An Evaluation Method for the Evaluation of Big Data based Streaming Analytic Clinical Decision Support Systems

by

Jonah Glass

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

Masters of Health Sciences in Health Informatics

Faculty of Health Sciences

University of Ontario Institute of Technology (Ontario Tech University)

Oshawa, Ontario, Canada

April 2020

THESIS EXAMINATION INFORMATION

Submitted by: **Jonah Glass**

Masters of Health Sciences in **Health Informatics**

Thesis title: An Evaluation Method for the Evaluation of Big Data based Streaming Analytic Clinical Decision Support Systems

An oral defense of this thesis took place on April 14, 2020 in front of the following examining committee:

Examining Committee:

Chair of Examining Committee	Dr. Manon Lemonde
Research Supervisor	Dr. Carolyn McGregor AM
Examining Committee Member	Dr. Andrew Miller
Thesis Examiner	Dr. Stephen Jackson

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

This thesis presents a methodology for evaluating a scalable clinical decision support systems (CDSS) that uses high frequency streaming physiological data using a holistic approach that includes the presence of population health indicators. The plan applies concepts and uses indicators suggested in the HOT-Fit framework, while applying the evaluation template developed by Public Health Ontario and uses an indicator structure described in York Region Public Health's Monitoring and Evaluation Framework.

The methodology is applied within the research to the implementation of the Artemis Platform at the McMaster Children's Hospital (MCH) Neonatal Intensive Care Unit (NICU). NICUs, have specific requirements relating to the use of clinical data and the implementation of new IT infrastructure. These requirements predicate the need for informative documentation that describes the utilization of the CDSS including a Privacy Impact Assessment (PIA), Threat and Risk Assessment (TRA), and a research and ethics proposal.

Keywords: decision support; health analytics; evaluation methods; metrics;

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 to lend this thesis to other institutions or individuals for the purpose of scholarly research. I further authorize University of Ontario Institute of Technology 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.

Jonah Glass

STATEMENT OF CONTRIBUTIONS

I hereby certify that I am the sole author of this thesis and that no part of this thesis has been published or submitted for publication. I have used standard referencing practices to acknowledge ideas, research techniques, or other materials that belong to others. Furthermore, I hereby certify that I am the sole source of the creative works and/or inventive knowledge described in this thesis.

DEDICATION

This thesis is dedicated to my grandparents. To Lucy, who was always excited by my work in healthcare with babies. To Essie, who always supports me in everything I do. To Dave, who was my number one fan and was always happy to see me.

ACKNOWLEDGEMENTS

First and foremost, thank you to Carolyn for taking me as a student in 2015 and giving me every opportunity to succeed. I have learned so much from you and am proud to have worked within the Health Informatics Lab at Ontario Tech. To all the other students and members of the lab, thank you for teaching me about neonates, astronauts, streaming data and so much more.

Thank you to my friends and colleagues in Western Australia who took me in for my 4 months as a visiting scholar. I learned so much about population health while there, and got to visit one of the most beautiful places in the world.

And finally thank you to my family, specifically my parents and sister who put up with 5 years of me working through this Thesis. My post-secondary education has been a lot of ups and downs over the years, but I’ve always had the support from my family to reach my potential.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS AND SYMBOLS

AM	Member of the Order of Australia
CAC	Queen’s University Centre for Advanced Computing
CDSS	Clinical Decision Support System
CRISP-DM	Cross Industry Standard Process for Data Mining
CSE	Communication Security Establishment
CW	Cognitive Walkthrough
ECG	Electrocardiogram
EFMI	European Federation for Medical Informatics
EHR	Electronic Health Record
EMR	Electronic Medical Record
FIPPA	Freedom of Information and Protection of Privacy Act
GEP-HI	Good Evaluation Practice for Health Informatics Systems
GUI	Graphical User Interface
HE	Heuristic Evaluation
HiREB	Hamilton Integrated Research Ethics Board
HIS	Health Information System
ICT	Information and Communications Technology
ILC	Information Life Cycle
ISO	International Standard Organization
IMIA	International Medical Informatics Association
LONS	Late Onset Neonatal Sepsis
MCH	McMaster Children’s Hospital
NICU	Neonatal Intensive Care Unit
ORION	Ontario Research and Innovation Optical Network
OTHIR	Ontario Tech Health Informatics Research Lab
PIA	Privacy Impact Assessment

PHIPA	Personal Health Information Protection Act
PHO	Public Health Ontario
RCMP	Royal Canadian Mounted Police
STARE-HI	Statement of Reporting Evaluations of Health Informatic Systems
TRA	Threat and Risk Assessment
UOIT	Ontario Tech University
UVON	Unified eValuation using ONtology
x_1	Cases defined as definite LONS by Artemis Platform algorithm
x_2	Cases defined as definite LONS by clinician(s)
y_1	Cases defined as probable LONS by Artemis Platform algorithm
y_2	Cases defined as probable LONS by clinician(s)
z_1	Cases defined as potential LONS by Artemis Platform algorithm
z_2	Cases defined as potential LONS by clinician(s)
q_1	Cases defined as without LONS by Artemis Platform algorithm
q_2	Cases defined as without LONS by clinician(s)

Chapter 1. Introduction

1.1 Introduction

This thesis presents an evaluation methodology and metrics for evaluating scalable clinical decision support systems (CDSSs) that use high frequency streaming physiological data analytics to support improved population health. The evaluation methodology uses a holistic approach that includes the presence of population health metrics, technical metrics, and implementation specific metrics. The methodology applies concepts and uses metrics suggested in the HOT-Fit framework (Yusof et al. 2008), while applying terminology and plan design from the Public Health Ontario (PHO) evaluation plan template (Public Health Ontario 2016), and using a hierarchal metric structure described in York Region Public Health's Monitoring and Evaluation Framework (Glass et al. 2018). The evaluation methodology demonstrates how implementation artifacts can be leveraged in the development of evaluation metrics.

The evaluation plan will be applied in this research within the context of a CDSS implementation in an Ontario neonatal intensive care unit (NICU). Healthcare facilities in Ontario have specific requirements relating to the use of clinical data and the implementation of new IT infrastructure. These requirements predicate the need for informative documentation that describe the utilization of the CDSS including a Privacy Impact Assessment (PIA), and Threat and Risk Assessment (TRA). The implementation was governed by a research study, from which approval documents were used to design potential high-level population health outcome metrics like morbidity and mortality in preterm infants. Population health approaches aim to improve the health of entire populations through collaborative, upstream, evidence based care (Public Health Agency of Canada 2013).

More than 1 in 10 babies are born premature, and approximately one million babies die annually due to complications related to premature birth. This is a global population health issue, present in both the developed and developing world (World Health Organization 2018). In Canada, the costs to care for individual premature infants are significant, and can extend for years after the infant leaves the hospital. Many infants develop complications resulting from prematurity (Johnston et al. 2014). The high risk patients in NICUs are extremely fragile and require around the clock care monitoring and care using a multitude of devices that output physiological data; however, the high frequency of these readings rendered almost all of the data unusable for analysis historically (Catley et al. 2009).

CDSSs are designed to impact clinical decision making at an individual patient level in real time. Berner identified two types of CDSSs used by clinicians in patient care. Knowledge-Based CDSSs are systems that assist in decision making by providing clinicians with information, allowing them to make a more informed decision. They use previous research entered as rules in text format for clinicians to review when making a decision. These systems aren't meant to make the decision for the clinicians. Non-Knowledge-Based CDSSs use machine learning and artificial intelligence to find and develop patterns in health data without expert input, and can provide both predictions and diagnosis. In some cases they have been more accurate in their diagnosis than clinicians (Berner 2007). CDSSs use varying network topologies / system architectures to achieve their goals. System topology describes design, and where components used in the system are located. Components for CDSSs may exist in or outside of the hospital network.

Big Data is referred to as having "4Vs" – volume, velocity, variety, and veracity (Kitchin and Mcardle 2016; Raghupathi and Raghupathi 2014). In healthcare, there are a large variety of datasets that collect physiological, demographic, pharmaceutical, care-history, and more datasets (Raghupathi and Raghupathi 2014). Of specific interest in this research is the sub group of Big Data from medical sensors such as the streams of various physiological data from medical sensors and demographic information from Electronic Medical Records (EMRs). This research focusses on physiological data streams as a form of Big Data.

The Artemis Platform is a CDSS that fits the classification of both non-knowledge and knowledge-based systems. It performs the non-knowledge functions of acquiring, analysing, and storing high frequency data, and uses machine learning to identify patterns. Through the classification of patterns observed by the platform, Artemis performs knowledge based system functions of predicting conditions and informing clinicians to assist with enhancing their decision making (McGregor et al. 2011).

1.2 Research Motivation

The motivation for this research is to understand how a CDSS that uses high frequency streaming physiological data, and is being implemented in multiple hospital environments, can be effectively evaluated. New CDSSs like the Artemis Platform may be able to help identify adverse conditions and medical symptoms through the use of big data analytics; however, a holistic evaluation is needed to validate and verify the system's capabilities and impact in real world environments where external factors can affect the utilization of the system.

The case study for this research is the development of an evaluation plan of the implementation of the Artemis Platform at the McMaster Children's Hospital (MCH) NICU. The MCH NICU is a collaborating partner with the Ontario Tech University Health Informatics Research Lab (OTHIR).

By developing an evaluation plan for the evaluation of a CDSS that uses artifacts specific to each implementation, while maintaining a hierarchical approach to population health metrics, further implementations will have a cohesive evaluation model that can easily be populated using site-specific artifacts. Organizations will have an informative model that evaluates population health, organizational, technical, and user-specific aspects of the CDSS.

1.2.1 Research Motivation within a Hospital Setting

Technology is becoming increasingly relevant in the health sector, specifically as it relates to creating efficiencies and improving patient outcomes in hospitals. In Ontario, the Excellent Care for All Act was introduced as legislation in 2010. This legislation focused on improving the quality of services in the healthcare environment and tied hospital funding to performance metrics (Government of Ontario 2017). CDSSs are a form of technology that are implemented in hospitals with the intention of assisting with or improving clinical decision making (Berner 2007). An evaluation of CDSSs will help determine their impact on the quality of service in the healthcare environment, and may be used as a launching point for introducing more technology in to hospitals to improve quality of care and hospital performance.

The Artemis Platform uses high frequency streaming data for the purpose of detecting adverse conditions and medical symptoms in critical care settings (McGregor 2013). The Artemis Platform changes clinical practices by utilizing physiological data that reflect healthy and disease states. Since this data is produced at high frequency and high volume, previously it was only useful when viewed in overview, that is, some time after it was acquired. The analytics within the Artemis platform can identify disease trends in physiological data before it is obvious in overview and provide ‘advance’ warning to clinicians (McGregor 2013). Through the early identification of conditions, the system has the potential to demonstrate benefits and increased efficiency in NICUs by reducing morbidity and mortality from the delays in timely diagnosis of treatable disease.

Implementing a system in a hospital is not as simple as suggesting that there may be clinical benefits. An evaluation of the system is required so that key decision makers and stakeholders are able to understand the system impact on care provided, stakeholder satisfaction, technical performance and security, and the impact of the system on the preterm birth population as a whole. There is a higher rate of morbidity and mortality in the preterm birth population. These rates can be reduced through the mitigation of morbidity and mortality from post-natal conditions developed by neonates in the NICU. The early identification of conditions may mitigate the population health problems (morbidity and mortality in neonates), while also reducing the longer term affects of neonatal medical conditions on morbidity and mortality.

1.3 Research Aims and Objectives

The aim of this research is to propose an evaluation methodology for the assessment of CDSSs that utilise high frequency Big Data to create evaluation metrics for CDSS implementations that use streaming Big Data, and move to develop a holistic evaluation plan that includes the impact of the CDSS on population health outcomes. A key benefit of the methodology is that it will support the inclusion and organization of implementation artifacts to inform the design of evaluation metrics also. Three research hypotheses are presented and addressed in this work:

1. *That an evaluation methodology can be developed that includes population health, technical, and algorithm specific metrics specific to the implementation of a high frequency streaming physiological data analytics.*
2. *The abovementioned methodology will integrate key implementation artifacts for the purpose of determining metrics.*
3. *That the methodology can be demonstrated within an instantiation to support the evaluation of a Big Data analytics based CDSSs within a NICU in Ontario*

1.4 Contribution to Knowledge

The contributions to knowledge within this thesis include:

- Design of an evaluation plan for CDSSs that includes population health metrics and depicts the use of implementation artifacts for the purpose of determining metrics and benchmarks
- Demonstrate the plan's connection to existing evaluation frameworks and methodologies
- Demonstrate the effectiveness of the plan through the development and partial completion of an evaluation for a CDSS implementation in a NICU setting

1.5 Research Method

This research was completed using a constructive research methodology. The constructive research method is an effective way to solve a specific problem while creating or enhancing a knowledgebase. A key element of the research method is to take existing knowledge and find ways to fill gaps or add missing information to help expand the existing knowledgebase (Dodig Crnkovic 2010; Kasanen, E., Lukka, K. Siitonen 1993). Using the constructive approach allows for flexibility in the development of the new evaluation methodology. The research design follows Kasanen et al.'s six step constructive approach (Kasanen, E., Lukka, K. Siitonen 1993), which includes:

Phase	Constructive Research	Evaluation Methodology Constructive Research
Phase 1	Find a practically relevant problem which has research potential	This research proposes an evaluation methodology for high frequency streaming analytic CDSSs. The methodology includes population health metrics that link the outcomes of the CDSS to an improvement in population health. It also proposes the use of implementation artifacts for the purpose of developing metrics and benchmarks for the evaluation plan.
Phase 2	Obtain a general and comprehensive understanding of the topic	This research includes a review of existing Health Information System (HIS) evaluation methodologies and suggested evaluation metrics as well as how the intricacies of a system like its system topology, and use of data recovery time affect the evaluation of the system. The research also includes an overview of key artifacts used in the implementation of a CDSS within a NICU for the purpose of research and improving clinical outcomes.
Phase 3	Innovate (i.e., construct a solution idea)	The research includes a methodology with the use of population outcomes to help define the impact of CDSSs that use high frequency streaming physiological data. The research also includes a methodology for including implementation artifacts as resources when developing metrics and benchmarks in the evaluation plan.
Phase 4	Demonstrate the solutions feasibility	The newly developed evaluation model is applied in a case study with the implementation of the Artemis Platform at the MCH NICU for the purpose of creating an evaluation plan.

Phase 5	Show theoretical connections and research contribution of the solution concept	The case study demonstrates how population outcomes are connected to the evaluation of a hospital specific CDSS, and how the artifacts used to implement the system help determine metrics and benchmarks within the evaluation plan.
Phase 6	Examine scope of applicability of the solution	This methodology can be used to evaluate other CDSS that are deployed in real-time that use physiological data. Once the research contribution has been demonstrated, discussion will follow regarding the scope if its application.

Table 1 Constructive Research Methodology

1.6 Thesis Overview

Chapter 2 presents the literature review, focusing on current HIS evaluation models applicable to the implementation of a CDSS and suggested evaluation metrics. The application domain for this research is the Artemis Platform implementation at the MCH NICU and is introduced in Chapter 3 with a discussion of the details of the implementation and key artifacts used to support the implementation process. Chapter 4 describes the development of the evaluation plan, defines the types of metrics that are components to an evaluation plan, and how implementation artifacts are used to define metrics and benchmarks. Chapter 5 demonstrates the new evaluation plan as it is used to evaluate the Artemis Platform implemented at the MCH NICU for the purpose of a medical study focused on late onset neonatal sepsis (LONS). Chapter 6 is a discussion of how the developed evaluation plan addresses the issues identified within the literature review. Chapter 7 concludes the thesis with an assessment of how the thesis has addressed the research aims and objectives, and the contribution to knowledge outlined in Chapter 1.

1.7 References

Berner, Eta S. 2007. *Clinical Decision Support Systems - Theory and Practice*. ed. Eta S. Berner.

Catley, Christina, Kathy Smith, Carolyn Mcgregor, and Mark Tracy. 2009. "Extending CRISP-DM to Incorporate Temporal Data Mining of Multi- Dimensional Medical Data Streams : A Neonatal Intensive Care Unit Case Study." : 0–4.

Dodig Crnkovic, Gordana. 2010. "Constructive Research and Info-Computational Knowledge Generation." *Model-Based Reasoning in Science and Technology*: 52–67.

Glass, Jonah et al. 2018. "Development of a Monitoring and Evaluation Framework for a Public Health Setting."

Government of Ontario. 2017. *Excellent Care for All Act*.

Johnston, Karissa M et al. 2014. "The Economic Burden of Prematurity in Canada." *BMC Pediatrics* 14(1): 1–10. BMC Pediatrics.

Kasanen, E., Lukka, K. Siitonen, A. 1993. "The Constructive Approach in Management Accounting Research." *Journal of Management Accounting Research* 5(August).

Kitchin, Rob, and Gavin Mcardle. 2016. "What Makes Big Data , Big Data ? Exploring the Ontological Characteristics of 26 Datasets." (June): 1–10.

McGregor, Carolyn. 2013. "Big Data in Neonatal Intensive Care." *Computer* 46(6): 54–59.

McGregor, Carolyn, Christina Catley, Andrew James, and James Padbury. 2011. "Next Generation Neonatal Health Informatics with Artemis." *Studies in Health Technology and Informatics* 169: 115–19.

Public Health Agency of Canada. 2013. "What Is the Population Health Approach." https://www.canada.ca/en/public-health/services/health-promotion/population-health/population-health-approach/what-population-health-approach.html#health.

Public Health Ontario. 2016. 117 Public Health *Evaluating Health Promotion Programs: Introductory Workbook*.

Raghupathi, Wullianallur, and Viju Raghupathi. 2014. "Big Data Analytics in Healthcare : Promise and Potential." : 1–10.

World Health Organization. 2018. "Preterm Birth." https://www.who.int/news-room/fact-sheets/detail/preterm-birth.

Yusof, Maryati Mohd, Jasna Kuljis, Anastasia Papazafeiropoulou, and Lampros K. Stergioulas. 2008. "An Evaluation Framework for Health Information Systems: Human, Organization and Technology-Fit Factors (HOT-Fit)." *International Journal of Medical Informatics* 77(6): 386–98.

Chapter 2. Literature Review

2.1 Introduction

This chapter presents research into literature used to inform how HISs are evaluated and which frameworks, methods, and components may be adaptable for use in the evaluation of a high frequency streaming data and big data analytics platform with the goal of being implemented as a public health intervention.

2.2 Background

HIS is a generic term defined as an information system utilized in the health sector with four key functions: data generation, compilation, analysis, and communication (World Health Organization 2008). Given the costs and potential impact, any implementation of a HIS requires an evaluation to determine the level of success of the implementation. HIS have many subtypes including CDSSs. CDSSs support health systems and providers as decision making aids through a variety of means including real-time alert systems and presenting clinical data analysis (Berner 2007).

HIS are becoming increasingly prevalent in the healthcare industry because they promise to assist in care, find efficiencies, and improve processes. As more HIS are researched, designed, developed, and implemented, researchers and organizations have developed a multitude of ways to evaluate them. Considering the amount of resources required to implement these systems, a proper evaluation is critical in determining whether the implemented system meets its objectives. Evaluation frameworks that have been developed specifically for HIS consist of many different components for designing, undertaking, or reporting an evaluation. They offer suggestions of what to evaluate, how to evaluate it and may be either process or outcome focused (Eivazzadeh et al. 2016).

In 2013, the International Medical Informatics Association (IMIA) Working Group on Technology Assessment and Quality Improvement and the European Federation for Medical Informatics (EFMI) Working Group on Assessment of Health Information Systems reported on efforts to promote the principal of evidence-based health informatics

(Rigby et al. 2013). In that report, the working groups noted that the move to an evidence-based approach was paramount, and discussed progress towards earlier recommendations needed to achieve an evidence-based approach in the evaluation of HIS (Rigby et al. 2013). These recommendations included:

- Guidelines for good evaluation practice should be made available
- Terms, concepts and guidelines for reporting on results of information and communications technology (ICT) assessment studies should be made available
- Evaluation networks should be established
- Appreciation of methods of evaluation should be part of health informatics curricula
- An open access repository about evaluation studies should be established

Through the endorsement of the Statement of Reporting Evaluations of Health Informatics Systems (STARE-HI), and the Guideline for Good Evaluation Practice for Health Informatics Systems (GEP-HI), the health informatics community have requirements and guidelines to follow for completing evaluations of health informatics information systems, and researchers have developed and utilized evaluation methods that meet STARE-HI standards. The purpose of this literature review is to examine these evaluation methods and to determine effective processes for evaluating specific factors of a CDSS implementation in a health informatics information system evaluation.

2.2.1 Big Data in Healthcare

Big Data is referred to as having "4Vs" – volume, velocity, variety, and veracity (Kitchin and Mcardle 2016; Raghupathi and Raghupathi 2014). In healthcare, there are a large variety of datasets that collect physiological, demographic, pharmaceutical, care-history, and more datasets (Raghupathi and Raghupathi 2014). Of specific interest in this research is the sub group of Big Data from medical sensors such as the streams of various physiological data from medical sensors and demographic information from Electronic Medical Records (EMRs). This research focusses on physiological data streams as a form of Big Data.

Wang presents a best practice approach for the use of Big Data analytics in healthcare. The model consists of five architectural layers: Data, Data Aggregation, Analytics, Information Exploration and Data Governance. The data layer consists of all the data used to provide insights. This includes both structured and unstructured datasets. The data aggregation layer is where data is digested, cleaned, and transformed into usable structured data for analytics. In the analytics layer, the data is processed and analyzed. Stream computing is an example of an analytics layer task, where high performance data is processed in real-time or near real-time for the purpose of detecting abnormalities. The information layer includes outputs like data visualizations derived from the analytics. This is the layer that most end-users interact with on a day-to-day basis. The data governance layer is the final layer, and consists of management policies like data security and privacy management. In healthcare, rigorous data rules and policies are used to protect sensitive clinical data and ensure proper patient care (Wang, Kung, and Anthony 2018). Within Wang's model, the evaluation components are limited to the data focus areas of data immediacy, data completeness, data accuracy and data availability. While these components in Big Data CDSSs should be evaluated, Wang's article does not discuss the security of the system, and most importantly does not discuss the evaluation of the clinical / population health outcomes achieved through the use of Big Data analytics. Wang's research did highlight the potential of improved quality and accuracy of clinical decision making through Big Data Analytics (Wang, Kung, and Anthony 2018). Extrapolating and demonstrating how this improved quality and accuracy affects health outcomes should be a major goal of any CDSS evaluation.

2.2.2 Clinical Decision Support Systems

CDSSs utilizing high frequency streaming data and big data analytics with the goal of being a public health intervention is an open research area. Current evaluation methods do not have the specific metrics required to perform an evaluation of the specific nuances within the deployment of Big Data analytics architectures for new approaches to clinical decision support in healthcare.

The Artemis Platform has been implemented at hospitals in both North America and China. It was implemented at MCH NICU in early 2017 as part of a clinical research study. As the Artemis Platform is implemented in more hospital NICUs, an evaluation method to determine the level of success is required.

To determine which evaluation metrics should be used for evaluating Big Data analytics based CDSSs that include cloud-based components, a literature review was performed focusing on previously defined evaluation categories and existing evaluation frameworks.

Existing evaluation frameworks have been developed for HIS; however, these evaluation frameworks may be too broad for a high frequency streaming big data analytics CDSS. CDSS are outcome focused information systems, which have the specific goals of assisting in decision making to improve patient outcomes (Berner 2009). CDSSs can be implemented as part of quality improvement initiatives for population health by impacting clinical decision making through the provision of clinical analytics as evidence. Examples of this include systems that detect medication errors, provide alerts based on specific criteria, such as increased heart rate, or use predictive algorithms to inform clinicians of patient needs. The Artemis Platform uses a predictive algorithm for the early detection of sepsis in neonates (McGregor et al. 2013). This differs from other HIS like EMR systems, which are focused on patient management, or general IT systems which are implemented to assist healthcare organizations in their overall day to day management (ordering, timetables, billing, resources). CDSSs also have different network topologies / system architectures, which may impact how the availability and security of the system is evaluated. The Artemis Platform uses a system topology with multiple components outside of the hospital network. Due to these differences, CDSS should be evaluated differently than other HIS.

One issue with evaluating CDSS is that, despite their differences from other types of HIS, current evaluation frameworks are applied to all types of HIS. The goals of this literature review were:

1. to quantify differences in evaluation components that can be applicable to a high frequency, big data analytics CDSS like the Artemis platform;
2. to review whether and how population health metrics are considered within an HIS evaluation and whether that can be applied to a CDSS;
3. to assess if and how a CDSSs topology is included within the evaluation

4. to determine how the recovery from downtime of a system like Artemis can be effectively included in an evaluation of the system's availability;
5. to determine whether or not deliverables involved in the implementation of a CDSS can be used to help perform the evaluation. Evaluations in their current practice are performed after the implementation is completed, but there may be an opportunity to utilize the deliverables that assisted in implementing the system to help develop the evaluation criteria for these systems as part of the implementation process.

2.3 Methods

A PubMed title/abstract search was undertaken to find literature describing the evaluation of HISs. The first search used the terms "Evaluation Framework" and "Health Information Systems", the second used "Evaluation Method*" and "Health Information Systems", and the searches were constrained to the date range 2011-2019. Eleven results were found in the first search, and twenty-eight in the second. One duplicate was removed, leaving thirty-eight articles. After reviewing abstracts, twenty-three articles were removed. Two articles were added based on recommendations and citations from the remaining articles. Therefore, seventeen articles are included in this literature review. The articles reviewed include the development and application of existing evaluation frameworks, as well as best practices and tools for measuring evaluation metrics. Figure 1 depicts the method for this literature review.

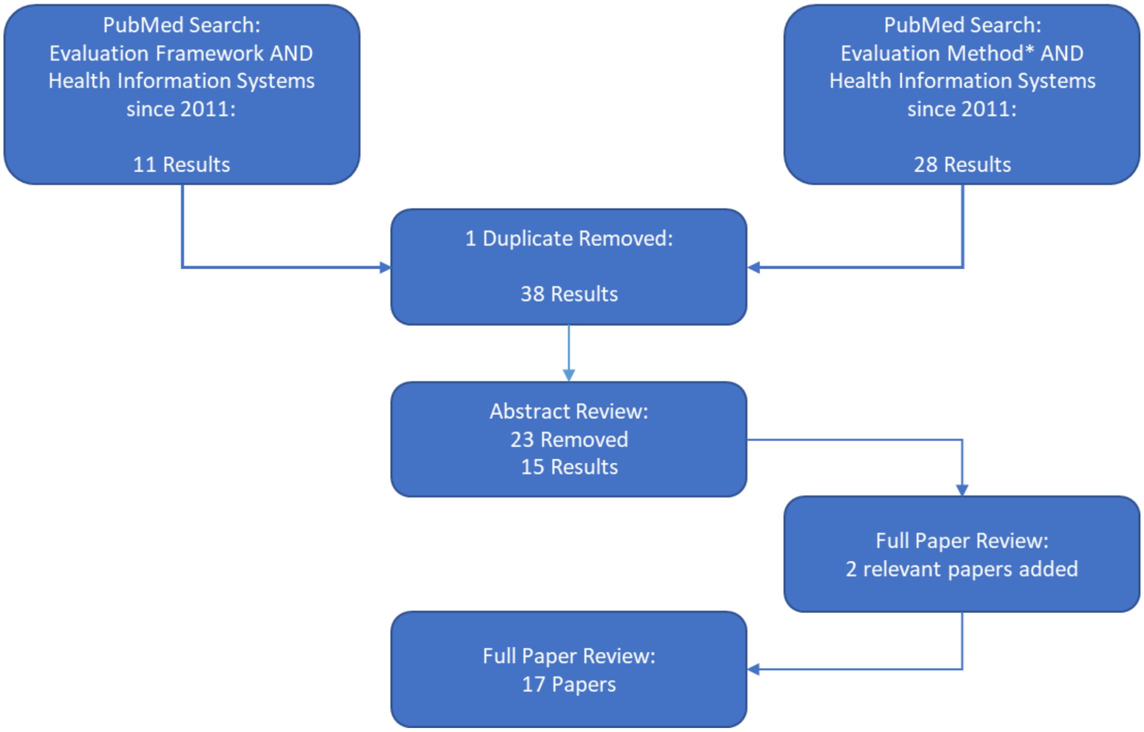

Figure 1 Literature Review Method

2.4 Results

Of the seventeen articles reviewed, eight were focused on reviewing existing literature and providing insight into specific evaluation factors. Some of these articles attempted to provide an understanding as to why evaluations fail, and to what were key factors determining the success or failure of a health information system implementation. Table 2 describes the eight articles where researchers have attempted to better understand effective evaluation methods and factors that are important to an effective evaluation.

The review process for the papers identified in the literature review was focused on key considerations described in the research questions.

The translation of HIS evaluation metrics and methodologies developed and used by others to evaluate a different high frequency streaming analytic CDSS is particularly relevant. To aid with this translation task, an evaluation of how the goals and expectations of authors influenced the objectives and methods in the paper is pertinent.

The types of metrics included in the articles were analyzed to develop understanding of what metrics are included in HIS evaluations. Of specific interest was population health and usability metrics, for which a separate column in the results table were maintained highlighting their use in the article. The use of technical component segmentation and system topology, as well as the inclusion of data recovery time were analyzed in columns because of their application to the development of an evaluation methodology for high frequency streaming analytics CDSSs. Also included was the analysis of whether implementation artifacts were involved in the development of the metrics or used in the methodology.

Article	Objective	Method	Types of Metrics Included	Population Health	Usability	Technical Component Segmentation / System Topology	Data Recovery	Use of Implementation Artifacts
Application of Usability Metrics in a Multi-User and Multimedia EHR Evaluation Framework(Kopanitsa, Tsvetkova, and Veseli 2012)	Review GUI usability evaluations for EHR Understand how they take into account different stakeholder / user groups	A literature review of relevant papers read by two researchers.	The article splits usability evaluation metrics in to three groups • Efficiency • Effective • Satisfaction Sample metrics are provided in each group.	The impact of the end user's use of the system is not considered	Universal usability metrics are needed for HIS interfaces.	Evaluation specific details about technical components are not included.	The evaluation proposes time to complete tasks as a metric Does not include impact of limited function	Using implementation artifacts is not listed as part of the evaluation tools section.
Towards an evaluation framework for information quality management (IQM) practises for health information systems – evaluation criteria for effective IQM practises(Mohammed and Yusof 2013)	Identify evaluation criteria that can influence the production of good information quality in HIS	Six frameworks and best practises were reviewed to identify evaluation criteria	The article focuses on the evaluation criteria needed to measure information quality in a HIS Criteria is separated into the sub factors presented in the HOT-fit framework (Human, Organization Technology) Some of the proposed metrics act more like a checklist, ensuring that the right components are in place to support an evaluation	The evaluation metrics included do not include a system's impact on population health or clinical outcomes.	Usability is listed as a system quality metric	The authors suggest reviewing system quality and service quality metrics System design is mentioned as part of the evaluation design.	System design's inclusion in evaluation planning could suggest the inclusion of data recovery	The authors have integrated the Information life cycle (ILC) Documentation from the 'plan' and 'acquisition' stages could be viewed as implementation artifacts used for the purpose of evaluating the information quality of the HIS.
Measuring value for money: a scoping review on economic evaluation of health information	Determine how components of economic evaluations have been included in HIS Provide guidance for future evaluations	A literature review of relevant papers that contain economic evaluation methods	The article focuses on economic evaluation methods only Some of the outcome metrics discovered apply to other evaluation factors	Some economic analysis that the authors found includes the clinical outcome of disease prevention Disease prevention can reduce the need	Usability metrics were not included	Evaluation specific details about technical components are not included	Not included	Historical costs and estimates are used as an input for cost analysis

systems(Bassi and Lau 2013)			including clinical outcomes Clinical outcome improvements are viewed through the lens of healthcare savings	for hospital care, saving a health system money				
Electronic immunization data collection systems: application of an evaluation framework(Heidebrecht et al. 2014)	Development and application of an evaluation framework	Uses the Centers for Disease Control and Prevention's Guideline for Evaluating Public Health Surveillance Systems as a guide	The article includes five attributes for evaluation • Simplicity • Flexibility • Quality • Timeliness • Acceptability	Despite evaluating a public health system, the impact of the system on population health is not explored	Usability metrics would fall under simplicity and acceptability based on the definitions provided in the article	The evaluation compared two different topologies Components within the topology were not evaluated	Not included	System implementation was not explored The systems already existed for many years
Determining of factors influencing the success and failure of hospital information systems and their evaluation methods: a systematic review(Sadoughi et al. 2013)	Identify evaluation methods for specific success / failure criteria for HISs	A literature review of HIS studies where success and failure factors were included	The literature reviewed by the authors found a variety of success factors. Specific metrics were not provided to connect to these success factors Subfactors provided by the authors suggest an opportunity for a variety of metrics	Clinical outcomes including mortality and morbidity were found in some of the articles (25%)	Usability was listed as a single metric in their review, with 43% of literature they found including a usability metric Metrics like satisfaction are listed separately	Evaluation specific details about technical components are not included	Not included	Using implementation artifacts is not listed as part of the evaluation tools section
Evaluation methods used on health information systems (HISs) in Iran and the effect of HISs on Iranian healthcare: a systematic review(Ahmadian, Salehi,	Study the methods used for the evaluation of HISs in Iran	Systematic search of papers evaluating HIS in Iran Data collection form used for extracting information 53 relevant evaluations included	Does not highlight metric types Highlights effects and objectives of implementation The most common objective of evaluation has been to demonstrate improvements in data quality	General public health metrics liked improved quality of care and improving disease management are included	Usability metrics were not included	Evaluation specific details about technical components are not included	Not included	Using implementation artifacts is not listed as part of the evaluation tools section. The authors found that all of the evaluations were done as summaries after the system's implementation

and Khajouei 2015)								No information about evaluation during development of the systems was included in the papers they reviewed. They found this to be inconsistent with other literature reviews
Integrating Methods to Evaluate Health Information Systems(Borim et al. 2015)	Identify the main aspects used to evaluate HIS	Literature review of HIS evaluations and software quality analyses	Specific metrics are not provided General questions are provided for different evaluation areas including: • Functionality • Maintainability • Information quality • Efficacy • Effectiveness • Usability • Availability	Not considered	The authors propose that usability answers 'will the system be easy to use'	Evaluation specific details about technical components are not included.	Not included	The authors propose that aspects of the system, and the aims of the system should be part of the evaluation design They include a literature review as an effective way to identify metrics for the study
Comparison of heuristic (HE) and cognitive walkthrough (CW) usability evaluation methods for evaluating health information systems(Khajouei, Esfahani, and Jahani 2016)	Understand differences between two usability evaluation methods	Compared the methods using a case scenario Goal of comparison was to determine the: Number of errors Severity of errors Coverage of errors	Usability metrics Six key attributes of usability	Not considered	Compares two different methods (HE/CW) Both are prospective evaluation methods	Not included Focus is on front-facing GUIs	Not included	Uses pre-created scenario documents based on available system information Scenarios can be created from implementation documents that describe scope

Table 2 Literature Review Results

2.4.1 Population Health and Clinical Outcome Evaluations

In Sadoughi et al.'s review of success factors, they found that an improvement in clinical performances measured through an improvement in patient outcomes was present in only 4 of the 16 evaluations reviewed (Sadoughi et al. 2013). The absence of patient outcome measurement in HIS evaluations may be because some HISs are not implemented for the purpose of directly improving patient outcomes. For CDSSs, which are implemented with the goal of improving clinical decision making and patient outcomes, altered clinical outcomes (performance) and patient outcome success factors should be more prevalent. Bassi and Lau also indicated that clinical outcomes could be used in evaluations, although their work views clinical outcomes through an economic lens (Bassi and Lau 2013) where improved patient outcomes could result in shorter hospital stays, fewer misdiagnosis, and less repeat uses of the hospital or other health services. For example, healthier babies discharged from the NICU may require less on-going services, saving health systems money. The premise of requiring less on-going services is an outcome of improved population health; however, the metric for improved population health is not the reduction of services, but the reasoning for the reduction – reduced morbidity and mortality through the earlier detection and treatment of a condition. The HOT-Fit framework included clinical outcomes to patient care and the population as potential net benefit metrics in their evaluation framework (Yusof et al. 2008) in addition to net benefit metrics focused on internal organization benefits.

2.4.2 Technical Evaluations

Sadoughi et al. also identified technical factors as a common evaluation area in their systematic review. Technical factors include sub-factors such as complexity, infrastructure, and response time, which depending on the system could be valid evaluation metrics. A more commonly evaluated sub factor, which should be almost universally applicable to HIS with a front-facing component was usability (Sadoughi et al. 2013). Usability, is defined by ISO 9241 -11 as "the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use."

2.4.2.1 Usability

The usability sub-factor of satisfaction was the most commonly discussed and evaluated factor in HIS literature, yet Sadhougi et al. noted that user satisfaction is still not as popular a criterion in the focus of HIS compared to information systems in general (Sadoughi et al. 2013).

To properly measure user satisfaction, Sadoughi et al. suggested that focus group interviews and questionnaires could be effective evaluation methods; however, Khajouei et al. suggested that a heuristic evaluation (HE) could be used as an effective way to identify problems related to user satisfaction. In their study of HE and cognitive walkthrough (CW) evaluations, they compared the two evaluation methods using a single HIS case with five independent evaluators. To perform their HE, evaluators went through ten scenarios and examined the conformity of the software to Nielsen 10 heuristic principles. Nielsen's heuristic principles, included in Figure 2, are guidelines that system and interface designers should follow when creating a user interface.

Nielsen's 10 Principles(Nielsen 1994)
1. *The system status should be visible, so that users know what is going on* 2. *The system should use language the users understand and are familiar with rather than system-oriented terms* 3. *Users should be able to easily exit functions and undo mistakes* 4. *There should be consistency in wording and actions across the system* 5. *The system should be designed to prevent errors from occurring in the first place, include confirmation options* 6. *Make objects and actions in the system visible with clear instructions* 7. *Allow for flexibility and efficiency for power-users that speed up interactions with the system* 8. *Do not include dialogue that is irrelevant, use a minimalist design* 9. *Provide error messages in plain language, and suggest solutions* 10. *Include help and searchable documentation*

Figure 2 Neilsen's Heuristic Principles

The development of these scenarios was performed in consultation with physicians and the designers of the system. Through meetings with the evaluators, identified problems were organized into groups based on usability attributes proposed by the International Standard Organization (ISO).

In their results, Khajouei et al. noted that of the 26 unique problems relating to satisfaction, 21 were identified by HE, and 5 by CW. It is possible that more satisfaction problems were found in both systems, but were removed from the results as they were duplicates. While their results suggest that HE may be an effective way to measure user satisfaction, the authors noted that overall HE was only able to identify 53% of the problems with the system, and that CW may be better for evaluating other factors (Khajouei, Esfahani, and Jahani 2016). HE would likely be more resource intensive than a questionnaire or survey, but does have the advantage of adhering to Nielsen and ISO standards, which have been heavily researched and represent a universal benchmark and approach. As a less resource intensive method, the USE questionnaire measures stakeholder opinions on a Likert rating scale in four categories: usefulness, ease of use, ease of learning, satisfaction (Lund 2001).

User satisfaction is only one of the sub-factors that was identified as part of the behavioural factor category identified in Sadoughi et al.'s systematic review; however, it was by far the most common sub factor identified. All of the behavioural factor category sub-factors were highly focused on the user and their opinions and interactions with the system. In evaluation frameworks, these are often categorized as human or user factors (Eivazzadeh et al. 2016; Yusof et al. 2008).

While HE is one effective way to measure usability, Khajouei et al. noted in their study that it could not provide a complete evaluation on its own (Khajouei, Esfahani, and Jahani 2016), and suggested that HE is best used for evaluating usability when experts have utilized similar systems (Khajouei, Esfahani, and Jahani 2016). The fact that HE does not identify a large portion of issues suggests that multiple evaluation practices could be used in evaluating usability. Kopanitsa et al. suggested a matrix like framework that can be used in evaluating usability for multiple stakeholders, although it is specific to evaluating a graphical user interface (GUI) (Kopanitsa, Tsvetkova, and Veseli 2012).

Kopanitsa et al. note three specific usability evaluation metrics for EHRs from ISO 9241 -11: efficiency, effectiveness, and satisfaction. They propose smaller performance metrics can be used in a testing approach with users when a GUI is available for use. They suggest that in the case of multiple GUIs and multiple user types, a matrix is developed which correlates a score-based value from the evaluation, a user impact value, and a device impact value to determine the overall usability of a system and its GUI (Kopanitsa, Tsvetkova, and Veseli 2012). This approach is an effective way of displaying results, but a specific method for performing evaluations is not explicitly stated by the authors.

2.4.2.2 Technical Verification

To evaluate other technical factors, Sadoughi et al.'s systematic review often suggests technical verification as an evaluation method. According to the IEEE Standard Glossary of Software Engineering Technology, verification as the process of determining whether or not the products of a given phase of the software development cycle fulfill the requirements established during the previous phase (Boemh and Barry W. 1984). This definition first suggests that technical verification can be performed regularly during the development and implementation of a new information system, and additionally explicitly states that effective requirements gathering is important in the development phase and in the evaluation of a new information system.

Boehm's article states that the key verification and validation criteria that can be used in the development phase of a new information system are completeness, consistency, feasibility, and testability (Boemh and Barry W. 1984). Fulfilling the testability criterion in the development of a new system can help prepare a system for testing and evaluation including having proper test criterion developed and having assurances in place in regards to specifics like privacy or accessibility (based on agreed upon values with the end user). While Boehm's article does not provide effective frameworks for evaluating completed software, the validation and two verification criteria that have been developed can be used as a benchmark for what an effective information system should be from a technical perspective.

2.4.2.3 Information Quality

Another technical factor that can be evaluated is information quality. Mohammed and Yusof's research noted that information quality management is a key practice in health information systems, and therefore should be involved in HIS evaluations (Mohammed and Yusof 2013); however, only half of the evaluations in Sadoughi et al.'s review included evaluations of this specific sub factor (Sadoughi et al. 2013). Mohammed and Yusof's literature review focused on information quality and quality management in HIS, and they found six frameworks that depicted quality management practices. They split the criteria from these frameworks into human, organization, and technology factors and further developed integrated IQM evaluation criteria. Their integrated framework suggests that IQM evaluation requires inputs from across the information life cycle, and have effectively developed a matrix that places key evaluation criteria in sections based on whether it is a human, organizational, or technical factors as well as where it exists within the information life cycle (Mohammed and Yusof 2013).

2.4.3 Component Segmentation and System Topology

While Mohammed and Yusof proposed that system topology should be considered in the evaluation design, no detailed metrics or examples were provided (Mohammed and Yusof 2013). Mohammed and Yusof's inclusion of topology refers more to the database design, and is not inclusive of systems that include components that exist both within and outside a hospital's network. In Yusof et al.'s HOT-Fit framework, the system is continuously treated as one unit, instead of a grouping of components that could potentially have their own metrics (Yusof et al. 2008). Researchers evaluating a system's GUI have separated that portion of the system from the rest for the purpose of evaluating usability (Khajouei, Esfahani, and Jahani 2016; Kopanitsa, Tsvetkova, and Veseli 2012); however, these evaluation methods ignore the other technical components of the system altogether.

2.4.4 Availability and Data Recovery Time

Availability is an important metric in the implementation and evaluation of a CDSS. Defined as being present or ready for use (Bhagwan, Savage, and Voelker 2003), it is associated with the system quality dimension of the HOT-Fit Framework(Yusof et al. 2008). The purpose of measuring availability is to determine how often the system can be used, and how often the system is inaccessible.

In two different reviews of factors that influence CDSS, Kilsdonk et al. found that availability had been measured as a system quality factor in CDSS evaluations (Kilsdonk, Peute, and Jaspers 2012); however, it's measurement occurred in less than 10 of 35 evaluations they reviewed (Kilsdonk, Peute, and Jaspers 2017). Sadhougi et al.'s review of HIS evaluation methods did not include any evaluations that measured availability (Sadoughi et al. 2013). Considering the difference in the types of HIS and CDSS that exist, it is not surprising that availability is not a commonly measured measure. Kilsdonk et al. noted in their gap analysis that service quality (of which availability is a part) had the fewest evaluation measures in their review.

Despite the fact that availability hasn't regularly been measured, it is imperative that it is included in the evaluation of any CDSS uses data in real-time since it is crucial that data is available for any decision making. In the case of the Artemis Platform, it is expected that data is consistently being collected from bedside monitoring devices for analysis and storage. In stating the expectation of consistent data collection, an availability metric is already inferred. In systems that aren't consistently used, availability may not be as important. Since the system is only expected to be used at a given point in time to meet a specific purpose, maintenance and down time can be managed around the schedule of the system use. Different expectations exist for systems where data is constantly being generated and changing with the expectation that it is being consistently utilized at all times.

It is unrealistic and impractical to expect 100 percent availability for 24 hours a day, 7 days a week. There are many technical reasons why a system may fail including network issues, software or hardware failure, and user misuse, which negatively affect the availability of a system. The challenge in proposing a consistently used system is determining an availability metric that takes in to account prospective issues and system maintenance times (Bhagwan, Savage, and Voelker 2003) and includes ways to mitigate or reduce the amount of down time that the system will have. What has not been considered in the literature, is partial system availability and its metrics. For example, if the monitoring portion of a system is available, but the analysis portion is not. An approach to measure this form of system's availability is a new research area.

2.4.5 Implementation Artifacts in Evaluations

The articles reviewed provide little detail about implementation artifacts and their use as they typically report on already completed HIS evaluations. Borim et al. did validate the importance of understanding the details of the system, the results aimed for, and to complete a literature review to characterize which metrics should be included (Borim et al. 2015). The system topology and the goals of using the system would be obtained through implementation documents and discussions with the stakeholders involved in the system implementation. Borim et al. also suggested completing a literature review to find what metrics and evaluation techniques were used for similar HIS. Other articles focus on what was evaluated, and not necessarily on how the metrics and benchmarks were determined. Mohammed and Yusof's review of IQM practices does mention the use of some implementation artifacts, as they note that documentation created in the different stages of the information life cycle (ILC) could be used in the planning of an evaluation proposal and methodology (Mohammed and Yusof 2013). Scenarios developed from training documentation and manuals for HE and CW are also applicable as implementation artifacts used for the purpose of determining HIS usability in an evaluation (Khajouei, Esfahani, and Jahani 2016). Overall, little detail is provided about specific implementation artifacts, like the privacy impact assessment (PIA), threat and risk assessment (TRA) and research proposal, and how they may impact the evaluation of a HIS.

2.4.6 Frameworks and Methods

There are many more factors that are included in evaluations outside of behavioural (or human/user), technical, and organizational factors; however, these three factors are large components of the prominent HOT-fit evaluation method (Yusof 2015; Yusof et al. 2008). HOT-fit, developed by Yusof et al. utilizes the IS success model to categorize evaluation factors, with the IT-Organization Fit Model to incorporate the concept of fit between specific evaluation factors (Yusof et al. 2008). HOT-fit utilizes eight interrelated dimensions (similar to factors), which influence each other. The importance of fit comes from research that showed that a lack of fit between new information systems and human/organizational factors was a major factor in the failure of information system implementations in the healthcare sector (Yusof et al. 2008). Figure 3 depicts the HOT-fit framework.

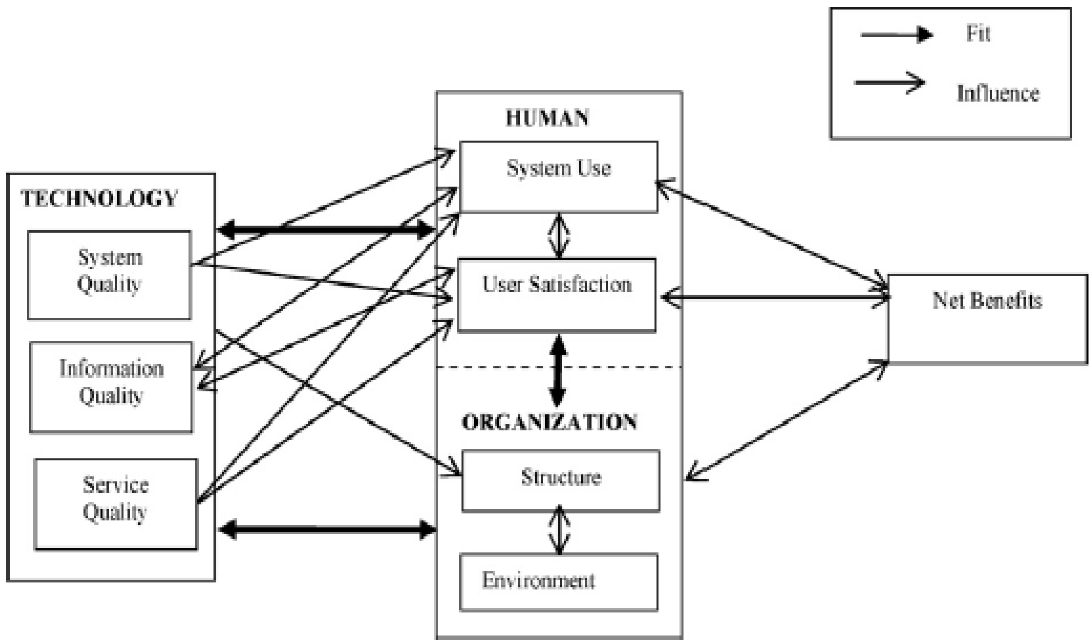

Figure 3 - HOT-fit Framework

The technology factor of HOT-Fit consists of three domains: system quality, information quality, and service quality. The system quality domain includes potential metrics for data accuracy and platform availability. The information quality domain includes reliability and timeliness. The service quality domain includes more soft-skill focused metrics including the usefulness of technical support and other follow up services. The human factor of HOT-Fit consists of two domains: system use and user satisfaction. The system use domain includes metrics about how the system is used by end users, and the user satisfaction domain includes metrics about their enjoyment and their opinions of the system. The organization factor includes the domains structure and environment. Metrics in the structure domain involve how the system is viewed by leadership, and the process for which the system was implemented. The environment domain includes metrics on the interactions with those external to the organization, such as governments, that may have involvement or interest in the system. The eighth domain is net benefits, which include metrics that demonstrate how the system has made improvements, be it in performance, workflow, or clinical outcomes.

The HOT-fit framework’s eight dimensions may fit with or influence each other. The model suggests that an evaluation of ease of learning (system quality), would influence overall satisfaction with the system (user satisfaction), which provides net benefits to the organization. Improved system quality also will be influential in top management support (structure), which may be driven by user satisfaction as well. The inter-relatability of these evaluation dimensions suggests that an effective information system must fit and influence strongly in all aspects of the evaluation to be considered a success; however, the HOT-fit framework does not define specific tools or methods to be used to complete an evaluation, but does propose that components related to the eight dimensions are necessary.

Eivazzadeh et al. developed a method entitled the Unified eValuation using ONtology (UVON) to assist in the selection of evaluation sub factors. They wrote that often evaluators either use evaluation frameworks that aren’t specific enough to their case, or develop their own frameworks that are too specific and are therefore only usable for that specific case (Eivazzadeh et al. 2016). The UVON method is meant to address what should be evaluated in a health information system. They propose that aspects, or sub factors from existing frameworks can be effectively managed through an ontology as a formal and computable way of capturing evaluation information effectively (Eivazzadeh et al. 2016).

As noted in section 2.2, The Statement on Reporting of Evaluation Studies in Health Informatics (STARE-HI) was developed as a guideline for how an evaluation study should be reported to the research community. The 38 elements included in the guideline give structure to evaluation reporting, although they do not assist in the determination of a framework to use, or areas to evaluate. Simply, the guideline is useful for organizing a study and reporting the outcomes (Brender et al. 2013).

2.5 Discussion

Researchers have continued to build upon and utilize the guidelines set out through STARE-HI and GEP-HI in developing best practices for evaluations, as well as new evaluation frameworks. Effective frameworks can cover a multitude of evaluation factors, although the existing frameworks do not necessarily suggest specific tools, methods, or activities that can be completed during a health information system evaluation. Some reviews (Ahmadian, Salehi, and Khajouei 2015; Borim et al. 2015; Sadoughi et al. 2013) found a large variety of evaluation methods, with questionnaires being the most common evaluation method. Tautologically, other researchers have looked into which methods are effective within an evaluation as a way to suggest how evaluations can be completed effectively (Bassi and Lau 2013; Khajouei, Esfahani, and Jahani 2016; Kopanitsa, Tsvetkova, and Veseli 2012; Mohammed and Yusof 2013). Borim et al. recommend completing reviews of evaluation metrics and methods before completing an evaluation as a way to hone in best practices and methods being used specific to the type of system being evaluated (Borim et al. 2015).

The challenge with those evaluation frameworks, and the metrics they include are that they are meant to be very general, and don't account for the intricacies of high frequency streaming data and public health interventions. The frameworks tend to be much more focused on the system use as opposed to the clinical or public health outcomes associated with using the system. Perhaps this is because many HIS are not implemented with a CDSS focus. While improved clinical outcomes were discussed as a potential evaluation metric that highlights the effectiveness of a new system or as a way to highlight the economic impact of the system, detailed metrics of how this impacts population health were not included in the evaluation frameworks.

Additionally, the evaluation frameworks do not account for details like system topology and data recovery time; two key aspects that can define a system like the Artemis Platform's use. Researchers have suggested evaluating system components like the GUI separate from the technical components or back-end database that define the system, but have not proposed evaluation metrics that take in to account each component's role in the system, as well as its location (be it in or outside the hospital network).

The Artemis Platform, and other CDSSs with a similar high frequency streaming Big Data approach are emerging, and the evaluation of the system needs to consider the uniqueness of these systems. As the systems are being implemented, there is an opportunity to use artifacts of system implementation to assist with the development of metrics for the evaluation. At a minimum, an understanding of system design is crucial to the development of an evaluation plan (Borim et al. 2015). Implementation artifacts help define the system topology, the security measures being taken place to ensure that patient data remains safe and that the system functions accurately, and provide definitions and objectives for how the system provides benefit to the health system and population as a whole.

2.6 Conclusion and Research Implications

There is no shortage of documented factors and metrics to consider when evaluating a CDSS, although most literature focuses on different ways to evaluate user satisfaction. Nevertheless, clinical outcomes have an acknowledged role in evaluating the success of a HIS. For CDSSs this will be more relevant because of the implication that CDSSs exist to enhance clinical care. For this research, the evaluation of the Artemis Platform should include metrics that assess the Platform’s impact on population and patient health.

The existing research reviewed provides little detail on how technical metrics should be managed. Technical metrics are often presented at a high level, with metrics being used to provide an overall assessment of the system. Component specific metrics should exist, as it allows evaluators to assess the strengths and weaknesses of system components like databases, network connections, and visualizations separately or as a whole. This is especially relevant in systems that have an architecture spanning multiple organizations.

Some of the metrics presented give credence to the idea of using implementation artifacts to support the construction of the evaluation plan. Artifacts created in the planning stage of the ILC could be used as part of the evaluation (Mohammed and Yusof 2013), while the evaluation scenarios built for HE and CW could be considered implementation artifacts based on the fact that they are developed using technical documents that support the systems implementation (Khajouei, Esfahani, and Jahani 2016). High-level knowledge of system design and evaluation metrics used for similar systems can also assist in the development of an evaluation plan (Borim et al. 2015).

Further research that focuses on CDSSs evaluations should consider how clinical and population outcomes are impacted by the implementation and use of a CDSS. For CDSSs being implemented for multiple purposes (for example having multiple algorithms for different conditions) or being implemented in multiple sites, evaluation metrics can be segmented to specific CDSS components for the purpose of providing more detail. An overall metric can still be used in the evaluation of the system, but segmenting component evaluators will lead to a more detailed and actionable evaluation. Using implementation artifacts to help set benchmarks and determine component segmentation is also an opportunity that can be explored further.

2.7 References

Ahmadian, Leila, Simin Salehi, and Reza Khajouei. 2015. "Evaluation Methods Used on Health Information Systems (HISs) in Iran and the Effects of HISs on Iranian Healthcare : A Systematic Review." *International Journal of Medical Informatics* 84(6): 444–53.

Bassi, Jesdeep, and Francis Lau. 2013. "Measuring Value for Money : A Scoping Review on Economic Evaluation of Health Information Systems." *Journal of the American Medical Informatics Association* 20(1): 792–801.

Berner, Eta S. 2007. *Clinical Decision Support Systems - Theory and Practice*. ed. Eta S. Berner.

Berner, Eta S. 2009. "Clinical Decision Support Systems : State of the Art." (09).

Bhagwan, Ranjita, Stefan Savage, and Geoffrey M Voelker. 2003. "Understanding Availability." *Peer to Peer Systems II*: 1–6.

Boemh, and Barry W. 1984. "Verifying and Validating Software Requirements and Design Specifications." (January).

Borim, Heloa Costa et al. 2015. "Integrating Methods to Evaluate Health Information Systems." *Context Sensitive Health Informatics: Many Places, Many Users, Many Contexts, Many Uses*.

Brender, J. et al. 2013. "STARE-HI – Statement on Reporting of Evaluation Studies in Health Informatics." : 331–58.

Eivazzadeh, Shahryar et al. 2016. "Evaluating Health Information Systems Using Ontologies." *JMIR Medical Informatics* 4(2): 1–13.

Heidebrecht, Christine L et al. 2014. "Electronic Immunization Data Collection Systems : Application of an Evaluation Framework."

Khajouei, Reza, Misagh Zahiri Esfahani, and Yunes Jahani. 2016. "Comparison of Heuristic and Cognitive Walkthrough Usability Evaluation Methods for Evaluating Health Information Systems." *Journal of the American Medical Informatics Association* (0): 1–6.

Kilsdonk, E, L W Peute, and M W M Jaspers. 2017. "International Journal of Medical Informatics Factors Influencing Implementation Success of Guideline-Based Clinical Decision Support Systems : A Systematic Review and Gaps Analysis." *International Journal of Medical Informatics* 98: 56–64. http://dx.doi.org/10.1016/j.ijmedinf.2016.12.001.

Kilsdonk, E, L W P Peute, and M W M Jaspers. 2012. "Factors Known to Influence Acceptance of Clinical Decision Support Systems."

Kitchin, Rob, and Gavin Mcardle. 2016. “What Makes Big Data , Big Data ? Exploring the Ontological Characteristics of 26 Datasets.” (June): 1–10.

Kopanitsa, Georgy, Zhanna Tsvetkova, and Hasan Veseli. 2012. “Application of Usability Metrics in a Multi-User and Multimedia EHR Evaluation Framework.” In *Quality of Life through Quality of Information*, IOS Press, 358–62.

Lund, By Arnold M. 2001. “Usability Interface Measuring Usability with the USE Questionnaire.” *Usability Interface* 8(2). http://www.stcsig.org/usability/newsletter/0110_measuring_with_use.html.

McGregor, Carolyn et al. 2013. “Real-Time Multidimensional Temporal Analysis of Complex High Volume Physiological Data Streams in the Neonatal Intensive Care Unit.” *Studies in Health Technology and Informatics* 192(1–2): 362–66.

Mohammed, Siti Asma, and Maryati Mohd. Yusof. 2013. “Towards an Evaluation Framework for Information Quality Management (IQM) Practices for Health Information Systems – Evaluation Criteria for Effective IQM Practices.” *Journal of Evaluation in Clinical Practise* 19: 379–87.

Nielsen, Jakob. 1994. “Ten Usability Heuristics.” http://intra.iam.hva.nl/content/1112/verdieping1/research_for_design/intro-en-materiaal/RfD-Heuristic-Evaluation.pdf.

Raghupathi, Wullianallur, and Viju Raghupathi. 2014. “Big Data Analytics in Healthcare : Promise and Potential.” : 1–10.

Rigby, M. et al. 2013. “Evidence Based Health Informatics : 10 Years of Efforts to Promote the Principle.” *IMIA Yearbook of Medical Informatics*: 34–46.

Sadoughi, Farahnaz, Khalil Kimiafar, Maryam Ahmadi, and Mohammad Taghi Shakeri. 2013. “Determining of Factors Influencing the Success and Failure of Hospital In-Formation System and Their Evaluation Methods : A Systematic Review.” *Iran Red Crescent Medical Journal* 15(12).

Wang, Yichuan, Leeann Kung, and Terry Anthony. 2018. “Technological Forecasting & Social Change Big Data Analytics : Understanding Its Capabilities and Potential Bene Fi Ts for Healthcare Organizations.” 126: 3–13.

World Health Organization. 2008. “Health Information Systems.” (June).

Yusof, Maryati Mohd. 2015. “A Case Study Evaluation of a Critical Care Information System Adoption Using the Socio-Technical and Fit Approach.” *International Journal of Medical Informatics* 84(7): 486–99.

Yusof, Maryati Mohd, Jasna Kuljis, Anastasia Papazafeiropoulou, and Lampros K. Stergioulas. 2008. “An Evaluation Framework for Health Information Systems: Human, Organization and Technology-Fit Factors (HOT-Fit).” *International Journal of Medical Informatics* 77(6): 386–98.

Chapter 3. Implementation of the Artemis Platform at the McMaster Children's Hospital Neonatal Intensive Care Unit

This chapter presents the scenario for which the development of a new evaluation plan was required. It includes descriptions about the artifacts used in Ontario-based implementations of IT systems at hospitals. After introducing the scenario, three key implementation documents used for the study are described.

3.1 Introduction

Beginning in 2016, researchers with OTHIR and the MCH NICU worked to implement the Artemis Platform within MCH as part of a research study. The implementation process included the creation of an ethics proposal, which governs the research study portion of the implementation, a TRA and a PIA.

3.1.1 The Ontario Tech Health Informatics Lab (OntarioTechHIR)

The OntarioTechHIR lab, led by Dr. Carolyn McGregor is a research lab focused on improving the understanding of temporal patterns in physiological data collected through sensors and devices. The goals of the lab include using physiological data and algorithms to provide decision support, and researching the impact of changes in physiological signals throughout activities and events (Ontario Tech HIR 2019). Some of the areas of research include astronaut and first responder health; however, the main research area for the lab is the health of preterm infants. Researchers at the OntarioTechHIR lab have received funding and support for numerous implementations of Artemis and other big data analytics research in North American hospitals, as well as internationally in China and Australia (McGregor 2008). One of the sources of funding for OntarioTechHIR is FedDev Ontario through the Health Ecosphere Innovation Pipeline. In a press release, FedDev Ontario noted that the aim of using Artemis, in collaboration with hospitals and industry partners, is to reduce the rate of mortality in premature babies, which represents a public health outcome goal (FedDev Ontario 2017).

3.1.2 The McMaster Children's Hospital Neonatal Intensive Care Unit

The MCH NICU is one of five NICUs in Ontario with a 3b level of care designation, which is the highest level of care for an Ontario NICU (Provincial Council for Maternal and Child Health 2018). As a 3b NICU, they provide care to neonates of any gestational age or weight, have mechanical ventilation support available, have a comprehensive range of consultants available, and have on site surgical capability (Provincial Council for Masternal and Child Health 2013).

The MCH NICU provides services to 22 regional hospitals (which have lower level NICUs that may feed in to MCH) within the LIHN 3 and LIHN 4 areas of Ontario. The NICU has 51 beds, as well as an additional 14 beds in an intermediate care nursery. They provide care to approximately 1500 infants annually. The MCH NICU considers research to be a very important part of the quality of care delivered at the hospital, and conduct multiple research studies within the unit with voluntary participation from patients. One of the clinics mandates is to maximize the development of potential infants (McMaster Children's Hospital 2019).

3.1.3 The Artemis Platform

The Artemis Platform is a cloud platform capable of analysing multiple physiological data streams in real-time. It was first deployed in the NICU at the Hospital for Sick Children in August 2009 as a tool to gather physiological data from bedside monitors and supported clinical studies on a variety of conditions including late-onset neonatal sepsis (LONS) (McGregor 2013). The Artemis Platform uses an application program interface (API) to ingest large amounts of physiological data from monitors at each bedside in a real time streaming cloud architecture. The data captured is fully de-identified in the hospital environment before being transmitted to the cloud, which exists outside the hospital environment (Inibhunu et al. 2019). Multiple algorithms analyze data during the transmission for physiological conditions before being stored in a standard database format. Data is accessible from the Artemis Platform databases for visualization. This architecture has since been deployed at both the MCH NICU and at Southlake Regional Hospital's NICU (Inibhunu et al. 2019).

Ontario implementations of Artemis leverage the use of the Ontario Research and Innovation Optical Network (ORION) (ORION 2019), a provincial research and education internet network capable of ultra-fast data transmission across Ontario (ORION 2020) to transmit data to the Centre for Advanced Computing (CAC) at Queen's University, where the cloud environment is located. The CAC is a highly secure high-performance computing environment (Centre for Advanced Computing 2020). The analysis and storage components of the Artemis Platform occur within the cloud environment at the CAC. Figure 4 depicts the Artemis Platform system topology. The Platform collects data from bedside monitors and transmits the data to a local server within the hospital. Data is then sent via ORION to the CAC for processing, transformation, analysis, storage, and visualization. These visualizations can be reviewed at the hospital and are the clinical deliverable from the Platform.

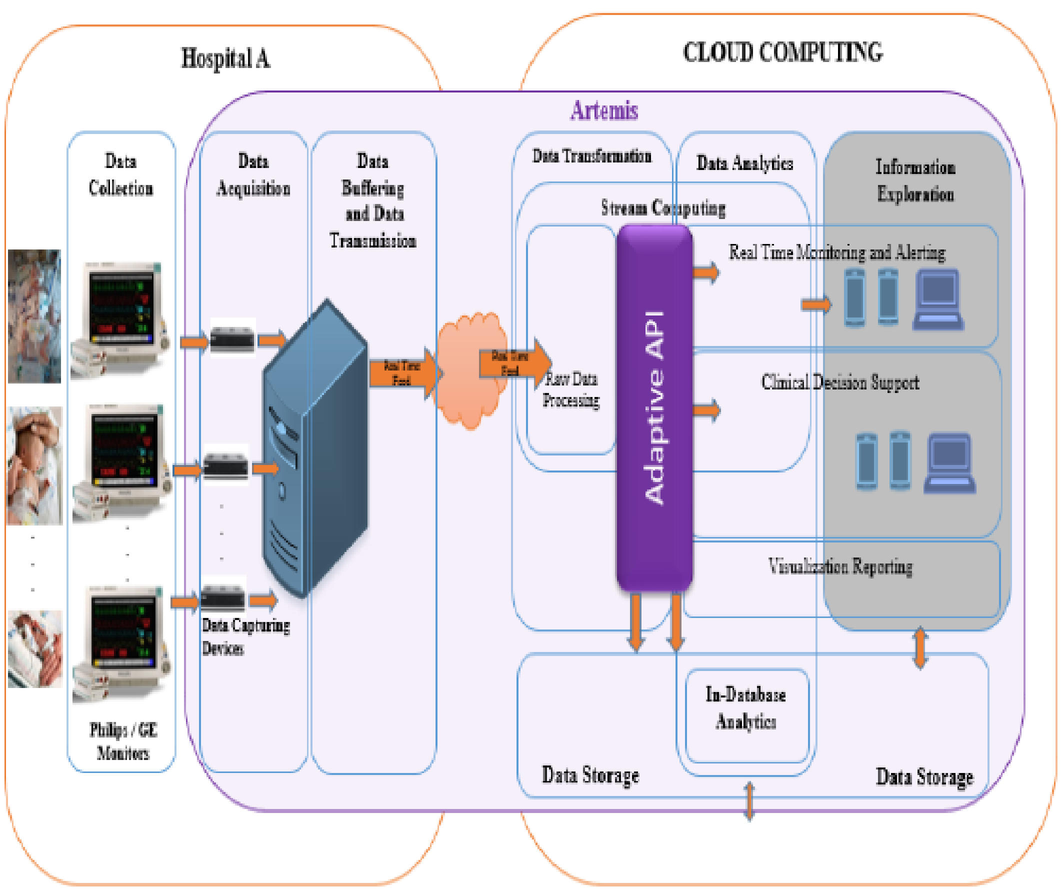

Figure 4 Artemis System Topology

3.1.4 Late Onset Neonatal Sepsis

Neonatal sepsis is defined by the presence of infection in the bloodstream, for which bacteria and viruses are frequently the cause. Researchers categorize neonatal sepsis within the first 3 days after birth as early onset, while any sepsis afterwards is categorized as late onset (Dong and Speer 2015; Stoll et al. 2002). Late onset neonatal sepsis (LONS) is especially prevalent in preterm infants and those with very low birth weight. LONS is more often connected with the environment after the birth of the child as opposed to early onset neonatal sepsis, which is usually connected to the maternal environment (Dong and Speer 2015).

LONS is a significant cause of morbidity and mortality in preterm infants, the risks of which are increased in those with extreme prematurity and very low birth weight (Stoll et al. 2002). In a research study in the United States, which analyzed data from over 6000 neonates, those with very low birth weight, researchers found that more than one in five developed LONS. Neonates that developed LONS had an increased rate of mortality when compared to very low birth weight neonates without LONS. Those with LONS also required a longer hospital stay (Stoll et al. 2002), were more likely to need invasive interventions, and less likely to begin breastfeeding (Dong and Speer 2015).

A timely and accurate diagnosis of LONS is especially important given the high morbidity and mortality rates associated with the condition. Blood culture testing, a definitive way to determine the presence of infection in the blood stream remains the defining diagnostic procedure; however, there is an inevitable time lag between drawing blood and confirming the infection (Dong and Speer 2015).

Researchers have focused on the early detection of LONS by utilizing physiological data, specifically the heart rate of the neonate (Coggins et al. 2016; Fairchild and Shea 2011; Griffin et al. 2003). In a cohort study completed within the United States, Griffin et al. found that reduced heart rate variability was present in some infants 12 to 24 hours before sepsis was clinically diagnosed in the patient. They proposed that a predictive model that used continuous, non-invasive monitoring of heart rate characteristics could be an effective strategy for improving patient outcomes in the NICU (Griffin et al., 2003).

Using the HerO monitoring system, researchers at Vanderbilt University retrospectively analyzed heart rate characteristics using a scoring system they developed to predict bloodstream infections in neonates consistent with LONS. They found that heart rate characteristic monitoring in clinical practice was uncertain, and that high scores in their predictive scoring system was not sensitive or specific to infection. They concluded that heart rate characteristic scores have limited ability to predict infection (Coggins et al. 2016).

Through the development of the Artemis platform, which is capable of analyzing and multiple streams of physiological data concurrently in real-time, Dr. McGregor et al. combined heart rate variability and respiratory rate variability in an algorithm to detect LONS (McGregor, Catley, and James 2012). The algorithm built within the Artemis platform has a high rate of accuracy in detecting LONS hours before clinical symptoms.

3.1.5 Artemis Platform Implementation at the MCH NICU

Proposals for the implementation of the Artemis platform by UOITHIR at the MCH NICU was submitted to the Hamilton Integrated Research Ethics Board (HiREB) and the UOIT Research Ethics Board. An ethics proposal was required because the Platform was being implemented as part of a validation research study and requires the use of identified patient demographic and physiological data. The Platform also transmits some of this data to a cloud platform outside of the hospital network for analysis.

The proposal submitted to HiREB focuses on the use of the late onset neonatal sepsis (LONS) algorithm within the Artemis platform (Pugh et al. 2018). The LONS algorithm, developed by Dr. McGregor and demonstrated through a case study with the Hospital for Sick Children, uses physiological data including heart rate and respiratory rate to predict the presence of sepsis within neonates(McGregor et al. 2013). Outlined in the proposal is the researchers plan to collect physiological data using Artemis from consenting patients, validate the LONS algorithm by comparing findings from Artemis to a panel of clinicians, and eventually implement the platform as a clinical decision support system capable of providing real-time data and support to clinicians (Pugh et al. 2018).

3.2 The Ethics Proposal

The Canadian Institutes of Health Research, partnering with the Natural Sciences and Engineering Research Council of Canada and Social Sciences and Humanities Research Council of Canada created a policy for researchers to help ensure that research is conducted in an ethical way specifically when the research involves humans (Canadian Institutes of Health Research, Natural Sciences and Engineering Research Council of Canada, and Canada 2014). Since CDSSs are expected to support clinicians with their decision making, and therefore impact patient care, an ethics proposal is a necessity.

Ethics proposals are required to contain key information to help decision makers and stakeholders ensure that the research being conducted is being conducted in a proper manner, and that risks and benefits have been balanced. The ethics proposal is also used to ensure that consent has been given voluntarily, as opposed to being received through methods like persuasion or improper incentives. Information around the privacy and confidentiality and data is also an important aspect of the ethics proposal, though more detailed documentation of information privacy and security would be instead highlighted in the PIA.

From a content perspective, the ethics proposal is a crucial document because it explains key details of the study including the objectives and the methods. This information is valuable in positioning the purpose of the study, its expected benefits, and provides criteria that can be interpreted to determine whether or not the study is successful or not.

3.3 The Threat and Risk Assessment

The Threat and Risk Assessment (TRA) (Communications Security Establishment and Royal Canadian Mounted Police 2007) is a Canadian document that is usually required as part of the implementation of information technology projects. The objective of preparing a TRA is to inform decision makers about the variety of threats and risks that are possible when implementing a new system, as well as the probability of the threats occurring and their overall impact (Tusikov and Fahlman 2008). Additionally, mitigating factors and preventative measures are proposed and discussed within the TRA to provide decision makers with complete information regarding the new implementation. The methodology for preparing a TRA was prepared in 2007 by the Communication Security Establishment (CSE) and Royal Canadian Mounted Police (RCMP), after they identified issues in the preparation of TRAs across government institutions (Communications Security Establishment and Royal Canadian Mounted Police 2007).

The Canadian Government's Security Risk Management policy outlines key deliverables of a TRA (Communications Security Establishment and Royal Canadian Mounted Police 2007):

1. Establish the scope of assessment and identify employees and assets to be safeguarded
2. Determine the threats to employees and assets in Canada and abroad, and assess the likelihood and impact of their occurrence
3. Assess the vulnerabilities based on the adequacy of safeguards and compute the risk
4. Implement additional safeguards, if necessary, to reduce risk to an acceptable level

Figure 5 Four-Step TRA Process (Communications Security Establishment and Royal Canadian Mounted Police 2007)

In the implementation of a new CDSS with components in an off-site cloud environment, a TRA is imperative for identifying technical measures related to security, as well as important performance metrics relating to the system's utilization of data. Since CDSSs utilize data as an input in supporting clinical decision making, risks to the integrity and availability of the data need to be addressed, and benchmarks should be set that can be used in the verification and validation of the system before implementation.

Through identifying risks and addressing vulnerabilities through the development of safeguards, the TRA promotes the effective governance of new IT systems through the lens of ensuring security. Developing one as part of the implementation of a CDSS is an important task that not only provides decision makers with the important information they need, but also can help in the validation, verification, and eventual validation of the system.

3.4 The Privacy Impact Assessment

Before an implementation of a CDSS that utilizes personal health information can begin, a privacy impact assessment (PIA) must be completed to ensure that the system is compliant with legislation. Implementations in Ontario are subject to the Freedom of Information and Protection of Privacy Act (FIPPA) as well as PHIPA (Personal Health Information Protection Act). The objective of the PIA is to identify how the system interacts with personal information, so that gaps in security can be identified and effectively resolved before the implementation of the system (Information and Privacy Commisioner of Ontario 2015).

By preparing the PIA, implementers of the CDSS are required to provide key information about how the system utilizes personal health information. This information includes biological data and biographical data (Information and Privacy Commisioner of Ontario 2015). While physiological data isn't specifically identified in The Information and Privacy Commissioner of Ontario's Privacy Impact Assessment Guide, data such as electrocardiogram readings (ECG) could eventually be used as identifying information and therefore should be including in a PIA as well (Biel et al. 2014).

In addition to the type of information being collected, the flow of information through the system must be considered. This includes how it is acquired, where data travels, and how it is stored. The retention and destruction of data must be considered in accordance with legislation. For each piece of personal health information, questions of collection, use, retainment, security, disclosure, and disposal are paramount (Information and Privacy Commisioner of Ontario 2015).

Upon the identification of relevant details, those completing the PIA must provide analysis on the privacy of the system. This includes whether there are risks to the privacy of personal health information, and the identification of safeguards that can be implemented to reduce or eliminate these risks. These risks can include the collection of irrelevant information, the use of information in an unauthorized manner, the failure to keep information secure, and more (Information and Privacy Commisioner of Ontario 2015). Solutions that minimize or fully address risks to personal health information are included in the PIA as well as an implementation plan.

By identifying risks to personal health information, and providing solutions that minimize or address these risks, the PIA promotes data security for a new CDSS. Identifying these privacy measures and data security standards is important for ensuring the system meets government standards and can be approved by an organization. Key information from the PIA includes the content and flow of data, as well as the standards for ensuring its security.

3.5 Conclusion and Next Steps

To implement a CDSS that uses personal health information from patients to support clinical decision making in Ontario, researchers need to complete key implementation artifacts – the ethics proposal, TRA, and PIA. All three of these documents are valuable in a system evaluation because they provide key information like goals, objectives, system topology, and other implementation details.

Chapter 4 presents the evaluation plan template for evaluating a high frequency streaming analytics CDSS being implemented with a research study within a hospital environment.

3.6 References

Biel, L, O Pettersson, L Philipson, and P Wide. 2014. “ECG Analysis : A New Approach in Human Identification.” (April).

Canadian Institutes of Health Research, Natural Sciences and Engineering Research Council of Canada, and and Social Sciences and Humanities Research Council of Canada. 2014. *Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans*. http://www.pre.ethics.gc.ca/eng/policy-politique/initiatives/tcps2-eptc2/Default/.

Centre for Advanced Computing. 2020. “What Is the CAC?” https://cac.queensu.ca/about_us/.

Coggins, Sarah A. et al. 2016. “Heart Rate CHaracteristic Index Monitoring for Blood Stream Infection in an NICU: A 3-Year Expierence.” 106(3): 201–7.

Communications Security Establishment, and Royal Canadian Mounted Police. 2007. *Harmonized Threat and Risk Assessment (TRA) Methodology*.

Dong, Ying, and Christian P. Speer. 2015. “Late-Onset Neonatal Sepsis:Recent Developments.” *Archives of Disease in Childhood: Fetal and Neonatal Edition* 100(3): F257–63.

Fairchild, Karen D, and T Michael O Shea. 2011. “Heart Rate Characteristics: Physiomarkers for Detection of Late- Onset Neonatal Sepsis.” *Clin Perinatol* 37(3): 581–98.

FedDev Ontario. 2017. “The Health Ecosphere Innovation Pipeline Is Improving Lives with New Healthcare Advances.” https://www.feddevontario.gc.ca/eic/site/723.nsf/eng/02416.html?OpenDocument.

Griffin, M. Pamela et al. 2003. “Abnormal Heart Rate Characteristics Preceding Neonatal Sepsis and Sepsis-like Illness.” *Pediatric Research* 53(6): 920–26.

Information and Privacy Commisioner of Ontario. 2015. *Planning for Success : Privacy Impact Assessment Guide*.

Inibhunu, Catherine et al. 2019. “Adaptive API for Real-Time Streaming Analytics as a Service.” *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*: 3472–77.

McGregor, Carolyn. 2008. “Ontario Tech Health Informatics Research.” hir.uoit.ca.

———. 2013. “Big Data in Neonatal Intensive Care.” *Computer* 46(6): 54–59.

———. 2013. “Real-Time Multidimensional Temporal Analysis of Complex High Volume Physiological Data Streams in the Neonatal Intensive Care Unit.” *Studies in Health Technology and Informatics* 192(1–2): 362–66.

McGregor, Carolyn, Christina Catley, and Andrew James. 2012. “Variability Analysis with Analytics Applied to Physiological Data Streams from the Neonatal Intensive

Care Unit." *2012 25th IEEE International Symposium on Computer-Based Medical Systems (CBMS)*: 1–5.

McMaster Children's Hospital. 2019. "Neonatal Nurseries." http://www.mcmasterchildrenshospital.ca/body.cfm?id=85&fr=true.

Ontario Tech HIR. 2019. "Ontario Tech HIR Twitter Page." twitter.com/uoithir.

ORION. 2019. "Saving Ontario's Most Vulnerable Infants." https://www.orion.on.ca/news-events/blog/saving-ontarios-most-vulnerable-infants/.

———. 2020. "Our Network."

Provincial Council for Masternal and Child Health. 2013. "Standardized Maternal and Newborn Levels of Care Definitions." : 1–12.

Provincial Council for Maternal and Child Health. 2018. *Level of Care Designation*. http://www.pcmch.on.ca/wp-content/uploads/2018/06/LOC-Website-List-June-2018-Update.pdf.

Pugh, James Edward et al. 2018. *Artemis Cloud Clinical Decision Support Framework - An Algorithm to Accurately Detect Late on Set Neonatal Sepsis*.

Stoll, B. J. et al. 2002. "Late-Onset Sepsis in Very Low Birth Weight Neonates: The Experience of the NICHD Neonatal Research Network." *Pediatrics* 110(2): 285–91. http://pediatrics.aappublications.org/cgi/doi/10.1542/peds.110.2.285.

Tusikov, Natasha, and Robert C. Fahlman. 2008. *Threat and Risk Assessments*.

Chapter 4. The Evaluation Methodology

This chapter presents the evaluation methodology as a contribution to the thesis. This evaluation methodology addresses the research questions of this thesis. The evaluation methodology, which is being developed specifically to manage the evaluation of population, technical, and algorithm specific metrics, leverages key implementation documents in the creation of metric definitions. The method in which new metrics are created allows for scalability both in technical metrics when there are implementations at more hospitals, and algorithm specific metrics when new algorithms added to the Platform.

4.1 Evaluation Methodology Construct

The evaluation methodology allows for the creation of evaluation objectives, questions, and metrics in a structure used in the PHO evaluation template. The methodology includes a pathway for creating population health metrics. As noted in Chapter 2, while specific HIS evaluation templates and frameworks exist, they are lacking in their measurement of the system's impact on population health. The evaluation methodology is a five-step process for evaluating a CDSS implemented as a health intervention in a hospital environment. The methodology is depicted in Figure 6.

4.1.1 Discover a Need – Implement CDSS as a Solution

The first step of the methodology (coloured in Figure 6 as red) is the discovery of a need where a CDSS is implemented as a health intervention solution. In this step, clinicians discover a population with a clinical need and propose a research study. This is documented in a research and ethics proposal. In parallel, an existing CDSS with an algorithm-based intervention is proposed as the system implementation for the research study. During the system implementation phase, system documentation, a threat and risk assessment, and privacy impact assessment are provided to hospital IT and management for approval.

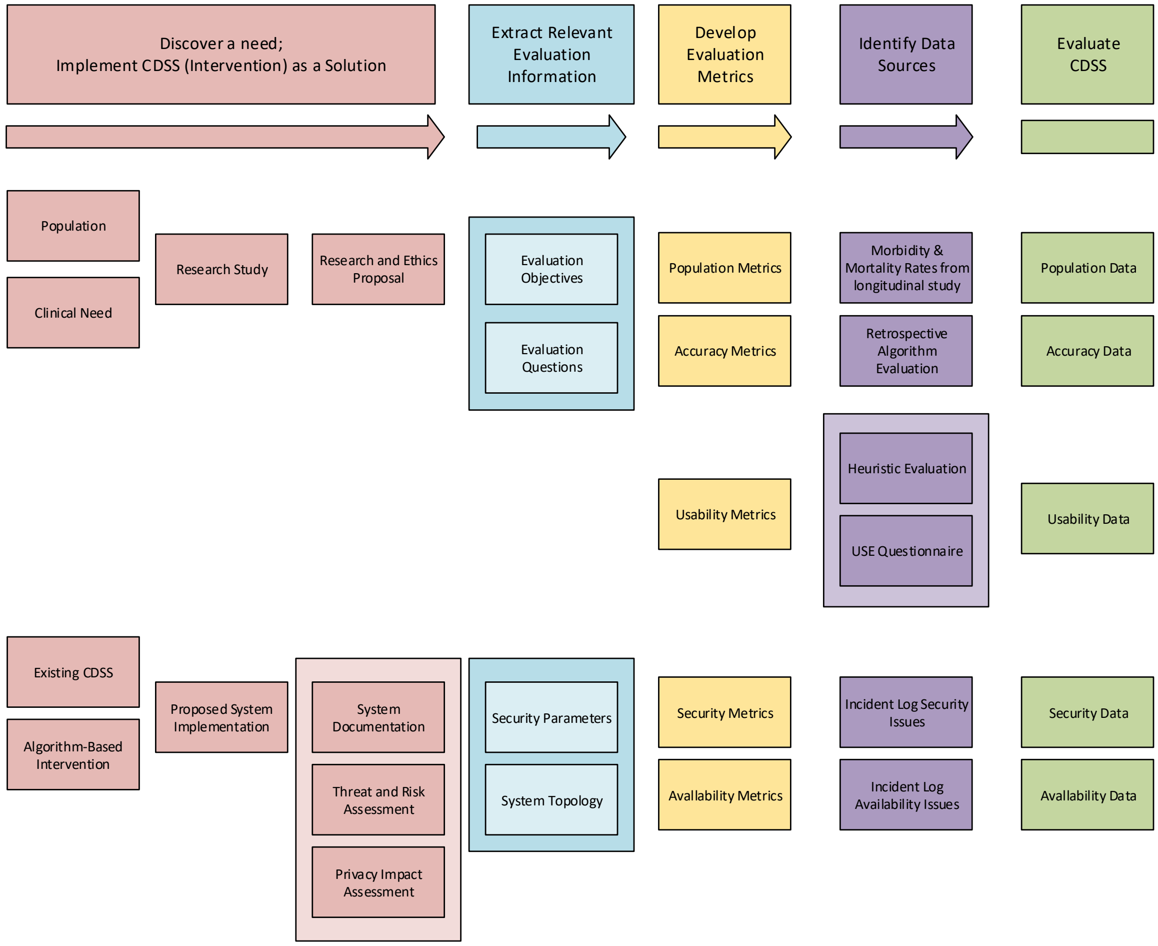

Figure 6 Evaluation Methodology

4.1.2 Extract Relevant Evaluation Information

The second step (coloured in Figure 6 as blue) is to extract the relevant evaluation information from the research and ethics proposal, threat and risk assessment, and privacy impact assessment. From the research and ethics proposal, the evaluation objectives and evaluation questions can be extracted. These are two components of the PHO evaluation template. PHO's evaluation plan template, shown in Figure 7 is relevant for use in a CDSS evaluation because the long-term goal of the CDSS implementation is a systematic change to the care practices in the hospital. The change includes being more proactive (or upstream) in care through the use of predictive analytics, as opposed to taking a reactive (or downstream) approach to care. The organized structure includes sections for objectives, evaluation questions, and metrics as well as the methods for collecting and analyzing each metric. From the system documentation, threat and risk assessment, and privacy impact assessment, the security parameters and system topology of the system can be extracted.

Objective(s)		•				
Evaluation Question(s)		•				
Code	**Metric(s)**	Data Collection Method	Data Source	Timeline	Roles and Responsibilities	Method of Data Analysis

Figure 7 PHO Data Collection Matrix

4.1.3 Develop Evaluation Metrics

The third step of the evaluation methodology (coloured in Figure 6 as yellow) is to develop the evaluation metrics for population, accuracy, usability, security, and availability. Information about the population, the clinical need, and the evaluation objectives help define the population health metrics for the evaluation. To address the efficacy of the intervention algorithm, accuracy and usability metrics are created. These metrics define how effective the intervention is. The security parameters and system topology are used in the creation of security and availability metrics. These metrics exist to define whether the system is performing and is capable of applying the algorithm intervention being used to improve population health. This hierarchal metric structure is similar to that of the YRPH Monitoring and Evaluation Framework(Glass et al. 2018), which classifies metrics into a hierarchal structure. The framework presents metrics in a way that support public health interventions. In the YRPH structure the population health objective governs the hierarchy, and all metrics are used as measurements of progress towards the population health objective.

4.1.4 Identify Data Sources

The fourth step of the methodology (coloured in Figure 6 as purple) is the identification of data sources. For each metric to be collected, evaluators will review different data sources, and may use a variety of data collection methods including questionnaires, quantitative analysis, and heuristic evaluations. This information is documented in the data collection matrix (Figure 7) as each metrics data collection method, data source, timeline, roles and responsibilities, and method of data analysis. and Section 4.2 describes the different types of metrics and their respective data sources.

4.1.5 Evaluate the CDSS

The fifth, and final step (coloured in Figure 6 as green) involves addressing the evaluation plan metrics with the relevant data collected during the evaluation of the system. Using the sources identified in step 4, evaluators analyze and address the metrics outlined in step 3.

4.2 Metric Types

In the evaluation methodology, five different metric subsets were developed in step 3. Each of these subsets have data sources identified in step 4, and are addressed in step 5. The following sections include descriptions of the metrics, the data sources for each, and how they can be addressed in the evaluation.

4.2.1 Population Metrics

Population metrics exist to measure whether or not the population is better off after an intervention (Friedman 2005). In the case of a CDSS evaluation, demonstrating a positive impact on population health is an important step because it proves the direct value of the CDSS to the healthcare system as a whole. Common metrics that depict the impact of the intervention on population metrics are morbidity and mortality rates. These metrics can be measured in a longitudinal study that depicts the impact of the algorithm on the population.

Population metrics should be developed as an extension and an actionable way of measuring the evaluation objectives previously determined. The population metrics are a way to operationalize objectives, and predict the impact of the CDSS.

A consideration with population health metrics is the sample size required to demonstrate a meaningful impact on the population. A single algorithm within a single system that focuses on only a single condition may not produce a large enough sample to demonstrate the solution’s impact on the population depending on the frequency of the condition occurring; however, a system with multiple algorithms, that can treat multiple conditions, or a system that has been implemented in multiple settings to increase its reach will have a larger impact on the overall health of the population. The evaluation plan accounts for that by having algorithm and technical metrics scalable. This means that both algorithm and technical metrics can be reused to support additional algorithms and additional architectures.

4.2.2 Accuracy Metrics

Accuracy is a concept that may need to be measured multiple times. In the case of accuracy, each algorithm needs to be measured and confirmed to meet a set accuracy threshold so that it can be reliably used.

An algorithm evaluation may not need to be performed in real time, and could instead use retrospective data for testing and fine-tuning. This approach supports the analysis of the algorithms prior to when they are available to clinicians to impact care. This approach may even be necessary in some cases, where the discovery and confirmation of conditions would occur much later without the algorithm. In those cases, real time analysis would not be feasible because clinicians would not be able to determine whether the system is correct until much later. Retrospective analysis enables the assessment of whether the algorithm satisfies the approved threshold to be deployed for use to impact care. This initial phase of evaluation is a necessary step in the deployment of any algorithm within the healthcare setting.

Suppose a scenario where after reviewing data from 50 patients, the Artemis Platform proposes 10 definite cases of LONS. In this example scenario, the clinician’s analysis is treated as 100 percent correct. After review, the clinicians determine that there were 10 cases of LONS; however, only 8 of them overlapped. While both determined that there were 10 cases of LONS, the algorithm is not 100% accurate. The algorithm both over and under estimated different cases. For the algorithm to be 100% accurate it would need to have the exact same list of definite LONS cases, with neither the algorithm or the clinicians having additional cases on their lists. Table 3 are results from different scenarios with an assumption of 100% clinician accuracy. The scenarios do not necessarily take into account sample size, and should be treated as an example for the purpose of depicting different cases of algorithm accuracy.

Scenario	Result
- Algorithm identifies exactly 12 cases of condition being present - Clinicians confirm 10 of the 12 cases only	- The algorithm is not 100% accurate. - The algorithm is over-representing the amount of cases with the condition present - This case is a false positive, as too many cases were identified
- Algorithm identifies exactly 8 cases of conditions being present - Clinicians confirm the 8 cases, but identify an additional 2 cases	- The algorithm is not 100% accurate. - The algorithm is under-representing the amount of cases with the condition present - This case is a false negative, as too few cases were identified
- Algorithm identifies exactly 10 cases of condition being present - Clinicians identify exactly 10 case of condition being present - Algorithm and clinician cases do not match	- The algorithm is not 100% accurate. - The algorithm is over-representing the amount of cases (it identified cases clinicians did not confirm) - The algorithm is under-representing the amount of cases (clinicians confirmed cases that the algorithm did not identify) - This case is both a false positive and false negative
- Algorithm identifies exactly 10 cases of condition being present - Clinicians confirm all 10 cases, and do not identify any additional cases	- The algorithm is 100% accurate

Table 3 Accuracy Metric Considerations

4.2.3 Usability Metrics

The system can have a high level of uptime, suitable security, and effective algorithms, but it also needs to be something that clinicians and other stakeholders are able to and willing to use as a decision support tool. Evaluating the usability of the user-facing components of the system generates an understanding of how the system will be used.

To complete a prospective evaluation on the usability of the CDSS a heuristic evaluation can be used to ensure that Nielsen’s 10 heuristic principles are addressed. This method is commonly used before the deployment of a user-facing components to ensure that the system meets the 10 heuristic principles. Heuristic evaluations use the opinions of user experience design experts, and; however, they omit the actual user's opinions

The researchers that developed the HOT-fit framework suggested perceived usefulness as a potential metric for determining user satisfaction with hospital IT systems (Yusof et al. 2008). While perceived usefulness can act as one metric for determining usability, additional feedback about the algorithm should be obtained from clinicians and other stakeholders as a way to obtain more fulsome information about the usability of the system. The USE questionnaire measures stakeholder opinions on a Likert rating scale in four categories: usefulness, ease of use, ease of learning, satisfaction (Lund 2001). Human factors evaluations, and other formal stakeholder observations can also be used to assess the system's capability to meet the needs of the clinicians (Whitefield, Wilson, and Dowell 1991). Collecting prospective data through a heuristic evaluation, and live/retrospective data using a USE questionnaire provides evaluators with more fulsome data on the use of the system.

4.2.4 Security Metrics

In order to implement a CDSS within a healthcare setting, standardized assessments of the impact of the CDSS within the context of treats, risks and privacy are required. Within Ontario, these are governed by the completion of TRA and PIA. These documents provide stakeholders with security protocols, and how any threats or breaches will be managed.

The process of approval of these documents is utilized within this evaluation template to evaluate system security. The approval suggests that key stakeholders find the system's security to be at a reasonable standard for collecting and analyzing personal health information. During the study utilizing the CDSS, the Incident Log provides details to determine whether there were any detected security incidents. While there may be undetected security incidents, the approval of the security documents suggests some level of reassurance that the system is secure, and that the team responsible for IT security is confident in the detection and handling of breaches. Evaluators review the Incident Log after the system has been used for a set amount of time and identify the amount, severity, and impact of security issues during the evaluation period.

4.2.5 Availability Metrics

Availability, or uptime, is defined the amount of time that the system is online and usable (Rance 2013). To measure this, an Incident Log should be maintained. The Incident Log contains information on each incident, including the components and beds affected, the total downtime, and the data recovery time in the case where data was buffered during the downtime and a catch up of data is required. Availability should be considered from the perspectives of availability for the user, the bed space, and the overall system availability.

While downtime is an important factor in the measurement of availability, another consideration for a CDSS is the data recovery time. CDSSs are designed to provide real-time analytics for use prospectively. During downtime of components within the CDSSs such as network failures, data acquisition failures, planned maintenance windows for server security patching, data may be queued for delivery after the CDSS comes back online. Data recovery can take a significant amount of time depending on the speed the data can be processed relating to real-time, and while that is occurring the CDSS cannot CDSS necessarily be considered to be providing real-time analytics.

The evaluation template is populated with multiple availability metrics, providing options for researchers and other stakeholders to provide a specific availability metric for different questions. Each measurement of availability differs in the way it answers the question of “how frequently was the system available to provide real-time analytics”. Different options for measuring availability and considerations for each option are included in Table 2.

Numerator	Denominator	Considerations
Sum of downtime minutes from each incident	Total minutes	- Treats the system as a singular unit - Does not include data recovery time - Downtime incidents that overlap only affect the numerator once - The metric is reported as uptime

		(which is 100% minus downtime)
Sum of downtime minutes and data recovery minutes from each incident	Total minutes	- Treats the system as a singular unit - Incidents where data recovery is affected count as system downtime - Downtime incidents that overlap only affect the numerator once - The metric is reported as uptime (which is 100% minus downtime)
The Individual sum of downtime minutes for each bed	Individual beds total minutes	- Each bed is treated as an individual unit - Incidents that affect the entire system are counted as downtime for each bed - Incidents that affect individual beds do not count towards the down time of unaffected beds - Downtime incidents that overlap only affect the numerator once - Ignores data recovery time - The metric is reported as percentage of beds at or above an agreed upon availability threshold
The Individual sum of downtime minutes and data recovery minutes for each bed	Individual beds total minutes	- Each bed is treated as an individual unit - Incidents that affect the entire system are counted as

		downtime for each bed - Incidents that affect individual beds do not count towards the down time of unaffected beds - Downtime incidents that overlap only affect the numerator once - Incidents where data recovery is affected count as system downtime - The metric is reported as percentage of beds at or above an agreed upon availability threshold

Table 4 Availability Metric Considerations

The Pareto principle states that 80% of the output in a given situation is produced by 20% of the input (Grosfeld-nir, Ronen, and Kozlovsky 2007). As a result, the range of metrics proposed enables the provision of details as to why there was a large variance in their values, and what the key issues are. Large portions of the downtime and data recovery time could be attributed to faulty wiring at a single bed space or collectively for all data acquisition for example. Sharing this information in an evaluation template is helpful as it allows stakeholders to target areas for improvement. The goal of using the hierarchal approach to monitoring and evaluation, is a scenario in which improvements each element impacting availability can be address, which leads to more patients receiving support from the CDSS intervention, and supports the overall objective of the intervention.

4.3 Conclusion

This chapter has presented the Evaluation Methodology proposed for the evaluation of streaming Big Data based CDSSs. Evaluating a public health intervention being implemented at a single hospital with a single algorithm involves an understanding of the vision for and objectives of the CDSS. While the sample size of one location and one algorithm isn't yet enough to determine the impact of the system on public health metrics, an evaluation template with a hierarchal structure can help highlight the connection of each implementation and algorithm to improving public health. The evaluation methodology demonstrated in this chapter is a structured approach that utilizes information from implementation documents including the research study, threat and risk assessment, privacy impact assessment, and incident log can provide evaluators with information to create metrics for population, availability, security, accuracy, and usability. This information, along with evaluation objectives and questions are used within the PHO evaluation template.

When measuring availability, it is important to consider the topology of the system. With a star topology system, evaluating the system as if it were many smaller units with equivalent topology can provide a more accurate depiction of the system's availability. It is also important to consider the system's purpose, and to account for things like data recovery time in availability metrics. For the system to be considered fully available, every component should be operational, and data should be up to date.

System security can be measured first by the acceptance and approval of key implementation documents, as well as through confirmation that the security has held up without any breeches or other security issues.

While availability, security, and accuracy can determine whether or not the system is capable, usability is imperative in determining whether the system will be used to its full potential. By evaluating for key user experience design principles, and interviewing stakeholders that use the system regularly, evaluators can get a more fulsome understanding of the system's use. This information is also helpful to developers, who can adjust or redesign the system to help it meet the user's needs.

The next chapter includes a demonstration of how the proposed evaluation template is utilized for the evaluation of the deployment of the Artemis Platform at MCH.

4.4 References

Friedman, Mark. 2005. *Trying Hard Is Not Good Enough*. 1st Editio. Trafford: FPSI Publishing.

Glass, Jonah et al. 2018. "Development of a Monitoring and Evaluation Framework for a Public Health Setting."

Grosfeld-nir, Abraham, Boaz Ronen, and N I R Kozlovsky. 2007. "The Pareto Managerial Principle : When Does It Apply ?" 45(10): 2317–25.

Lund, By Arnold M. 2001. "Usability Interface Measuring Usability with the USE Questionnaire." *Usability Interface* 8(2). http://www.stcsig.org/usability/newsletter/0110_measuring_with_use.html.

Rance, Stuart. 2013. *Defining Availability In The Real World*.

Whitefield, Andy, Frank Wilson, and John Dowell. 1991. "A Framework for Human Factors Evaluation." *Behaviour & Information Technology* 10(1): 65–79.

Yusof, Maryati Mohd, Jasna Kuljis, Anastasia Papazafeiropoulou, and Lampros K. Stergioulas. 2008. "An Evaluation Framework for Health Information Systems: Human, Organization and Technology-Fit Factors (HOT-Fit)." *International Journal of Medical Informatics* 77(6): 386–98.

Chapter 5. Case Study

This chapter instantiates the method in the case study context of the deployment of a streaming Big Data based clinical decision support system, Artemis within neonatal intensive care. In this chapter, metrics are presented for use in the Artemis Platform evaluation at the MCH NICU. When applicable, the results of evaluating each metric are included. For metrics that were not able to be measured, details on how they can be measured in the future are included.

5.1 Introduction

The Artemis Platform evaluation is a work in progress, with portions having being completed during the implementation of the system for the collection of retrospective data in March to September 2018. The key implementation artifacts like the TRA and PIA were developed for the implementation and approved in 2017. The intervention objective, evaluation questions, availability, and security metrics are presented in this chapter. Procedures to measure accuracy and usability are included in the evaluation. The results of the accuracy and usability evaluation are the subject of a separate clinical study that is outside the scope of this research. A data collection matrix demonstrates all of the information for the evaluation plan organized in a format that matches the PHO evaluation template (Public Health Ontario 2016).

5.2 Discover a Need – Implement a CDSS as a Solution

The MCH NICU services the neonate population, and had a clinical need to find more effective ways to identify LONS. Through the improved identification of conditions like LONS they hoped to reduce the rates of morbidity and mortality within the population. A research study was proposed by Dr. Edward Pugh to implement the Artemis Platform within the MCH NICU. A research and ethics proposal identified the purpose of the implementation, while Artemis Platform IT staff provided necessary system documentation including a PIA and TRA for approval of the implementation within the MCH NICU.

5.3 Extract Relevant Evaluation Information

To determine the intervention objectives as part of the first phase of the evaluation template for the Artemis Platform information was gathered about the long-term goals and objectives of implementing Artemis as a health intervention. The aim was to develop an understanding of Artemis as a health intervention through meetings with its creator, Dr. Carolyn McGregor, as well as key clinical stakeholders that plan on implementing Artemis. Dr. Edward Pugh was the clinical champion of Artemis for the Neonatal Intensive Care Unit, McMaster Children's hospital and Dr. Hugh Dawkins was the Director or Rare Disease in the Western Australian Department of Health. Dr. Hugh Dawkins shared the importance of precision public health, and the impact of up-stream approaches to healthcare. Dr. Dawkins believed that a platform like Artemis, which focuses on neonatal intensive care, is very much an up-stream approach because it aims to help and assist the youngest of humans, and can help limit long-term health issues. Healthy newborns have the opportunity to provide a positive impact on society for many years to come. Dr. McGregor and Dr. Pugh echoed the importance of Artemis as a public health intervention, as it aims to reduce the morbidity and mortality rate of neonates.

In the case of Artemis, algorithm improvements and technical improvements all support the idea of reducing the rates of morbidity and mortality by improving access to or use of the system. In the YRPH structure the intervention objective governs the hierarchy, and all metrics are used as measurements of progress towards the public health objective (in the case of Artemis, reduced morbidity and mortality).

In the research and ethics proposal for the MCH implementation, the authors state the evaluation (research) question as whether Artemis can operate and function in a large-scale NICU. The effectiveness of the operation and function are not explicitly stated, but it is clear that for Artemis to succeed as a public health intervention, it will need to be technically effective (available, secure), and the algorithms that provide the decision support will need to be effective (accurate). It will also need to be something that clinicians are willing and able to use (usable). These four metrics (availability, security, and accuracy, usability) all need to be at appropriate thresholds for the implementation to be considered successful. A successful implementation does not necessarily mean that the

Platform is successful as an intervention, but will allow for further research that determines the impact of the Platform.

5.4 Develop Evaluation Metrics, Identify Data Sources, and CDSS Evaluation

The population metrics are high level metrics that depict the impact of the Platform. The metrics for availability and security help stakeholders understand if the system is capable of providing service, and the frequency in which service can be provided, but they do not represent the effectiveness of the system in providing decision support. Accuracy metrics assess the ability of the algorithm to assist in clinical decision making. Usability metrics assess whether clinicians and other stakeholders will be able to use the system in a way that allows for improved decision making. Clinicians need to be aware of the level of accuracy the algorithm provides in detecting conditions, including the chances of false positives or negatives. Additionally, the algorithm needs to be presented in a way that clinicians consider it to be usable as a decision support tool.

5.4.1 Population Metrics

While Artemis as a public health intervention is an intriguing opportunity, there are currently limited implementations. For Artemis to be defined as an effective public health intervention, an evaluation of a single algorithm, at a single hospital, for a relatively short time period is a small sample. While the metrics can be set to measure the impact of Artemis on morbidity and mortality in neonates, it will take a larger sample to effectively determine the impact of the Platform and the long-term benefits brought on by the reduction of morbidity and mortality in neonates. Some considerations of how Artemis may be evaluated as a public health intervention are:

1. Comparing between hospital sites that are using Artemis and those that aren't. Limitations could include the different acuity levels, care practices, and population area that the hospital serves.
2. Comparing within a hospital site for neonates that are connected to Artemis for health monitoring compared to those that aren't. There may be fewer limitations in this study, but it will also include a smaller sample.

5.4.2 Accuracy

After collecting physiological data from beds used by neonates using the Artemis Platform at MCH, researchers will be able to retroactively apply the developed LONS algorithm to neonate datasets for the purpose of identifying cases with LONS. In applying the algorithm, they will define x_1 cases as definite LONS from the sample of n cases, y_1 cases as probable LONS, z_1 cases as potential LONS, q_1 cases as without LONS.

Clinicians will review the same cases, and categorize them as definite (x_2), probable (y_2), potential (z_2), and without (q_2) LONS. The clinician review includes complete diagnostic data from the patient. For the purpose of evaluating the algorithm, the clinician's review is treated to be correct, and the algorithm's goal should be to match it.

The evaluation of the algorithm's accuracy was not completed as part of this thesis; however, the accuracy metrics included in the thesis match the process researchers and clinicians outlined within the research study document as their plan for measuring accuracy.

5.4.2.1 AC1 – Percentage of Cases Identified as Definite Sepsis by the Artemis Platform Algorithm Confirmed by Clinician(s)

To measure the percentage of false positives, where the algorithm identifies a case as definite sepsis and clinicians disagree, x_1 is included as the denominator. The numerator is calculated as the number of cases that are included in both x_1 and x_2.

The same algorithm could be applied with y, z, and q; however, the priority first should be to match the algorithm to clinician reviews for definite cases.

5.4.3.2 AC2 – Percentage of Cases Identified as Definite Sepsis by Clinician(s) Confirmed by the Artemis Platform Algorithm

To measure the percentage of false negatives, where the clinician(s) identifies a case as definite sepsis and the algorithm does not, x_2 is included as the denominator. The numerator is calculated as the number of cases that are included in both x_1 and x_2.

The same algorithm could be applied with y, z, and q; however, the priority first should be to match the algorithm to clinician reviews for definite cases.

5.4.3.3 AC3 – Inter-Rater Reliability between the Artemis Platform LONS Algorithm and Clinicians

Cohen's Kappa score is used to measure the level of agreement between two raters reviewing the same *n* cases into mutually exclusive categories. It involves the probability of randomly agreeing compared to how frequently the two raters agree(Mchugh 2012).

With four categories, and two reviewers (under the assumption that clinician(s) are grouped as one reviewer even if there are more than one). In this format, *x, y, z,* and *q* are all used in the evaluation of the algorithm. A matrix is used and each individual case is placed in a cell based on both reviewer's opinions. Figure 8 depicts the matrix that uses Cohen's Kappa score.

Shaded cells indicate agreement		Clinicians			
		X_2	y_2	z_2	q_2
Algorithm	X_1	x_1 and x_2	x_1 and y_2	x_1 and z_2	x_1 and q_2
	y_1	y_1 and x_2	y_1 and y_2	y_1 and z_2	y_1 and q_2
	z_1	z_1 and x_2	z_1 and y_2	z_1 and z_2	z_1 and q_2
	q_1	q_1 and x_2	q_1 and y_2	q_1 and z_2	q_1 and q_2

Figure 8 Level of Agreement using Cohen's Kappa

A potential addition to the matrix would be to increase weighting on cases of definite sepsis to count for more, meaning that the agreements for definite cases are more important than the agreement for probable, potential, or without sepsis.

5.4.3 Usability

Usability metrics assess whether clinicians and other stakeholders will be able to use the system in a way that allows for improved decision making. Some usability evaluations can be completed pre-implementation using heuristic principles through a cognitive walkthrough or heuristic evaluation. Post-implementation evaluations can be completed using interviews, focus groups, and questionnaires to get individual feedback.

The evaluation of the algorithm's usability was not completed as part of this thesis; however, the creation of evaluation plan metrics will assist in the development of the data visualization for the LONS algorithm within the Artemis Platform.

5.4.3.1 US1 – Number of Unresolved Usability Issues Identified via Pre-Implementation Usability Evaluations

The clinician/nurse facing portion of the Artemis Platform will be an alert system and data visualization that depict a neonate's physiological data over a period of time, and is meant to assist in the identification of conditions. The LONS algorithm will have its own data visualization for each neonate. This visualization differs from that of the physiological data monitors due to its advanced ability to show adjusted timeframes with a higher level of detail.

During the development of the data visualizations for the Artemis Platform, heuristic principles should be considered to ensure that it is an effective tool for clinicians and nurses. A heuristic evaluation will be completed by researchers and evaluators to ensure that Nielsen's heuristic principles have been followed. This evaluation can be repeated as improvements are made to the visualization pre-implementation.

5.3.3.2 US2 – Perceived Usefulness, Ease of Use, Ease of Learning, Satisfaction measured Post-Implementation

The USE Questionnaire is a Likert rating scale tool that will be provided as a survey to clinicians, nurses, and other stakeholders using the Artemis Platform. These results can be categorized as a way to understand user opinions of the Platform's usefulness (does it provide a helpful service), ease of use (is it easy to use this service), ease of learning (is it easy to learn how to use the service), and satisfaction (are they happy with the service). Upon implementation of the real-time Platform and data visualization to support the use of the LONS algorithm, stakeholders will be asked to complete the questionnaire. Feedback will also be gathered through interviews and focus groups with stakeholders using the system.

5.5 Technical Metrics

During the deployment of Artemis within the NICU at MCH an incident log was kept to record any form of system downtime or other incident. The incident log tool contained six months of data from the Artemis implementation. Specifically, the tool contains information on each incident, including the components impacted, beds affected, the total downtime, and the data recovery time. The Incident Log is an effective tool for monitoring both the availability and the security of the Platform.

5.5.1 Availability

The availability of the Artemis Platform at McMaster Children's Hospital (MCH) was evaluated by reviewing all incidents within incident log and analyzing the downtime, reason for the downtime, the components and number of beds affected, and data recovery time. The Artemis Platform was evaluated between the period of March 1, 2018 and September 30, 2018. A period of 213 days, or 306,720 minutes. During the period, 38 incidents that affected system availability were listed in the incident log. Data recovery time consists of the time it takes for the system to catch-up and begin including any missing data that was not transmitted during downtime.

The summary table includes the relevant incident log data for each of the four metrics. Figure 8 includes the key data points for the availability of the system. Table 4 includes the calculation for each metric. Detailed information for each metric is included after Table 4.

Total System Minutes: 306,720
Total Incidents: 38
Total Beds: 51

Total Minutes of Documented Unplanned Downtime (as a single system): 10,158
Total Minutes of Documented Planned Downtime (as a single system): 126
Total Minutes of Documented Downtime (as a single system): 10,284
Total Minutes of Data Recovery Time (as a single system): 33,025

Minimum Minutes of Downtime for each bed: 905
Minimum Minutes of Data Recovery Time for each bed: 33,025

Figure 9 Key Data Points

Metric	Numerator	Denominator	Result
AV1 – Availability of Artemis as a Single System	Total Minutes (306,720) minus Total Minutes of Documented Downtime (10,284) equals 296,436 minutes	Total Minutes (306,720)	296,436 / 306,720 = 0.966 = 96.6% Availability
AV2 – Availability of Artemis as a Single System – with Data Recovery Time Included	Total Minutes (306,720) minus Total Minutes of Documented Downtime (10,284) minus Total Minutes of Data Recovery Time (33,025) equals 263,411 minutes	Total Minutes (306,720)	263,411 / 306,720 = 0.859 = 85.9% Availability
AV3 – Percentage of Artemis Enabled Beds Maintaining Availability of 99.5% or More	Each bed's availability percentage is calculated using the same formula as AV1. Count of beds with availability percentages greater than 99.5% equals 45	Total Beds (51)	45 / 51 = 88.2% of beds with availability greater than 99.5%
AV4 – Percentage of Artemis Enabled Beds Maintaining Availability of 99.5% or More – with Data Recovery Time Included	Each bed's availability percentage is calculated using the same formula as AV2. Count of beds with availability percentages greater than 99.5% equals 0	Total Beds (51)	0 / 51 = 0% of beds with availability greater than 99.5%

Table 4 Availability Metric Details

5.5.1.1 AV1 – Availability of the Artemis Platform as a Single System

Using a definition of platform availability that proposes the entire system is performing and that there are no planned or unplanned outages, then the entire system's availability is measured as the number of minutes the system was available divided by the total minutes the system was meant to be available.

Out of the 306,720 minutes that the system could have been available, there were 10,284 minutes of documented downtime, and 296,436 minutes of uptime. This equates to an availability percentage of 96.6%.

Of the 10,284 minutes of downtime, only 126 minutes were planned. 10,158 minutes were unplanned. If the Platform had no unplanned downtime, then the system availability percentage would have been 99.96%. This metric can be considered the technical availability of the system as the system and its components are considered functioning.

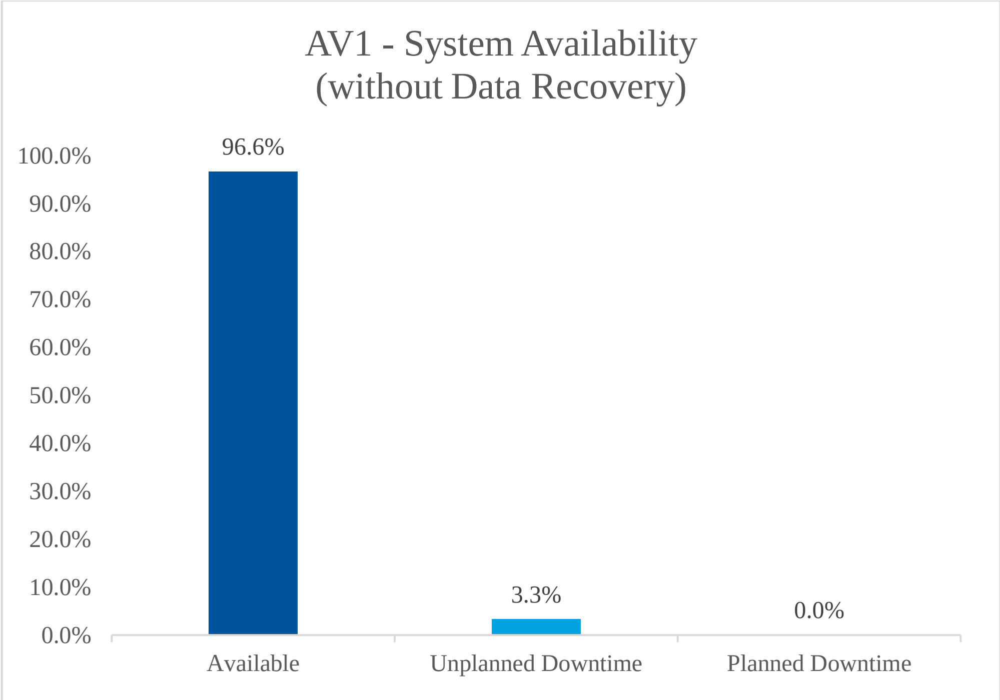

Figure 10 Availability Metric AV1 Chart

5.5.1.2 AV2 – Availability of the Artemis Platform as a Single System – with Data Recovery Time Included

Using a definition of platform availability that proposes the entire system is performing and that there are no planned or unplanned outages, and that any data recovery time incurred from outages means that the system is unavailable, then the entire system's availability is measured as the number of minutes the system was available divided by the total minutes the system was meant to be available.

Out of the 306,720 minutes that the system could have been available, there were 43,309 minutes of documented downtime plus data recovery time, and 263,411 minutes of uptime. This equates to an availability percentage of 85.9%. This metric can be considered clinical availability, as it defines how frequently the Platform is able to perform clinical functions effectively with all retrospective data included.

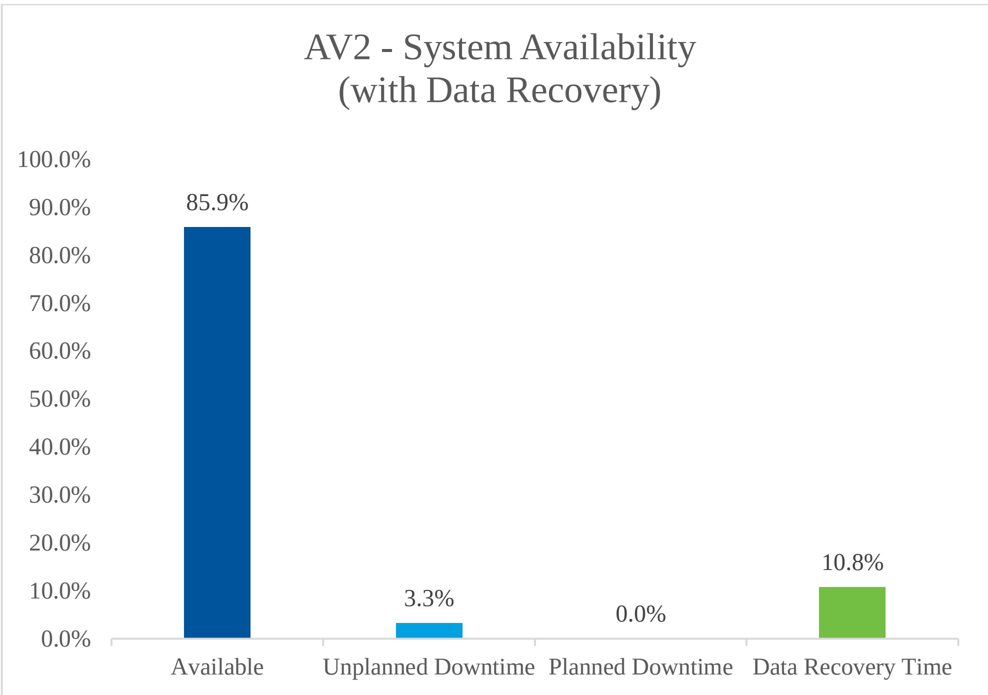

Figure 11 Availability Metric AV2 Chart

5.5.1.3 AV3 – Percentage of Artemis Platform Enabled Beds Maintaining Availability of 99.5% or More

Using a definition of platform availability that considers the Artemis system topology, each bed's availability is calculated separately using the same formula as if each were a single system. This method leverages the detailed information available in the incident log. If shared system components are not available, then all bed's availability will be negatively affected. Each bed's availability is then compared to a set threshold (99.5%).

There were 905 minutes of downtime that affected every bed, meaning the maximum uptime was 305,815 minutes out of 306,720 minutes. This is equivalent to 99.7% availability. 45 out of 51 beds (88.2%) had availability greater than 99.5%. The minimum availability for a single bed was 300,412 minutes, which is the equivalent to 97.9% availability.

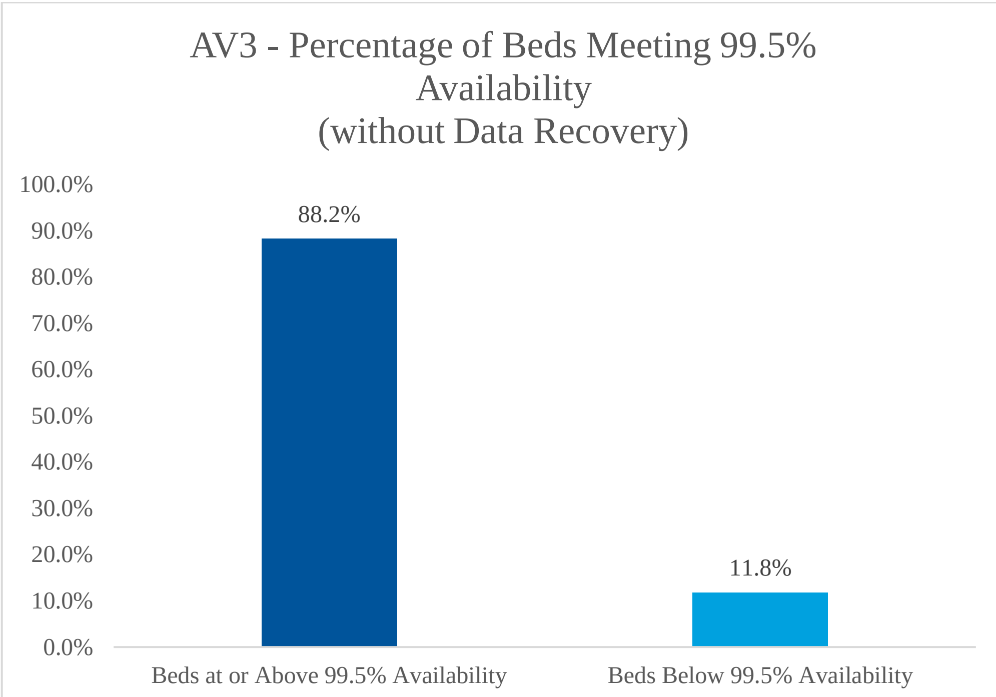

Figure 12 Availability Metric AV3 Chart

5.5.1.4 AV4 – Percentage of Artemis Platform Enabled Beds Maintaining Availability of 99.5% or More – with Data Recovery Time Included

Using a definition of platform availability that considers the Artemis system topology, each bed's availability is calculated separately using the same formula as if each were a single system. This method leverages the detailed information available in the incident log. If shared system components are not available, then all bed's availability will be negatively affected. Each bed's availability is then compared to a set threshold (99.5%).

There were 905 minutes of downtime that affected every bed, and an additional 33,025 minutes of data recovery time that affected every bed, for a total downtime of 33,930 minutes. The maximum uptime was 88.9%. None of the 51 beds had a greater availability than the 99.5% threshold.

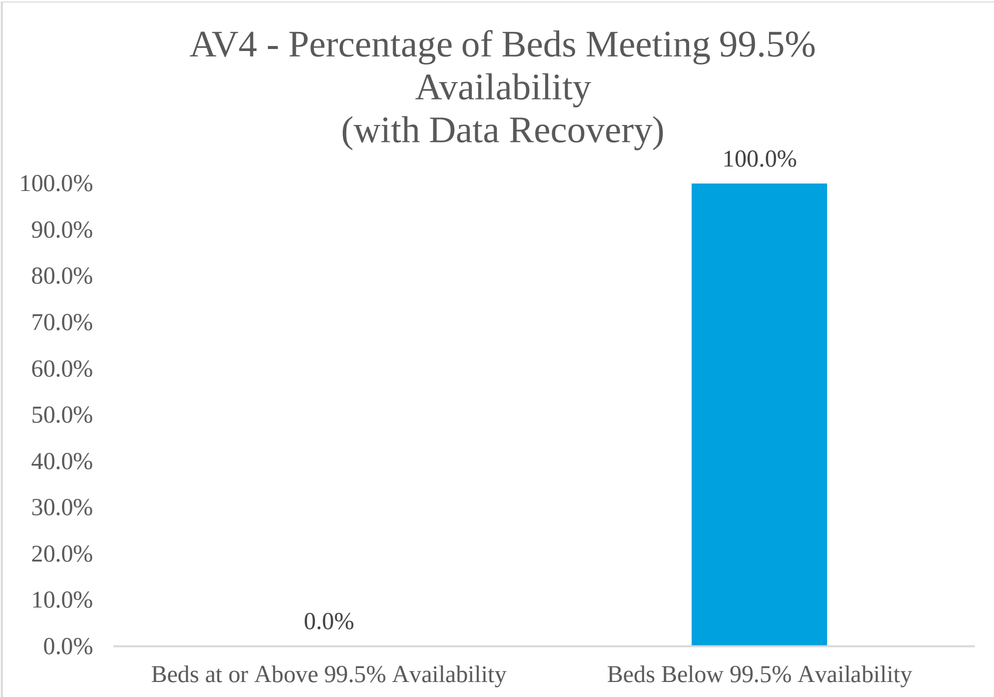

Figure 13 Availability Metric AV4 Chart

5.5.2 Security

The security of the Artemis Platform at McMaster Children's Hospital (MCH) was evaluated by reviewing the TRA for approval, as well as the Incident Log for all incidents documented as security issues. The Artemis Platform was evaluated between the period of March 1, 2018 and September 30, 2018.

5.5.2.1 SE1 – Acceptance of Security Parameters

The final Threat and Risk Assessment, dated December 1st 2017 contains a list of all assets (hardware, infrastructure, software, and data) used by the Artemis Platform, a list of threats and vulnerabilities, and safeguards in place to protect from these issues. While multiple high impact issues were identified as risks, the likelihood of any security issues were all stated to be low because of the use of encryption, firewalls, and a whitelist IP system. The Threat and Risk Assessment was approved by the CIO and CMIO at MCH.

5.5.2.2 SE2 – Total Security Incidents / Breeches

To determine whether there were any breeches, the incident log maintained by IT staff was reviewed. While there were 38 incidents during the six-month period of March 1st to September 30th 2018, none of the incidents involved or put at risk the security of the system or any data within it. There were no data breeches recorded during the time period either.

5.6 Data Collection Matrix

The data collection matrix summarizes the objective, evaluation questions, and metrics used for the study, as well as how they will be measured. The data collection matrix is constructed based on the PHO evaluation template.

<table>
<tr><td colspan="2">Objective(s)</td><td colspan="5">Does the Artemis Platform, and the algorithms maintained within it reduce the rates of morbidity and mortality</td></tr>
<tr><td colspan="2">Evaluation Question(s)</td><td colspan="5">Can the Artemis Platform be developed, operate on a large-scale using hospital and cloud infrastructure to support research in a busy NICU
Can physiological data used prospectively in an algorithm for LONS identification help reduce the amount of prescribed medication</td></tr>
<tr><td>Code</td><td>Metric(s)</td><td>Data Collection Method</td><td>Data Source</td><td>Timeline</td><td>Roles and Responsibilities</td><td>Method of Data Analysis</td></tr>
<tr><td>AV1</td><td>Availability percentage of Artemis as a singular system</td><td>Documentation of all incidents including the amount of downtime minutes</td><td>Incident Log</td><td>March – September 2018</td><td>Ontario Tech U IT Staff will document all incidents</td><td>Sum of down time from incidents divided by total time

*Must account for overlapped downtime</td></tr>
<tr><td>AV2</td><td>Availability percentage of Artemis as a singular unit, with data recovery time considered</td><td>Documentation of all incidents including the amount of downtime minutes and data recovery minutes</td><td>Incident Log</td><td>March – September 2018</td><td>Ontario Tech U IT Staff will document all incidents</td><td>Sum of down time from incidents and data recovery minutes divided by total time

*Must account for overlapped downtime</td></tr>
<tr><td>AV3</td><td>Percentage of beds maintaining availability equal to above 99.5%</td><td>Documentation of all incidents including the amount of downtime minutes and beds affected</td><td>Incident Log</td><td>March – September 2018</td><td>Ontario Tech U IT Staff will document all incidents</td><td>For each bed: sum of down time from incidents divided by total time.</td></tr>
</table>

						For platform: sum of beds with availability equal to or above 99% divided by total beds. *Must account for overlapped downtime
AV4	Percentage of beds maintaining availability equal to above 99.5%, with data recovery time considered	Documentation of all incidents including the amount of downtime minutes, data recovery minutes, and beds affected	Incident Log	March – September 2018	Ontario Tech U IT Staff will document all incidents	For each bed: sum of down time and data recovery time from incidents divided by total time. For platform: sum of beds with availability equal to or above 99% divided by total beds. *Must account for overlapped downtime
SE1	Acceptance of security parameters	Signatures and approvals from key stakeholders accepting potential security risks and measures taken to minimize risks	Privacy Impact Assessment, Threat and Risk Assessment	March 2018	Ontario Tech U Researchers and IT Staff will prepare documents for approval, MCH executives will approve documents	Boolean (Yes/No) of security parameters accepted
SE2	Total security incidents or breeches; response to security incidents or breeches.	Documentation of all security incidents and breeches	Incident Log	March – September 2018	Ontario Tech U IT Staff will document all incidents	Count of security incidents and breeches
AC1	Percentage of cases identified as definite sepsis by algorithm confirmed by clinician review (measurement of false positives)	Artemis Platform to collect physiological data from neonates at each bed	Artemis Platform DB2 with heart rate and respiratory rate values	March – September 2018	Ontario Tech U Health Informatics Lab will provide data to clinicians, as well as results from the algorithm's analysis	Count of definite sepsis cases found by the algorithm and confirmed by clinicians divided by count of definite sepsis cases found by the algorithm
AC2	Percentage of cases identified as definite sepsis by clinicians also identified by algorithm (measurement	Artemis Platform to collect physiological data from neonates at each bed	Artemis Platform DB2 with heart rate and respiratory rate values	March – September 2018	Ontario Tech U Health Informatics Lab will provide data to clinicians, as well as results from the algorithm's analysis	Count of definite sepsis cases found by clinicians and confirmed by the algorithm divided by count of definite sepsis

	of false negatives)					cases found by the algorithm
AC3	Inter-rater reliability between Artemis Platform algorithm and Clinicians	Artemis Platform to collect physiological data from neonates at each bed	Artemis Platform DB2 with heart rate and respiratory rate values	March – September 2018	Ontario Tech U Health Informatics Lab will provide data to clinicians, as well as results from the algorithm's analysis	Using Cohen's kappa coefficient to determine the level of inter-rater reliability using all cases involved in analysis
US1	Number of unresolved usability issues identified via heuristic evaluation and cognitive walkthrough	Cognitive Walkthrough and Heuristic Evaluation	Artemis Platform data visualizatio n	Pre-implementatio n	Evaluator to complete usability evaluations and provide feedback to Ontario Tech U Health Informatics Lab	Number of cases left unresolved at the implementation phase
US2	Usefulness, ease of use, ease of learning, and satisfaction of the Artemis Platform	Survey /Questionnaire	USE Questionna ire	Training and post-implementatio n	Front-line stakeholders (clinicians, nurses)	Using a Likert scale the percentage of stakeholders that strongly agree, agree with positive statements describing the Artemis Platform

Table 5 Data Collection Matrix

5.7 Conclusion

The Artemis Platform evaluation template includes evaluation objectives, questions, and four different metric categories – availability, security, accuracy, and usability. Where possible, analysis has been performed for each metric. For metrics where analysis was not yet possible, metrics are proposed and procedures are recommended to evaluate the metrics. Technical metrics can be applied to each implementation of the Platform at new hospitals, whereas algorithm metrics can be applied each time a new algorithm is being deployed.

The next chapter is a discussion of how the evaluation plan and case demonstrates an effective way to evaluate the Artemis Platform and other similar CDSSs. The discussion confirms that the goals of the literature review, outlined in Chapter 2, have been met.

5.8 References

Mchugh, Mary L. 2012. “Lessons in Biostatistics Interrater Reliability : The Kappa Statistic.” *Biochemia Medica*: 276–82.

Public Health Ontario. 2016. 117 Public Health *Evaluating Health Promotion Programs: Introductory Workbook.*

Chapter 6. Discussion

This chapter presents a discussion of the case study demonstration of the evaluation methodology developed for the Artemis Platform LONS algorithm implementation at the MCH NICU within the context of the literature presented in Chapter 2.

6.1 Introduction

The completion of the Artemis Platform evaluation plan includes the use of evaluation objectives and questions, which define the population health outcomes expected once the system has reached a critical mass of implementations. The Artemis Platform aims to improve the diagnosis process for LONS, a condition that if not treated effectively leads to increased morbidity and mortality in an already delicate population. The evaluation plan leverages the use of the implementation objectives in the creation and measurement of new metrics. Considerations like data recovery are also very important in the evaluation of systems that use streaming analytics.

6.2 Population Health Outcomes

An upstream approach to healthcare means preventing issues before they happen, creating a healthier society that uses the health system less. A downstream approach means treating people only when they are sick.

The Artemis Platform is a downstream approach since it assists in the detection of already existing diseases; however, the premise of treating neonates effectively and solving health issues earlier so that they require less ongoing care and have a potentially healthier life is an upstream approach. While the Artemis Platform does not exist to prevent neonates from entering the NICU (John Newnham's work in preventing pre-term birth through population health interventions reduces the frequency in which neonates enter the NICU(Newnham et al. 2017), it aims to assist in the clinical care of neonates, and reduce their rates of morbidity and mortality. This is especially important considering how small and young neonates are. In a 2017 meeting with Dr. Dawkins, a Western Australian health researcher and former government executive, he stated that each neonate has the opportunity to provide many more good years to society. Reducing their need for ongoing care greatly impacts the health system in a positive way.

Implementing Artemis at one hospital provides benefit, but if the Platform were used universally across hospitals than the capability to use real-time monitoring for neonatal conditions would be evolutionary. With that in mind, the evaluation aims to evaluate the Artemis Platform as its creators and stakeholders aim to use it – as a clinical decision support system population health intervention. Hence the inclusion of population health metrics is important in describing the impact the Artemis Platform could have on neonatal care.

In the evaluations of health information systems in Chapter 2, there were some discussions on clinical outcomes; however, more evaluations are completed to demonstrate the system's usability. In most cases the hospital systems evaluated were EMRs, billing systems, or other back-end systems not designed to assist in clinical decision making. One reason for this might be because of the healthcare environment it is being implemented in. The Excellent Care for All Act fosters an approach to improved clinical outcomes as a funding model (Government of Ontario 2017). This may differ from other health systems that run in a for-profit structure. Not all of the papers reviewed were not from the Canadian healthcare system.

The goal of a CDSS is to assist and improve clinical decision making. When evaluating these systems, it is important to consider the outcomes of that assistance and improvement, and how that affects population health as a whole.

6.3 Evaluation Plan Metrics

There are many metrics categories of HIS evaluations. Chapter 2 included a review of metrics that could be used for the evaluation of CDSSs like the Artemis Platform. The Artemis Platform LONS algorithm implementation at the MCH NICU included metrics for availability, security, accuracy, and usability. The metrics are common for assessing both the system's capability and the user's experience with systems and are inline with evaluation metrics listed in the HOT-Fit framework(Yusof et al. 2008). Instead of focusing on organizational factors like the impact of the system on organizational efficiencies or costs, the evaluation plan for the Artemis was focused on the population health objectives and outcomes of implementation the Platform.

6.4 Implementation Artifacts

The development of the evaluation plan for the Artemis Platform occurred before the Platform had begun being implemented at the MCH NICU. Implementing a CDSS at a hospital, especially one where portions of the Platform exist outside of the hospital environment, required numerous approvals from hospital stakeholders, and multiple documents that highlighted the security and privacy of information being used and analyzed by the CDSS.

These documents represent an opportunity for evaluators, as they are helpful in the creation of evaluation metrics. The medical study proposal and research and ethics board approval documents provided key insights into the goals and objectives of implementing the Platform. These two documents were valuable for interpreting the Platform's objectives, as well as how the researchers planned for the accuracy of the implemented algorithm to be evaluated.

The PIA and TRA were also used in the creation of the evaluation plan. Since these documents contain information about all of the Platform components, they help with understanding the system topology, which was especially relevant in the creation of the availability metric. They also contain a list of security risks and the efforts taken to minimize these risks. The approval of these documents shows that the level of security assured by the stakeholders implementing the system has been approved by the hospital's key decision makers, and also provides a template for how security of the Platform can be measured. The incident log maintained by the Platform IT staff is then used to measure events that impact the availability and the security of the system.

6.5 System Topology

The Artemis Platform has a topology similar to a star network connected to a point-to-point network. In this topology each bed has its own portion of the system, but they share a connection to the Vines Server, which is also the connection to ORION and to the Queen's CAC. What makes this topology important is in how metrics like availability and security are considered. Often the system's availability is calculated as the amount of time the system is usable by stakeholders; however, the Artemis Platform's evaluation showed that this isn’t completely accurate. Within the MCH NICU one bed had significantly more issues with wiring than every other bed; however, the other beds were not affected by these issues because of the system topology. If a single bed is not functioning, but the other fifty are, metrics need to be able to depict that those fifty beds were available while the single bed was not functioning. a metric that considers how many beds met an acceptable availability threshold meets this purpose. This metric allowed for beds to have shared downtime when main hubs connected to all beds were not available, as well as individual downtime for issues like wiring.

6.6 Data Recovery Time

One of the goals of the Artemis Platform is to provide continuous data to clinicians. This is relevant in the LONS algorithm, which requires a constant flow of data from the bedside monitors. Without the flow of data, the accuracy of the algorithm may be compromised. Another goal of the Platform is to provide both up-to-date and relevant historic data to clinicians at an individual patient level. When the system is impacted by downtime, the capability to provide data current is impacted. It was clear from the Artemis Platform incident log that even once the Platform became available again, queued data could take hours to be transmitted and analyzed. The Platform was effectively up and running, but the algorithm and presented data were not complete or up to date. In this case, the platform may not be considered fully available. This needs to be further explored with clinicians using the Platform, who can confirm the impact of data recovery time on their clinical decision making.

6.7 Conclusion

This chapter has presented a discussion of some of the key points evaluators need to consider in the evaluation of high frequency streaming analytic platforms that use physiological data. It includes which evaluation metrics should be used in a CDSS, and makes a case for the importance of population health metrics being included in the evaluation plan. Implementation artifacts are helpful tools in developing the evaluation plan and used for availability and security metrics. For algorithms that require the use of high frequency streaming analytics in real time, the measurement of data recovery time and its impact on availability also must be included. Data recovery time is important because when the system's data is not up to date, the algorithms are not able to function at their full capacity.

6.8 References

Government of Ontario. 2017. *Excellent Care for All Act*.

Newnham, John P. et al. 2017. "Reducing Preterm Birth by a Statewide Multifaceted Program : An Implementation Study." *American Journal of Obstetrics and Gynecology* 216(5): 434–42. http://dx.doi.org/10.1016/j.ajog.2016.11.1037.

Yusof, Maryati Mohd, Jasna Kuljis, Anastasia Papazafeiropoulou, and Lampros K. Stergioulas. 2008. "An Evaluation Framework for Health Information Systems: Human, Organization and Technology-Fit Factors (HOT-Fit)." *International Journal of Medical Informatics* 77(6): 386–98.

Chapter 7. Conclusion

As a conclusion to this thesis, this chapter outlines how the research aims and objectives were addressed, as well as the contribution to knowledge from this thesis.

7.1 Addressing the Research Aim and Objectives

In the introduction, the aim and objectives of the research were outlined as:

1. *That an evaluation methodology can be developed that includes population health, technical, and algorithm specific metrics specific to the implementation of a high frequency streaming physiological data analytics.*
2. *The abovementioned methodology will integrate key implementation artifacts for the purpose of determining metrics.*
3. *That the methodology can be demonstrated within an instantiation to support the evaluation of a Big Data analytics based CDSSs within a NICU in Ontario*

To address the first objective, an evaluation methodology was developed for the implementation of a high frequency streaming analytics CDSS. The evaluation plan included the measurement of population health objectives, as well as technical and algorithm metrics. These metrics are specific to high frequency streaming physiological data analytics platforms as they enable the measurement of platform specific metrics like data recovery time.

The second objective was addressed through the inclusion of the research study documents as a way to inform the evaluation plan objectives. The PIA and TRA were included and assisted in the development of metrics for availability and security.

The third objective were resolved through the creation of a scalable evaluation plan used for the implementation of the Artemis Platform at the MCH NICU. While the evaluation was not completed, a plan exists to collect all metrics that inform the success of the system use. The expectation is that the Platform will be implemented in multiple hospital settings, each of which can utilize the same technical metrics. This is especially helpful in Ontario where the Platforms will share some technical components at the Queen’s Centre for Advanced Computing. Each algorithm can be measured separately using the same analytics for accuracy and usability. In the case of an implementation where the platform has multiple algorithms, the accuracy and usability metrics can be repeated for each algorithm being implemented.

7.2 Contribution to Knowledge

This thesis contributes to knowledge the design of an evaluation plan for CDSS that includes population health considerations and depicts the use of implementation artifacts. The evaluation plan leverages the existence of already existing frameworks and methodologies including PHO, YRPH, and the HOT-Fit Framework (Glass et al. 2018; Public Health Ontario 2016; Yusof et al. 2008). The plan was used to partially evaluate the implementation of the Artemis Platform CDSS within the MCH NICU, but can be used to support multiple implementations and multiple analytics.

7.3 Limitations and Future Research

A major limitation of availability metrics in HIS evaluations today is the consideration of how data recovery or buffering time should be measured, and how limited function of real-time analytic systems affect the ability to impact care. Instead of only considering a system to be either available or not available, an important point is whether or not the ability to access retrospective data is imperative to the system's function. The Artemis Platform had numerous incidents where data being sent to the Platform's analytics engine was delayed. When these incidents occur, it can take days for the Platform to catch up and ingest all of the buffered data. As a real-time alert system, some functionality will still be usable during this buffering, but visualizations and algorithms that use retrospective analysis have their accuracy greatly impacted. While this paper suggested that measuring the system as individual bed units to reduce the impact of downtime events that do not impact every bed, further research could include developing individual availability metrics for each activity the system performs.

As noted in Chapter 5, another limitation of this thesis was that accuracy and usability metrics were proposed but not measured. These items are the subject of a separate clinical study outside the scope of this research.

Further research into CDSS evaluation methodologies may be able to show the scalability of the methodology by applying it to multiple implementations of the same Platform and multiple algorithms within the same Platform. The CDSS evaluation methodology can be applied to the implementation of the Artemis Platform at Southlake Regional Health Centre, an Ontario based hospital using a similar system topology to the MCH implementation. As new algorithms are implemented, accuracy and usability metrics used for LONS can be repurposed as well.

7.4 Conclusion

This thesis presented an evaluation plan and metrics for evaluating scalable clinical decision support systems (CDSSs) that use high frequency streaming physiological data analytics to support improved population health. The evaluation plan used a holistic approach that includes the presence of population health metrics, technical metrics, and implementation specific metrics. The evaluation plan applied concepts and used metrics suggested in the HOT-Fit framework (Yusof et al. 2008), while applying terminology and plan design from the Public Health Ontario (PHO) evaluation plan template (Public Health Ontario 2016), and used a hierarchal metric structure described in York Region Public Health's Monitoring and Evaluation Framework (Glass et al. 2018). The methodology described in the evaluation plan also demonstrated how implementation artifacts can be leveraged in the development of evaluation metrics.

The evaluation plan was applied in this research within the context of a CDSS implementation in an Ontario neonatal intensive care unit (NICU). The implementation was governed by a research study, from which approval documents were used to design potential high-level population health outcome metrics.

7.5 References

Glass, Jonah et al. 2018. "Development of a Monitoring and Evaluation Framework for a Public Health Setting."

Public Health Ontario. 2016. 117 Public Health *Evaluating Health Promotion Programs: Introductory Workbook.*

Yusof, Maryati Mohd, Jasna Kuljis, Anastasia Papazafeiropoulou, and Lampros K. Stergioulas. 2008. "An Evaluation Framework for Health Information Systems: Human, Organization and Technology-Fit Factors (HOT-Fit)." *International Journal of Medical Informatics* 77(6): 386–98.

BIBLIOGRAPHY

Ahmadian, Leila, Simin Salehi, and Reza Khajouei. 2015. "Evaluation Methods Used on Health Information Systems (HISs) in Iran and the Effects of HISs on Iranian Healthcare : A Systematic Review." *International Journal of Medical Informatics* 84(6): 444–53.

Bassi, Jesdeep, and Francis Lau. 2013. "Measuring Value for Money : A Scoping Review on Economic Evaluation of Health Information Systems." *Journal of the American Medical Informatics Association* 20(1): 792–801.

Berner, Eta S. 2007. *Clinical Decision Support Systems - Theory and Practice*. ed. Eta S. Berner.

Berner, Eta S. 2009. "Clinical Decision Support Systems : State of the Art." (09).

Bhagwan, Ranjita, Stefan Savage, and Geoffrey M Voelker. 2003. "Understanding Availability." *Peer to Peer Systems II*: 1–6.

Biel, L, O Pettersson, L Philipson, and P Wide. 2014. "ECG Analysis : A New Approach in Human Identification." (April).

Boemh, and Barry W. 1984. "Verifying and Validating Software Requirements and Design Specifications." (January).

Borim, Heloa Costa et al. 2015. "Integrating Methods to Evaluate Health Information Systems." *Context Sensitive Health Informatics: Many Places, Many Users, Many Contexts, Many Uses*.

Brender, J. et al. 2013. "STARE-HI – Statement on Reporting of Evaluation Studies in Health Informatics." : 331–58.

Canadian Institutes of Health Research, Natural Sciences and Engineering Research Council of Canada, and and Social Sciences and Humanities Research Council of Canada. 2014. *Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans*. http://www.pre.ethics.gc.ca/eng/policy-politique/initiatives/tcps2-eptc2/Default/.

Catley, Christina, Kathy Smith, Carolyn Mcgregor, and Mark Tracy. 2009. "Extending CRISP-DM to Incorporate Temporal Data Mining of Multi- Dimensional Medical Data Streams : A Neonatal Intensive Care Unit Case Study." : 0–4.

Centre for Advanced Computing. 2020. "What Is the CAC?" https://cac.queensu.ca/about_us/.

Coggins, Sarah A. et al. 2016. "Heart Rate CHaracteristic Index Monitoring for Blood Stream Infection in an NICU: A 3-Year Expierence." 106(3): 201–7.

Communications Security Establishment, and Royal Canadian Mounted Police. 2007. *Harmonized Threat and Risk Assessment (TRA) Methodology*.

Dodig Crnkovic, Gordana. 2010. "Constructive Research and Info-Computational

Knowledge Generation.” *Model-Based Reasoning in Science and Technology*: 52–67.

Dong, Ying, and Christian P. Speer. 2015. “Late-Onset Neonatal Sepsis:Recent Developments.” *Archives of Disease in Childhood: Fetal and Neonatal Edition* 100(3): F257–63.

Eivazzadeh, Shahryar et al. 2016. “Evaluating Health Information Systems Using Ontologies.” *JMIR Medical Informatics* 4(2): 1–13.

Fairchild, Karen D, and T Michael O Shea. 2011. “Heart Rate Characteristics: Physiomarkers for Detection of Late- Onset Neonatal Sepsis.” *Clin Perinatol* 37(3): 581–98.

FedDev Ontario. 2017. “The Health Ecosphere Innovation Pipeline Is Improving Lives with New Healthcare Advances.” https://www.feddevontario.gc.ca/eic/site/723.nsf/eng/02416.html?OpenDocument.

Friedman, Mark. 2005. *Trying Hard Is Not Good Enough*. 1st Editio. Trafford: FPSI Publishing.

Glass, Jonah et al. 2018. “Development of a Monitoring and Evaluation Framework for a Public Health Setting.”

Government of Ontario. 2017. *Excellent Care for All Act*.

Griffin, M. Pamela et al. 2003. “Abnormal Heart Rate Characteristics Preceding Neonatal Sepsis and Sepsis-like Illness.” *Pediatric Research* 53(6): 920–26.

Grosfeld-nir, Abraham, Boaz Ronen, and N I R Kozlovsky. 2007. “The Pareto Managerial Principle : When Does It Apply ?” 45(10): 2317–25.

Heidebrecht, Christine L et al. 2014. “Electronic Immunization Data Collection Systems : Application of an Evaluation Framework.”

Information and Privacy Commisioner of Ontario. 2015. *Planning for Success : Privacy Impact Assessment Guide*.

Inibhunu, Catherine et al. 2019. “Adaptive API for Real-Time Streaming Analytics as a Service.” *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*: 3472–77.

Johnston, Karissa M et al. 2014. “The Economic Burden of Prematurity in Canada.” *BMC Pediatrics* 14(1): 1–10. BMC Pediatrics.

Kasanen, E., Lukka, K. Siitonen, A. 1993. “The Constructive Approach in Management Accounting Research.” *Journal of Management Accounting Research* 5(August).

Khajouei, Reza, Misagh Zahiri Esfahani, and Yunes Jahani. 2016. “Comparison of Heuristic and Cognitive Walkthrough Usability Evaluation Methods for Evaluating Health Information Systems.” *Journal of the American Medical Informatics Association* (0): 1–6.

Kilsdonk, E, L W Peute, and M W M Jaspers. 2017. “International Journal of Medical Informatics Factors Influencing Implementation Success of Guideline-Based Clinical Decision Support Systems : A Systematic Review and Gaps Analysis.” *International Journal of Medical Informatics* 98: 56–64. http://dx.doi.org/10.1016/j.ijmedinf.2016.12.001.

Kilsdonk, E, L W P Peute, and M W M Jaspers. 2012. “Factors Known to Influence Acceptance of Clinical Decision Support Systems.”

Kitchin, Rob, and Gavin Mcardle. 2016. “What Makes Big Data , Big Data ? Exploring the Ontological Characteristics of 26 Datasets.” (June): 1–10.

Kopanitsa, Georgy, Zhanna Tsvetkova, and Hasan Veseli. 2012. “Application of Usability Metrics in a Multi-User and Multimedia EHR Evaluation Framework.” In *Quality of Life through Quality of Information*, IOS Press, 358–62.

Lund, By Arnold M. 2001. “Usability Interface Measuring Usability with the USE Questionnaire.” *Usability Interface* 8(2). http://www.stcsig.org/usability/newsletter/0110_measuring_with_use.html.

McGregor, Carolyn. 2008. “Ontario Tech Health Informatics Research.” hir.uoit.ca.

———. 2013. “Big Data in Neonatal Intensive Care.” *Computer* 46(6): 54–59.

———. 2013. “Real-Time Multidimensional Temporal Analysis of Complex High Volume Physiological Data Streams in the Neonatal Intensive Care Unit.” *Studies in Health Technology and Informatics* 192(1–2): 362–66.

McGregor, Carolyn, Christina Catley, and Andrew James. 2012. “Variability Analysis with Analytics Applied to Physiological Data Streams from the Neonatal Intensive Care Unit.” *2012 25th IEEE International Symposium on Computer-Based Medical Systems (CBMS)*: 1–5.

McGregor, Carolyn, Christina Catley, Andrew James, and James Padbury. 2011. “Next Generation Neonatal Health Informatics with Artemis.” *Studies in Health Technology and Informatics* 169: 115–19.

Mchugh, Mary L. 2012. “Lessons in Biostatistics Interrater Reliability : The Kappa Statistic.” *Biochemia Medica*: 276–82.

McMaster Children’s Hospital. 2019. “Neonatal Nurseries.” http://www.mcmasterchildrenshospital.ca/body.cfm?id=85&fr=true.

Mohammed, Siti Asma, and Maryati Mohd. Yusof. 2013. “Towards an Evaluation Framework for Information Quality Management (IQM) Practices for Health Information Systems – Evaluation Criteria for Effective IQM Practices.” *Journal of Evaluation in Clinical Practise* 19: 379–87.

Newnham, John P. et al. 2017. “Reducing Preterm Birth by a Statewide Multifaceted Program : An Implementation Study.” *American Journal of Obstetrics and Gynecology* 216(5): 434–42. http://dx.doi.org/10.1016/j.ajog.2016.11.1037.

Nielsen, Jakob. 1994. “Ten Usability Heuristics.” http://intra.iam.hva.nl/content/1112/verdieping1/research_for_design/intro-en-materiaal/RfD-Heuristic-Evaluation.pdf.

Ontario Tech HIR. 2019. “Ontario Tech HIR Twitter Page.” twitter.com/uoithir.

ORION. 2019. “Saving Ontario’s Most Vulnerable Infants.” https://www.orion.on.ca/news-events/blog/saving-ontarios-most-vulnerable-infants/.

———. 2020. “Our Network.”

Provincial Council for Masternal and Child Health. 2013. “Standardized Maternal and Newborn Levels of Care Definitions.” : 1–12.

Provincial Council for Maternal and Child Health. 2018. *Level of Care Designation*. http://www.pcmch.on.ca/wp-content/uploads/2018/06/LOC-Website-List-June-2018-Update.pdf.

Public Health Agency of Canada. 2013. “What Is the Population Health Approach.” https://www.canada.ca/en/public-health/services/health-promotion/population-health/population-health-approach/what-population-health-approach.html#health.

Public Health Ontario. 2016. 117 Public Health *Evaluating Health Promotion Programs: Introductory Workbook.*

Pugh, James Edward et al. 2018. *Artemis Cloud Clinical Decision Support Framework - An Algorithm to Accurately Detect Late on Set Neonatal Sepsis*.

Raghupathi, Wullianallur, and Viju Raghupathi. 2014. “Big Data Analytics in Healthcare : Promise and Potential.” : 1–10.

Rance, Stuart. 2013. *Defining Availability In The Real World*.

Rigby, M. et al. 2013. “Evidence Based Health Informatics : 10 Years of Efforts to Promote the Principle.” *IMIA Yearbook of Medical Informatics*: 34–46.

Sadoughi, Farahnaz, Khalil Kimiafar, Maryam Ahmadi, and Mohammad Taghi Shakeri. 2013. “Determining of Factors Influencing the Success and Failure of Hospital In-Formation System and Their Evaluation Methods : A Systematic Review.” *Iran Red Crescent Medical Journal* 15(12).

Stoll, B. J. et al. 2002. “Late-Onset Sepsis in Very Low Birth Weight Neonates: The Experience of the NICHD Neonatal Research Network.” *Pediatrics* 110(2): 285–91. http://pediatrics.aappublications.org/cgi/doi/10.1542/peds.110.2.285.

Tusikov, Natasha, and Robert C. Fahlman. 2008. *Threat and Risk Assessments*.

Wang, Yichuan, Leeann Kung, and Terry Anthony. 2018. “Technological Forecasting & Social Change Big Data Analytics : Understanding Its Capabilities and Potential Bene Fi Ts for Healthcare Organizations.” 126: 3–13.

Whitefield, Andy, Frank Wilson, and John Dowell. 1991. “A Framework for Human Factors Evaluation.” *Behaviour & Information Technology* 10(1): 65–79.

World Health Organization. 2008. “Health Information Systems.” (June).

———. 2018. “Preterm Birth.” https://www.who.int/news-room/fact-sheets/detail/preterm-birth.

Yusof, Maryati Mohd. 2015. “A Case Study Evaluation of a Critical Care Information System Adoption Using the Socio-Technical and Fit Approach.” *International Journal of Medical Informatics* 84(7): 486–99.

Yusof, Maryati Mohd, Jasna Kuljis, Anastasia Papazafeiropoulou, and Lampros K. Stergioulas. 2008. “An Evaluation Framework for Health Information Systems: Human, Organization and Technology-Fit Factors (HOT-Fit).” *International Journal of Medical Informatics* 77(6): 386–98.