

**A Structured Approach to Assessing the Suitability of Clinical Gold Standards for
use in Computational Decision Support Algorithm Development.**

by

J. Edward V. Pugh BSc, MBBS, FRCPC

A thesis submitted to the
School of Graduate and Postdoctoral Studies in partial
fulfillment of the requirements for the degree of

Master of Health Science in Health Informatics

Faculty of Health Sciences

University of Ontario Institute of Technology (Ontario Tech University)

Oshawa, Ontario, Canada

January 2022

© J. Edward V. Pugh, 2022

THESIS EXAMINATION INFORMATION

Submitted by: **J. Edward. V. Pugh**

Master of Health Science in Health Informatics

Thesis title: A Structured Approach to Assessing the Suitability of Clinical Gold Standards for use in Computational Decision Support Algorithm Development.

An oral defense of this thesis took place on June 25th, 2021 in front of the following examining committee:

Examining Committee:

Chair of Examining Committee	Dr Manon Lemonde
Research Supervisor	Professor Carolyn McGregor AM
Examining Committee Member	Dr Paul Yelder
Thesis Examiner	Dr Amandeep Sidhu

The above committee determined that the thesis is acceptable in form and content and that a satisfactory knowledge of the field covered by the thesis was demonstrated by the candidate during an oral examination. A signed copy of the Certificate of Approval is available from the School of Graduate and Postdoctoral Studies.

ABSTRACT

Advanced decision support diagnostic algorithm use in medicine is being hampered by the problem that high quality evidence is not available to support adoption. The strongest evidence for a diagnostic tool is obtained from accuracy comparison trials to a true gold standard.

Clinical Gold standards are the best tests available for a condition and frequently have high false positive rates.

This thesis presents and tests a novel methodology for reviewing clinical gold standards. Two test cases were used one for neonatal spells comparing it to an unmodified clinical gold standard, the second for late onset neonatal sepsis using a modified clinical gold standard.

Before embarking on any algorithm testing a detailed review of the clinical gold standard is needed with an expert panel of clinicians. Most clinical gold standards will need modification before being used as a comparator for advanced clinical decision support systems.

Keywords: Validation; Clinical gold standard; Neonatal Spells; Late onset neonatal sepsis

AUTHOR'S DECLARATION

I hereby declare that this thesis consists of original work of which I have authored. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I authorize the University of Ontario Institute of Technology (Ontario Tech University) to lend this thesis to other institutions or individuals for the purpose of scholarly research. I further authorize University of Ontario Institute of Technology (Ontario Tech University) to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research. I understand that my thesis will be made electronically available to the public.

The research work in this thesis that was performed in compliance with the regulations of Research Ethics Board of UOIT, Hamilton Integrated Research Ethics Board and The Research Ethics Board of the Hospital for Sick Children Toronto

UOIT REB: 12-083

Hospital for Sick Children REB: 1000036505

Hamilton Integrated Research Ethics Board: 3859 and 4833

J. E. V. Pugh

STATEMENT OF CONTRIBUTIONS

Parts of this work described in chapter 3 and 4 were contributed by other team members as listed below:-

1. Pugh JE, Thommandram A, Ng E, et al. Classifying Neonatal Spells Using Real-Time Temporal Analysis of Physiological Data Streams – Algorithm Development. In: Journal of Critical Care.; 2012:Accepted for publication.
2. Pugh JE, Thommandram A, McGregor C, Eklund M, James A. Classifying neonatal spells using real-time temporal analysis of physiological data streams—verification tests. J Crit Care. 2013;28(6):e40-e41. doi:10.1016/j.jcrc.2013.07.036
3. Choi, Y., Bressan, N., James, A., Pugh, E., McGregor, C., (2013), “Design of temporal analysis of neonatal vagal spells at different gestational ages using the Artemis' framework”, Journal of Critical Care, 28 (1) p. e4-5
4. Thommandram A, Eklund JM, McGregor C, Pugh JE, James AG. A rule-based temporal analysis method for online health analytics and its application for real-time detection of neonatal spells. Proc - 2014 IEEE Int Congr Big Data, BigData Congr 2014. 2014:470-477. doi: 10.1109/BigData.Congress. 2014.74
5. Thommandram A, Pugh JE, Eklund JM, McGregor C, James AG. Classifying Neonatal Spells Using Real-Time Temporal Analysis of Physiological Data Streams: Algorithm Development. In: 2013 IEEE Point-of-Care Healthcare Technologies (PHT), Bangalore, India: 2013, p. 240–3. <http://ieeexplore.ieee.org/document/6461329> 10.1109/PHT.2013.6461329

The spells algorithm have been published and presented at conference in the following publications. Specifically for the Spells Algorithm the design and pseudocode were my original work, however, the coding was completed by Anirudh Thommandram.

The LONS algorithm has been developed by Dr. McGregor and is appropriately referenced in the work.

The novel sepsis flow diagrams for Definite, Probable and Possible sepsis were developed jointly with members of the division of Paediatric Infectious diseases and members of the division of Neonatology, McMaster Children’s Hospital, Hamilton

Ontario and myself and are currently being written up for publication. The diagrams were prepared jointly between myself and Geoff Travis – McMaster Neonatal Research Co-ordinator. The Data for all cases for Sepsis in the Artemis database was gathered jointly by myself, Dr. Hawash Al-Onazi and Dr. Bassim Elattal (neonatal fellows in the division of neonatology).

ACKNOWLEDGEMENTS

I would like to first express my extreme gratitude to Dr. Carolyn McGregor, my primary supervisor, for her thoughtful mentorship, encouragement, understanding and patience through this process. Among all the scientists with whom I have worked, Dr. McGregor, has had the greatest impact on my development as a clinical scientist. She has imparted her genuine enthusiasm and love of computer science and data analysis and has provided me the unique guide and support to become one of the few neonatologists working in medical informatics in Canada. In addition, Dr. McGregor has worked tirelessly to create and develop the LONS algorithm which this thesis uses as a test of the methodology.

I would also like to thank Paul Yelder, for his support in preparing this thesis and for his teaching and encouragement during the courses of this Masters. A unique thank you has to be said to Anirudh Thommandram for his programming expertise in converting the spells algorithm from my original pseudocode to a real functioning algorithm and for his support over the many nights we tested it at the Hospital for Sick Children. I must also thank Dr. Andrew James for his initial support in the development of the spells algorithm and for his continued interest in my career. The spells algorithm testing would not have been possible without the sleep lab team at the Hospital for Sick Children led by Dr. Indra Narang.

The testing of the LONS algorithm would not have been possible without the amazing team from medical engineering and HITS at Hamilton Health Sciences and for their readiness to accept new technology in their hospital. Geoff Travis our fantastic

neonatal research co-ordinator at McMaster Children's hospital is also owed a debt for consenting the large number of children and support with completing all the documentation for HITS and the Hamilton integrated research ethics board.

The team from the Health Informatics Research lab and support team from Ontario Tech are also owed a huge thank you for keeping the Artemis platform up and running and collecting data with minimal downtime.

I would also like to thank my colleagues from the Division of Neonatology and Division of Infectious Diseases at McMaster Children's Hospital for their support in reviewing the cases of sepsis that did not fall within the flow diagrams.

To my family, and especially my wife Dr. Alene Toulany, thank you for your love, kindness, unconditional support, and for cheering me on from start to finish (which included the birth of two gorgeous children, buying our first home, starting two full-time academic jobs, and a major house renovation!). To my lovely children, Charles, and Alexandra, thank you for being my reminder to play. You are both growing up too fast and are amazing children.

TABLE OF CONTENTS

THESIS EXAMINATION INFORMATION.....	ii
ABSTRACT.....	iii
AUTHOR’S DECLARATION.....	iv
STATEMENT OF CONTRIBUTIONS.....	v
ACKNOWLEDGEMENTS.....	vii
TABLE OF CONTENTS.....	ix
LIST OF TABLES	xii
LIST OF FIGURES	xiii
LIST OF ABBREVIATIONS AND SYMBOLS	xiv
Chapter 1: Introduction	1
1.1 Artemis Platform	3
1.2 Research Motivation.....	6
1.3 Defining a gold / golden standard	6
1.3.1 Neonatal Spells	7
1.3.2 Late onset Neonatal Sepsis	8
1.4 Aims and objectives of the research.....	9
1.4.1 Primary objective	9
1.4.2 Secondary objective	9
1.5 Research Question.....	9
1.6 Scope of the thesis.....	10
1.7 Research Method.....	10
1.8 Thesis Overview	13
Chapter 2: Background.....	14
2.1 Medical computational algorithm testing.....	14
2.1.1 Gold and Golden Standard.....	15
2.2 Neonatal Spells.....	17
2.2.1 Apnoea.....	18
2.2.2 Computer detection.....	19

2.2.3 Polysomnography	20
2.2.4 Patterns of change of physiological variables during neonatal spells	21
2.2.5 Significance of increased spells.....	23
2.3 Heart rate variability and sepsis	24
2.4 Conclusion and Implications for Thesis of Background	26
Chapter 3: Methods	27
3.1 Research construction.....	27
3.2 Methodology	27
3.3 Methodology applied to Neonatal Spells	29
3.3.1 Algorithm Development	30
3.3.2 Verification	30
3.3.3 Pre-Validation and Refinement.....	30
3.3.3.1 Patient Selection for Validation	31
3.3.3.2 Ethical Considerations	31
3.4 Methodology applied to LONS – Late onset neonatal Sepsis.....	32
3.4.1 Clinical gold standard modification	32
3.4.2 Algorithm Development	37
3.4.3 Verification	37
3.4.4 Pre-Validation and Refinement.....	38
3.4.5 Artemis Cloud Database at McMaster Establishment	38
3.4.6 Recruitment.....	38
3.4.7 Data Collection	39
3.4.8 Data Analysis	39
3.5 Comparison of studies	40
Chapter 4: Informatics Data Analysis and Clinical Correlation.....	42
4.1 Spells Algorithm Development	42
4.1.2 Processing the RI waveform	42
4.1.3 Filtering the RI waveform to remove artefact.....	42
4.1.4 Breath finding algorithm.....	43
4.1.5 Absolute change	43
4.1.6 Relative change algorithm for heart rate and blood oxygen saturation	44
4.1.7 Combining the Relative and absolute alarms.....	46

4.1.8	Spells Algorithm Verification	46
4.1.9	Clinical Validation	48
4.2	Validation of Neonatal Sepsis algorithm.....	49
4.3	Comparison of real-world tests	51
Chapter 5: Discussion		52
5.1	Limitations.....	53
5.2	Future work	53
Chapter 6: Conclusion of Thesis.....		55
Chapter 7: References		56

LIST OF TABLES

Table 1: Constructive Research Outline	12
Table 2: Varying definitions of apnoea between studies	18
Table 3: Studies describing patterns of spells	22
Table 4: Clinical algorithm refinement suggestions	33
Table 5: Classification Matrix HRV/RRV ¹⁰²	37
Table 6: Types of Spell ¹⁰³	46
Table 7: Sepsis classification	50

LIST OF FIGURES

Figure 1: Types of Neonatal Spell ^{10,31,55,56}	17
Figure 2: Clinical gold standard assessment	27
Figure 3: Definite sepsis ¹⁰¹	34
Figure 4: Probable sepsis ¹⁰¹	35
Figure 5: Possible Sepsis ¹⁰¹	36
Figure 6: Absolute Change ¹⁰³	44
Figure 7: Relative Change Algorithm ¹⁰³	45
Figure 8: Data Viewer ¹⁰³	47
Figure 9: Verification Study Results ¹⁰³	48

LIST OF ABBREVIATIONS AND SYMBOLS

LONS – Late Onset Neonatal Sepsis

WIHRI – Women’s and Infants Hospital Rhode Island

NICU – Neonatal Intensive Care Unit

HRV – Heart Rate Variability

RRV – Respiratory Rate Variability

UOIT – University of Ontario Institute of Technology

CSF – Cerebrospinal Fluid

SickKids – Hospital for Sick Children Toronto

PICU – Paediatric Intensive Care Unit

ECG – Electrocardiogram

SOSCIP - Southern Ontario Smart Computing Innovation Platform

CAC - Centre for Advanced Computing

ORION - Ontario Research and Innovation Optical Network

AAP – American Academy of Pediatrics

AASM – American Academy of Sleep Medicine

REB – Research Ethics Board

Resp - Respiratory

CRP – C-reactive Protein

WBC – White Blood Cell count

CMV – Cytomegalovirus

CoNS – Coagulase Negative Staphylococcus Aureus

HR – Heart Rate

RI – Respiratory Impedance

Sat – Blood oxygen saturation level

Chapter 1: Introduction

With increasing amounts of computational algorithms being developed there is a clear need in the medical community for a strong methodology for testing these algorithms in the real world before adoption. Algorithm development is the first step and there are a variety of techniques that can be used for this. Verification is the second step and ensures that the algorithm functions the way it is expected to in controlled environments.

Validation is the last and most important step in developing an algorithm to assess the performance in the real world setting, in a safe way, to process real world data with environmental, data quality issues and human variation. ^{1,2}.

A gold standard is the test that is used as the exemplar of quality and correctness³. The need for gold standards to give accurate information is key in all areas of development and testing of a computer algorithm. However, the need for strong gold standards has been around before computing. Many paper clinical algorithms have been trained and evaluated the same way in the medical field⁴⁻⁶. The simplest gold standard is the laboratory test with a binary positive or negative outcome. However, because of human variation, variation in sampling, variation in testing even this may not be a perfect gold standard. In the case of a low risk disease, missing the odd case or over diagnosing a few cases may be better than the current process ⁴. But in the high-risk situation, missing a case may result in death. In addition, while over treating of a group may be the norm, unnecessary treatment can also result in significant morbidity.

In the validation step some researchers have avoided the idea of having a perfect gold standard by moving into interventional trial methodologies such as randomized control trials⁷. These trials work well where there is a forced intervention. However, in many of these studies it has been a diagnostic warning system with no forced intervention⁸. As a result, that study design does not produce the evidence necessary for strong validation. The correct methodology is a metanalysis of several trials against an agreed validated

gold standard. So, it is key that finding or validating a strong gold standard is part of any diagnostic algorithm development process⁹.

This thesis presents a methodology for developing and reviewing gold standards for computer algorithm development in high-risk, high impact contexts. It is demonstrated within the field of neonatal intensive care (NICU). The thesis outlines the development of two algorithms for deployment within a clinical decision support system known as the Artemis Platform. In the first algorithm for detecting neonatal spells, the problems of utilizing an accepted gold standard in medicine is presented. In the second, it shows how clinicians refined the commonly held clinical gold standard to enable the development of an advanced algorithm for late onset neonatal sepsis.

Neonatal spells are cardiorespiratory events where one or a combination of vital signs (heart rate, respiratory rate or blood oxygen saturation) fall outside the normal range either as a result of prematurity or a pathological event¹⁰. The spells algorithm was based on clinically known patterns to type spells in real time using feature extraction from continuous physiological bedside data. The available and only accepted gold standard was polysomnography¹¹. This had huge limitations as it cannot be performed on infants receiving ventilatory support, which is the way most infants are supported who are having neonatal spells. In addition, the output of a polysomnogram is a clinical impression of events detected in 15 channels of data¹². When examining the basis of this form of spells classification, there is significant human variation despite detailed rules in the timing duration and typing of events. Each event classification is based on a single observer impression of the data based around a wide set of rules¹².

Late onset Neonatal sepsis (LONS) is a serious medical condition where an infant has a bacterial infection in their blood stream or cerebrospinal fluid (CSF) that has started beyond 72 hours of life¹³. Prior sepsis algorithms used a single data stream of heart rate with derived heart rate variability (HRV) at its core. This approach is described in the

medical informatic literature and has been shown to be useful in identifying late onset neonatal sepsis ¹⁴. The key issue has not been the development of these algorithms or the verification that they work on selected test data. The key issue has been in their validation because late onset sepsis is more difficult to define, because not all cases are culture positive. Most authors have used the definition of sepsis as culture positive sepsis and clinical sepsis as defined by the situation where the baby was treated by the clinician for sepsis with a course of antibiotics that exceeded three days ⁷. Frequently this clinical impression in hindsight turns out to have been caused by something other than sepsis ^{13,15,16}. In this thesis a process is described to review the medical gold standard to determine whether the granularity of data and accuracy of data supplied is sufficient to meet the needs of the proposed algorithm. This method proposes that this step be undertaken at the outset of the research study process rather than at the data analysis phase.

1.1 Artemis Platform

Artemis is a high frequency, multisource, real time, health analytics platform developed through a collaboration between the University of Ontario Institute of Technology (UOIT now Ontario Tech University) and the IBM T.J Watson Research Center. It was first piloted for a case study for late onset neonatal sepsis at the Hospital for Sick Children, Toronto (SickKids) in 2009¹⁷. Artemis has been designed specifically for use in neonatal and paediatric intensive care units (NICUs and PICUs), however, its architecture would be expandable to many other areas of medicine. Artemis acquires high frequency data from bedside monitors including electrocardiogram (ECG), respiratory impedance (RI), blood pressure (BP), and pulse wave plethysmography along with the monitor derived values of heart rate, blood pressure, respiratory rate, and blood oxygen saturation (sats)¹⁸. This information can be processed and returned in real-time or stored for later analysis and research. The early iterations of the Artemis platform have been implemented to support clinical research studies within NICUs at SickKids, the Women, Infants Hospital Rhode Island (WIHRI)¹⁹ and the Children's Hospital of Fudan University (Fudan). The

instance at WIHRI enabled the testing of Artemis provision through a simplified cloud infrastructure. The WIHRI implementation used data rates up to 0.02Hz, the Fudan implementation used a locally buffered delayed cloud implementation system and did not process any data in real time²⁰. The implementation of the new Artemis Cloud at McMaster Children's Hospital enables high frequency data to be used in real time for analysis at much higher speeds²¹.

Artemis is capable of connecting to multiple medical devices used within the NICU, and converts data created by these devices to a data stream for transfer ¹⁷.

In 2010, an early iteration of the Artemis Cloud began collecting data from bedside monitors at WIHRI. In this implementation, the Artemis Cloud acquired data from the bedside every minute for storage and analysis. The platform and its data were used as part of a larger, consented study on neonatal instability ²². Artemis was used to support a Late onset neonatal sepsis (LONS) study at WIHRI to help determine the association between low HRV and LONS ²³. Within the first twelve months of implementation at WIHRI, the Artemis cloud maintained 100% service availability ²⁴. To support the implementation of the Artemis Cloud in Fudan, real time data analysis was not possible due to the hospital's infrastructure. Instead, all of the data collected was transferred to the Artemis Cloud once per day so that it could be used in retrospective analysis. A local server based version of Artemis was initially deployed in the NICU at SickKids in 2009, and has collected data from hundreds of patients to support clinical studies on late-onset neonatal sepsis, apnea of prematurity, anemia of prematurity, retinopathy of prematurity, and pain ²².

Through a partnership with the Southern Ontario Smart Computing Innovation Platform (SOSCIP) and the Centre for Advanced Computing (CAC) at Queen's University, Artemis Cloud has been developed to provide a robust real time cloud to provide Health Analytics as a Service²¹. The McMaster University Children's Hospital was the first site providing data to the new Artemis Cloud and formed the first real-time test of this new

infrastructure that has capability of storing and processing multiple channels of data provided at up to 1000hz. The Neonatal Intensive Care Unit at the McMaster Children's Hospital transfers all the data from the 51 bedside medical monitors and has the potential in the future to connect the ventilators, incubators, infusion pumps and other electronic medical devices to the cloud, where this de-identified data will be processed and stored. The new Artemis cloud leverages the secure data network of the Ontario research and innovation optical network (ORION) for connecting to the McMaster NICU proving a fast and reliable secure network to transfer de-identified high frequency data ²⁵.

The most recent instance of Artemis at McMaster children's hospital and Southlake regional hospital captures full frequency data from bedside monitors and has it linked with clinical data in a fully consented cohort of patients allowing it to be mined for a broad spectrum of conditions. At McMaster data is captured from all 51 level 3 beds²⁶.

One of the key steps in moving the platform forward from an innovative data capture platform to an advanced decision support system for clinical use is the development of well validated algorithms. The process of providing strong evidence to support the use of a computational advanced decision support algorithm is best achieved by comparison with a gold standard⁹. This process relies on the clinical gold standard having a high specificity and sensitivity. However, most times in medicine the clinical gold standard is a safe standard and specificity, and sensitivity may be quite low. This thesis sets out to describe a methodology that can be applied to clinical diagnostic problems and used to assess and refine a clinical standard to make it suitable for validation. This thesis provides a pilot assessment of this methodology by applying it to two algorithms for two conditions commonly seen in the neonatal intensive care unit. With future study this methodology may form a basis for the future validation of many new advanced decision support algorithms in medicine.

1.2 Research Motivation

The original research proposed to validate a novel algorithm for neonatal spells, but during the course of the progression of the initial design and development it became apparent that real-world validation in medicine held significant challenges.

In addition, at the same time the initial design of the spells algorithm was being performed, other members of the Artemis project team were working on sepsis. The real-world validation and information kept proving to be the roadblock because there were no clear gold standards for the neonatal conditions we were studying.

Upon further investigation, many other groups, particularly those working on the problem of automated image recognition for radiology, have found the same issue as noted earlier with spells annotation that there is significant inter observer variation in each report. What may look like a pneumonia to one radiologist may look like lung collapse to another ²⁷. This has led many image recognition groups to move away from using artificial intelligence to diagnose x-rays and move to the easier tasks of collating x-ray image banks, improved voice recognition for radiology reports and enhancing areas of images for radiologist review ²⁸.

These challenges with the methods of diagnosis motivated the need to revisit the diagnostic process of late onset neonatal sepsis with the goal to establishing a new gold standard or a methodology to develop a significantly improved clinical standard.

1.3 Defining a gold / golden standard

The challenge with the gold standard is that the naming is misleading. A better naming would be the golden standard. The gold standard refers to the time when so many pounds

of a currency were equal to that weight in gold. There is no ambiguity between a one-pound note being equal to the value of one pound (lbs) of gold. Since exiting the gold standard, the value of pound is no longer so straight forward.

In science, other gold standards are also used such as the definition of a meter which initially was a platinum bar with markings, then many platinum iridium bars with markings held in various parts of the world and now a meter is defined by interference measures of the speed of light.

In medicine the term gold standard has not been applied with such rigor as in science and the term golden standard has frequently been usurped by gold standard. The golden standard is the best test currently available for diagnosing a condition³. However, if you want to improve on the accuracy or even meet the accuracy of the current best test, you need a true outcome to learn from. In medicine, this is frequently performed by clinical acumen and so is open to massive variability dependent on the individual clinicians²⁷. When the outcome is common, the best-known method is to use large populations and population statistics to iron out individual biases. This is of little use when you are utilizing a rare event or individual events for the purposes of training a computational algorithm. For this type of investigation, a better gold standard is required.

1.3.1 Neonatal Spells

Neonatal spells are a common occurrence in all neonatal intensive care units (NICUs) and affect nearly all infants born <1000g²⁹. An increase in the frequency or severity of the neonatal spell may be an early indication of the development of sepsis, seizures, lung collapse or other serious pathology. Most research has focused on understanding the cause of these complex clinical events and how to treat them¹⁰.

Currently neonatal units rely on the clinical skills of bedside staff to detect, type, and note increased frequency of neonatal spells. Overall this technique has been shown to produce

a serious underestimation of events³⁰. The only technique that has proven reliable to determine the various forms of neonatal apnea is polysomnography. Polysomnography employs multiple sensors to assess infants breathing, movement, heart rate, and neuronal electrical activity while being actively observed using either video or a dedicated technician during sleep. In an environment where infants have spells due to the immaturity of their respiratory centers alone this is impractical to use in the neonatal clinical environment except for the most challenging of cases^{10,12}.

In the past neonatal clinicians have attempted to type spells manually using long printed strips of waveforms, however, this is labor intensive and is not practice with the current clinical work load³¹. A few researchers have looked at how to better detect and type spells using manual and automated interpretation of single streams of physiological data³¹⁻³³. One or two of these researchers have combined a detailed analysis of a single stream with threshold alarms that are commonly used in the NICU³². Unfortunately, these techniques have yet to get beyond the research stage.

1.3.2 Late onset Neonatal Sepsis

Premature infants are very susceptible to infectious pathogens³⁴. Early diagnosis of sepsis is important because infants are often diagnosed only when seriously ill which decreases the probability for prompt, complete recovery with antibiotic therapy³⁵. Diagnosing neonatal sepsis is a challenging problem because the signs are often nonspecific^{36,37} with no clinical or biochemical markers that comprise an accepted 'gold standard' for sepsis detection³⁸. LONS, which is the focus of this study, occurs in approximately 10% of neonates and 25% of very low birth weight infants hospitalized in NICUs³⁷. In 2001, Griffin and Moorman concluded that patients that developed sepsis had reduced heart rate variability (HRV) and short heart rate decelerations for up to 24 hours preceding clinical deterioration³⁹. These findings indicate that subtle changes, which may not be apparent

through manual recordings at regular intervals, can be important in detecting the onset of sepsis in neonates ²³.

1.4 Aims and objectives of the research

To develop a methodology for reviewing the current medical golden standard and assessing whether it provides a suitable outcome to develop and test a computational diagnostic algorithm. To perform pilot testing of this methodology on the two clinical problems of sepsis and neonatal spells.

1.4.1 Primary objective

Develop a methodology for reviewing the accepted golden standard and assessing whether it provides sufficiently accurate output for use to develop computational algorithms.

1.4.2 Secondary objective

Apply this methodology for the Development and testing of preliminary algorithms for identifying late onset Neonatal sepsis and identifying and typing Neonatal spells.

1.5 Research Question

Can a standard methodology be used to assess the accepted medical golden standard to make it suitable for computational diagnostic algorithm development?

Can this method be applied to identifying late onset Neonatal sepsis and identifying and typing Neonatal spells?

1.6 Scope of the thesis

The scope of this thesis is to propose and pilot a new methodology to assess the medical golden standard before algorithm development as a means to make sure the standard is strong enough to train, verify and validate a new computational algorithm. This methodology will be applied to the examples of neonatal spells and late onset neonatal sepsis as test cases. The methodologies that can be used to strengthen a medical golden standard will be examined. The potential implications of these methods and the likely cost impacts will be discussed.

The Artemis team led by McGregor will go on and further develop the Neonatal algorithms for spells and late onset neonatal sepsis that have been used as the test cases. The validation of the neonatal spells algorithm and the development of the late onset sepsis algorithm will be discussed in future work.

1.7 Research Method

Constructive Research methodology was used in this thesis to examine the problem of defining gold standards for validating medical early warning and diagnosis systems. For the purposes of this thesis it resulted in the development of a proposed methodology for examining and refining the clinical gold standard and adapting to make it useful for training and validating a clinical early warning system⁴⁰.

The constructive approach divides the research process into the following phases ⁴¹:

- 1. Find a practically relevant problem, which also has research potential.*
- 2. Obtain a general and comprehensive understanding of the topic.*
- 3. Innovate, i.e., construct a solution idea.*

4. *Demonstrate that the solution works.*
5. *Show the theoretical connections and the research contribution of the solution concept.*
6. *Examine the scope of applicability of the solution.*

A summary of the Constructive Research Phases as they relate to the creation of the contribution of this thesis is presented in the table below.

Constructive Research Phases	Artemis Constructive Research
Find a practically relevant problem which also has research potential	The development and adoption of advanced decision support diagnostic algorithms in medicine is being hampered by the problem that high quality evidence is yet to be available to support their use
Obtain a general and comprehensive understanding of the topic	<p>Disciplinary understanding of medical computational algorithm testing</p> <p>Disciplinary understanding for the diagnosis of neonatal spells</p> <p>Disciplinary understanding for the diagnosis of late onset neonatal sepsis</p> <p>Disciplinary understanding of the current state of research for physiological data behaviours prior to the clinical suspicion of Late Onset Neonatal Sepsis (LONSs) as an initial clinical test case.</p> <p>Disciplinary understanding of the current state of research for physiological data behaviours during various spells as an initial clinical test case.</p>

Innovate, i.e., construct a solution idea.	Methodology to assess the medical golden standard before algorithm development as a means to make sure the standard is strong enough to train, verify and validate a new computational algorithm
Demonstrate that the solution works.	<p>Test solution within the context of a case study for neonatal spells at The Hospital for Sick Children, Toronto, Ontario</p> <p>Test solution within the context of a case study for late onset neonatal sepsis at McMaster Children’s Hospital, Hamilton, Ontario</p>
Show the theoretical connections and the research contribution of the solution concept.	The informatics contribution of this thesis is the pre-study investigation framework of the golden standard for a computational algorithm to be trained, verified, or validated against
Examine the scope of applicability of the solution.	Proposed methodology is applicable to assess the medical golden standard before algorithm development for other medical conditions within and beyond neonatology.

Table 1: Constructive Research Outline

1.8 Thesis Overview

The following chapters of this thesis will lay out a literature review of what is known about gold standards and training data in computational algorithms in the informatics literature. The clinical background to late onset neonatal sepsis and what is known about heart rate variability as an early sensitive marker are then presented. The patterns observed and medical knowledge of feature detection in neonatal spells is outlined. The methodology is detailed and the development of the two algorithms and their real-world testing against a standard is then described. The gold standard in the case of spells and a novel adapted gold standard in the case of LONS are presented. The result section outlines the two real world tests of these algorithms. The final chapter will undertake a discussion and proposal of a methodology for assessing and adapting gold standards. It will also layout future work that should be completed before this is adopted as a standard practice for computational algorithm development in the medical field.

Chapter 2: Background

This chapter presents background context in support of this thesis as part of phase 2 of the Constructive Research process. The background is divided into three sections. The first section reviews what is known about advanced decision support system testing, and the development of evidence for clinical use. This section makes the case for the need for the development of strong clinical gold standards and provides the background for the methodology used in this thesis.

The latter two sections in this background provide the necessary background for the two cases studies that provided the testing ground for the methodology. Section two describes the literature to support the development of an algorithm for detecting and typing neonatal spells. Section three describes the background behind the Late Onset Neonatal Sepsis algorithm and the relevant physiological and clinical challenges in diagnosing sepsis.

2.1 Medical computational algorithm testing

Medical computational algorithms are developed by utilising one of the many methodologies available such as: - rule based algorithms, supervised learning, or unsupervised learning for example. To assess these algorithms, the first step is verification. Verification is the process of taking a test set of data and confirming the algorithm is working well to detect events as designed. This step has been demonstrated in many studies and frequently is one of the ways algorithms are tested before being released⁴²⁻⁴⁵.

Validation is a key step for a clinical algorithm to gain acceptance in the medical field⁴⁶. However, as demonstrated through the case studies within this thesis, this is frequently one of the most challenging steps. Without this step being completed many algorithms

will never be adopted for clinical use. Some researchers have tried to do this by using standard intervention trial methodology of using randomised control trials. On the surface this seems like a secure methodology however, it is designed to test an intervention. If there is a forced intervention or forced change in management this may be an appropriate study design. Without a forced intervention the evidence such a study provides is limited. The Australian National Institute of clinical studies produced levels of evidence for various types of studies⁹ and this provides a good framework for deciding appropriate validation steps for different methodologies. The most common realm for clinical algorithms is early warning and diagnosis. For diagnostic systems the highest level of evidence is provided using a study of test accuracy by comparing it with an independent blinded comparison with a valid reference standard, among consecutive patients with a defined clinical presentation. The key piece for researchers is finding a valid reference standard. One of the biggest issues that clinical standards are rarely gold and more frequently golden³.

2.1.1 Gold and Golden Standard

Classen wrote in a commentary in the British Medical Journal in 2005 about the gold and the golden standard in medicine. Far too often what is accepted as a gold standard turns out to be a golden standard. The gold standard is a certain measure where there is almost no chance of the comparator being in question. An example of a common gold standard is the measure of 1 meter which is defined using a particular laser light travelling in a vacuum and a known speed of light. This is a unified standard and there is very little question of the length of one meter. More frequently in medicine we end up with a golden standard this is defined as the best standard available in the world and does not in fact denote a standard that is the next best thing to a fact. If a new test becomes accepted as the best way of defining something that will then become the gold standard³.

The golden standard is often a form of satisficing which is commonplace in medicine. Satisficing is defined as the good enough solution. In medicine this is often the

acceptable standard e.g. a treatment that does not incur too much risk if you are wrong but does not over treat everything so that the treatment causes harm ⁴⁷. Many authors argue that satisficing is not an acceptable standard in medicine ⁴⁸, however, when information is sparse it may be the only possibility ⁴⁷.

A comparison to a gold standard or the current method is one of the key pieces of evidence of effectiveness of a new monitor ⁴⁹. When that standard or method is very subjective or is very weak this can be challenging to know which is providing the right outcome.

One of the Epidemiological ways of overcoming this is to use population statistics to reduce noise in the outcome. By utilising a large population and large amount of data individual variations can be ruled out. One of the most effective uses of this has been in the development of self-driving vehicles⁵⁰. As a means of programming artificial intelligent computers to drive software programmers have started to take the sensor data and have the computer predict what the human would do in this situation. If you had only one human training the computer how to drive it may always make the same mistake or drive “poorly” by taking a large population the decisions are better and it removes the inter observer variability. If it is a common test in medicine this works well such as x-rays with a few hundred shot and reported with a high level of accuracy each day ⁵⁰. For much rarer events where variability in reporting is higher, these high volume statistical approaches do not work as well and a improving the clinical gold standard is the only option^{28,51}.

2.2 Neonatal Spells

The term neonatal spell indicates that the infant has had a cardiorespiratory event that has either been clinically noted by observation or more commonly has been noted due to a threshold value being breached and the infant's monitor alarming. The common neonatal alarm limits are a respiratory pause of 20 seconds, a heart rate less than 100 beats per minute⁵² or a blood oxygen saturations below a threshold commonly between 88 and 85 percent⁵³. The less frequent clinical observation is due to the fact that most infants in the NICU are kept in a darkened environment to promote neurodevelopment⁵⁴ and are not continuously observed. Neonatal spells include apneic events as well as other cardiorespiratory events (Figure 1).

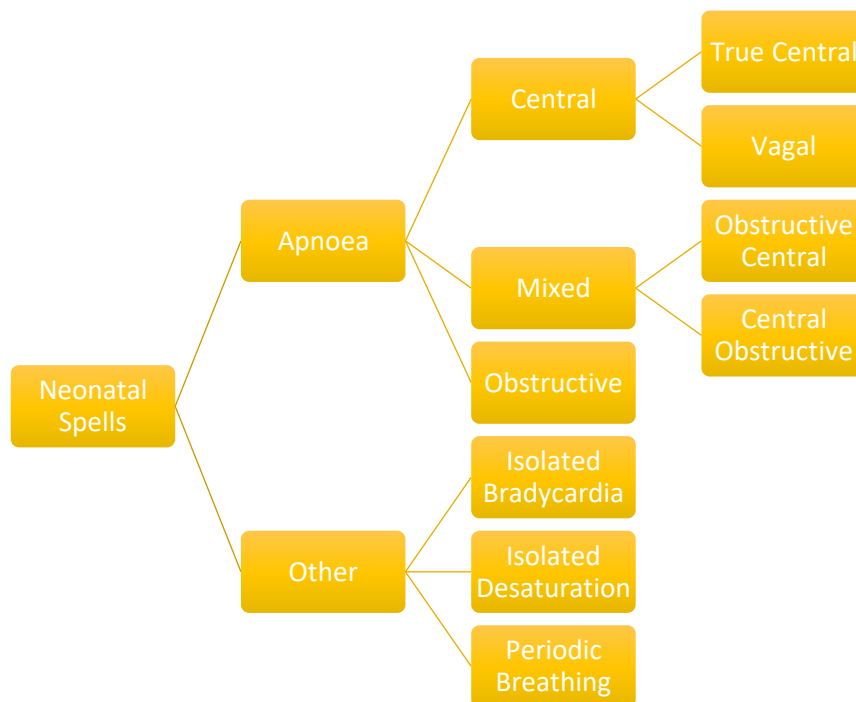


Figure 1: Types of Neonatal Spell^{10,31,55,56}

2.2.1 Apnoea

Apnoea of infancy has been defined by the American Academy of Paediatrics as an unexplained episode of cessation of breathing for 20 seconds or longer, or a shorter respiratory pause associated with bradycardia, cyanosis, pallor and or marked hypotonia⁵⁵. This definition of Apnoea is complex and difficult for automated monitoring to detect so the majority of studies use alternate definitions (Table 1).

Author	Year	Respiratory pause duration	Change in heart rate	Change in saturation
Menon et al ⁵⁶	1985	>20 seconds, short apnea > 3sec	<100	Clinical cyanosis
Finer et al ³¹	1992	>15 s	Fall by 20%	Fall of 5%
Poets et al ⁵⁷	1993	>4s	Fall by 33% for >4s	Less than 80%
Lee et al ³²	2012	>20 s or >10 s if fall in HR or sat	<100	Less than 80%
Zagol et al ⁵⁸	2012	> 10 s	<100	Less than 80%
Belal et al ⁵⁹	2011	>15 s	<80	decline
Girling et al ⁶⁰	1972	>30s	-	-
Daily et al ⁶¹	1969	>20s	-	-
Di Fiorie et al ⁶²	2010	>10s	<80	<85
Finer et al ²⁹	2006	>20s or >10s if fall in HR or sat	<80	<80-85
Mathew et al ⁶³	1982	>20s or <20s if fall in HR	<100	-
Upton et al ⁶⁴	1992	>3s	-	-
Miller et al ⁶⁵	1985	>5s	-	-
Slocum et al ⁶⁶	2009	>15s but also pauses of 10 and 5s	<85	<85
Kurlak et al ⁶⁷	1994	>20s or <20s if fall in HR or sat	<100	<90
Thach et al ⁶⁸	1979	>20s	-	-
Hayes et al ⁶⁹	2005	>10s decrease in airflow to 10% of previous	-	-
Suichies et al ⁷⁰	1989	>6s	5% fall from baseline	-

Table 2: Varying definitions of apnoea between studies

Definitions for apnoea have been influenced and restricted by the functionality and limitations of the monitoring devices used within the studies. The studies on apnoea have used a vast array of monitoring devices. The most common device used to monitor respiration is the chest and abdominal respiration band and nasal airflow sensors (table 2). It was not until the 1980ies that respiratory impedance plethysmography became standard on bedside monitors. Blood oxygen saturation monitoring also became

commercially available and commonly used at this time⁷¹. Since then most studies have used pauses in respiration with associated falls in heart rate or saturation however, in the majority this has not been a fall from baseline but rather a fall through a hard threshold value³⁰.

2.2.2 Computer detection

There have been only a few attempts at computer detection of apnoea reported in the literature. There is a clear pattern in these publications with the majority attempting to detect neonatal apnoea using a single line of physiological data. Finer et al., was one of the earliest to report on this. Finer et al. described a system that used computers to flag prolonged respiratory pauses that were associated with falls in heart rate of more than 20% and falls in blood oxygen saturations by more than 5%. These events were then manually reviewed. The goal of the study was to see if from the ECG signal change, commonly seen associated with respiratory pauses, it was possible to determine heart rate changes that would identify obstructive apnea. Overall it was concluded that nasal airflow monitoring was still required to identify obstructive apnoea³¹.

Belal et al. 2011 used three statistical methods on retrospective recorded data to look for pauses in respiration, heart rate and blood oxygen saturation, the overall technique appeared promising taking in the complex definition of apnoea however, the validation was against an unvalidated automated mark up of human physiological data and all episodes of apnoea were classed as apnoea of prematurity and no attempt was made to distinguish the type of apnoea⁵⁹.

Lee et al. in 2012 went back to a single stream approach and combined this with threshold breaches of heart rate and blood oxygen saturation to detect apnoea. A complex statistical method was used on the respiratory impedance trace to remove ECG artifact. The data was then further smoothed and respiratory pauses were detected. In

this study a respiratory pause of 20 seconds alone or a pause of 10 seconds associated with a fall in the heart rate below 100 and fall in saturations below 80 were used. Again in this study only apnoea of prematurity was considered and there was no attempts to determine the type of apnoea³².

Fairchild et al. used a system of analysis of respiratory impedance, desaturation and heart rate variability to identify apnea and had good correlation with nurse identified apnoea, bradycardia and desaturation. However, in these studies no attempt was made to identify what types of apnoea were occurring^{32,45,58,72}.

2.2.3 Polysomnography

With sufficient information the problem of detecting and accurately typing neonatal spells is possible. The process involves a clearly defined set of rules. These rules, however, do not consider the full AAP definition of apnea. With regard to respiratory pauses the American Academy of Sleep Medicine Manual looks for a pause in breathing equivalent to the time for 2 missed breaths. There are also very strict rules on blood oxygen saturation changes looking for a 3% fall and a recovery within 1% of the baseline. However, bradycardia is only defined as a fall in heart rate below 60 and does not consider the relative change described in the definition of infantile apnoea. The overall reason for this seems to be that the rules are defined for all children¹² and not specifically neonates. Currently there is no agreed standard for the changes in heart rate associated with the varying forms of neonatal apnoea⁷³. Nearly all authors writing on the topic of neonatal apnea describe the heart rate changes associated with the varying forms of apnoea^{31,74-76}. These limitations in the AASM manual offer up fluidity in the definition and area for reporter interpretation.

2.2.4 Patterns of change of physiological variables during neonatal spells

As previously noted, the modern-day bedside monitoring that is available in NICUs has mostly only been available since the early to mid 1980ies. As a result, many of the studies that describe the patterns do so with a varying array of monitoring equipment and sensors Table 3.

Paper	Age range	No. subjects	Purpose	Type of spell	Streams Used	Description of pattern of spell
Finer et al, 1992 ³¹	Neonates	47	Classify apnea	Central Mixed Obstructive	ECG, Sat, RI	Words and graphical
Poets et al, 1993 ⁵⁷	Neonates	80	Pattern of change	Not classified	Nasal air flow, Sat, ECG	Words
Sale et al, 2010 ⁷⁴	Neonates	Review	Review	Central, Mixed, Obstructive, Periodic	ECG, Nasal air flow, RI, Sat, Abdominal band	Graphical
Beuchee et al, 2007 ⁷⁷	Neonatal Lambs	7	Vagal apnea	Vagal	ECG, BP, Chest and Abdo bands, Sats	Graphical
Cohen et al, 1986 ⁷⁶	Neonates	22	Investigate airway occlusion	Central, Central Mixed, Obstructive Mixed, Obstructive	Chest and abdo band, Nasal airflow, Eosophageal pressure, mask pressure, EEG, EOG, Chin EMG	Graphical
Di Fiore et al, 2001 ⁷⁸	Neonates	68	Comparison of apnea events of those on monitoring and those off monitoring	Apena, Brady Desats	Chest and abdo expansion Sat, ECG,	Words
Girling et al, 1972 ⁶⁰	Neonates	8	Changes in HR, BP, PP during apnea	Apnea, Brady, Periodic breathing	ECG, RI, Saline pressure transducer	Words, Graphical
Kurz, 1999 ⁷⁹	Neonates	13	Influence of CPAP on Apnea	Central, mixed, obstructive apnea, bradycardia,	Nasal air flow Polygraphic – not defined	Words – not clearly defined
Martin et al 2005 ⁸⁰	Neonates	Review	Pathways of Apnea			N/A

Mathew et al 1982 ⁶³	Neonates	9	Investigate pharyngeal airway obstruction in apnea	Central, Mixed, and obstructive	ECG, Chest and abdo expansion, pharyngeal pressure, respiratory effort, nasal airflow	Graphical
Menon et al, 1985	Neonates	10	Investigate gastroesophageal reflux and apnea	Central, Mixed, Obstructive	Nasal airflow, Oral EtCO ₂ , abdo band, pharyngeal pH	Words, Graphical

Table 3: Studies describing patterns of spells

There are two types of central apnoea: vagal and true central as noted in Figure 1. The vagal form involves a cessation of respiration followed by an almost immediate fall in heart rate and a rapid decline in blood oxygen saturation⁸¹. The true central form involves a cessation of respiratory movement followed by a slow fall in heart rate and then a decline in blood oxygen saturation. Central apnoea may also occur without changes in heart rate or saturation³¹.

Obstructive apnoea is the least common of the three types of neonatal apnoea. These are characterised by continued efforts by the infant to breath but with no airflow as the airway is obstructed¹⁰. It is depicted on bedside monitors by an increase in respiratory variability an increased amplitude on the respiratory impedance trace and an increase in heart rate and a gradual fall in blood oxygen saturation³¹.

There are two distinct types of mixed apnoea as noted in figure 1: (a) central obstructive — central apnea leading to obstructive apnoea and (b) obstructive central — obstructive apnoea leading to central apnoea. The features of central obstructive apnoea are an initial cessation of respiration, a gradual decline in heart rate and a fall in blood oxygen saturation. This is followed by sudden breathing movements followed by a more rapid decline in blood oxygen saturation and a further fall in heart rate. The obstructive central events show an increased amplitude and respiratory variability on the respiratory impedance trace with a rapid fall in blood oxygen saturation and a static or accelerating heart rate then there is a cessation of respiration a gradual decline in heart rate and a further fall in blood oxygen saturation⁶⁸.

An isolated bradycardia is a fall of the heart rate below a defined threshold that is not associated with changes in respiratory rate or blood oxygen saturation. Isolated desaturations are falls in the blood oxygen saturation that are not associated with a change in the respiratory rate or heart rate. Periodic breathing is where there are more than 3 episodes of central apnoea lasting more than 3 seconds separated by <20 seconds of normal breathing¹².

2.2.5 Significance of increased spells

Causes of spells could be any of the following, airway obstruction, impaired oxygenation, temperature instability, infection, neurological disorders, metabolic disorders, abdominal disorders, congenital heart disease, arrhythmia, maternal drugs and side effects of drugs⁸².

It is not known which types of spells are associated with the various clinical causes. It has so far not been possible due to the difficulty of recognizing and typing neonatal spells to be able to associate each type with the various pathological conditions.

2.3 Heart rate variability and sepsis

For approximately 2 decades the association between altered beat-to-beat heart rate variability and neonatal clinical deterioration has been reported ³⁹. It has been postulated and researched by Moorman et al. that these changes in heart rate variability are associated with inflammatory process secondary to sepsis ^{14,35}. In 2005 Griffin et al. published a study of infants with blood culture positive sepsis that strongly associated changes in heart rate characteristics with late onset neonatal sepsis. The same study also correlated the immature to mature neutrophil count (white blood cells) ratio, hyperglycaemia and acidosis with late onset neonatal sepsis ¹⁴. A significant amount of the early work on these algorithms focused on looking at the changes in heart rate characteristics monitoring at the time of definite sepsis i.e. bacteria seen in blood culture or cerebral spinal fluid ^{35,36,83,84}. It did not look at changes in heart rate characteristics at other times during an infant's stay, leading to limited analysis of what these changes mean other than sepsis ⁸⁵.

Late onset neonatal sepsis is often difficult to clinically diagnose as many other conditions present in a similar manner. When an infant appears unwell, with skin mottling, increased neonatal spells, acidosis, poor perfusion one of the most life-threatening things that could be the cause is sepsis. Given the speed with which septic infants can deteriorate and die the physician is obligated to perform all necessary investigations and commence treatment rapidly. Once all the investigations return negative, the clinical changes seen within 24 hours of commencing the antibiotics may be so convincing that this was sepsis that the child is diagnosed with culture negative clinical sepsis. With this approach a considerable number of cases will be treated as sepsis and may not have had sepsis. The reason for this is that many of the treatments performed to try and keep the child alive through presumed sepsis may also treat the other possible causes e.g. lung collapse; that made the child look unwell. This makes a clear-cut case for a better method or an additional method of accurately identifying septic infants^{8,13,14,35,36,39,42,43,83-89}.

Heart rate variability can be calculated in a number of ways, most of these have utilised a statistical approach, using complexity calculations, frequency analysis, time domain analysis and fractal analysis⁹⁰⁻⁹². These statistical approaches have been required to handle the large amounts time aligned physiological values generated by modern medical monitoring. One of the drawbacks of the statistical analysis over analytic time sequence⁹³ methodologies is many of them make use of small chunks of data ignoring the linear time relationship of this data^{94,95}. Heart rate variability is controlled by a mix of hemodynamic variation and the autonomic nervous system, it is unclear how much each plays a part in the activities of daily living⁹⁶. Despite this, given the time sequence nature of heart beats, using statistics that ignore the sequential data loses important information from the data⁹⁴. Despite this loss of the temporal nature of the heart rate by using small chunks of data they have demonstrated promising clinical utility in various areas of medicine in research. Currently there has been little use of them clinically because of the relatively poor sensitivity of these algorithms in real world situations⁸⁵.

The promising nature of these algorithms on heart rate, despite poor sensitivity, has led other researchers to apply similar techniques to other time-sequence physiological data. The areas that have been investigated include heart rate, blood pressure, saturation plethysmography and blood oxygen saturation. Applying variability analysis to these has also proven successful in studies investigating various clinical problems including late onset neonatal sepsis in test cases^{92,97}. It has been seen that combining various biomarkers may improve the sensitivity of variability analysis in many areas of medicine particularly late onset neonatal sepsis^{87,98}.

One remaining challenge for developing a strong algorithm for identifying sepsis is the need to have a better standard to compare new algorithms. Up until now most studies have used either blood culture positivity or clinical sepsis i.e., the clinician thought the child was septic. As described in paragraph two of this section, this is not always

reliable. This limits the training opportunity because we are working towards a golden standard and not a gold standard^{8,99}.

2.4 Conclusion and Implications for Thesis of Background

It is clear that to have strong evidence a gold standard comparator is needed. This is hard to achieve in many situations because the clinical gold standard is not sufficiently accurate for algorithm development and testing. A methodology will need to be developed to assess the adequacy of the gold standard and methods to refine the gold standard will need to be found.

The background goes on to provide the literature review for the development and testing of two clinical detection algorithms for neonatal spells and Late onset neonatal sepsis. These two clinical case studies will provide a testing ground for the novel methodology developed in this thesis for testing and real-world validation of advanced clinical detection algorithms.

Chapter 3: Methods

3.1 Research construction

This chapter presents the research construction and methodology for developing and reviewing gold standards for computer algorithm development in high-risk, high impact contexts. The 3rd phase of the Constructive Research method is used to Innovate and create the proposed methodology.

3.2 Methodology

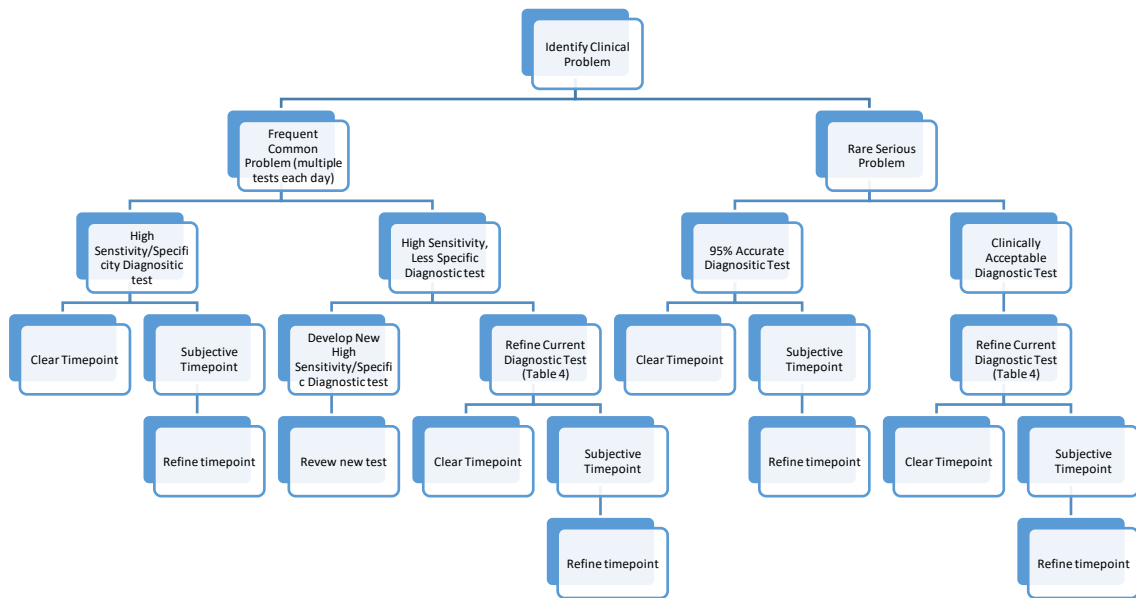


Figure 2: Clinical gold standard assessment

Once a clinical problem is identified it needs to be classified either into a *Frequent Common Problems* where multiple tests are performed each day or *Rare Serious Problems, with far fewer examples*. This classification is useful because in the case of a *Frequent Common Problem* a reasonably sensitive and specific diagnostic test can remove a few false positive or false negatives by the fact that the vast majority of tests are

correct. In the case of *Rare Serious Problems*, because these are rare the test needs to be highly accurate as one false result will have a large effect on the outcome.

Frequent Common Problems then need their diagnostic test sensitivity and specificity assessing. In the case of a highly sensitive and specific test this should be a good diagnostic test. But because of the frequency of events, it is necessary to further review this test for the temporal accuracy to correctly match up events between two diagnostic tools. The other possibility is that the test has a good sensitivity but poor specificity in this instance it is necessary to refine the clinical diagnostic test. Methods to refine the clinical diagnostic test are shown in Figure 2 and Table 4. Following this it is necessary to move onto the final process of looking at the temporal accuracy in order to correctly compare events.

Rare Serious Problems, can be divided into those with a highly accurate test >95% of events are detected accurately with few false positives and those with more common and more practical reasonable safe clinical diagnostic tests. For those conditions with a highly accurate test it is also necessary to then look at the temporal relationship of the diagnosis. In the case of *Rare Serious Problems*, because the time point is further apart between events the need for high temporal accuracy may be less. For those *Rare Serious Problems* that have a more practical reasonable safe clinical diagnostic test a review of the standard will need to be undertaken and modifications will need to be made as outlined in the diagram x and table y. Following this process the temporal accuracy also needs to be looked at to help align events between the two diagnostic systems.

In this project two algorithms were designed, one for neonatal spells and one for late onset neonatal sepsis (LONS). For each algorithm we used the same methodology of design, verification and validation.

Before commencing, each clinical golden standard was reviewed using the Methodology for problem classification above:-

Neonatal Spell events were frequent and a common problem happening multiple times each day. The test was high sensitivity and specificity and was based around a well-defined set of rules. The time point appeared clear so was thought to be a good standard for comparing the algorithm without modification.

LONS was a rare but very serious event. The diagnostic test produced an acceptable clinical standard with reasonable specificity, but large numbers of false positives, and poor sensitivity. There was a clear enough temporal time point so a group was assembled and a hybrid refinement strategy was reviewed as described below.

3.3 Methodology applied to Neonatal Spells

The research construction for the spells algorithm was broken into three distinct phases. The first phase was the algorithm development that enabled the development of a family of algorithms for the detection and classification of neonatal spells. The second phase was the verification phase to verify that the algorithm was providing the expected output. The third phase was an initial validation phase, to perform an initial test to determine how well the algorithm performed with real world collected data. To perform the validation, the output of the algorithms during this last phase were compared with the results of formal sleep studies that use polysomnography, the current, accepted clinical gold standard for identifying and classifying neonatal spells.

3.3.1 Algorithm Development

The medical team with the support and assistance of the computer science team, led by Dr. Carolyn McGregor at the Ontario Tech University, Oshawa, developed the algorithms for recognizing and characterizing the type of each neonatal spell. These algorithms were run on the Artemis framework allowing reports to be generated detailing the number, time, frequency, duration and type of neonatal spells seen within the captured physiological data.

3.3.2 Verification

Two medically trained assessors reviewed the algorithm output and the raw data to confirm that all of the medically relevant events were detected by the algorithm and classified according to the algorithm rules.

3.3.3 Pre-Validation and Refinement

Ten neonates were enrolled in this phase of the study. Each neonate had normal bedside monitoring of heart rate, blood oxygen saturation and respiratory rate. These high-fidelity data streams were obtained and captured from the bedside medical devices. The neonate also underwent a concurrent polysomnographic, sleep study. The physiological data recorded by Artemis for the study subjects was be stored for further analysis.

A comparison was made between the Artemis output and the polysomnographic data. Our goal was to achieve a 90% concordance in recognizing apnoea as well as identifying the types of apnoea.

3.3.3.1 Patient Selection for Validation

A targeted selection approach was used to recruit neonates identified by the most responsible neonatologist on service in the NICU to be having frequent spells or those referred to respiratory medicine with the problem of recurrent spells.

Any neonate admitted to the NICUs in the study or any infant referred to respiratory medicine with the problem of recurrent spells was eligible to participate in this study. We required that the neonates were capable of breathing for the duration of the study without the need for respiratory support. This requirement to breath without respiratory support was necessary as the American Academy of Sleep Medicine rules for scoring and identifying apnoea during polysomnography stipulates that the technique can only be scored for a subject not on any respiratory support¹⁰⁰. The most responsible staff physician caring for the neonate made this assessment. Study participants will be recruited from the SickKids NICU and the SickKids Sleep Laboratory.

Neonates with active infection, necrotising enterocolitis, congenital cardiovascular malformation, cardiac arrhythmia, sudden acute deterioration, or those requiring respiratory support were not eligible for enrollment in the study.

3.3.3.2 Ethical Considerations

Research ethics board approval for this study was provided by The Hospital for Sick Children (#1000036505) and Ontario Tech University (#12-083). Each subject was consented through surrogate consent from their parent/guardian. The parents of eligible neonates were approached by Dr. Pugh provided with information about the study. Explicit, informed consent was obtained from the parents of the neonates enrolled in the study.

All identifiable patient data was stored securely in accordance with both federal and provincial privacy and confidentiality of personal health information legislation.

The Artemis output was not communicated to the neonates' most responsible physicians or other clinicians involved in the neonates' care. The Artemis data was not documented in the infants' health records.

The neonates in the study underwent an additional testing process of polysomnography that they would not normally undergo. The results of the sleep study were communicated to that neonate's most responsible physician by Dr. Indra Narang within two weeks of the study being performed.

3.4 Methodology applied to LONS – Late onset neonatal Sepsis

3.4.1 Clinical gold standard modification

For the LONS algorithm the clinical team reviewed the gold standard as shown in figure 2. A simultaneous process of improving the standard needs to occur as the algorithm is being designed and verified so that validation by accuracy testing can be the next step.

A series of methodologies to improve clinical gold standard were discussed with the expert review panel

Methodology	Advantages	Disadvantages
Develop new gold standard	<ol style="list-style-type: none">1. Strong validity2. Purpose designed3. Maintains blinding of algorithm and gold	<ol style="list-style-type: none">1. Time consuming2. Requires a pre-study of new methodology3. Research Costs

	standard results until final comparison.	4. Bias introduction
Use clinical gold standard with retrospective review of outcome by group of expert clinicians	<ol style="list-style-type: none"> 1. Multiple expert opinions improves accuracy 2. Accepted standard so easier to interpret 	<ol style="list-style-type: none"> 1. Increased time 2. Larger clinical team 3. Large clinical data set required 4. More challenging REB
Give clinical output of algorithm to clinical assessor of gold standard so they can be compared in real time	<ol style="list-style-type: none"> 1. Lesser evidence given by validation 2. Quick to see if algorithm is working 	<ol style="list-style-type: none"> 1. High chance of bias 2. Difficult to assess real clinical outcome 3. Require future studies for true validation

Table 4: Clinical algorithm refinement suggestions

For LONS we used a hybrid methodology of having a series of flow diagrams for reducing the number of cases that needed review for the common clear-cut cases and then used a process of expert panel retrospective review of each case that did not fit one of the common algorithms¹⁰¹.

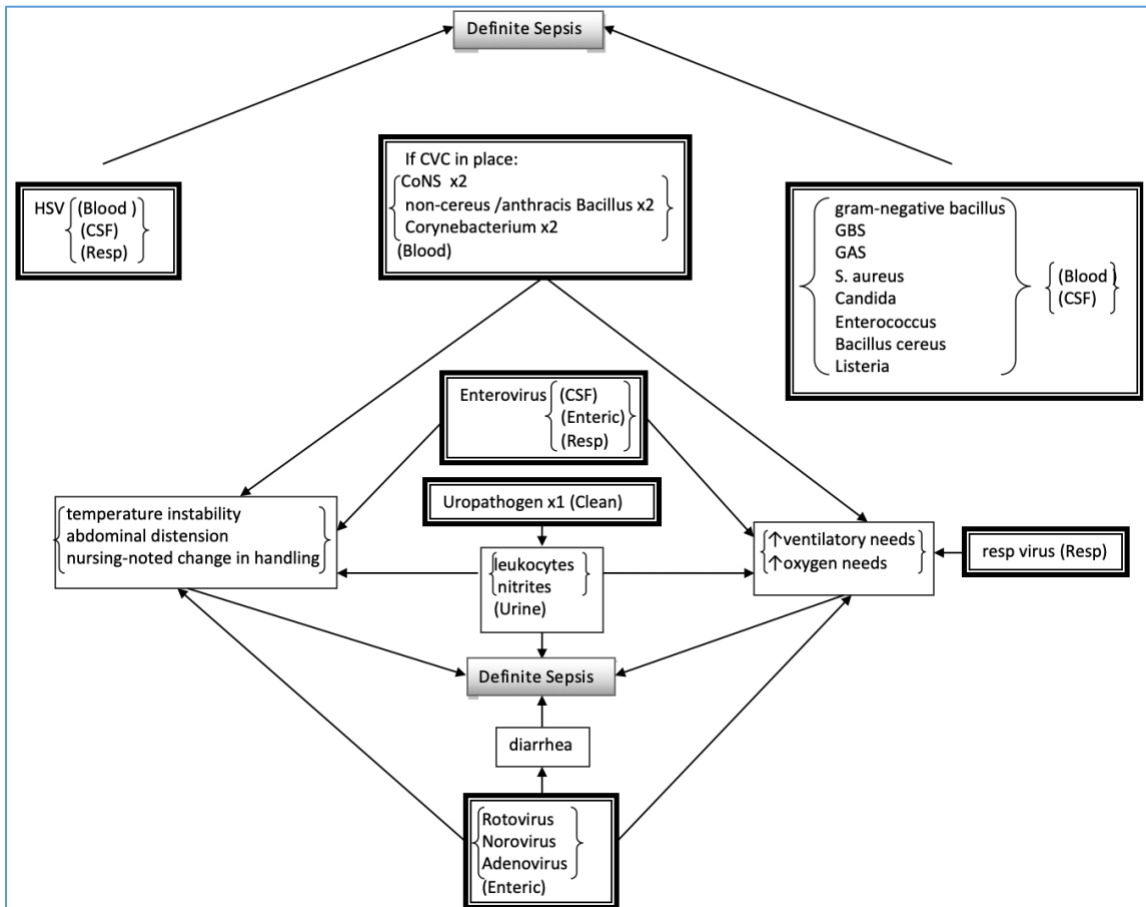


Figure 3: Definite sepsis¹⁰¹

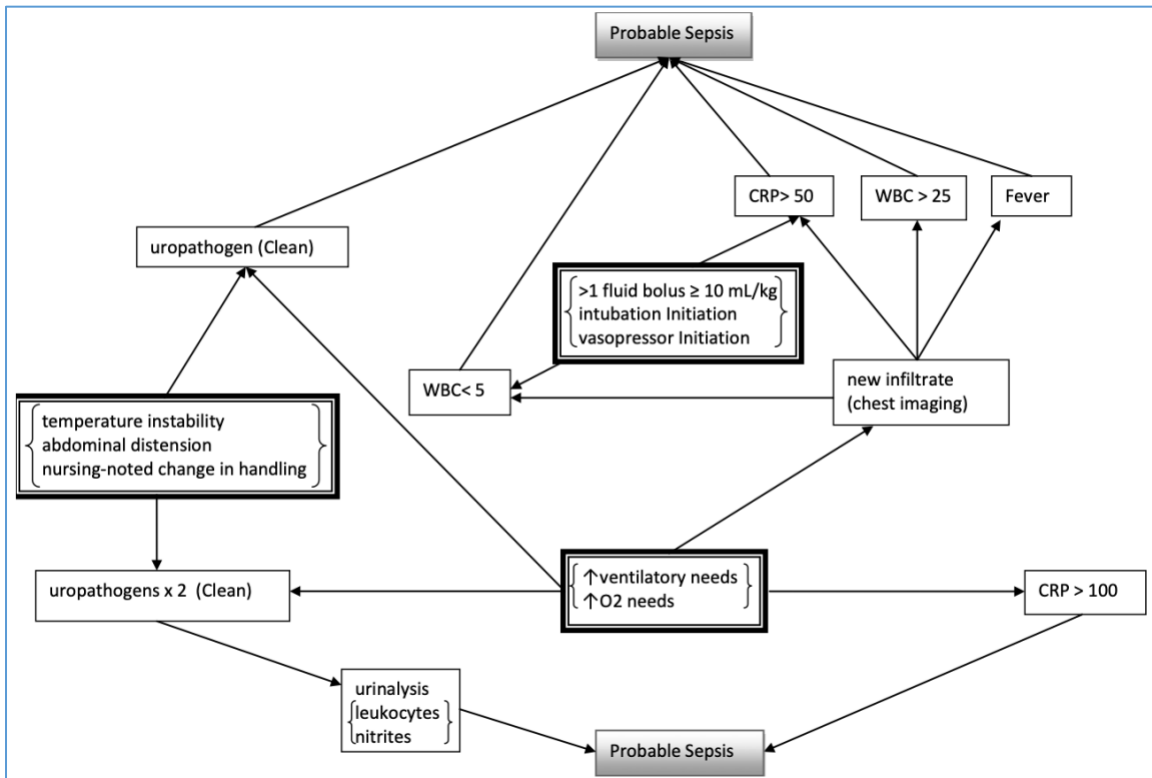


Figure 4: Probable sepsis¹⁰¹

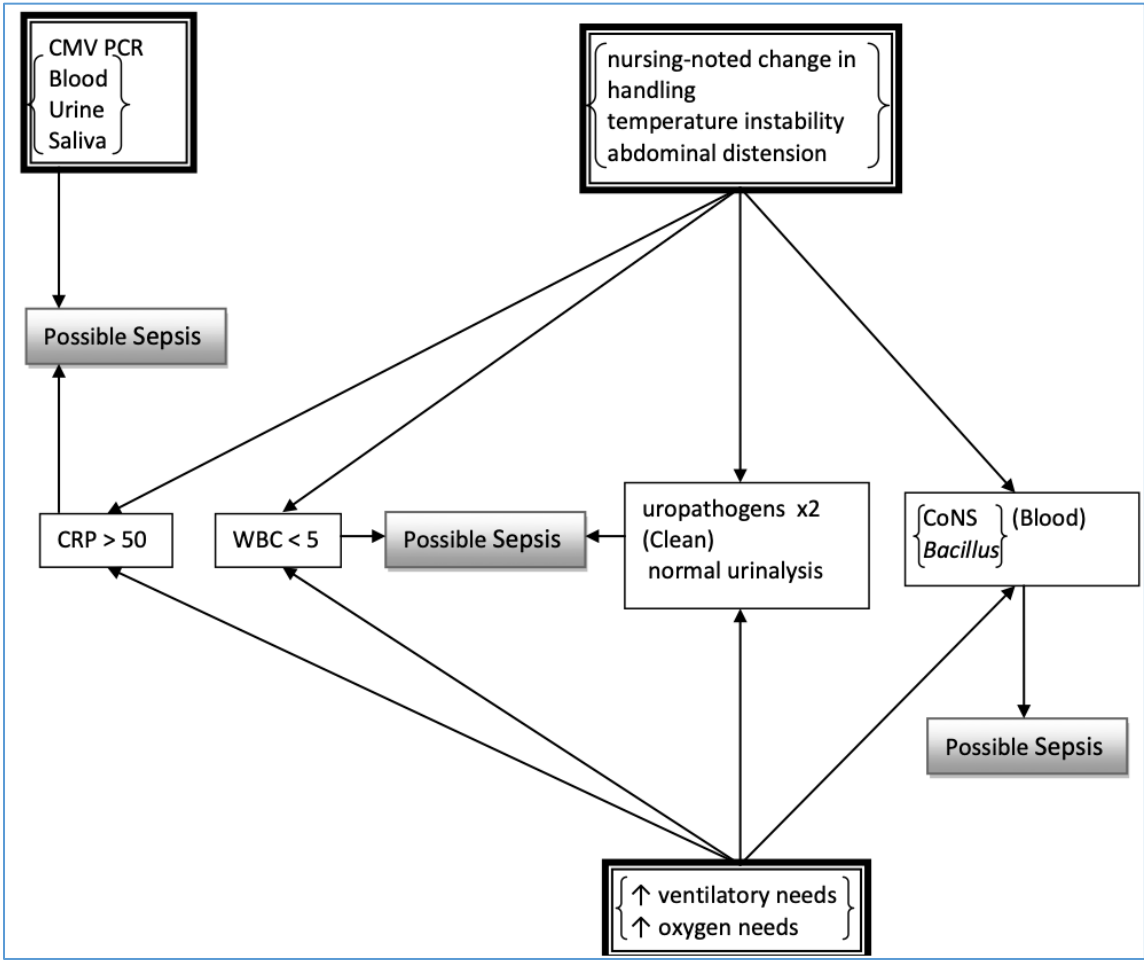


Figure 5: Possible Sepsis¹⁰¹

3.4.2 Algorithm Development

The algorithm utilized for LoNS was created through prior research¹⁰². That work proposed associations between trends in HRV and RRV and the clinical situations are presented in Table 5.

	HRV Baseline	HRV Shift Low
RRV Baseline	Negative on classifications for LONS, surgery, and narcotics or other drugs	Patients with proven LONS (LONS +)
RRV Shift Low	Uncommon	Post-operative patients receiving narcotics or other drugs (Surgery + Narcotics) Non post-operative patients receiving narcotics or other drugs (Narcotics)

Table 5: Classification Matrix HRV/RRV¹⁰²

3.4.3 Verification

Verification of the LoNS algorithm was performed during the same prior research as noted in the prior algorithm development section¹⁰²

3.4.4 Pre-Validation and Refinement

To perform pre-validation and refinement, this thesis presents results of a clinical study performed at McMaster Children's Hospital NICU. Further details on that study are presented below.

3.4.5 Artemis Cloud Database at McMaster Establishment

A database of physiological data was established that started capturing high frequency physiological data from the medical monitors of 53 Neonatal Intensive Care Beds at McMaster Children's hospital for the period of March 2018 – Jan 1 2019^{21,26}. In addition, this database also allowed for manual gathering and some electronic download of the pertinent electronic medical record data required for the projects. The details of this database are beyond the scope of this thesis. That database was used to gather all the information and provide consented approved data for the following project. The database and secondary use of data are covered by relevant approvals from the institutional research ethics boards of the Hamilton institutional Research Ethics Board (#3859 and #4833) and the Ontario Tech Research Ethics board (#14736 and #15536)

3.4.6 Recruitment

This study utilised the secondary usage of de-identified physiological data captured in the Artemis Cloud Database at McMaster. As per the approved protocol for the database, parents/guardians of infants admitted to the NICU were approached for consent to enroll their infants in the database and use of that data for this and subsequent studies. All infants admitted to the NICU were eligible to be enrolled in the database and in this study. Infants' data was only excluded from the analysis if the initial database consent is not obtained or is withdrawn by the parents/guardians. All infants received the current standard treatment in the NICU. The data gathered was not used to influence clinical management during this current study.

The inclusion criteria for the study

Infants who were:

1. Admitted to the NICU
2. With parents/guardians who have provided signed informed consent for the McMaster Children's Hospital Artemis Cloud Database.

Exclusion Criteria

1. Consent for McMaster Children's Hospital Artemis Cloud Database was not obtained or was withdrawn.

3.4.7 Data Collection

Artemis Cloud was utilized to support the acquisition, collection, transmission, real-time processing, storage and analysis on wave form and numeric physiological data streams combined with supporting clinical information including laboratory results and observations. Artemis captures physiological data streams continuously from NICU bedside devices in real-time.

3.4.8 Data Analysis

The acquired cardiorespiratory physiological information including heart rate and respiratory rate were analysed as individual signals and together.

- 1) The heart rate variability (HRV) and respiration rate variability (RRV) technique proposed in (McGregor et al, 2012) to determine HRV and RRV scores and assess the risk factor of LONS based on hard thresholds was utilized.

- 2) The algorithm output was compared against the definitions of sepsis shown in figures 3, 4 and 5 one of definite, probable and possible sepsis. These definitions were developed by a group of three paediatric infectious diseases physicians and one neonatologist. This system was developed using consensus expert opinion. If there are episodes where an infant receives antibiotics for more than 72 hours which does not fall into any of the categories below, a group of three independent neonatologist and three ID specialists will meet (Minimum quorum two neonatologists and two ID specialists) and adjudicate each case retrospectively using the full information available and long-term outcome individually to categorize the suspected sepsis. A consensus outcome of the group was used to determine the overall outcome. The possible outcomes were, definite sepsis, probable sepsis, possible sepsis and not sepsis, or other named cause e.g. pneumonia, cellulitis etc. The rate of clinically treated late onset neonatal sepsis that does not meet any of the LONS definitions provided in Appendix 1 will be calculated.

The assessment of the methodology will be demonstrated by assessing the percentage of cases of definite sepsis that were detected by the algorithm.

3.5 Comparison of studies

The spells study developed an algorithm and compared to a known medical gold standard for spells. A quantitative assessment to assess the success of this technique.

The LONS study had a detailed assessment of the standard for comparison prior to commencing the study. This study had a custom designed gold standard designed to give

the best data for development of an algorithm. A similar quantitative assessment was used for this study to assess the effectiveness of this technique.

Chapter 4: Informatics Data Analysis and Clinical Correlation

4.1 Spells Algorithm Development

Apnoea is defined by the American Academy of Pediatrics as a pause in breathing for more than twenty seconds or a pause in breathing of any duration that is associated with cyanosis, pallor, bradycardia, or marked hypotonia

From the detailed literature review of what was known about neonatal spells a series of features were identified that could be seen in the derived heart rate, blood oxygen saturations and the high frequency trace of the respiratory impedance waveform. In this thesis chapter the medical rules that were used to develop the pseudocode are described. The coding itself is beyond the scope of this thesis.

4.1.2 Processing the RI waveform

The RI waveform was captured at 65.8 Hz from the bedside monitor. It was decided from careful observation that the RI trace was highly accurate however, the monitor derived respiratory rate was frequently inaccurate. A re-interpretation of respiratory pauses was needed for accurate detection of spells.

4.1.3 Filtering the RI waveform to remove artefact

The RI waveform was assessed while observing a small series of variable gestation infant breathing in the NICU, yawns, reflux, bearing down when passing stools, crying and other events were noted to cause artifactual peaks on the RI trace that were not associated with the baby taking a breath. The amplitude of the waveforms associated with breaths was significantly different with the events that were seen to generate much smaller artifactual peaks. As a result, a simple cut filter was designed that assessed the amplitude of the waveform over a period of 10 seconds and any peaks that were less than 10% of this




amplitude were eliminated. On reviewing the effect of this by visual interpretation of the respiratory impedance waveform this was found to work quite well to get an accurate breath to breath output from the respiratory impedance trace.

4.1.4 Breath finding algorithm

The newly filtered respiratory impedance trace then used a peak finding algorithm to mark each breath. The breath-to-breath interval was calculated and averaged over 3 breaths. This rolling average was used to see when breath to breath pauses exceeded twice the average breath length. This then marked the start of an isolated respiratory pause. The breath-to-breath average was locked in until regular breathing returned and 3 new breaths had occurred.

4.1.5 Absolute change

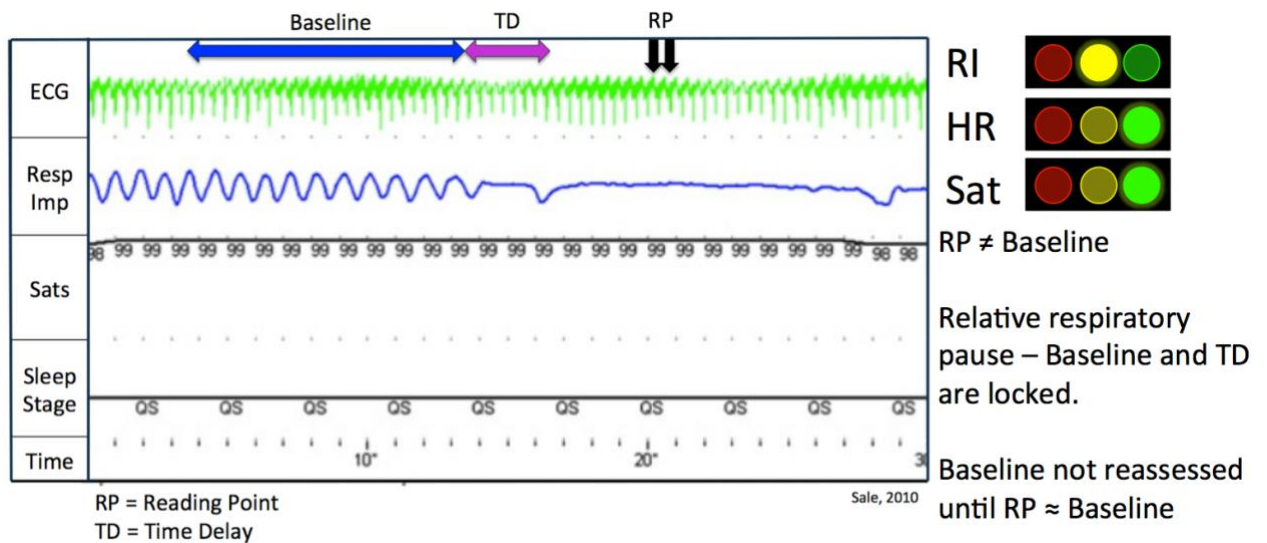
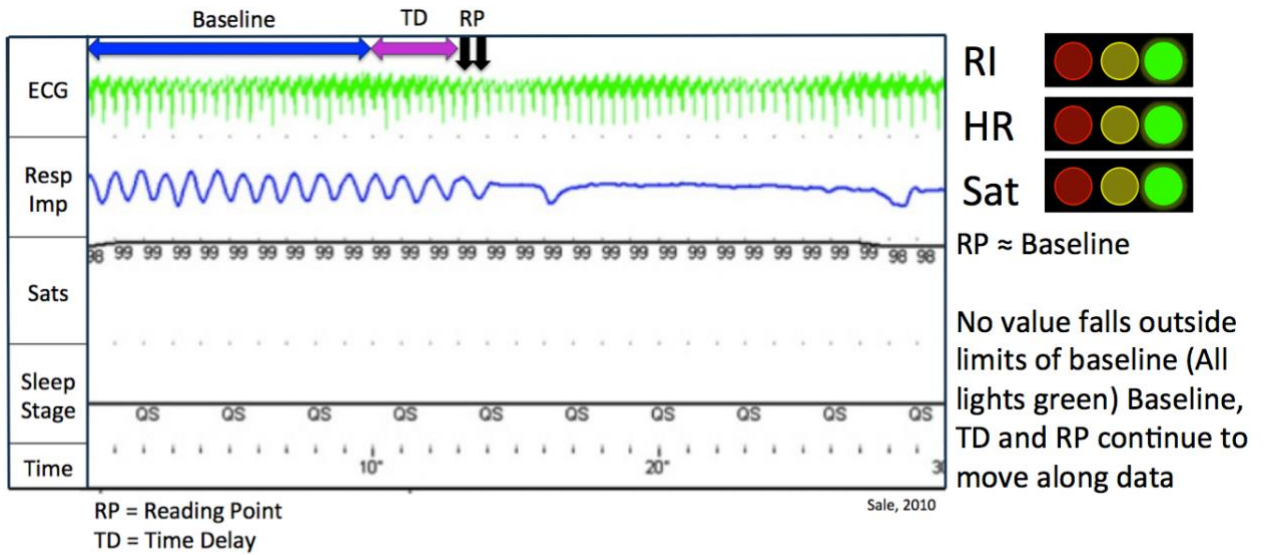
A group of fixed threshold numbers for heart rate, respiration and blood oxygen saturation are triggered an immediate alert if breached. A respiratory pause of 20 seconds or more is defined as an apnoea and causes an alert. For this we used the breath finding algorithm and if a pause exceeded 20 seconds an apnoea was logged. If HR or blood oxygen saturations fell below the thresholds shown in Fig 6, a bradycardia or desaturation was defined. If they occurred together the pattern was analyzed and a spell type was defined.

RI		Respiratory pause for ≥ 20 sec
HR		HR < 100 preterm, HR < 80 term
Sat		Sat < 85 preterm, Sat < 92 term

*Figure 6: Absolute Change*¹⁰³

4.1.6 Relative change algorithm for heart rate and blood oxygen saturation

The monitor derived heart rate and blood oxygen saturation are accurate interpretations of their physiological measures. A relative change algorithm was created with a reading window, a time delay and a baseline moving window that averaged the baseline value. These sliding windows ran through the data in real time unless a fixed percentage change or a hard value were breached Figure 7.






- RI  Respiratory pause for > 2 breaths
- HR  HR falls more than 10% of baseline
- Sat  Sat falls more than 10% of baseline

Figure 7: Relative Change Algorithm¹⁰³

4.1.7 Combining the Relative and absolute alarms

A final combining classifier algorithm was used that assesses the sequential pattern of events. If events did not fall in one of the sequential patterns in Table 1, the episode was logged as unclassified.

Type of spell	Fall from baseline			Recovery to baseline	
	HR	RR	Sats	RR	HR
Central	2	1	3	4	5
Vagal	1	1	2	3	3
Obstructive	1(Incr.)	-	2		3
Obstructive Central	1(Incr.)	3	2	4	5
Central Obstructive	2	1	3	4	5 (Fall)
Desaturation	-	-	Absolute	-	-
Bradycardia	Absolute	-	-	-	-

Table 6: Types of Spell¹⁰³

4.1.8 Spells Algorithm Verification

The process of verification is determining that the algorithm does what it is meant to do with real world data as assessed by a human evaluator against the design parameters of the algorithm. For this thesis a single human evaluator used a custom written viewer to review simultaneously the time aligned raw physiological tracing against the algorithm marked events. A single infant's trace was reviewed for a period of 24 hours which gave ample examples of the algorithm managing artifactual events such as handling, diaper changes, feeds etc.

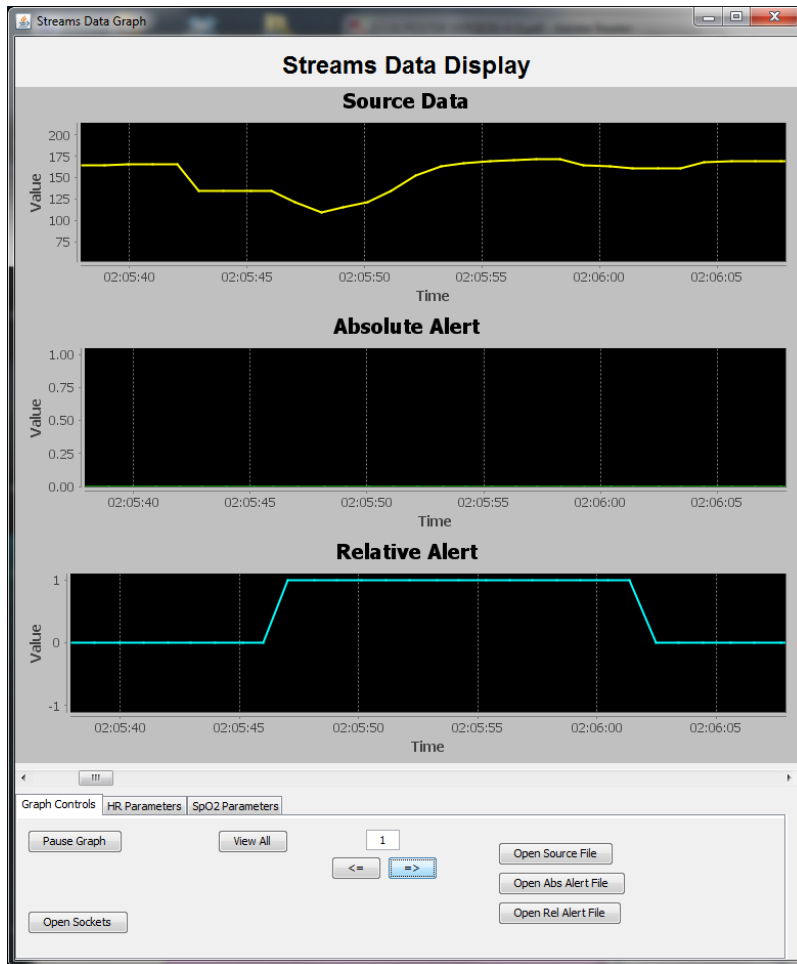
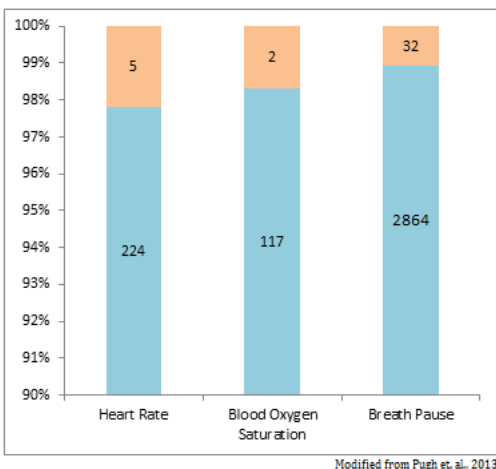


Figure 8: Data Viewer¹⁰³

A 24 hour convenient sample of data from a neonatal patient who was reported to be experiencing spells was used to evaluate the accuracy of the event detection algorithms described above. The data was processed with these algorithms and then the results were reviewed for accuracy.



Algorithm Specificity - 98.8%

Algorithm Sensitivity - 100%

*Figure 9: Verification Study Results*¹⁰³

4.1.9 Clinical Validation

After the algorithm that had been verified it was compared to the medical gold standard of polysomnography. Infants who were not on respiratory support who were known to have spells were selected and consented and given an overnight polysomnography with simultaneous standard medical monitoring with the Artemis algorithm working in a time synchronized manner. The output of these two methods were compared.

It had been our initial intention to study 20 infants with Artemis and overnight polysomnography. However, because of limitations of availability of technologists and polysomnography equipment for research a much lower number were studied.

In reviewing the data and with the technical glitches we had obtaining information and capturing the sleep studies out of the seven studies completed – there were two good overnight recordings. In the recordings there were a total of 521 events detected by Artemis and 473 events detected by the polysomnography. Temporal correlation with a 18-22

second time difference for all events was noted. The event duration from Artemis was on average 20% -50% higher than the sleep study and the typing of events did not correlate. In the polysomnography the longest central apnea was 22 seconds with the majority being 4-5 seconds. There were a total of 6 mixed apneas and 7 obstructive apneas all with durations less than a few seconds.

On manual review of the Artemis data with the scoring sleep technologist it became apparent that there is a highly subjective nature to their scoring as not all 15 information channels pick up the same information. The duration of events from one sleep technologist was much shorter than the other. It was clear that this did not impact the clinical output of the sleep studies however, it did make it somewhat difficult to utilize this medical gold standard to further refine the algorithm.

There were significant difficulties in recruitment and obtaining data and as a result, it was decided to pause the spells work and not collect further clinical studies. The work on the spells algorithm will be continued beyond this thesis when a real-time output of the algorithm is available for the sleep technologist to review and a markup system is developed to confirm correct identification during the sleep study.

4.2 Validation of Neonatal Sepsis algorithm

Data was gathered using the newly established Artemis Cloud database at McMaster Children's Hospital. Over the 9 months study duration 239 neonatal patients were enrolled.

In this cohort of patients there were 111 episodes of infants having suspected late onset neonatal sepsis (a clinical picture of sepsis that occurred beyond the third day of life). These infants received lab work and blood cultures when they appeared unwell and were commenced on empiric antibiotic therapy. They may also have received, urine, CSF or

sputum cultures as was clinically indicated. In previous studies all of these infants would have been defined as having sepsis. A far more rigorous approach was used in this study to determine true sepsis as described in the methods section earlier in this thesis. This yielded the following data set.

Sepsis Type	No. of Cases
Definite	32
Probable	17
Possible	21
Other	41

Table 7: Sepsis classification

For this thesis results presentation focuses on the Definite cases alone. In future work we will look at the Probable cases where there is a more than 50% likelihood that the infant was septic. The Possible and Other cases are clinical cases of suspected sepsis that with review are likely to have not been due to a bacterial or viral infection.

Of the cases of Definite sepsis 19 of these were present in the blood stream and or the cerebral spinal fluid (meningitis) the remaining were present in the urine. None of the cases were attributed to viruses. Of the 19 infants 3 had missing data. 12 out of 16 cases were detected by the algorithm. The 4 not detected were older more mature infants.

These 12 cases that were detected all showed up in the data more than 6 hours before sepsis was clinically suspected. Overall there was a strong correlation with the cases of definite sepsis.

4.3 Comparison of real-world tests

The first study had a medical gold standard that was based around a series of rules and should have provided unambiguous results. What became clear as the study was being completed was that there was a high degree of subjectivity in the application of these rules as 15 continuous channels of data were being gathered and it was a more subjective decision on type of apnea as a result of which information stream was weighted the most by the reading technologist. In addition, the clinical question for the patients who were being sleep studied for clinical reasons and not for research alone, impacted the reading of the studies.

The second study built on previous work by McGregor et al. that demonstrated the need to have a strong gold standard as nearly all work in the field of LONS was limited by the ambiguity between lab diagnoses sepsis and clinical sepsis. Using this far more robust approach a clear strong result was obtained and we will now be able to continue our work to further refine the algorithm to achieve the goal of 100% sensitivity with a high specificity as would be needed by a system designed to detect sepsis.

Chapter 5: Discussion

In the burgeoning world of data analytics in medicine, as a means to provide bedside early warning systems of clinical change, it is necessary to have strong testing techniques, in order for these systems to be accepted. It is relatively easy to take simulated or selected high quality noise free data and get good results^{35,83}. It is far more complex to develop a real-world testing routine for new algorithms. With this goal and to avoid developing new techniques many people have tried to use the same paradigms as are used for clinical trials of therapeutics. Unfortunately using such trials positions them at the lower end of the hierarchy of evidence. The use of a randomized control trial for diagnostic test may sound like a strong methodology but when looked at critically it is not the right choice for a study where intervention is optional. When looking at choosing the right methodology the National Health and Medical Research Council of Australia has developed levels of evidence for different type of research questions. In most cases of early warning systems this falls under diagnosis where a systematic review of many independent studies of test accuracy provide the highest level of evidence.

As the highest level evidence is provided by a systematic review of the comparison of diagnostic system against a reference standard it becomes increasingly important to review the strength of the current reference standards or if one does not exist develop one early or even before embarking on designing an algorithm. Without a robust reference standard the evidence will not be there for a real world test of the system. Without a robust validation, real world test, for medical professionals adoption of the tool will be hampered and utilization may never occur. For much of medicine these reference standards are designed to be used clinically and as such are not built for 100% accuracy as over diagnosis is often safer than not treating. The focus for most clinical reference standards is safety so there is always a focus on higher sensitivity than specificity.

For the case of spells because of the temporal association required to accurately identify events the human variation in the reporting seemed to be a huge challenge. It was not

possible from the independent standard to identify the events with certainty. The data set was also much smaller than had been planned as a result of equipment availability. The total number of events from the clinical standard and the computer algorithm were close but that was the closest correlation obtainable. Without the certainty we were comparing an algorithm detected event with an event in the polysomnography it was impossible to determine how to refine the algorithm.

For the case of late onset neonatal sepsis, previous studies have used culture positive sepsis combined with clinical sepsis (those infants thought by the clinician to be unwell and treated with antibiotics) as the reference standard. When you analyze the results of the improved reference standard (Assumption: Definite and probable sepsis cases are true sepsis) you find that this previous standard has only a 44% positive predictive value. If you either train or compare an algorithm for finding sepsis using culture positivity and clinical treatment with antibiotics you are in building an algorithm with an inherent poor predictive value.

5.1 Limitations

Both case studies were performed at a single center's this makes it difficult to conclude about the relative accuracy of clinical diagnoses. There may be wide variation center to center depending on diagnostic practices and clinical experience. In addition, both studies had different temporal time points and accuracy of time point comparison was not so critical in the LONS study as compared with the spells study.

5.2 Future work

For the LONS Study in future work additional factors will be put into the algorithm with adjustments to the algorithm to improve the specificity and sensitivity. Once complete a real-world test of the algorithm will be implemented where data will be supplied to the

clinician to act on. This will still use accuracy as the validated end point as it will be clinically hard to make intervention the endpoint because of the need for clinicians to override the algorithm when clinical concern arises.

For spells, a review of the output of the polysomnography by 3 clinicians will be required to complete the study as designed. An alternative way to do the initial evaluation of the algorithm providing lesser clinical evidence is to provide the algorithm output in an unblinded way to the technician at the bedside and then see if agreement can be achieved. If this works well returning to the original study design with 3 assessors would provide a strong validation of the algorithm as a second step.

This has been a pilot of this process for refining a clinical gold standard. Future studies with this process will be required to determine any modifications to this methodology to encompass all eventualities.

Chapter 6: Conclusion of Thesis

This thesis presented a methodology for developing and reviewing gold standards for computer algorithm development in high-risk, high impact contexts. The thesis outlined the development of two algorithms for deployment within a clinical decision support system known as the Artemis Platform within the context of neonatal intensive care.

This thesis tested the proposed methodology to achieve a functional gold standard to validate an algorithm. In this thesis we worked through the development, verification and clinical validation of two algorithms. Both of these algorithms achieved strong practical verification, using test real world data. In the process of validating the spells algorithm a computerized detection algorithm it became apparent that clinical gold standards have such high variability validation was almost impossible. With this in mind the LONS algorithm validation step was looked at with great care by a wider group of clinicians. Using a consensus retrospective approach, a strong clinical validation set to test the algorithm against was developed. From this process it became apparent that looking at this as a step in designing future studies would be key to generating successful outcomes. With future testing of this methodology to look closer at the validation step it may be possible to develop a pathway to make future studies successful.

Returning to the original research question “Can a standard methodology be used to assess the accepted medical golden standard to make it suitable for computational diagnostic algorithm development?”. It is clear with enough rigor and focus on the accuracy and temporal nature of the standard it is possible to utilise the method laid out in table 2 to assess and hence modify a clinical gold standard; making it suitable to validate advanced decision support algorithms. This is a pilot test with two algorithms and minor modifications may be required to meet all eventualities before including this in a standard methodology.

Chapter 7: References

1. Giannoulatou, E., Park, S. H., Humphreys, D. T. & Ho, J. W. K. Verification and validation of bioinformatics software without a gold standard: A case study of BWA and Bowtie. *BMC Bioinformatics* **15**, 1–8 (2014).
2. Lee, I. *et al.* High-Confidence medical device software and systems. *Computer (Long. Beach. Calif.)* **39**, 33–38 (2006).
3. Claassen, J. A. H. R. The gold standard: not a golden standard. *BMJ* **330**, 1121 (2005).
4. Mayaud, P. *et al.* Validation of a WHO algorithm with risk assessment for the clinical management of vaginal discharge in Mwanza, Tanzania. *Sexually Transmitted Infections* **74**, (1998).
5. Eriksen, B. O., Hoff, K. R. S. & Solberg, S. Prediction of acute renal failure after cardiac surgery: Retrospective cross-validation of a clinical algorithm. *Nephrol. Dial. Transplant.* **18**, 77–81 (2003).
6. Jakobsson, C. *et al.* Validation of a clinical algorithm to identify low-risk patients with pulmonary embolism. *J. Thromb. Haemost.* **8**, 1242–1247 (2010).
7. Moorman, J. R. *et al.* Mortality Reduction by Heart Rate Characteristic Monitoring in Very Low Birth Weight Neonates: A Randomized Trial. *J. Pediatr.* **159**, 900–906 (2011).
8. Fairchild, K. D. *et al.* Septicemia mortality reduction in neonates in a heart rate characteristics monitoring trial. *Pediatr. Res.* **74**, 570–575 (2013).
9. National Health and Medical Research Council. Levels of evidence and recommendation grading. *Emerg. Dep. stroke transient Ischaem. attack care bundle Inf. Implement. Packag.* 46–47 (2009).
10. Abu-shaweesh, J. M. & Martin, R. J. Apnea of prematurity : past , present and future. *Clin. Perinatol.* **2**, 63–73 (2005).
11. Di Fiore, T. Use of Sleep Studies in the Neonatal Intensive Care Unit. *Neonatal Netw. J. Neonatal Nurs.* **24**, 23–30 (2005).
12. AASM. *The AASM manual for the scoring of sleep and associated events: rules,*

terminology and technical specifications. (2012).

13. Fairchild, K. D. *et al.* Septicemia mortality reduction in neonates in a heart rate characteristics monitoring trial. *Pediatr. Res.* **74**, 570–575 (2013).
14. Griffin, M. P., Lake, D. E. & Moorman, J. R. Heart rate characteristics and laboratory tests in neonatal sepsis. *Pediatrics* **115**, 937–941 (2005).
15. Yapicioglu, H., Ozlu, F. & Sertdemir, Y. Are vital signs indicative for bacteremia in newborns?. *J. Matern. Fetal. Neonatal Med.* **28**, 2244–2249 (2015).
16. Kumar, N., Akangire, G., Sullivan, B., Fairchild, K. & Sampath, V. Continuous vital sign analysis for predicting and preventing neonatal diseases in the twenty-first century: big data to the forefront. *Pediatr. Res.* **87**, 210–220 (2020).
17. Blount, M. *et al.* Real-Time Analysis for Intensive Care. *IEEE Eng. Med. Biol. Mag.* (2010).
18. McGregor, C. "Big Data in Critical Care Using Artemis", in *Signal Processing and Machine Learning for Biomedical Big Data.* (CRC Press, 2018).
19. McGregor, C., Catley, C., James, A. & Padbury, J. Next generation neonatal health informatics with Artemis. *Stud. Health Technol. Inform.* **169**, 115–9 (2011).
20. Kamaleswaran, R. *et al.* Cloud framework for real-time synchronous physiological streams to support rural and remote Critical Care. in *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems* 473–476 (2013). doi:10.1109/CBMS.2013.6627844
21. McGregor, C. *et al.* Health Analytics as a Service with Artemis Cloud: Service Availability(). *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. IEEE Eng. Med. Biol. Soc. Annu. Int. Conf.* **2020**, 5644–5648 (2020).
22. McGregor, C. Big Data in Neonatal Intensive Care. *Computer (Long Beach, Calif).* **46**, 54–59 (2013).
23. Mcgregor, C., Catley, C., Padbury, J. & James, A. Abstract 21. *J. Crit. Care* **28**, e11–e12 (2012).
24. A Cloud Computing Framework for Real-time Rural and Remote Service of Critical Care.
25. Izaddoost, A. & McGregor, C. Enhance Network Communications in a Cloud-

based Real-time Health Analytics platform Using SDN. 3–6 (2016).

26. Inibhunu, C. *et al.* Adaptive API for Real-Time Streaming Analytics as a Service. *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. IEEE Eng. Med. Biol. Soc. Annu. Int. Conf.* **2019**, 3472–3477 (2019).
27. Swingler, G. H. Observer variation in chest radiography of acute lower respiratory infections in children: A systematic review. *BMC Med. Imaging* **1**, (2001).
28. Lakhani, P. *et al.* Machine Learning in Radiology: Applications Beyond Image Interpretation. *J. Am. Coll. Radiol.* **15**, 350–359 (2018).
29. Finer, N. N., Higgins, R., Kattwinkel, J. & Martin, R. J. Summary proceedings from the apnea-of-prematurity group. **117**, S47-51 (2006).
30. Muttitt, S. C., Finer, N. N., Tierney, A. J. & Rossmann, J. Neonatal Apnea: Diagnosis by Nurse Versus Computer. *Pediatrics* **82**, 713–720 (1988).
31. Finer, N. N., Barrington, K. J., Hayes, B. J. & Hugh, A. Obstructive, mixed, and central apnea in the neonate: Physiologic correlates. *J. Pediatr.* **121**, 943–950 (1992).
32. Lee, H. *et al.* A new algorithm for detecting central apnea in neonates. *Physiol. Meas.* **33**, 1–17 (2012).
33. Zagol, K. *et al.* Anemia, apnea of prematurity, and blood transfusions. *J. Pediatr.* **161**, 417-421.e1 (2012).
34. Ganatra, H. A., Stoll, B. J. & Zaidi, A. K. M. International perspective on early-onset neonatal sepsis. *Clin. Perinatol.* **37**, 501–523 (2010).
35. Griffin, M. P. *et al.* Abnormal heart rate characteristics preceding neonatal sepsis and sepsis-like illness. *Pediatr. Res.* **53**, 920–926 (2003).
36. Griffin, M. P., Lake, D. E., O’Shea, T. M. & Moorman, J. R. Heart rate characteristics and clinical signs in neonatal sepsis. *Pediatr. Res.* **61**, 222–227 (2007).
37. Beck-Sague, C. M. *et al.* Bloodstream infections in neonatal intensive care unit patients: results of a multicenter study. *Pediatr Infect Dis J* **13**, 1110–1116 (1994).
38. Riedemann, N. C., Guo, R.-F. & Ward, P. A. The enigma of sepsis. *J. Clin. Invest.* **112**, 460–467 (2003).

39. Griffin, M. P. & Moorman, J. R. Toward the Early Diagnosis of Neonatal Sepsis and Sepsis-Like Illness Using Novel Heart Rate Analysis. *Pediatrics* **107**, 97–104 (2001).
40. Kasanen, E., Lukka, K. & Siitonen, A. The constructive approach in management accounting research. *J. Manag. Account. Res.* **5**, 243–264 (2003).
41. Kasanen, E., Lukka, K. & Siitonen, A. The constructive approach in management accounting research. *J. Manag. Account. Res.* **5**, 243 (1993).
42. Sullivan, B. A. & Fairchild, K. D. Predictive monitoring for sepsis and necrotizing enterocolitis to prevent shock. *Semin. Fetal Neonatal Med.* **20**, 255–261 (2015).
43. Fairchild, K. D. Predictive monitoring for early detection of sepsis in neonatal ICU patients. *Curr. Opin. Pediatr.* **25**, 172–179 (2013).
44. Stone, M. L. *et al.* Abnormal heart rate characteristics before clinical diagnosis of necrotizing enterocolitis. *J. Perinatol.* **33**, 847–850 (2013).
45. Fairchild, K. D. & Lake, D. E. Cross-Correlation of Heart Rate and Oxygen Saturation in Very Low Birthweight Infants: Association with Apnea and Adverse Events. *Am. J. Perinatol.* **35**, 463–469 (2018).
46. Bakken, S. & Ruland, C. M. Translating clinical informatics interventions into routine clinical care: how can the RE-AIM framework help?. *J. Am. Med. Inform. Assoc.* **16**, 889–897 (2009).
47. Carter, B. S. & Leuthner, S. R. Decision making in the NICU--strategies, statistics, and 'satisficing'. *Bioethics forum* **18**, 7–15 (2002).
48. Grais, R. F. & Juan-Ginera, A. Vaccination in humanitarian crises: Satisficing should no longer suffice. *Int. Health* **6**, 160–161 (2014).
49. Laurent, R. T. S. Evaluating Agreement with a Gold Standard in Method Comparison Studies Author (s): Roy T . St . Laurent Published by : International Biometric Society Stable URL : <https://www.jstor.org/stable/3109761> REFERENCES Linked references are available on JSTOR . **54**, 537–545 (2019).
50. Agrawal, A., Gans, J. & Goldfarb, A. *Prediction Machines: The Simple Economics of Artificial Intelligence*. (Harvard Business Review Press, 2018).
51. Challen, R. *et al.* Artificial intelligence, bias and clinical safety. *BMJ Qual. Saf.* **28**, 231–237 (2019).

52. Aggarwal, R., Singhal, A., Deorari, A. & Paul, V. Apnea in the Newborn. *Indian J. Pediatr.* **68**, 959–962 (2001).
53. Schmidt, B. *et al.* Survival Without Disability to Age 5 Years After Neonatal Caffeine Therapy for Apnea of Prematurity. *JAMA J. Am. Med. Assoc.* **307**, 275–282 (2012).
54. Brandon, D. H., Holditch-Davis, D. & Belyea, M. Preterm infants born at less than 31 weeks' gestation have improved growth in cycled light compared with continuous near darkness. *J. Pediatr.* **140**, 192–9 (2002).
55. AAP Committee on Fetus and Newborn. Apnea, Sudden Infant Death Syndrome, and Home Monitoring. *Pediatrics* **111**, 914–917 (2003).
56. Menon, A. P., Schefft, G. L. & Thach, B. T. Apnea associated with regurgitation in infants. *J. Pediatr.* **106**, 625–629 (1985).
57. Poets, C. F., Stebbens, V. A., Samuels, M. P. & Southall, D. P. The relationship between bradycardia, apnea, and hypoxemia in preterm infants. *Pediatr. Res.* **34**, 144–147 (1993).
58. Zagol, K. *et al.* Anemia, apnea of prematurity, and blood transfusions. *J. Pediatr.* **161**, 417-421.e1 (2012).
59. Belal, S. Y., Emmerson, A. J. & Beatty, P. C. W. Automatic detection of apnoea of prematurity. *Physiol. Meas.* **32**, 523–42 (2011).
60. Girling, D. J. Changes in heart rate, blood pressure, and pulse pressure during apnoeic attacks in newborn babies. *Arch. Dis. Child.* **47**, 405–10 (1972).
61. Daily, W. J. R., Klaus, M., Belton, H., Meyer, P. & Daily, J. R. APNEA IN PREMATURE INFANTS : MONITORING , INCIDENCE , HEART RATE CHANGES , AND AN EFFECT OF ENVIRONMENTAL TEMPERATURE. *Pediatrics* **43**, 510–518 (1969).
62. Di Fiore, J., Arko, M., Herynk, B., Martin, R. & Hibbs, a M. Characterization of cardiorespiratory events following gastroesophageal reflux in preterm infants. *J. Perinatol.* **30**, 683–7 (2010).
63. Mathew, O. P., Roberts, J. L. & Thach, B. T. Pharyngeal airway obstruction in preterm infants during mixed and obstructive apnea. *J. Pediatr.* **100**, 964–968 (1982).

64. Upton, C. J., Milner, A. D. & Stokes, G. M. Response to external obstruction in preterm infants with apnea. *Pediatr. Pulmonol.* **14**, 233–238 (1992).
65. Miller, M. J., Carlo, W. A. & Martin, R. J. Continuous positive airway pressure selectively reduces obstructive apnea in preterm infants. *J. Pediatr.* **106**, 91–94 (1985).
66. Slocum, C., Arko, M., Di Fiore, J., Martin, R. J. & Hibbs, A. M. Apnea, bradycardia and desaturation in preterm infants before and after feeding. *J. Perinatol.* **29**, 209–212 (2009).
67. Kurlak, L. O., Ruggins, N. R. & Stephenson, T. J. Effect of nursing position on incidence, type, and duration of clinically significant apnoea in preterm infants. **71**, F16-9 (1994).
68. Thach, B. T. & Stark, A. R. Spontaneous neck flexion and airway obstruction during apneic spells in preterm infants. *J. Pediatr.* **94**, 275–281 (1979).
69. Hayes, J., Mousa, H., Woodley, F. W. & Metheney, M. Testing the Association Between Gastroesophageal Reflux and Apnea in Infants. *J. Pediatr. Gastroenterol. Nutr.* **41**, 169–177 (2005).
70. Suichies, H. E. *et al.* Skin blood flow changes during apneic spells in preterm infants. *Early Hum. Dev.* **20**, 155–163 (1989).
71. Severinghaus, J. W. Historical Development of Oxygenation Monitoring BT - Pulse Oximetry. in (eds. Payne, J. P. & Severinghaus, J. W.) 1–18 (Springer London, 1986).
72. Moorman, J. R. *et al.* Beyond neonatal heart rate variability : predictive cardiorespiratory monitoring for apnea of prematurity. 1–4
73. Elder, D. E., Campbell, A. J. & Galletly, D. Current definitions for neonatal apnoea: Are they evidence based? *J. Paediatr. Child Health* 1–9 (2013). doi:10.1111/jpc.12247
74. Sale, S. M. Neonatal apnoea. *Best Pract. Res. Clin. Anaesthesiol.* **24**, 323–336 (2010).
75. Read, D. J. C. & Henderson-Smart, D. J. Regulation of Breathing in the Newborn During Different Behavioral States. *Annu. Rev. Physiol.* **46**, 675–685 (1984).
76. Cohen, G. & Henderson-Smart, D. J. Upper airway stability and apnea during

- nasal occlusion in newborn infants. *J. Appl. Physiol.* **60**, 1511–1517 (1986).
77. Beuchee, A. *et al.* Prolonged dynamic changes in autonomic heart rate modulation induced by acid laryngeal stimulation in non-sedated lambs. **91**, 83–91 (2007).
 78. Di Fiore, J. M. *et al.* Cardiorespiratory Events in Preterm Infants Referred for Apnea Monitoring Studies. *Pediatrics* **108**, 1304–1308 (2001).
 79. Kurz, H. Influence of nasopharyngeal CPAP on breathing pattern and incidence of apnoeas in preterm infants. *Biol. Neonate* **76**, 129–133 (1999).
 80. Martin, R. J. & Abu-Shaweesh, J. M. Control of breathing and neonatal apnea. **87**, 288–295 (2005).
 81. Thach, B. T. Maturation and transformation of reflexes that protect the laryngeal airway from liquid aspiration from fetal to adult life. *Am. J. Med.* **111 Suppl**, 69S–77S (2001).
 82. Bhatia, J. Current Options in the Management of Apnea of Prematurity. *Clin. Pediatr. (Phila)*. **39**, 327–336 (2000).
 83. Griffin, M. P. *et al.* Heart rate characteristics: Novel physiomarkers to predict neonatal infection and death. *Pediatrics* **116**, 1070–1074 (2005).
 84. Goldstein, B. Heart rate characteristics in neonatal sepsis: A promising test that is still premature. *Pediatrics* **115**, 1070–1072 (2005).
 85. Coggins, S. A. *et al.* Heart rate characteristic index monitoring for bloodstream infection in an NICU: a 3-year experience. *Arch. Dis. Child. Fetal Neonatal Ed.* **101**, F329–F332 (2016).
 86. Cuestas, E., Rizzotti, A. & Aguero, G. [Heart rate variability analysis: a new approach in clinical research methodology for neonatal sepsis]. *Arch. Argent. Pediatr.* **109**, 333–338 (2011).
 87. Srinivasan, L. & Harris, M. C. New technologies for the rapid diagnosis of neonatal sepsis. *Curr. Opin. Pediatr.* **24**, 165–171 (2012).
 88. Mani, S. *et al.* Medical decision support using machine learning for early detection of late-onset neonatal sepsis. *J. Am. Med. Inform. Assoc.* **21**, 326–336 (2014).
 89. Bohanon, F. J. *et al.* Heart rate variability analysis is more sensitive at identifying neonatal sepsis than conventional vital signs. *Am. J. Surg.* **210**, 661–667 (2015).

90. Ahmad, S. *et al.* Continuous multi-parameter heart rate variability analysis heralds onset of sepsis in adults. *PLoS One* **4**, (2009).
91. Lake, D. E., Richman, J. S., Pamela Griffin, M. & Randall Moorman, J. Sample entropy analysis of neonatal heart rate variability. *Am. J. Physiol. - Regul. Integr. Comp. Physiol.* **283**, 789–797 (2002).
92. Seely, A. J. E. & Macklem, P. T. Complex systems and the technology of variability analysis. *Crit. Care* **8**, (2004).
93. Amess, P., Rabe, H. & Wertheim, D. Visual assessment of heart rate variability patterns associated with neonatal infection in preterm infants. *Early Hum. Dev.* **134**, 31–33 (2019).
94. Goldstein, B. & Buchman, T. G. Heart Rate Variability in Intensive Care. *J. Intensive Care Med.* **13**, 252–265 (1998).
95. Stein P.K., P. & Kleiger R.E., M. D. Insights from the Study of Heart Rate Variability. *Annu. Rev. Med.* **50**, 249–261 (1999).
96. Awad, M. *et al.* Early denervation and later reinnervation of the heart following cardiac transplantation: A review. *J. Am. Heart Assoc.* **5**, 1–21 (2016).
97. Yapıcıoğlu, H., Özlü, F. & Sertdemir, Y. Are vital signs indicative for bacteremia in newborns? *J. Matern. Neonatal Med.* **28**, 2244–2249 (2015).
98. Bekhof, J., Reitsma, J. B., Kok, J. H. & Van Straaten, I. H. L. M. Clinical signs to identify late-onset sepsis in preterm infants. *Eur. J. Pediatr.* **172**, 501–508 (2013).
99. Hicks, J. H. & Fairchild, K. D. Heart rate characteristics in the NICU: What nurses need to know. *Adv. Neonatal Care* **13**, 396–401 (2013).
100. AASM. The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications.
101. Pernica J, Pugh J. E. V., Khan S, Fulford M, Twiss J, Williams C, El Helou S, Alonazi H, Ellattal B, M. C. *A novel retrospective methodology to improve the clinical accuracy in the diagnosis of Late onset Neonatal Sepsis.* (2021).
102. McGregor, C., Catley, C. & James, A. Variability analysis with analytics applied to physiological data streams from the neonatal intensive care unit. *Proc. - IEEE Symp. Comput. Med. Syst.* 1–5 (2012). doi:10.1109/CBMS.2012.6266385

103. Thommandram, A., Eklund, J. M., McGregor, C., Pugh, J. E. & James, A. G. A rule-based temporal analysis method for online health analytics and its application for real-time detection of neonatal spells. *Proc. - 2014 IEEE Int. Congr. Big Data, BigData Congr. 2014* 470–477 (2014). doi:10.1109/BigData.Congress.2014.74