blood SpO_2 levels'. This step can be achieved by driving the Data Preparation, Modelling and Evaluation phases through the CRISP-TDM_0 and tested during iterative explanatory data mining activity.

6.2 Reflection of what is known in the clinical literature about the risk factors for ROP

From the clinical literature review presented in chapter 3, prior known risk factors for the onset of ROP are excessive blood oxygen saturation levels and prolonged oxygen therapy and these have been found consistently relating to the disease onset [40], [41], [42]. Studies designed to find the effect of the duration of administering supplemental oxygen and the occurrence of ROP by Patz et al. [47], Kinsay et al. [48], [49] and Lanman et al. [50] have shown that premature infants treated with oxygen regimens of longer duration and higher concentration experienced significantly more ROP than controls.

In the study aimed to detect the efficacy of graded oxygen saturation (SpO_2) targets to apply as a preventive measure in ROP, Arora et al. [56] found that implemented graded oxygen saturation targets resulting in a significant decrease in ROP, laser surgery rates and a trend towards lower mortality. They concluded that use of graded oxygen saturation targets based on postmenstrual age (PMA) would become more beneficial rather than adhering to low or high saturation targets during the two different phases of ROP as detailed in chapter 3.

In order to find the influence of arterial oxygen fluctuation in ROP in very low birth weight infants, York et al. [43] launched a retrospective study of 231 infants weighing < 1500 g at birth. The fluctuations in partial pressure of dissolved arterial oxygen (PaO_2) were expressed as coefficient of variation (CoV) for each infant. They investigated the relationship between CoV at three intervals and the risk of developing threshold ROP. The odds ratio (OR) of developing threshold ROP versus pre-threshold ROP was associated with an increase in the CoV during the oxygen therapy. Hence they concluded that VLBW infants experiencing fluctuation of arterial oxygen tension are at a higher risk of threshold ROP.

The ELGAN study [64] was designed to test whether preterm neonates who had a blood gas derangement on at least two of the first three postnatal days are at increased risk for more severe ROP than the infants who were not. A neonate was considered to be exposed if the blood gas measure was in the highest or lowest quartile for GA on at least two of the first three postnatal days. They concluded that repeatedly high blood concentrations of oxygen, (PaO_2) and carbon dioxide ($PaCO_2$) and low blood pH are associated with increased risk of severe ROP [64].

In addition there are literature regarding studies that found very low birth weight (VLBW) and general immaturity making important independent contributions to the overall incidence of severe ROP [42], [43].

Chen et al. aimed to identify and to explore the interrelationships between prominent risk factors for ROP and found that oxygen associated ROP risk was more prominent among infants of 23-25 weeks gestational age, while infection associated ROP risk was higher among infants born at 28-29 weeks. The study further suggested that neonatal sepsis, oxygen exposure, and low gestational age are not only independently associated with a significantly increased risk of ROP but also interact beyond additive and even multiplicative patterns [25]. Nevertheless, asphyxia and apnoea were found to be high risk factors for ROP [42], [44].

However, the current understanding of ROP is that the progression of disease is multifactorial and the other risk factors are multiple gestations, intracranial haemorrhages, anaemia, prolonged mechanical ventilation, and multiple blood transfusions [40], [42], [65], [66]. A recent report suggests that elevated glucose may also play a role [40]. In a prior research Termote et al. [63] found that surfactant replacement therapy (SRT) could increase the risk of ROP onset.

De Jonge et al. [65] stated that oxidative injury may contribute to the development of ROP and hypothesised bilirubin may be a physiologically important antioxidant that can act for protecting the eyes from ROP. But they found no definite association between bilirubin levels and severe ROP. Yeo et al. [66] found an association of prolonged phototherapy and low peak serum bilirubin concentrations with severe visual loss attributable to ROP but suggested the findings should be interpreted with caution until the evidence is reinforced in other patient populations.

6.3 Discussion based on the performed clustering of study subjects

As previously mentioned in chapter 5, illustrations of the TA-based Process mining results of the subjects 3 and 8 shared a closer similarity in the visualisation and they were clustered together. These two subjects experienced the most severe ROP outcomes among the subjects of the study cohort and were of comparatively lower maturity, i.e. GA 26 and 25 weeks respectively. The reflection of ROP risk factors on them showed they had experienced most of the aforesaid adverse clinical events which correlate with severe transitions in ROP.

The association of ROP with RDS can be discussed in relation to the prior research by Termote et al. [63] who found that surfactant replacement therapy (SRT) could increase the risk of ROP onset. In this study population, among those with a GA ≥ 27 weeks, there were two instances with no onset of RDS. Out of the remaining four cases that encountered RDS, only one participant had received surfactant therapy. This is in contrast to the participants in clusters 2 and 3 with a GA ≤ 26 weeks all of whom encountered RDS and received surfactant (Tables 6.1 and 6.2). All these had ROP advancing in stages from stage 1 onwards. This research showed RDS occurring in all subjects with a comparatively lower GA ≤ 26 weeks at birth and in each case surfactant had been administered as well. They all acquired ROP ranging from stage 1 to stage 3.

The association of jaundice and ROP can be discussed referring to the work done by De Jonge et al. [65] who stated that oxidative injury may contribute to the development of ROP and hypothesised bilirubin may be a physiologically important antioxidant that can act for protecting the eyes from ROP. Likewise, Yeo et al. [66] found an association of prolonged phototherapy and low-peak serum bilirubin concentrations with severe visual loss attributable to ROP. In this study among the participants with a GA ≤ 26 weeks, out of the total three subjects, all of them had developed ROP, had jaundice and had undergone phototherapy. Nevertheless out of the six study participants of cluster 3 whose GA were ≥ 27 weeks at birth, five had developed jaundice and each one of them had undergone phototherapy (Tables 6.1 and 6.2). Both De Jonge et al. [65] and Yeo et al. [66] suggested their findings should be interpreted with caution until the evidence is reinforced in other patient populations. Likewise this study results as well need to be validated by running the study in a larger patient cohort.

Among those with a GA ≥ 27 weeks at birth (cluster 3), there were two instances where we encountered sepsis. These neonates did not develop ROP. However those that belonged to cluster 1 with a GA ≤ 26 weeks, all neonates had experienced sepsis and developed ROP (Tables 6.1 and 6.2).

Even though the study participants who developed ROP had all undergone blood transfusions and this is in contrast to the only 33.33% who did not develop ROP (Table 6.2), the association of blood transfusions to ROP onset needs to be validated through a possible larger future study.

We were not able to determine whether GA at birth below 27 weeks and low birth weight were stronger correlations to the onset of ROP. ROP risk parameters like RDS, surfactant administration, jaundice, application of phototherapy and neonatal sepsis etc. would require application on a larger study cohort.

Subject	GA	BW	Sex	RDS	SRT	Jaundice	PTX	Sepsis	Blood Tx	Over oxygenation	Under oxygenation	ROP
1	31	810	M	No	No	Yes	Yes	No	No	98.92%	0.175%	0
2	29	1210	M	Yes	No	Yes	Yes	Yes	Yes	97.86%	0.155%	0
3	26	810	F	Yes	Yes	Yes	Yes	Yes	Yes	51.74%	16.98%	3
4	28	1080	M	Yes	No	Yes	Yes	No	No	88.895%	3.005%	0
5	31	1400	F	No	No	No	No	No	No	78.34%	2.885%	0
6	27	746	F	Yes	Yes	Yes	Yes	Yes	No	89.456%	1.938%	0
8	25	780	M	Yes	Yes	Yes	Yes	Yes	Yes	40.66%	22.07%	3
9	26	1010	M	Yes	Yes	Yes	Yes	No	Yes	74.12%	7.25%	1
10	27	955	M	Yes	No	Yes	Yes	No	Yes	78.44%	6.44%	0

Table 6.1: Clustering of subjects based on some of the known risk factors for ROP (colour coded for the convenient differentiation of clusters)

	Occurrence for 6 subjects who did not develop ROP	%	Occurrence for 3 subjects who developed ROP	%
GA ≤ 26 weeks	0	0	3	100
Male	4	66.66	2	66.66
Female	2	33.33	1	33.33
RDS	4	66.66	3	100
SRT	1	16.66	3	100
Appearance of jaundice	5	83.33	3	100
Phototherapy given	5	83.33	3	100
Sepsis	2	33.33	2	66.66
Blood transfusions given	2	33.33	3	100

Table 6.2: Some of the subject criteria and some major risk factors for ROP; clustered to show the association with ROP

Chapter 7: Conclusions of the thesis

7.1 Thesis Purpose

This thesis has presented a systematic approach to find a solution for a broader medical challenge which is the limitation in the area of knowledge discovery in physiological data streams. This limitation can be viewed as emanating from the existing but still overlooked requirement for harnessing and using the valuable medical information in physiological data streams that originate in patient monitoring devises. A solution was demonstrated through a case study that utilised the medical informatics based standardised approach, the CRISP-TDM_0 methodology, to assess the relationship of high fidelity blood oxygen saturation data to the onset of ROP.

7.2 Thesis Objectives

7.2.1 Demonstrating the capability of the standardised informatics methodology to manage, report and perform the study

This thesis is built on the previous research work that proposed the CRoss Industry Standard Process for Data Mining, CRISP-TDM_0 methodology that merges CRISP-TDM and CRISP-DM_0 aspects as the standardised informatics methodology. Hence the primary objective is to demonstrate the capability of standardised informatics methodology to managing, reporting and performing retrospective clinical research using physiological data streams from the domain of neonatal intensive care.

Within this thesis, CRISP-TDM_0 methodology as the standardised approach for managing clinical research using medical data streams was demonstrated by the close interrelationship of the steps of CRISP-TDM_0 which enables one step to flow to the next. It provides a structure to plan out the remainder of the research based on the changed pathway in the event that changes were required.

The CRISP-TDM_0 methodology as a standardised approach for reporting clinical research using medical data streams was demonstrated by adhering to the steps proposed by CRISP-TDM_0 and reporting the information required by the phases as research progressed as demonstrated in chapter 4.

7.2.2 Demonstrating the standardised informatics methodology using a clinical research study designed from the domain of neonatal intensive care

The applicability of the CRISP-TDM_0 approach for performing this clinical research study was demonstrated through the results of a more detailed analysis of the blood SpO_2 level paradigms performed using the Artemis platform and comparing those values to the remaining ROP stages. To detect the blood SpO_2 levels second by second for each neonate, the multidimensional patient oriented data mining approach of Artemis was utilised through the instantiation of CRISP-TDM_0 methodology within the Artemis platform that supported this clinical research study. This enabled the discovery of physiological stream behaviours that may represent earlier condition onset than those currently used in evidence-informed practices. This aspect will be further detailed in the section which immediately follows.

7.2.3 The instantiation of this research design within the Artemis platform

The first secondary objective of this thesis was to demonstrate the instantiation of this design within the Artemis platform. A more detailed analysis of the blood oxygen saturation level paradigms using the Artemis platform was conducted and demonstrated that the standardised CRISP-TDM_0 approach within the knowledge discovery component of Artemis enabled us to follow the phases of CRISP-TDM_0 and this led to greater detail provided about the study rather than traditional approaches to performing and reporting the research results using medical data streams. Hence, we could demonstrate CRISP-TDM_0 as the standardised approach.

7.2.4 Developing a risk assessment score for the onset of ROP

Establishing a risk assessment score for the development of ROP based on the existing ROP stages (stages 0-5), which is the second one of the secondary objectives, in the case of this study was not achieved, and what was performed was simply an assessment of the correlation between SpO_2 and other vindicators with the onset of ROP. This was due to the very low number of study participants. However at this stage this aspect can be theoretically explained by referring to the clustering analysis of the eventual nine study participants according to the visualisations of the temporally abstracted behaviours of SpO_2 values. These patterns can then be correlated to the development of ROP of each individual patient. The next step is correlating the developed ROP stage and exposed oxygenation level to the GA.

A risk score can be given to each of the following parameters ranging from 1-10. If an individual's pattern of SpO_2 variation tallies with the already established patterns that are associating with advanced ROP a higher score is given. Likewise, a higher score is given for a lower GA and vice versa. This can guide us through the process of predicting a risk score for future development of ROP especially if this study will be replicated and applied within the context of a larger patient cohort.

7.2.5 Developing a real-time algorithm for the monitoring of blood oxygen saturation levels

Achieving this third of the secondary objective would lead us to minimise the risk for the development of ROP by allowing us to monitor the blood oxygen saturation levels to assist with oxygen administration. This should involve transition of the algorithm to run in real-time within the online part of Artemis and this work is currently in progress.

7.3 Could this research study answer the formulated research questions?

We could also find answers to the research questions that were formulated in chapter 1 as shown below.

Question 1: Can extensions to CRISP-TDM enable the complete management, reporting and completion of a clinical research study involving physiological streaming data that incorporates null hypothesis testing?

Answer: Yes

Explanation: As already discussed in Chapters 2 and 4, this could be achieved by running this research through the phases that contained extensions proposed to the CRISP-DM model [2], [4], [22] to form CRISP-TDM model because the analysis by Catley et al. revealed these phases of the CRISP 1.0 model were limited in their ability to describe certain issues [2], [4]. Those extensions involved phase 1, to incorporate the 'clinical objective' enabling support for clinical investigation and exploratory data mining on clinically relevant and population-based information; phase 2 to incorporate temporal aspects of multidimensional data of the clinical investigation through clinical study; phase 4 to apply relevant details in IDA knowledge management and IDA system integration providing support for integrating techniques, such as DM and TA, in a closed or open loop workflow and this is in contrast to that of CRISP 1.0 which concentrates on applying several DM techniques to arrive at one which offers the best results rather than providing support for integrating techniques, such as DM and TA, in a closed or open

loop workflow; phase 5 to assess the process mining of temporal abstraction based results of the clinical investigation being explored and phase 6 to incorporate the provision of the methodology for describing system storage, which is the key in TA-based systems where both the raw data and TAs must be archived for future use.

In addition, this research applied scientific method driven approach based on the cycle of observation - hypothesis formation - experimentation for the generation of results of data mining in the SpO_2 data streams and formulating a null hypothesis. As this needs to involve data mining tools and methodologies that conduct both 'exploratory' and 'confirmatory' data mining [9], [28], [29]. The CRISP-TDM_0 methodology that was used in this study enabled the use of exploratory data mining as a tool to find unknown patterns or relationships to create hypotheses through business understanding, data understanding and data modelling phases. An alternative null hypothesis was then created and a clear statement of the null hypothesis was identified as 'There is no correlation between ROP stages and detected blood SpO_2 levels'. However the testing of the null hypothesis that can be achieved through data preparation, modelling and evaluation phases during iterative explanatory or confirmatory data mining activity was not carried out due to the limited number of study participants. By adhering to the steps of the phases of the CRISP-TDM_0 methodology, we were able to manage this research using medical data streams because of the close interrelationship of the steps which enables one step to flow to the next. In the event that changes are required, it provided a structure to plan out the remainder of the research based on the changed pathway. Likewise we were able to report the information required in the clinical research by adhering to the steps proposed by extended CRISP-TDM_0 methodology.

Question 2: Is there a correlation between the blood oxygen saturation (SpO_2) levels of immature newborn infants, born at less than 32 weeks gestational age, during the first two weeks of life and the severity of retinopathy of prematurity occurring before 35 completed weeks postmenstrual age?

Answer: Yes

Explanation: subjects 3 and 8 of cluster 1were of GA $\leq$ 26 weeks. They both had some episodes each hour of over and under oxygenation and clinically detected sinister progression in ROP severity occurred before 35 weeks completed postmenstrual age. The progression in ROP

plotted according to the PMA for the study subject 3 is as shown in the table 7.1 below. For the clinical correlation, Artemis SpO_2 data were available till this neonate reached 35 weeks PMA.

PMA at assessment	ROP
33 weeks	Bilateral stage 3
34 weeks	Bilateral stage 3 (non-progressive)
35 weeks	Bilateral l stage 3, zone 2 with pre-plus on right eye
36 weeks	Bilateral stage 3, zone 2 with bilateral pre-plus disease, especially R eye
36 weeks, 3 days	Bilateral stage 3, zone 2; right sided plus disease, left sided pre-plus
37 weeks	ROP stabilized after LASER therapy

Table 7.1: Progression in ROP according to PMA of subject 3

The progression in ROP plotted according to the PMA for the study subject 8 is as shown in the table 7.2 below. For the clinical correlation, the Artemis SpO_2 data were available only up to 33 weeks PMA and this is a technical issue this study encountered.

PMA at assessment	ROP
31 weeks	Right eye, stage 0, zone 1; left eye, stage 1, posterior zone 2
32 weeks	Bilateral stage 2, zone 2
34 weeks	Bilateral stage 2, posterior zone 2
35 weeks	Right stage 2, zone 2; left stage 3, zone 2 with plus disease
35 weeks and 4 days	Right stage 3, zone 2 with plus disease; left stage 3 zone 2 with pre-plus disease

Table 7.2: Progression in ROP according to PMA of subject 8

Cluster 2 comprises subject 9 with a GA of 26 weeks. This neonate had episodes of over oxygenation with an average value that is more marked than that of cluster 1 and little less marked than that of cluster 3 The average value of under oxygenation too falls in between that of Clusters 1 and 3.

The progression in ROP plotted according to the PMA for the study subject 9 is as shown in the table 7.3 below. For the clinical correlation, the Artemis SpO_2 data were available for the entire period.

PMA at assessment	ROP
31 weeks	Left stage 1, zone 2; right stage 0, zone 2
32 weeks	Left stage 1, zone 2; right stage 0, zone 2
34 weeks	Bilateral stage 1, zone 2
36 weeks	Left stage 1, zone 2; right stage 0, zone 3 and then gradually improved.

Table 7.3: Progression in ROP according to PMA of subject 9

Cluster 3 that comprised study subjects 1, 2,4,5,6, and 10 of GA ≥ 27 weeks showed the most predominant over oxygenations. When assessed clinically, none of them developed ROP. The details are given in the table 7.4.

Subjects	PMA at assessment	ROP	Remarks
Subject 1	33 weeks	not detectable	Artemis SpO_2 data were available for correlation with the clinical assessments
Subject 2	34 weeks	not detectable	Artemis SpO_2 data were available for correlation with the clinical assessments
	35 weeks	not detectable	
Subject 4	32 weeks	not detectable	Artemis SpO_2 data were available for correlation with the clinical assessments
	34 weeks	not detectable	
Subject 5	33 weeks	not detectable	Artemis SpO_2 data were available for correlation with the clinical assessments
Subject 6	31 weeks	not detectable	Artemis SpO_2 data were available for correlation with the clinical assessments
	32 weeks	not detectable	
	34 weeks	not detectable	
	36 weeks	not detectable	
Subject 10	31 weeks	not detectable	Artemis SpO_2 data were available through the period except during the window from10/04/2010 to 26/04/2010 and this was due to technical reasons
	33 weeks	not detectable	
	35 weeks	not detectable	
	37 weeks	not detectable	

Table 7.4: Progression in ROP according to PMA of subjects 1, 2, 4, 5, 6, and 10

The above details show a correlation between SpO_2 levels of immature neonates of the study cohort born at less than 32 weeks GA, during the first two weeks of life and the severity of ROP

occurring before 35 completed weeks PMA. However, this would yield more valuable out come when applied to a larger patient cohort.

Question 3: Is it possible to calculate the probability at postnatal ages three days, seven days, ten and fourteen days for the severity of retinopathy of prematurity occurring before 35 completed weeks postmenstrual age?

Answer: No

Explanation: The available effective data in this study were not sufficient to formulate an answer to the above question and this was a limitation of this research. However, there is potential for using the correlative factors that were found in this research to be used as markers to watch for at three days, seven days, ten, fourteen etc. when this research is conducted on a larger patient cohort.

7.4 Conclusions

This study attempted to find an elucidation to the limitations that were in existence in the area of knowledge discovery in physiological data streams that pertained to neonatal critical care. This was sought through a single case study that employed a medical informatics based standardised approach which is CRISP-TDM_0 methodology.

The idea of this pilot study was to correlate the effects of neonatal blood SpO_2 detected in a high fidelity manner to the occurrence of ROP. This was done through an array of possibilities that operated either as clinical risk factors or informatics analysis based results.

The study participants born at ≥ 27 weeks GA had experienced the highest over oxygenation but did not develop ROP. The visualisation of overall distribution of the ROP hourly scores based on TA based results of SpO_2 of this cluster shared a similarity in the pattern, hence these subjects 1, 2, 4, 5, 6 and 10 were clustered together in cluster 3. In contrast, subjects 3 and 8 who were ≤ 26 weeks GA showed comparatively less exaggerated over and under oxygenations and had the most severe ROP progression i.e. stage 3. The visualisation of overall distribution of the ROP hourly scores based on TA based results of SpO_2 showed a similarity in the pattern shared by these subjects and they were considered together in cluster 1. Subject 9, with GA 26 weeks, showed a unique pattern of ROP hourly scores based on TA based results of SpO_2 and is considered separately in cluster 2. However, to consider the very fine contribution of over and

under oxygenation to the behaviour of ROP, this study would require to be carried out on a larger patient cohort with a robust statistical analysis of data.

The two subsequent sub-clustering attempted on birth weight (BW) and gender did not correlate with the initial clustering. This is not too surprising. GA is a more robust marker of immaturity than birthweight; there is currently no evidence to indicate that the neonate's sex influences the development of ROP. Details on another sub-clustering that was attempted based on known potential risk factors for the causation of ROP namely; sepsis, RDS, surfactant administration and blood transfusions are discussed below.

When the effect of sepsis on ROP is considered in relation to this study, among those with a GA $\geq$ 27 weeks at birth (cluster 3), there were 2 subjects one with sepsis. These two cases were considered as not having ROP as their clinically detected ROP never progressed beyond stage 0. However, those that belonged to cluster 1 with a GA $\leq$ 26 weeks, all had experienced sepsis and higher progression in ROP. However, the numbers of participants are limited, hence this finding is insufficient to conclude that sepsis as a contributor to ROP occurring in neonates born at GA $\leq$ 26 weeks.

According to Termote et al., there is an association of ROP with surfactant administration [63]. When the effect of RDS on ROP in relation to this study is considered, among those with GA $\geq$ 27 weeks (cluster 3), there were three instances of RDS but only one patient among them had received surfactant therapy. The remaining two patients in this cluster did not have RDS. None of the subjects in this cluster developed ROP. This is in contrast to the scenario observed in participants of clusters 1 and 2 with a GA $\leq$ 26 weeks where all of them had RDS, received surfactant therapy and developed ROP progressing from stage 1 to 3. However to better correlate the onset of ROP with surfactant therapy and RDS and fitting them within the initial clustering, this research would require to be conducted on a larger population.

However, this study could not determine whether GA at birth below and above 27 weeks was a stronger correlation to ROP. Clinical parameters such as onset of RDS and surfactant administration, neonatal sepsis, developing jaundice and phototherapy and blood transfusions too showed no correlation to the initial clustering within the purview of this study, however, these

have a potential to yield valuable outcome when applied within the context of a larger patient cohort.

7.5 Strengths and significance of this research

- The most significant strength of this study was CRISP-TDM_0 methodology that enabled the complete management, reporting and completion of clinical research.
- Another significant strength was the Artemis platform's ability to support clinical research to quantify the relationship between retinal exposure to oxygen and ROP more precisely and accurately as we gathered higher fidelity oxygen saturation data that provided a better representation of retinal exposure to oxygen during a defined period of time than intermittent spot readings.
- Artemis allows detailed information to be extracted from the raw data such as; ranges, times, frequency, hourly approaches and outliers which in turn gives us an insight to what is occurring in a particular time, for any particular length of time. This unique feature aids the tracking of hypoxaemia and hyperoxaemia and provides a better understanding of the current level of oxygenation. This detailed data analysis provided a more comprehensive data set inference than just a single value, in a single time frame which, would lack the connectivity and relevance of the oxygen saturation.
- Analyzing SpO_2 fluctuations in an hourly dataset can provide extensive information to the healthcare personnel with a great potential to improve adjustments in titrating the transcutaneous blood oxygen saturation levels at the bedside.
- Artemis has the remarkable ability to accurately reflect the activity of the hour and convert copious amounts of data into visual charts and graphs, a better means of conveyance of information which can be passed on to another health professional at the end of the shift. This acts as a valuable clinical and educational tool with a potential to discover other relevant causative factors.

7.6 Advantages of this study

- This research has presented a structured and standardised approach for reporting of retrospective studies using medical data streams within the context of an NICU environment.

- This research illustrates a methodology to assist clinicians to isolate consistent behaviours across multiple data sets.
- The patterns so generated through comparison and alignment of multiple abstractions could indicate an onset of a potentially detrimental condition, ROP, in a premature neonate.
- The adapted extensions to CRISP-DM, CRISP-TDM_0 are aimed at generating clinical rules based on correlations across multiple streams which could be regarded as suitable for evidence based medical practice.

7.7 Potential limitations of the methodology and weaknesses of this study

- There are confounding variables both known and unknown, hence the need arises to perform adjustments for these confounding variables.
- We considered only one physiological stream, blood oxygenation saturation level (SpO_2), in our study due to the work to date on the association of ROP and SpO_2. However Artemis has the potential to process multiple data streams of more than one patient simultaneously and we did not employ that capability in this study.
- The measurements of certain medical parameters had not taken place as frequently as we expected, e.g. haemoglobin estimations and consequently there were limitations in the data analysis and clinical correlation process.
- The ultimate number of subjects in this research was limited to only nine, hence clinical investigative support for null hypotheses driven confirmatory data mining was not carried out.
- The results of the temporally abstracted behaviours of SpO_2 levels of in target, above and below target ranges captured for each neonate revealed fully or partially unrecorded Artemis data and this was due to technical reasons with the medical devices and the limited Artemis bed spaces. Infants were be moved in and out of Artemis bed spaces during their NICU admission. As a result of these data gaps, we encountered difficulties during the process of clinical correlation.
- This being a study based on retrospective clinical data capture, there were instances where there was periodical non availability of clinical data especially those that correspond to patient follow-ups.

- At the time of conducting this study in the NICU of The Hospital for Sick Children, Toronto, every NICU bed was not connected to the Artemis platform.
- We were lacking clinical documentations detailed in the electronic health record about follow-up assessments either due to the fact that these assessments occurred either in the referring hospital or another community hospital. The results of these assessments were simply not readily available for this research.

7.8 Intended future work

Future research will be designed in order to counteract the above mentioned limitations of this study.

- In future work, we will refine the Artemis platform to incorporate the knowledge learned from CRISP-TDM_0 modelling to guide new knowledge acquisition in real-time. This will enable newly discovered knowledge to be translated and incorporated into the Artemis platform so that these new rules can be executed in real-time.
- We will employ Artemis to full capacity to process mine multiple data streams on more than one patient so that we will be able to analyse the effects of variation in other physiological data streams and correlate the combined effect to the outcome in order to come to a more rational conclusion.
- We will connect beds within the SickKids NICU to the Artemis platform.
- We will devise the needed modifications to the software system to capture every moment of data without losing them.
- We will enroll more patients to the study.
- We will perform a statistical analysis on the data in a future extension to this study.
- By means of recruiting more study subjects, analysing more data sets of blood SpO_2 and correlating to the outcome, ROP, we will set the objectives of this study to devise more suitable blood oxygenation levels according to the post-natal age.
- A future development of this study would further provide the integration of the method to test the null hypothesis to validate the proposed hypotheses.
- The adapted extensions to CRISP-DM, CRISP-TDM_0 will aid in generating clinical rules based on correlation of across multiple streams. This can be applied within the perspective of evidence based medicine practice.

- In future work, we will refine Artemis to incorporate the knowledge learned from CRISP-TDM$_0$ modelling to guide new knowledge acquisition in real-time. This will enable newly discovered knowledge to be translated and incorporated into Artemis so that these new rules can be executed in real-time.
- We will employ Artemis' full capacity to process mine multiple data streams of more than one patient so that we will be able to analyse the effects of variation in other physiological data streams and correlate the combination to the outcome in order to come to a more rational conclusion.

Chapter 8: References

1. Griffin, P., & Moorman, R. (2001). Toward the early diagnosis of neonatal sepsis and sepsis-like illness using novel heart rate analysis. Pediatrics, vol. 107, no. 1, pp. 97-104

2. Catley, C., Smith, K., McGregor, C., James, A., & Eklund, J. M. (2010). A Framework to Model and Translate Clinical Rules to Support Complex Real-time Analysis of Physiological and Clinical Data. IHI'10. Arlington, Virginia, USA.: 2010 ACM

3. Catley, C., Smith, K., McGregor, C., & Tracy, M. (2009, August). Extending CRISP-DM to incorporate temporal data mining of multidimensional medical data streams: A neonatal intensive care unit case study. In *Computer-Based Medical Systems, 2009. CBMS 2009. 22nd IEEE International Symposium on* (pp. 1-5). IEEE.

4. McGregor et al. 2011 European Federation for Medical Informatics. Next Generation Neonatal Health Informatics with Artemis User Centred Networked Health Care A. Moen et al. (Eds.) IOS Press, 2011

5. Finer, N., & Leone, T. (2009). Oxygen saturation monitoring for the preterm infant: the evidence basis for current practice. *Pediatric research*, *65*(4), 375-380.

6. VanderVeen, D. K., Mansfield, T. A., & Eichenwald, E. C. (2006). Lower oxygen saturation alarm limits decrease the severity of retinopathy of prematurity. *Journal of American Association for Pediatric Ophthalmology and Strabismus*, *10*(5), 445-448.

7. SUPPORT Study Group of the Eunice Kennedy Shriver NICHD Neonatal Research Network. (2010). Target ranges of oxygen saturation in extremely preterm infants. *The New England Journal of Medicine*, *362*(21), 1959.

8.Hagadorn, J. I., Furey, A. M., Nghiem, T. H., Schmid, C. H., Phelps, D. L., Pillers, D. A. M., & Cole, C. H. (2006). Achieved versus intended pulse oximeter saturation in infants born less than 28 weeks' gestation: the AVIOx study. *Pediatrics*, *118*(4), 1574-1582.

9. Heath, J., McGregor, C., (2010) CRISP-DM0: A method to extend CRISP-DM to support null hypothesis driven confirmatory data mining. Advances in Health Informatics Conference, May, pp 96-101, 2010.

10. McGregor, C. (2013). Big data in neonatal intensive care. *Computer*, *46*(6), 54-59.

11. Thommandram, A., Eklund, J. M., McGregor, C., Pugh, J. E., & James, A. G. (2014, June). A rule-based temporal analysis method for online health analytics and its application for real-time detection of neonatal spells. In *Big Data (BigData Congress), 2014 IEEE International Congress on* (pp. 470-477). IEEE.

12. Zikopoulos, P., Parasuraman, K., Deutsch, T., Giles, J., & Corrigan, D. (2012). *Harness the power of big data The IBM big data platform*. McGraw Hill Professional.

13. Cirelli, J., Graydon, B., McGregor, C., James, A., (2013) Analysis of Continuous Oxygen Saturation Data for Accurate Representation of Retinal Exposure to Oxygen in the Preterm Infant, Stud Health Technol Inform.; 183:126-31.

14. Podraza, W., Podraza, H., Jezierska, K., et al. (2012). EEG, brain maturation, and the development of retinopathy of prematurity. *Journal of Maternal-Fetal and Neonatal Medicine*, *25*(11), 2381-2384.

15. Chen, J., & Smith, L. E. (2007). Retinopathy of prematurity. *Angiogenesis*, *10*(2), 133-140.

16. Finer, N., & Leone, T. (2009). Oxygen saturation monitoring for the preterm infant: the evidence basis for current practice. *Pediatric research*, *65*(4), 375-380.

17. Askie, L. M., Henderson-Smart, D. J., Irwig, L., & Simpson, J. M. (2003). Oxygen-saturation targets and outcomes in extremely preterm infants. *New England Journal of Medicine*, *349*(10), 959-967.

18. Hartnett, M. E., & Lane, R. H. (2013). Effects of oxygen on the development and severity of retinopathy of prematurity. *Journal of American Association for Pediatric Ophthalmology and Strabismus*, *17*(3), 229-234.

19. Leeb, K., Jokovic, A., Sandhu, M., & Zinck, G. (2006). Intensive Care in Canada. Healthcare Quarterly, vol. 9, pg. 32-33

20. McGregor, C., & Smith, K. P. (2009). A Survey of Physiological Monitoring Data Models to Support the Service of Critical Care. 33rd IEEE International Computer Software and Applications Conference (pp. 104-109). Seattle, Washington, USA.: COMPSAC2009

21. Zhang, Y. (2007) Real-time Development of Patient-specific Alarm Algorithms for critical care. IEEE EMBS conference

22. Catley, C., Smith, K., McGregor, C., & Tracy, M. (2009). Extending CRISP-DM to incorporate temporal data mining of multidimensional medical data streams: A neonatal intensive care unit case study. 22nd IEEE International Symposium on Computer-Based Medical Systems, 2009 (pp. 1-5). Albuquerque, NM: IEEE

23. McGregor, C. P. (2013). *U.S. Patent No. 8,583,686*. Washington, DC: U.S. Patent and Trademark Office.

24. Sharek, P. J., Horbar, J. D., Mason, W., Bisarya, H., Thurm, C. W., Suresh, G., et al. (2006). Adverse Events in the Neonatal Intensive Care Unit: Development, Testing, and Findings of an NICU-Focused Trigger Tool to Identify Harm in North American NICUs. PEDIATRICS - Official Journal of the American Academy of Pediatrics, 1332-1340.

25. Chen, M., Çitil, A., McCabe, F., Leicht, K. M., Fiascone, J., Dammann, C. E., & Dammann, O. (2010). Infection, oxygen, and immaturity: interacting risk factors for retinopathy of Prematurity. *Neonatology*, *99*(2), 125-132.

26. Holmström, G., Broberger, U., & Thomassen, P. (1998). Neonatal risk factors for retinopathy of prematurity - a population-based study. *Acta Ophthalmologica Scandinavica*, *76*(2), 204-207.

27. Seiberth V., Linderkamp O. (2000) Risk Factors in Retinopathy of Prematurity; A Multivariate Statistical Analysis, Ophthalmologica;214:131-135

28. Ji, W., Naguib, R. N. G., & Ghoneim, M. A. (2003) Neural network-based assessment of prognostic markers and outcome prediction in bilharziasis-associated bladder cancer. Information Technology in Biomedicine, IEEE Transactions on, 7(3), 218-224. DOI:10.1109/TITB.2003.813796

29. Portney, L. G., & Watkins, M. P. (2000) Statistical measures of reliability. Foundations of clinical research: applications to practice, 2, 557-586.

30. McGregor, C. (2011, June). A cloud computing framework for real-time rural and remote service of critical care. In *Computer-Based Medical Systems (CBMS), 2011 24th International Symposium on* (pp. 1-6). IEEE.

31. McGregor, C., Catley, C., & James, A. (2011, July). A process mining driven framework for clinical guideline improvement in critical care. In *Proceedings of the Learning from Medical Data Streams Workshop. Bled, Slovenia (July 2011).*

32. Catley, C., Smith, K., McGregor, C., James, A., & Eklund, J. M. (2011). A Framework for Multidimensional Real-Time Data Analysis: A Case Study for the Detection of Apnoea of Prematurity. *International Journal of Computational Models and Algorithms in Medicine (IJCMAM)*, *2*(1), 16-37

33. Thommandram, A., Pugh, J. E., Eklund, J. M., McGregor, C., & James, A. G. (2013, January). Classifying neonatal spells using real-time temporal analysis of physiological data streams: Algorithm development. In Point-of-Care Healthcare Technologies (PHT), 2013 IEEE (pp. 240-243). IEEE.

34. Pugh, J. E., Thommandram, A., McGregor, C., Eklund, M., & James, A. (2013). Classifying neonatal spells using real-time temporal analysis of physiological data streams—verification tests. Journal of Critical Care, 28(6), e40-e41.

35. Naik, T., Bressan, N., James, A., & McGregor, C. (2013). Design of temporal analysis for a novel premature infant pain profile using Artemis. *Journal of Critical Care*, *28*(1), e4.

36. Naik, T., Thommandram, A., Fernando, K. E., Bressan, N., James, A., & McGregor, C. (2014, May). A Method for a Real-Time Novel Premature Infant Pain Profile Using High Rate, High Volume Physiological Data Streams. In *Computer-Based Medical Systems (CBMS), 2014 IEEE 27th International Symposium on* (pp. 34-37). IEEE.

37. Allen, M. C., Cristofalo, E., & Kim, C. (2010). Preterm birth: Transition to adulthood. *Developmental Disabilities Research Reviews*, *16*(4), 323-335.

38. Gucuyener, K. (2012). 173 Clues for the Neurodevelopmental Prognosis of the High Risk Preterm and Term Newborns. *Archives of Disease in Childhood*, *97*(Suppl 2), A50-A50.

39. Aranda, J. V., Beharry, K. D., Canal, D., Valencia, G. B., & Brunken, W. (2012) Molecular and clinical pharmacology of retinopathy of prematurity - Can a lifetime of blindness be prevented? Journal of Maternal-Fetal and Neonatal Medicine, 25(S2), 1

40. Visser, L., Singh, R., Young, M., Lewis, H., & McKerrow, N. (2013). Guideline for the prevention, screening and treatment of retinopathy of prematurity (ROP): guideline. *South African Medical Journal*, *103*(2), 116-125.

41. Dennery, P. A. (2010). Oxygen administration in the care of neonates: a double-edged sword. *Chinese Medical Journal (English Edition)*, *123*(20), 2938.

42. Jorge, E. C., Jorge, E. N., & El Dib, R. P. (2013). Early light reduction for preventing retinopathy of prematurity in very low birth weight infants. *Status and date: New search for studies and content updated (no change to conclusions), published in*, (8).

43. York, J. R., Landers, S., Kirby, R. S., Arbogast, P. G., & Penn, J. S. (2004). Arterial oxygen fluctuation and retinopathy of prematurity in very-low-birth-weight infants. *Journal of perinatology*, *24*(2), 82-87.

44. Hua, S. D., Chen, Y. Q., Dong, J. Y., Kong, X. Y., & Feng, Z. C. (2009). Screening and risk factors analysis of retinopathy of prematurity. *Zhonghua er ke za zhi. Chinese journal of pediatrics*, *47*(10), 757-761.

45. Stenson, B. J. (2013). Oxygen targets for preterm infants. *Neonatology*, *103*(4), 341-345.

46. Hartnett, M. E., & Lane, R. H. (2013). Effects of oxygen on the development and severity of retinopathy of prematurity. *Journal of American Association for Pediatric Ophthalmology and Strabismus*, *17*(3), 229-234.

47. Patz, A., Hoeck, L. E., & De La Cruz, E. (1952). Studies on the effect of high oxygen administration in retrolental fibroplasia. I. Nursery observations. *American journal of ophthalmology*, *35*(9), 1248-1253.

48. Kinsey, V. E., Jacobus, J. T., & Hemphill, F. M. (1956). Retrolental fibroplasia: cooperative study of retrolental fibroplasia and the use of oxygen. *AMA archives of ophthalmology*, *56*(4), 481-543.

49. Kinsey, V. E., Arnold, H. J., Kalina, R. E., Stern, L., Stahlman, M., Odell, G., ... & Patz, A. (1977). PaO_2 levels and retrolental fibroplasia: a report of the cooperative study. *Pediatrics*, *60*(5), 655-668.

50. Lanman, J. T., Guy, L. P., & Dancis, J. (1954). Retrolental fibroplasia and oxygen therapy. *Journal of the American Medical Association*, *155*(3), 223-226.

51. Anderson, C. G., Benitz, W. E., & Madan, A. (2004). Retinopathy of prematurity and pulse oximetry: a national survey of recent practices. *Journal of perinatology*, *24*(3), 164-168.

52. Sola, A., Chow, L., & Rogido, M. (2005). Retinopathy of prematurity and oxygen therapy: a changing relationship. *An Pediatr (Barc)*, *62*(1), 48-63.

53. STOP-ROP Multicenter Study Group. (2000). Supplemental Therapeutic Oxygen for Prethreshold Retinopathy of Prematurity (STOP-ROP), a randomized, controlled trial. I: primary outcomes. *Pediatrics*, *105*(2), 295-310.

54. Triven Bashambu, M., Bhola, M., & Walsh, M. (2012). Evidence for oxygen use in preterm infants. *Acta Paediatrica*, *101*(s464), 29-33.

55 Chen, M. L., Guo, L., Smith, L. E., Dammann, C. E., & Dammann, O. (2010). High or low oxygen saturation and severe retinopathy of prematurity: a meta-analysis. *Pediatrics*, *125*(6), e1483-e1492.

56 Arora, V., Cayabyab, R., Durand, M., & Ramanathan, R. (2013, January). GRADED OXYGEN SATURATION TARGETS FOR PREMATURE INFANTS IN RELATION TO OUTCOMES. In *JOURNAL OF INVESTIGATIVE MEDICINE* (Vol. 61, No. 1, pp. 205-205). 530 WALNUT ST, PHILADELPHIA, PA 19106-3621 USA: LIPPINCOTT WILLIAMS & WILKINS.

57. Kaur, C., Rathnasamy, G., & Ling, E. A. (2013). Roles of activated microglia in hypoxia induced neuroinflammation in the developing brain and the retina. *Journal of Neuroimmune Pharmacology*, *8*(1), 66-78.

58. International Committee for the Classification of Retinopathy of Prematurity. (2005). The international classification of retinopathy of prematurity revisited. *Archives of Ophthalmology*, *123*(7), 991

59. Jefferies, A. L. (2010). Retinopathy of prematurity: Recommendations for screening. *Paediatrics & child health*, *15*(10), 667.

60. Wilson, C. M., Wong, K., Ng, J., Cocker, K. D., Ells, A. L., & Fielder, A. R. (2012). Digital image analysis in retinopathy of prematurity: a comparison of vessel selection methods. *Journal of American Association for Pediatric Ophthalmology and Strabismus*, *16*(3), 223-228.

61. Reynolds, J. D., Dobson, V., Quinn, G. E., Fielder, A. R., Palmer, E. A., Saunders, R. A., ... & Tung, B. (2002). Evidence-based screening criteria for retinopathy of prematurity: natural history data from the CRYO-ROP and LIGHT-ROP studies. *Archives of ophthalmology*, *120*(11), 1470-1476.

62. Koh HC, Tan G. (2005) Data mining applications in health care. J Health Inf Manag. 2005 Spring; 19 (2):64-72.PMID: 15869215

63. Termote, J. U. M., Schalij-Delfos, N. E., Wittebol-Post, D., Brouwers, H. A. A., Hoogervorst, B. R., & Cats, B. P. (1994). Surfactant replacement therapy: a new risk factor in developing retinopathy of prematurity? *European journal of pediatrics*, *153*(2), 113-116

64. Hauspurg, A. K., Allred, E. N., Vanderveen, D. K., Chen, M., Bednarek, F. J., Cole, C., ... & Dammann, O. (2010). Blood gases and retinopathy of prematurity: the ELGAN Study. *Neonatology*, *99*(2), 104-111.

65. DeJonge, M. H., Khuntia, A., Maisels, M. J., & Bandagi, A. (1999). Bilirubin levels and severe retinopathy of prematurity in infants with estimated gestational ages of 23 to 26 weeks. *Journal of Pediatrics*, *135*(1), 102-104.

66. .Yeo, K. L., Perlman, M., Hao, Y., & Mullaney, P. (1998). Outcomes of extremely premature infants related to their peak serum bilirubin concentrations and exposure to phototherapy. *Pediatrics*, *102*(6), 1426-1431.

67. Balladini, J., Rozas, C., Frati, F. E., Vicente, N., & Orlandi, C. (2015). Big data analytics in intensive care units: challenges and applicability in an Argentinian hospital. *Journal of Computer Science & Technology*, *15*.

68. Baby, K., & Ravikumar, A. Big Data: An Ultimate Solution in Health Care.

69. Flynn, J. T., & Bancalari, E. (2000). On "Supplemental Therapeutic Oxygen for Prethreshold Retinopathy of Prematurity (STOP-ROP), a randomized, controlled trial. I: Primary outcomes" J AAPOS 2000 Apr:4(2):65-6.