

**IMPROVING THE TIMELINESS, ACCURACY, AND COMPLETENESS OF
MORTALITY REPORTING USING FHIR APPS AND MACHINE LEARNING**

A Dissertation
Presented to
The Academic Faculty

By

Ryan Alan Hoffman

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
Wallace H. Coulter Department of Biomedical Engineering

Georgia Institute of Technology
Emory University

August 2021

Copyright © Ryan Alan Hoffman 2021

**IMPROVING THE TIMELINESS, ACCURACY, AND COMPLETENESS OF
MORTALITY REPORTING USING FHIR APPS AND MACHINE LEARNING**

Approved by:

Dr. May Dongmei Wang, PhD, Advisor
Wallace H. Coulter Department of
Biomedical Engineering
*Georgia Institute of Technology and
Emory University*

Dr. Cassie S. Mitchell, PhD
Wallace H. Coulter Department of
Biomedical Engineering
*Georgia Institute of Technology and
Emory University*

Dr. Wilbur A. Lam, MD, PhD
Wallace H. Coulter Department of
Biomedical Engineering
*Georgia Institute of Technology and
Emory University*

Dr. Kevin O'Donnell Maher, MD
Department of Pediatrics
Emory University School of Medicine

Dr. Nikhil Kumar Chanani, MD
Department of Pediatrics
Emory University School of Medicine

Date Approved: April 28, 2021

If you torture the data long enough, it will confess to anything.

Attributed to Ronald H. Coase, "How Should Economists Choose?"

ACKNOWLEDGMENTS

I wish to thank my advisor, Professor May Wang, for many years of wise advice and patient supervision. I thank all of the members of my thesis committee for their time, support, and invaluable feedback. To my fellow Wang lab members over the years, too numerous to list here, I thank you for your collaboration, mentorship, and friendship.

Reaching this point would not have been possible at all without the personal support of many people, including my wonderful and supportive family. Mom, Dad, Reid, and Maisie, thank you for your support and motivation throughout this journey, and I love you all. Joey, Sydney, Hannah, and all of my friends from Georgia Tech and before, thank you for your all of your support and encouragement.

For their generous support, I thank: the Georgia Institute of Technology; Emory University; the National Institutes of Health, including through a T32 traineeship under Prof. Greg Gibson; the National Science Foundation; the Centers for Disease Control and Prevention; Children's Healthcare of Atlanta; and Shriners Hospitals for Children.

TABLE OF CONTENTS

Acknowledgments	iv
List of Tables	xi
List of Figures	xiv
List of Abbreviations	xvi
Summary	xviii
Chapter 1: Introduction	1
1.1 Specific Aims and Overview	1
1.1.1 Part 1: Mortality Reporting Informatics	1
1.1.2 Part 2: Using FHIR to Enable Public Health Informatics	2
1.1.3 Part 3: App Improvement and Usability Testing	2
1.2 Background	2
Part 1: Mortality Reporting Informatics	7
Chapter 2: Philosophy on Machine Learning for Mortality Reporting Decision Support	8
2.1 Preface	8
2.2 Supervised vs. Unsupervised Machine Learning for Mortality Reporting	8

2.2.1	Background	8
2.2.2	Data and Machine Learning Problems in Mortality Reporting	12
2.2.3	Choice of Unsupervised Techniques	14
2.2.4	Parameter Selection	14
2.3	Philosophy of Overall Project Design	15
2.3.1	Potential Benefits	15
2.3.2	Potential Pitfalls	16
2.3.3	Ethical Considerations	16
Chapter 3: Foundations of Rule Mining		19
3.1	Introduction	19
3.2	Association Rule Mining	19
3.2.1	Rule Metrics	21
3.2.2	Frequent Itemset Mining Approaches	25
3.2.3	Generating Association Rules	29
3.2.4	Case Study: ARM with Apriori Algorithm	29
3.3	Sequential Rule Mining	30
3.3.1	Comparison To ARM	32
3.3.2	Sequential Pattern Mining Approaches	35
3.3.3	Generating and Scoring Sequential Rules	38
3.4	Application to Biomedical Informatics	40
Chapter 4: Improving Validity of Cause of Death on Death Certificates		41
4.1	Preface	41

4.2	Abstract	41
4.3	Introduction	42
4.4	Experimental and Computational Details	46
4.4.1	Data Sources: Expert Knowledge Base	46
4.4.2	Data Sources: Death Certificate Data	46
4.4.3	Deriving Frequent COD Patterns from Death Certificates	47
4.4.4	Deriving Rules from Expert Knowledge	48
4.5	Results and Discussion	49
4.5.1	Results from SRM Analysis	49
4.5.2	Results from Comparing Rules from SRM with Expert Knowledge	49
4.6	Conclusions	50
4.7	Acknowledgements	54
Chapter 5: Advanced Rule Mining for Mortality Reporting Decision Support . .		56
5.1	Introduction	56
5.2	Data	56
5.2.1	Mortality Data	56
5.3	Methods	57
5.3.1	Computing Environment and Tools	57
5.3.2	PrefixSpan SPM	58
5.3.3	Sequential Rule Mining	58
5.4	Results and Discussion	60
5.4.1	Sequential Pattern Mining	60

5.4.2	Sequential Rule Mining	62
5.5	Conclusions and Future Work	69
Part 2: Using FHIR to Enable Public Health Informatics		72
Chapter 6: FHIR Profiling for Mortality Reporting		73
6.1	Preface	73
6.2	Problem Statement	73
6.3	Background	74
6.3.1	FHIR	74
6.3.2	FHIR Profiling	76
6.3.3	FHIR Package Registries	78
6.4	Mortality Reporting Profiling	78
6.5	FHIR - VR DAM Mappings	81
Chapter 7: Prototyping Mortality Reporting with FHIR		82
7.1	Preface	82
7.2	Abstract	82
7.3	Introduction	83
7.4	Web Application Design	84
7.4.1	Application Features	85
7.4.2	Illustrative Synthetic Data	86
7.4.3	Interface Design	86
7.5	Sequential Pattern Mining Analytics	86

7.5.1	Data	88
7.5.2	SPM Background and Related Work	88
7.5.3	SPM Problem Formulation	89
7.5.4	SPM Methodology	90
7.5.5	Pattern Mining Results	90
7.6	Conclusions and Future Work	92
7.7	Acknowledgements	92
Part 3: App Improvement and Future Directions		94
Chapter 8: Intelligent Mortality Reporting with FHIR		95
8.1	Preface	95
8.2	Abstract	95
8.3	Introduction	96
8.4	Web Application Design	97
8.5	Representing Death Certificate Data in FHIR	99
8.6	Sequential Pattern Mining Analytics	102
8.7	Conclusions and Future Work	105
8.8	Acknowledgements	105
Chapter 9: Conclusions		108
9.1	Future Work	108
9.1.1	Application Platforms	108
9.1.2	Mortality Pattern Mining	108

9.2	Author’s Perspectives	109
9.2.1	Data Harmonization and FHIR	109
9.2.2	Maximizing Adoption and Value	112
9.3	Closing	113
Appendices		114
Appendix A: Additional Tables and Results		115
Appendix B: Death Certificate Data Mapping		131
Appendix C: Comparison of Normalization Algorithms for Cross-Batch Color Segmentation of Histopathological Images		140
Appendix D: A High-Resolution Tile-Based Approach for Classifying Biologi- cal Regions in Whole-Slide Histopathological Images		154
Appendix E: Publications List		166
Appendix F: Copyright Statements		169
References		183
Vita		198

LIST OF TABLES

2.1	Excerpt from Fisher’s Iris Data Set	9
3.1	MSWeb Data Set - Size	19
3.2	MSWeb Data Set - Item Set Sample	20
3.3	MSWeb Data Set - Data Sample	20
3.4	Support Counting Example	23
3.5	Rule Metric Example	26
3.6	Case Study: Frequent Itemsets	30
3.7	Case Study: Association Rules	32
3.8	Case Study: Rule Metrics and Ranking	33
4.1	Size of Frequent Patterns from SRM Analysis	49
4.2	Top 10 Invalid Rules of Length 2	51
4.3	Top 10 Invalid Rules of Length 3	52
4.4	Top 10 Valid Rules of Length 2	53
5.1	NVSS Data Set Summary	57
5.2	Length of Frequent Sequences, PrefixSpan with minsupport= 10^{-3}	62
5.3	Top 15 Frequent 1-Sequences, PrefixSpan with minsupport= 10^{-3}	64
5.4	Top 15 Frequent 2-Sequences, PrefixSpan with minsupport= 10^{-3}	65

5.5	Top 10 Frequent 3-Sequences, PrefixSpan with minsupport= 10^{-3}	66
5.6	Length of Frequent Sequences, PrefixSpan with minsupport= 10^{-4}	69
5.7	Length of Frequent Sequences, PrefixSpan with minsupport= $2 \cdot 10^{-5}$	69
5.8	Length of Frequent Sequences, PrefixSpan with minsupport= 10^{-5}	70
5.9	Top Sequential Rules, Ordered by Lift	70
5.10	Top Sequential Rules, Ordered by Support	71
5.11	Top Sequential Rules, Ordered by Confidence	71
7.1	Highly Supported Prototype Rules	91
7.2	Prototype Rule Count by Length	92
8.1	NCHS Death Record Layout	103
8.2	Frequent Sequence Count of Different Lengths	105
8.3	Rules from 2012 NCHS Mortality Data	106
A.1	Top 15 Frequent 1-Sequences, PrefixSpan with minsupport= 10^{-4} , not included in minsupport= 10^{-3}	116
A.2	Top 15 Frequent 2-Sequences, PrefixSpan with minsupport= 10^{-4} , not included in minsupport= 10^{-3}	117
A.3	Top 10 Frequent 3-Sequences, PrefixSpan with minsupport= 10^{-4} , not included in minsupport= 10^{-3}	118
A.4	Top 10 Frequent 4-Sequences, PrefixSpan with minsupport= 10^{-4} , not included in minsupport= 10^{-3}	119
A.5	Top 4 Frequent 5-Sequences, PrefixSpan with minsupport= 10^{-4} , not included in minsupport= 10^{-3}	120
A.6	Top 15 Frequent 1-Sequences, PrefixSpan with minsupport= $2 \cdot 10^{-5}$, not included in minsupport= 10^{-4}	121

A.7	Top 15 Frequent 2-Sequences, PrefixSpan with minsupport= $2 \cdot 10^{-5}$, not included in minsupport= 10^{-4}	122
A.8	Top 10 Frequent 3-Sequences, PrefixSpan with minsupport= $2 \cdot 10^{-5}$, not included in minsupport= 10^{-4}	123
A.9	Top 10 Frequent 4-Sequences, PrefixSpan with minsupport= $2 \cdot 10^{-5}$, not included in minsupport= 10^{-4}	124
A.10	Top 10 Frequent 5-Sequences, PrefixSpan with minsupport= $2 \cdot 10^{-5}$, not included in minsupport= 10^{-4}	125
A.11	Top 15 Frequent 1-Sequences, PrefixSpan with minsupport= 10^{-5} , not included in minsupport= $2 \cdot 10^{-5}$	126
A.12	Top 15 Frequent 2-Sequences, PrefixSpan with minsupport= 10^{-5} , not included in minsupport= $2 \cdot 10^{-5}$	127
A.13	Top 10 Frequent 3-Sequences, PrefixSpan with minsupport= 10^{-5} , not included in minsupport= $2 \cdot 10^{-5}$	128
A.14	Top 10 Frequent 4-Sequences, PrefixSpan with minsupport= 10^{-5} , not included in minsupport= $2 \cdot 10^{-5}$	129
A.15	Top 10 Frequent 5-Sequences, PrefixSpan with minsupport= 10^{-5} , not included in minsupport= $2 \cdot 10^{-5}$	130
B.1	VR DAM to FHIR Resource Mapping	132
C.1	Accuracy Comparison of Normalization Methods	150
C.2	Performance Comparison of Normalization Methods	152
D.1	Contribution of Each Feature Class to Each Classifier Model	161
D.2	Cross-Validation Accuracy of Each Classifier Model	162
D.3	Correlation Coefficients for Slide-Level Classification Models	163

LIST OF FIGURES

2.1	Fisher’s Iris Data Visualization	10
2.2	Classification Illustration	10
2.3	Clustering Illustration	12
2.4	Example Death Certificate	13
2.5	Overall Project Design	18
3.1	Apriori Scala Implementation	31
4.1	Example Invalid Multi-Part Rule	50
5.1	Design and Methods Diagram	60
5.2	Support Parameter Sensitivity	63
5.3	Rule Mining Parameter Sensitivity	67
5.4	Rule Mining Results, Lift ≥ 10	68
6.1	FHIR Patient Resource Specification Excerpt	75
6.2	2003 US Standard Certificate of Death	79
6.3	First Draft of Resource-Level Mapping	81
7.1	Proposed Infrastructure for Death Reporting Application	85
7.2	Prototype Application User Interface	87

8.1	Proposed Infrastructure for Death Reporting Application	98
8.2	Web Application User Interface	99
8.3	FHIR Resource-Level Mapping Overview	101
9.1	FHIR Version Support Fragmentation	111
C.1	Normalization Algorithm Candidates	144
C.2	Clustering Performance Comparison	147
C.3	Segmentation Performance Comparison	151
D.1	Typical Whole-Slide Image	157
D.2	WSI Quality Control	158
D.3	WSI Color Normalization	158
D.4	Nested Cross-Validation Flowchart	160
D.5	External Validation	164

LIST OF ABBREVIATIONS

API	Application Programming Interface
ARM	Association Rule Mining
AWS	Amazon Web Services
AWS EMR	Elastic MapReduce
BHI	Biomedical and Health Informatics
CDA	Clinical Document Architecture
CDC	Centers for Disease Control and Prevention
CMS	Center for Medicare and Medicaid Services
COD	Cause(s) of Death
DSTU	Draft Standard for Trial Use
EHR	Electronic Health Record
EMR	Electronic Medical Record
FDA	U.S. Food and Drug Administration
FHIR	Fast Healthcare Interoperability Resources
GLM	Generalized Linear Model
HHS	US Department of Health and Human Services
HIPAA	The Health Insurance Portability and Accountability Act of 1996
HL7	Health Level Seven
IRB	Institutional Review Board
NCHS	National Center for Health Statistics
NIH	National Institutes of Health
NSF	National Science Foundation

NVSS National Vital Statistics System

PHI Protected Health Information

PHR Personal Health Record

R4 FHIR Release 4

SPM Sequential Pattern Mining

SRM Sequential Rule Mining

STU Standard for Trial Use

SVM Support Vector Machine

USA United States of America

SUMMARY

There are approximately 56 million deaths per year world-wide, with millions happening in the United States. Accurate and timely mortality reporting is essential for gathering this important public health data in order to formulate emergency response to epidemics and new disease threats, to prevent communicable diseases such as flu, and to determine vital statistics such as life expectancy, mortality trends, etc. However, accurate collection and aggregation of high-quality mortality data remains an ongoing challenge due to issues such as the average low frequency with which physicians perform death certification, inconsistent training in determining the causes of death, complex data flow between the funeral home, the certifying physician and the registrar, and non-standard practices of data acquisition and transmission. We propose a smart application for medical providers at the point-of-care which will use Fast Healthcare Interoperability Resources (FHIR) to integrate directly with the medical record, provide the practitioner with context for the death, and use machine learning techniques to enable the reporting of an accurate and complete causal chain of events leading to the death.

CHAPTER 1

INTRODUCTION

There are approximately 56 million deaths per year world-wide [1], with 2.6 million happening in the United States [2]. Accurate and timely mortality reporting is essential for gathering this important public health data in order to formulate emergency response to epidemics and new disease threats, to prevent communicable diseases such as flu, and to determine vital statistics such as life expectancy, mortality trends, etc. However, accurate collection and aggregation of high-quality mortality data remains an ongoing challenge due to issues such as the average low frequency with which physicians perform death certification (on the order of 1-2 times a year), inconsistent training in determining the causes of death, complex data flow between the funeral home, the certifying physician and the registrar, and non-standard practices of data acquisition and transmission [3, 4].

Fast Healthcare Interoperability Resources (FHIR) is a new HL7 interoperability standard for Electronic Health Record (EHR) systems, and SMART-on-FHIR is a framework that enables third-party FHIR app development that can work “out of the box” [5]. In this proposal, we outline a plan to develop advanced informatics tools for mortality reporting decision support, as well as to implement a FHIR application that can deliver these tools in a useful way at the point of care. With these methods, we aim to improve the timeliness, accuracy, and completeness of mortality reporting.

1.1 Specific Aims and Overview

1.1.1 Part 1: Mortality Reporting Informatics

Mortality reporting data is one of the most valuable public data sets available today. However, issues with data quality persist. In Part 1, we examine historical mortality reporting

data to quantitatively establish the need for more accurate and complete mortality reporting solutions. Building from this work, we use sequential rule mining techniques to develop decision support systems capable of closing this gap by proposing potential causal linkages between conditions in a decedent’s medical records. The current state of mortality reporting data quality and validity is covered in Chapter 4, and Chapter 5 outlines a novel approach to rule mining for clinical decision support in mortality reporting.

1.1.2 Part 2: Using FHIR to Enable Public Health Informatics

FHIR is a new standard from HL7 which enables high degrees of interoperability between medical systems. This standard design, and standard interfaces for it, make FHIR apps an ideal tool for delivering public health informatics solutions. In Part 2, we have developed a complete working prototype FHIR web application for mortality reporting, to demonstrate the value of FHIR in this field. The first version of the prototype app is presented in Chapter 7. In addition, we have developed and published mappings between existing vital records data standards and FHIR data models, enabling the development of FHIR apps that work within existing public health data workflows. These profiles are outlined in Chapter 6.

1.1.3 Part 3: App Improvement and Usability Testing

Finally, Part 3 introduces improved versions of the mortality reporting app. Chapter 8 shows an improved version of the prototype web app presented in Chapter 7, and we discuss ongoing improvements to the application and its future potential in Chapter 9.

1.2 Background

The global “big data” revolution is already having transformative impacts on healthcare in the United States and around the world. Driven by the manifest value of having health data easily accessible, and further reinforced by “meaningful use” and similar regulations, healthcare providers and stakeholders have largely taken to electronic storage and dissemi-

nation of records. With global expenditures for health exceeding \$6.5 trillion USD in 2010, accounting for over 10% of the total global GDP, there is a clear need to optimize healthcare practices and reduce health-related costs [6, 7]. The great challenge of health informatics is transforming and integrating this daunting pool of diverse information into generalizable knowledge and actionable insights that can be used to directly address these concrete objectives: improving patient outcomes, advancing the quality of healthcare science, and reducing the costs associated with healthcare.

The first recognizable Electronic Medical Record (EMR) and Electronic Health Record (EHR) systems were developed as early as the 1960s. One of the first such systems, the Problem-Oriented Medical Information System (PROMIS) was introduced in the 1970's [8]. The problem-oriented medical record philosophy on which it was based became the foundation of the "SOAP note" method for making patient notes, which is still in ubiquitous use today [9]. Immediately apparent was the opportunity and imperative to integrate medical knowledge with the more basic information retrieval functions of the EHR, through methods that would come to be called clinical decision support systems [10, 11]. Adoption of EHR systems in the United States accelerated rapidly following the passage of the HITECH Act, under which the Center for Medicare and Medicaid Services (CMS) introduced an EHR Incentive Program. This program provided financial incentives based on the concept of "meaningful use" of EHRs, outlining a list of key ways that EHR systems could be used to enhance patient care beyond the paper systems that they were replacing.

Personal Health Record (PHR) and patient portal access are extensions of EHR systems designed to provide patients direct access to their health history. In practice, the term "patient portal" generally refers to a patient-facing interface for accessing hospital- or physician-controlled EMR data. By contrast, PHRs are typically patient-controlled, on a system of the patient's choosing, and not tied to any one healthcare institution. Both can be considered part of the broader "personal health" initiative, in which it is thought that outcomes and engagement can be improved by providing patients with more detailed informa-

tion about their health, and both face some similar challenges in adoption and meaningful implementation.

Patient portals provide patients an interface to view the data stored in their provider's EHR. This gives patients the ability to detect and fix errors, as well as to more cooperatively engage in the process of healthcare decision making [12, 13]. However, the low medical literacy of patients and the perceived difficulty of navigating the portal software itself remain significant obstacles to patient [14]. Providers have also expressed concerns about sharing patient notes, regulatory requirements for correcting data, and intellectual property questions relating to the health data's ownership [15, 13]. In spite of these concerns, initial studies on outcome improvement driven by patient access to medical records appear positive [16] and adoption of portals and PHRs appear to be on pace to satisfy Stage 3 meaningful use targets before 2020 [13].

Mobile technologies first came into medical practice in the 1990's with the advent of PDAs, which could be used to provide clinicians with both medical record data and reference materials [17]. The terms "mobile health" and "mHealth" were coined in the early 2000's to acknowledge the transformative power of mobile technology to enable novel health applications. With the introduction of iOS and Android-based smartphones in 2007-8, the availability and computational power of mobile devices accelerated rapidly [18]. Varshney et. al. identified four major themes of mHealth research: extending the reach of healthcare to previously inaccessible environments, providing decision support, tracking and preventing chronic conditions, and accelerating emergency care [19].

The idea of integrating clinical care data with medical knowledge to enable decision support is as old as the earliest electronic medical record systems themselves, and remains an area of intense research today. With the proliferation of interoperable EHR systems and connected instruments and departments, the amount of data stored in US EHR systems alone is estimated to exceed 150 exabytes [20] (1 exabyte = 10^9 gigabytes). With such large volumes of data collected, clinical informatics is no longer just a way to add value

to existing data stores – it is becoming an ever more essential tool for condensing and evaluating data at the point-of-care.

Despite this clear and compelling benefit, significant institutional and technological hurdles exist in the condensation and integration of the already extant multimodal biomedical data that has the potential to enable advanced biomedical and health informatics. Diverse methods for enabling data interoperability have been proposed and implemented to varying degrees. Messaging-based systems such as DICOM and Health Level Seven’s (HL7) v2 standard are widely used around the world, and enable the standardized exchange of short, structured healthcare data transactions. Document-based systems like HL7 Clinical Document Architecture (CDA) add a level of abstraction to these systems, representing healthcare “documents” as more holistic collections of data and interpretation. An illustrative example of a clinical document that can be readily encoded as a CDA file is an imaging report – the document may include medical images, the interpretation of those images made by the radiologist, and the patient and provider information necessary to establish its context.

Messaging systems are limited by their small message size and structured layout, and document-centric models are limited by the static and cryptic nature of their large, flat files. HL7’s Fast Healthcare Interoperability Resources (FHIR) (pronounced “fire”) is an attempt to answer these concerns. FHIR’s approach is to provide both extensible data models (called FHIR Resources) as well as a set of RESTful APIs by which to modify those resources [21]. The use of modern web-based APIs is designed to make FHIR application development accessible to any web developer, as well as solve the issue of static, unchanging files that has limited CDAs usefulness. EHR vendors are using the FHIR standard to provide “app store”-like capabilities to integrate 3rd party, interoperable, and novel tools into their EHR systems [5].

Fully integrated, clinician-facing informatics tools are the state-of-the-art way to bring informatics research advances directly to biomedical practice. Such systems can be seam-

lessly integrated into current clinical practices, lowering the barriers to entry and minimizing the degree of clinician "buy-in" required before potential benefits can begin to be realized. What's more, such intimate access to subject-matter experts can provide insights that feed back into the application, using knowledge capture to improve informatics performance [22] and enabling research-clinician collaboration to improve patient care [23, 24, 25].

PART 1
MORTALITY REPORTING INFORMATICS

CHAPTER 2

PHILOSOPHY ON MACHINE LEARNING FOR MORTALITY REPORTING DECISION SUPPORT

2.1 Preface

In this chapter, the author presents his personal philosophy, insights, advice, and opinions on the design, pros, cons, and ethical implications of specific machine learning techniques as applied to public health and mortality data reporting. For the author’s personal philosophy and insights into the broader field of data science, public health data science, and the future of the field, please see Chapter 9.

2.2 Supervised vs. Unsupervised Machine Learning for Mortality Reporting

2.2.1 Background

Machine learning techniques can be broadly divided into three categories: (1) supervised learning, (2) unsupervised learning, and (3) reinforcement learning [26]. Reinforcement learning is an agent-based approach that is well suited to dynamic, simulation-based problems, but is not as directly applicable to the analysis of large, retrospective data sets such as are found in mortality reporting (see Section 5.2). So, we will confine ourselves to considering supervised and unsupervised approaches.

The fundamental difference between supervised and unsupervised machine learning techniques is whether some “ground truth” reference is needed. In supervised machine learning, the training data is labeled, tagged, or linked to some outcome measure of interest. The learning that takes place is then optimizing the ability to predict that outcome, based on the input data. Unsupervised methods do not require such labels, and instead discover patterns inherent to unlabeled data.

Table 2.1: Excerpt from Fisher’s Iris Data Set

Species	Sepal L	Sepal W	Petal L	Petal W
Setosa	5.1	3.5	1.4	0.2
Setosa	4.9	3.0	1.4	0.2
Versicolor	7.0	3.2	4.7	1.4
Versicolor	6.4	3.2	4.5	1.5
...			...	

2.2.1.1 Supervised Learning

The classic example of a problem well-suited to supervised machine learning is classification. In this problem, the data available for training the model is such that each data sample consists of one or more measures, and each sample is also labeled with its class membership. The challenge is then to train a classifier that, given new, unlabeled test sample, it can predict the new sample’s class membership. Consider as an example an excerpt from the Fisher’s Iris data set [27], excerpted in Table 2.1 and visualized in Figure 2.1. For each of the flowers sampled, four size measurements were taken. Each sample also includes the species label. The task of the classifier trained on the Fisher’s iris data set would then be to predict the species label of a new, unlabeled flower data sample. Regression is another type of supervised machine learning technique where the output measure is a continuous value to be predicted, rather than a set of categories or classes.

A typical example of supervised classification is the logistic regression, or the Generalized Linear Model (GLM) with a logit link function [28]. The logit function (Equation 2.1) is the inverse of the standard logistic function (Equation 2.2):

$$\text{logit}(p) = \log \left(\frac{p}{1-p} \right) \quad (2.1)$$

$$\text{logistic}(p) = \frac{1}{1 + e^{-p}} \quad (2.2)$$

In the Generalized Linear Model (GLM), a transformer called a “link” function is used

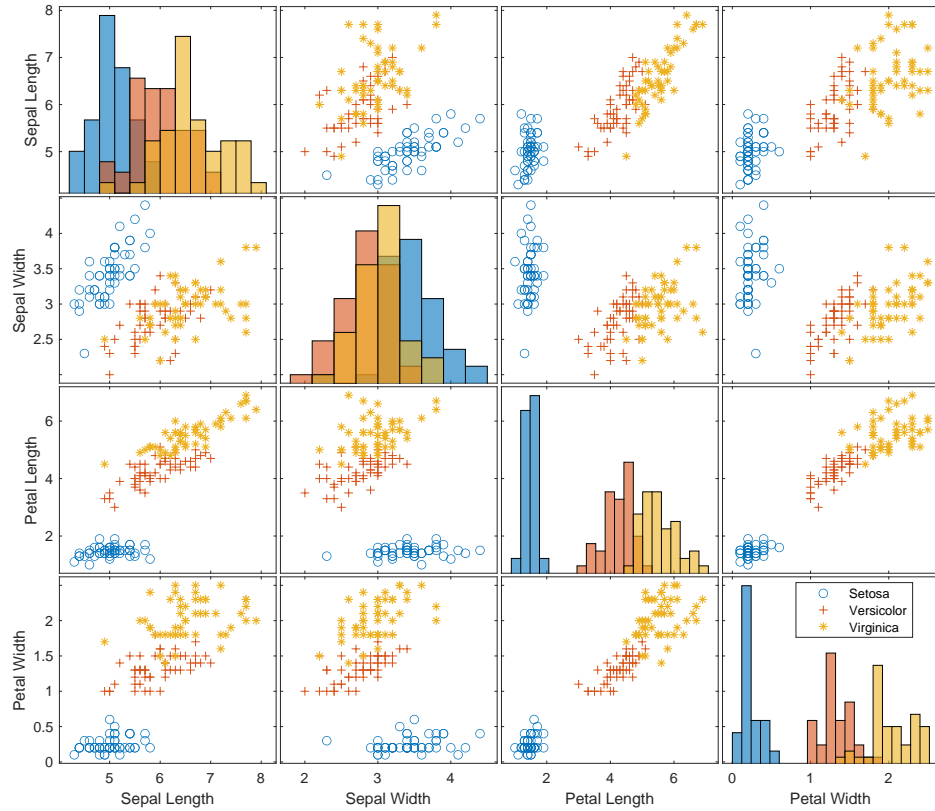


Figure 2.1: The four measures of the Fisher's iris data set are plotted in a 4×4 matrix, where the scatter plot in position (m, n) shows the m th measure on the vertical axis and the n th measure on the horizontal axis. On the diagonal, histograms show the univariate distributions of that measure for each iris species.

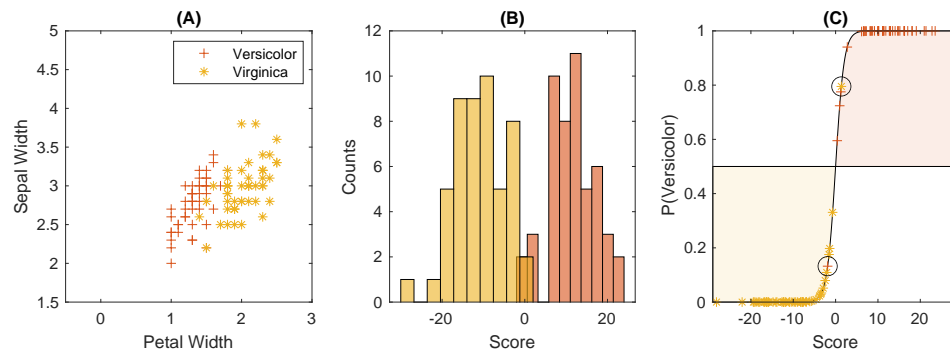


Figure 2.2: (A) The Fisher's iris data set is illustrated with two measures and plotted with each of two species having a different marker. (B) After training a generalized linear model with a logit link function, the two classes are visualized as histograms with the "Score" on the horizontal axis being the logit, or log-odds, of Versicolor class membership. (C) The log odds are transformed using the logistic function, so that the estimated probability of Versicolor class membership is shown on the vertical axis.

to map the response variables of a linear regression to some other space, so that a linear regression can be used to perform complex classification tasks. Intuitively, if the response variable is constrained, e.g. a probability of class membership $p \in [0, 1]$, it follows that linear responses to the model’s features could produce nonsensical results, e.g. negative probabilities. By using the link function to map a linear response space to the non-linear response variable, it is possible to create powerful and flexible classifiers using the computationally simple linear regression as a foundation. This process, a logistic regression using a logit link and assuming a binomial distribution, is illustrated in Figure 2.2.

The conceptual inverse of this method, where it is the feature space rather than the response variable that is transformed, is the foundation of the kernel methods, such as the famous Support Vector Machine (SVM) [29].

2.2.1.2 *Unsupervised Learning*

Unsupervised machine learning, on the other hand, is used when the corpus of data available is not “labeled” in a way conducive to predictive analysis. These techniques seek to identify patterns in the input data, in the hope that the structure of these patterns themselves will be of value.

As classification is the archetypal supervised machine learning problem, clustering is an excellent example of unsupervised learning. In this problem, the unlabeled training data must be segregated into some number of clusters based on its distribution, with the notion that samples in one cluster may be similar to one another. There are a great variety of clustering algorithms, one of the most straightforward being the naive k -means algorithm [30], in which: (1) k candidate cluster centers are randomly initialized; (2) the samples are grouped with the nearest cluster by some distance metric; (3) the cluster centers are updated to reflect the centroids of their assigned samples; and (4) steps 2-3 repeat until a solution has converged. To illustrate the potential value of unsupervised methods like k -means clustering, Figure 2.3 illustrates the algorithm successfully recovering reasonable

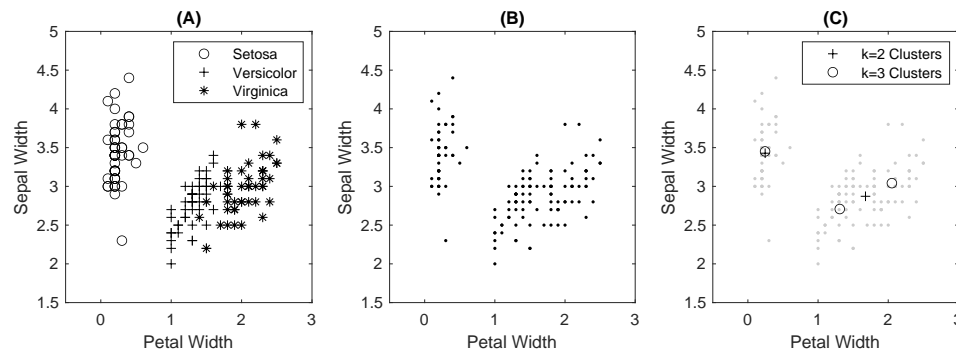


Figure 2.3: (A) The Fisher's iris data set is reduced to two measures and plotted with each species having a different marker. (B) As in an unsupervised problem the label information is withheld, leaving only the patterns and distributions inherent to the data. (C) The results of k -means clustering of the two-dimensional data, using a Euclidean distance metric with $k = 2$ and $k = 3$, are shown.

estimates of the true class centroids, using only the inherent distribution of the data.

2.2.2 Data and Machine Learning Problems in Mortality Reporting

The choice of machine learning techniques is driven not just by the outputs of the algorithm (i.e. what insights it could potentially provide), but by its inputs as well (i.e. what data is available to drive it). A detailed overview of the concrete data sets to be used is given in Section 5.2, but in general for public health, and particularly in mortality reporting, the data overwhelmingly consists of large repositories of public access data. These data sets often extend far back in time, pre-dating Electronic Medical Record (EMR) systems or the internet. This temporal coverage is a strength in and of itself, however it necessarily limits the form and complexity of the data to such as would have been possible to collect on a large scale without computer aid.

In mortality reporting, all of the data collection and reporting centers around such a reporting mechanism, pre-dating the internet: the death certificate. An example of a completed death certificate is seen in Figure 2.4. This example is typical of the information included on a death certificate for multiple years and jurisdictions. It includes: the decedent's name and location, family history and contact information, the location of the death,

BROOKLYN

Certificate of Death

Certificate No. **25366**

DEC 24 PM 4:16

1. NAME OF DECEASED (Print)		EDDIE		SCHNEIDER	
		First Name		Last Name	
PERSONAL PARTICULARS (To be filled in by Medical Examiner.)		MEDICAL CERTIFICATE OF DEATH (To be filled in by Medical Examiner. See page 2.)			
2 USUAL RESIDENCE: (a) State <u>N.Y.</u> (b) <u>Queens</u> (c) Town or City <u>Brooklyn</u> (d) No. <u>32-50 - 93d Street, Jackson Heights St.</u> (e) Length of residence or stay in City of New York immediately prior to death <u>Life</u>		16. PLACE OF DEATH: (a) NEW YORK CITY: (b) Borough <u>Brooklyn</u> (c) Name of Hospital or Institution <u>Flatbush Avenue &</u> (d) If elsewhere than in hospital or own residence, specify character of place of death, no: hotel, office, store, street, tenement, etc. <u>Deep Creek</u>			
3 SINGLE, MARRIED, WIDOWED, OR DIVORCED (write the word) <u>Married</u>		17 DATE AND HOUR OF DEATH (Month) (Day) (Year) (Hour) <u>December 23d 1940 P.M.</u>			
4 WIFE HUSBAND of <u>Gretchen</u>		18 SEX <u>Male</u> 19 Color or Race <u>White</u> 20 Approximate Age <u>29</u>			
5 DATE OF BIRTH OF DECEASED (Month) (Day) (Year) <u>October 20th 1911</u>		21 I hereby certify (a) that in accordance with Sections 878-2.0 and 878-3.0 of the Administrative Code for the City of New York, I went to, and took charge of the dead body at <u>Kings County Morgue</u>			
6 AGE <u>29</u> yrs. mos. da. hrs. or min.		this <u>24th</u> day of <u>December</u> 19 <u>40</u>			
7 OCCUPATION A Trade, profession, or particular kind of work, no sponsor, sawyer, bookkeeper, etc. B Industry or business in which work was done, as silk mill, cannery, bank, etc. <u>Aeroplane Pilot</u>		(b) that I examined the body and investigated the circumstances of this death, and I further certify from the investigation, (complete autopsy)* (partial autopsy)* (incision)* and examination, (c) that, in my opinion, death occurred on the date and at the hour stated above and resulted from (natural causes)* (accident)* (suicide)* (homicide)* (undetermined circumstances, pending further investigation)*, and (d) that the causes of death were: <u>Crushed Chest & Abdomen;</u> <u>Hemothorax & Hemoperitoneum:-</u> <u>in aeroplane crash.</u>			
8 BIRTHPLACE OF DECEASED (State or country) <u>U. S.</u> 9 How long in U. S. (if of foreign birth)					
10 IF DECEASED WAS VETERAN, NAME WAR					
11 NAME OF FATHER OF DECEASED <u>Emil</u>					
12 BIRTHPLACE OF FATHER (State or country) <u>Germany</u>					
13 MAIDEN NAME OF MOTHER OF DECEASED <u>Inga Petersen</u>					
14 BIRTHPLACE OF MOTHER (State or country) <u>Norway</u>					
15 SIGNATURE OF INFORMANT <u>GRETCHEN SCHNEIDER</u> RELATIONSHIP TO DECEASED <u>WIFE</u> ADDRESS <u>32-50-93RD ST. JACKSON HGT'S</u>		M. E. Case No. <u>4418</u>		Signed <u>Richard B. [Signature]</u> Approved <u>James A. [Signature]</u> (Cross out terms that do not apply.)	
22 PLACE OF BURIAL OR CREMATION <u>Fairview-bro. N.Y.</u>		DATE OF BURIAL OR CREMATION <u>Dec. 27, 1940</u>		PERMIT NUMBER <u>2383</u>	
23 FUNERAL DIRECTOR <u>New York Funeral Service</u>		ADDRESS <u>148 E. 24th St</u>		CITY OF NEW YORK	
BUREAU OF RECORDS		DEPARTMENT OF HEALTH			

Figure 2.4: A typical example of a completed death certificate. In this case, for a death occurring New York City in 1940. Public domain, retrieved from [31].

a sequence of causes which resulted in death, and certification by a responsible official. Of note is that death certificates do not generally contain information about the decedent's broader medical history, nor will it generally contain medical conditions which the certifier did not choose to include in the causal chain of death. There are exceptions to this rule, where additional questions have been added to death certificates to record specific risk factors, such as the decedent's smoking habits or (if they are female) whether they were recently pregnant [32]. However, these few special cases represent only a vanishingly small share of all possible medical histories.

2.2.3 Choice of Unsupervised Techniques

Driven primarily by the availability of large repositories of public health data amenable to unsupervised analysis methods, we engaged with our collaborators at the CDC to explore potential ways to add value to the mortality reporting process using unsupervised methods. One of the principal quality issues with mortality reporting data is completeness [33], and this is a problem well-suited to unsupervised methods such as collaborative filtering recommendation systems [34] or association rule mining techniques [35, 36].

Association Rule Mining (ARM) was chosen as the foundational technology for this project for several reasons: (1) the availability of suitable data for such analysis; (2) the technique's applicability to the general problem of recommendation; and (3) the method's adaptability, by way of Sequential Pattern Mining (SPM) and Sequential Rule Mining (SRM), to the temporally ordered sequences inherent to the reporting of the chain of causes of death. These concepts are explained in detail in Chapter 3.

2.2.4 Parameter Selection

In any machine learning problem, a key design decision is identifying which parameters ought to be optimized, and which ought to be fixed or constrained. Constraining the problem in at least some dimensions is a critical design choice that enables effective optimiza-

tion of the remaining parameters. In this specific application, we utilize the deep domain knowledge and subject matter expertise of our public health collaborators to constrain certain parameters. For example, in several chapter of this work various pattern mining and rule mining techniques are employed. For setting minimum support thresholds, we consulted with our collaborators at the Centers for Disease Control and Prevention (CDC) and National Center for Health Statistics (NCHS), based on what they would find interesting in their own experience. With the National Vital Statistics System (NVSS) multiple causes of death data set containing approximately 2.5 million records per year, this equates to a minimum support threshold of 0.00002, or $2 \cdot 10^{-5}$.

2.3 Philosophy of Overall Project Design

2.3.1 Potential Benefits

There are approximately 56 million deaths per year world-wide, with millions happening in the United States. The Federal government’s central listing of causes of death, referred to as the NCHS Multiple Causes of Death dataset, is one of the most valuable datasets in public health. It is used to justify large-scale policy and spending decisions, to target public health interventions, to track the efficacy of those interventions, to formulate emergency response to epidemics and new disease threats, to prevent communicable diseases such as flu, and to determine vital statistics such as life expectancy and mortality trends. However, as demonstrated by the analysis in Chapter 4, this data is demonstrably very messy and likely suffers from significant incompleteness and inaccuracy. Improvements to the quality of this dataset would make a real and immediate impact on medical practice and public policy across the United States.

The specific technical choices made in the design of this project, specifically FHIR-based interoperability and unsupervised pattern mining machine learning, have specific advantages for solving this problem. As outlined in detail in Chapter 6, FHIR strikes a delicate balance between firm, structured specification (necessary for interoperability) and flexible

customization that does not break this interoperability (to support a wide variety of users and build adoption momentum) using its profiling mechanism. This design echoes many of the challenges in public health data reporting. Multiple levels of jurisdiction with their own policies and procedures exist across the United States. For these multiple jurisdictions to readily adopt one data standard, it must be inherently flexible to easily adapt to their existing policies and practices. However this “structured flexibility” does not compromise the fundamental compatibility of the data structures, making true nation-scale interoperability achievable.

The project’s overall goals, key methods, and a few specific deliverables are shown in Figure 2.5

2.3.2 Potential Pitfalls

An inherent limitation in unsupervised machine learning methods is the more complex task of demonstrating their usefulness. Unlike supervised systems, which can express their quality in easily digestible, concrete metrics like predictive accuracy, the fundamentally exploratory nature of unsupervised methods does not lend itself to such easy analysis. Rather, it is on the data scientist to follow the data exploration to the point of generating insights outside the machine learning pipeline or, as is more applicable to this particular work, to apply the results of the analysis by creating a useful product or tool.

2.3.3 Ethical Considerations

Ethical and regulatory considerations are critical in any biomedical research where data is derived from human beings. Even though the subjects of this research are all deceased, there is still an equal ethical burden to protect the privacy of subjects’ Protected Health Information (PHI). In fact The Health Insurance Portability and Accountability Act of 1996 (HIPAA) continues to protect the PHI of the deceased for 50 years after their death.

To that end, research in this work was conducted on fully deidentified data wherever

practicable. Data was kept according to Category II - Category IV standards, as defined by the Georgia Institute of Technology's Data Access Policy [37]. Georgia Tech Institutional Review Board (IRB) clearance was sought when appropriate, and portions of this work and related or supporting projects have been governed by approved protocols including H18383, H16110, H16432, H15207, H13330, and H15073.

Beyond the issues of data integrity and privacy, there are additional ethical concerns to be reckoned with where biomedical decision support systems are concerned. Any software system that influences the care of a patient by their doctors, either directly or indirectly, must hold itself to high standards for quality.

A decision support tool is in this way ethically analogous to a medical device, which has led the U.S. Food and Drug Administration (FDA) to considering classifying clinical decision support tools as such. Over the past several years these guidelines have been relaxed, dramatically reducing the number of algorithms and tools subject to this regulation under the Cures Act [38]. This has the benefit of reducing regulatory burden for many small, innovative applications that might not otherwise make it to market. Only those applications responsible for image processing, signal processing, or otherwise making treatment suggestions on the basis of its own analysis would be subject to medical device regulations.

However, even where these specific U.S. Food and Drug Administration (FDA) regulations may not apply, or may not apply any longer, it remains incumbent upon all biomedical informaticists to hold themselves to equally high professional standards for quality and accuracy.

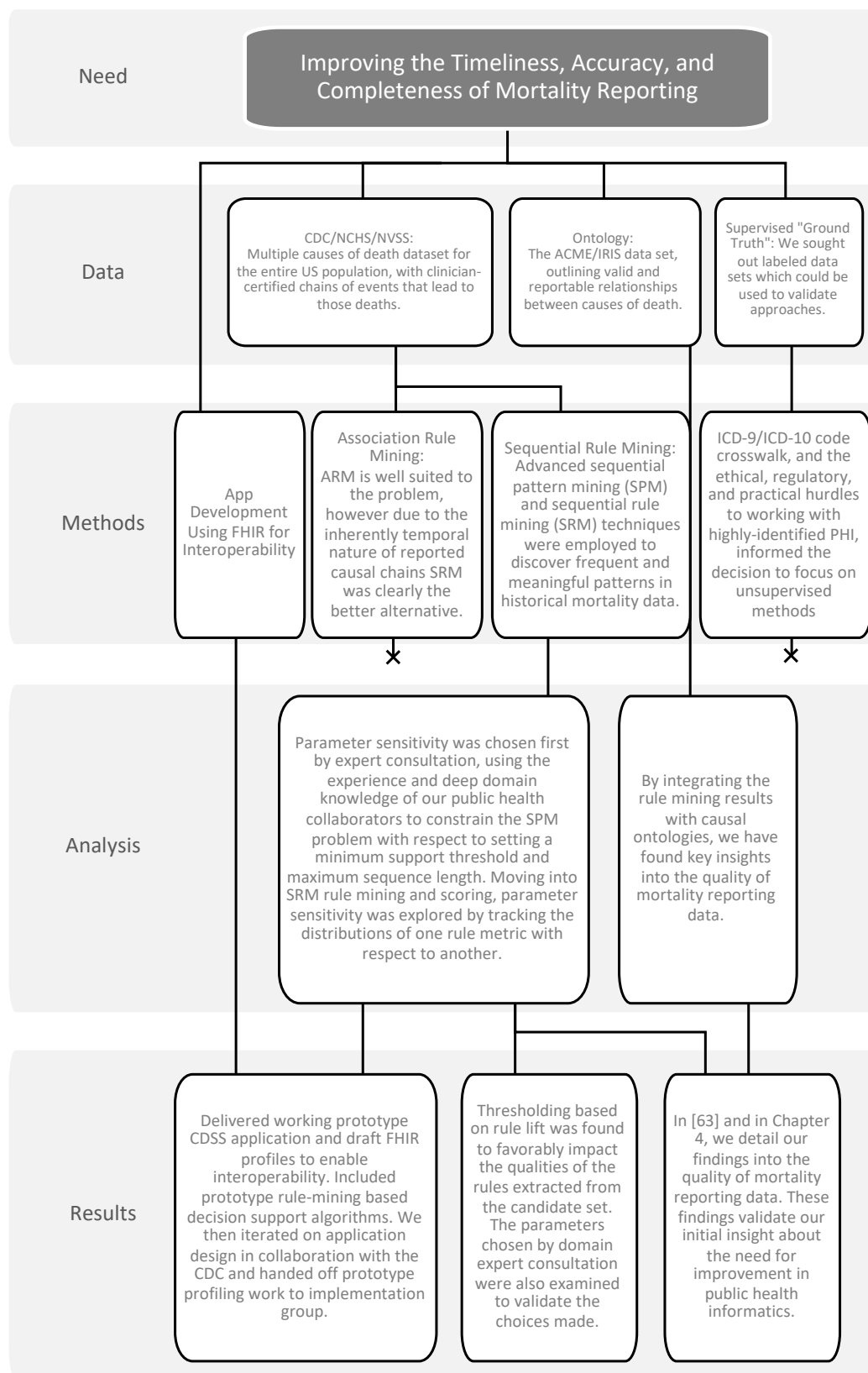


Figure 2.5: The project's overall goals, key methods, and a few specific deliverables.

CHAPTER 3

FOUNDATIONS OF RULE MINING

This chapter presents an overview of and introduction to the use of rule mining algorithms.

3.1 Introduction

For explanation and illustration purposes, we will use the MSWeb dataset [39], available from the UC Irvine Machine Learning Repository [40]. MSWeb is a transactional dataset sampled from one week of Microsoft web server logs, tracking which sections of the microsoft.com website (i.e. which of the many products and services offered on the site) that each of 37711 users visited [41, 42]. This makes it a valuable input for recommendation systems, such as those based on collaborative filtering or association rule mining, the latter of which will be detailed in this chapter. Table 3.1 shows the size of this dataset, and Table 3.2 and Table 3.3 show excerpts of the data. The excerpt in Table 3.3 will also be used for illustration purposes throughout this chapter.

3.2 Association Rule Mining

Association Rule Mining (ARM) is the process of discovering rules of the form $(X \rightarrow Y)$ from a list of transactions T , where each transaction t_n consists a set of items from the item

Table 3.1: The size of the MSWeb transaction dataset [39]. This data is parsed into one transaction for each distinct user, consisting of a set of one or more visits to sections of the microsoft.com web site.

Subset	Items	Users	Visits
Training	294	32711	98654
↳ Table 3.3	36	28	79
Testing	294	5000	15191
Total	294	37711	113845

Table 3.2: A sample of the itemset extracted from the MSWeb dataset [39]. Here the items represent "vroots", sub-directories of the microsoft.com web site corresponding to various products and services.

Item ID	Item Name	vroot
1282	home	/home
1004	Microsoft.com Search	/search
1205	Hardware Supprt	/hardwaresupport
1031	MS Office	/msoffice
1003	Knowledge Base	/kb
1118	SQL Server	/sql
1054	Exchange	/exchange
1009	Windows Family of OSs	/windows
...

Table 3.3: A sample of log data extracted from the "training" file of the MSWeb dataset [39]. Each transaction, separated by horizontal lines, represents the set of various "vroots" of the microsoft.com web site visited by a particular user.

User	Items	User	Items	User	Items	User	Items
10167	1017	10177	1008	(ctd.)	1003	10188	1038
	1071		1016		1018		1053
10168	1102		1011	10184	1008	10189	1004
	1103		1040		1007	10190	1035
10169	1001	10178	1079		1018		1008
10170	1008		1018	10185	1035		1076
	1009	10179	1038		1030		1017
10171	1030		1009		1037		1003
	1017		1026		1055		1018
10172	1061		1041		1009		1034
	1004		1034		1017	10191	1104
10173	1025	10180	1082		1004		1004
	1026		1062	10186	1077	10192	1025
10174	1034		1083		1017		1026
10175	1008	10181	1030		1031	10193	1008
10176	1020		1017		1040		1069
	1001		1004		1001	10194	1008
	1003	10182	1079		1003		1017
	1018	10183	1008	10187	1017		1034
	1004		1035		1096

set I . The antecedent itemset X is then said to be associated with the consequent itemset Y by the association rule $(X \rightarrow Y)$.

$$X \subseteq I, Y \subseteq I, |I| = K$$

$$T = (t_n) = (t_1, t_2, \dots, t_N), t_n \subseteq I$$

Two key areas of interest in the mining of association rules are the ranking of rules by certain metrics and the efficient mining of rules and frequent itemsets.

3.2.1 Rule Metrics

The most common and important metrics in the mining and analysis of association rules are support and confidence. The support of an itemset $X \subseteq I$ is defined as the proportion of transactions t in the data set T which contain the itemset.

$$\text{supp}(X) = \frac{|\{n \in \mathbb{N}_{\leq |T|}^* : X \subseteq t_n\}|}{|T|} \quad (3.1)$$

Of nearly equal importance is the definition of confidence in an association rule $\text{conf}(X \rightarrow Y)$, which is defined as the proportion of transactions containing X that also contain Y , i.e. the conditional probability of the consequent Y occurring in rules where the antecedent X has already been found. Thus confidence always falls in the interval $[0, 1]$, with a value of 1 meaning that wherever the antecedent X is found, the consequent Y is always also found.

$$\text{conf}(X \rightarrow Y) = \frac{\text{supp}(X \cup Y)}{\text{supp}(X)} \quad (3.2)$$

The lift of a rule $\text{lift}(X \rightarrow Y)$, sometimes also called its interest, divides confidence by $\text{supp}(Y)$ to measure the independence of the co-occurrence of X and Y . A lift value of 1 suggests that the antecedent and consequent are statistically independent, while $\text{lift} > 1$

indicates a positive co-occurrence between X and Y and $\text{lift} < 1$ indicates negative co-occurrence or substitution between X and Y . However, this definition gives lift the property of being completely non-directional, i.e. $\text{lift}(X \rightarrow Y) \equiv \text{lift}(Y \rightarrow X)$. Depending on the particularities of the data and the intended application of the mined rules, this may or may not be desirable.

$$\text{lift}(X \rightarrow Y) = \frac{\text{conf}(X \rightarrow Y)}{\text{supp}(Y)} = \frac{\text{supp}(X \cup Y)}{\text{supp}(X) \cdot \text{supp}(Y)} \quad (3.3)$$

In [43], the authors defined a new rule metric, conviction, to combine desirable properties of both confidence and lift. Conviction of a rule $\text{conv}(X \rightarrow Y)$ is defined in Equation 3.4 as the inverse of the lift of the opposite of the rule, $X \nrightarrow Y$, the case where X occurs and Y does not. Conviction thus has the properties of being infinite when a rule is always true (confidence of 1), incorporating both $\text{supp}(X)$ and $\text{supp}(Y)$, and having directionality (i.e. $\text{conv}(X \rightarrow Y) \neq \text{conv}(Y \rightarrow X)$).

$$\begin{aligned} \text{conv}(X \rightarrow Y) &= \text{lift}(X \nrightarrow Y)^{-1} \\ &= \frac{\text{supp}(X) \cdot \text{supp}(\neg Y)}{\text{supp}(X \wedge \neg Y)} \\ &= \frac{\text{supp}(X) \cdot (1 - \text{supp}(Y))}{\text{supp}(X) - \text{supp}(X \cup Y)} \end{aligned}$$

Applying (3.1) and (3.2):

$$\text{conv}(X \rightarrow Y) = \frac{1 - \text{supp}(Y)}{1 - \text{conf}(X \rightarrow Y)} \quad (3.4)$$

Leverage [44] is defined as the difference between the probability of finding the antecedent X and consequent Y together, and the probability with which one would expect to find them together if they were independent. Positive values thus suggest a stronger co-occurrence than would be expected by chance. By comparing to Equation 3.3, leverage can be understood simply the difference between the numerator and denominator of lift,

Table 3.4: Lists of the transactions supporting either the antecedent or the consequent of the example association rule ($\{1018\} \rightarrow \{1008\}$). For brevity, in each table only the matching item is shown.

(a) Support for $X = \{1018\}$		(b) Support for $Y = \{1008\}$	
User	Item	User	Item
10176	1018	10170	1008
10178	1018	10175	1008
10183	1018	10177	1008
10184	1018	10183	1008
10190	1018	10184	1008
		10190	1008
		10193	1008
		10194	1008

whereas lift expresses this information as a ratio. Depending on the specific application, the difference of leverage may be more useful than the unitless ratio of lift; for example, a leverage threshold can be multiplied by the size of the data set $N = |T|$ to express the threshold in terms of absolute occurrences.

$$\text{lvrg}(X \rightarrow Y) = \text{supp}(X \cup Y) - \text{supp}(X) \cdot \text{supp}(Y) \quad (3.5)$$

Many other such metrics for ranking itemset or rule saliency have been proposed, using such varied techniques as χ^2 values, the Jaccard coefficient [45], entropy [46], and mutual information [47]. Additional metrics can be defined on an ad hoc basis, to have whatever properties are applicable and valuable based on the structure of the underlying data and the intended application.

3.2.1.1 Examples

Select examples will be worked out by in detail, to illustrate the metric calculation process. For these examples, the metrics will be computed only for the subset of 28 transactions shown in Table 3.3.

Consider the association rule ($\{1018\} \rightarrow \{1008\}$). From Table 3.3, each transaction

containing the antecedent or consequent is selected and inserted into Table 3.4. Five transactions are found supporting $\{1018\}$, and eight transactions are found supporting $\{1008\}$.

$$\begin{aligned}\text{supp}(\{1018\}) &= \frac{5}{28} \approx 0.178571 \\ \text{supp}(\{1008\}) &= \frac{8}{28} \approx 0.285714\end{aligned}$$

Comparing the unique identifiers of users (i.e. transactions) contributing support counts to the antecedent and consequent in Table 3.4, there are three transactions in common. Now the support of the combined itemset can be calculated.

$$\text{supp}(\{1008, 1018\}) = \frac{3}{28} \approx 0.107143$$

With support counts for the antecedent, consequent, and their union, many more metrics can now be calculated.

$$\begin{aligned}\text{conf}(\{1018\} \rightarrow \{1008\}) &= \frac{\text{supp}(\{1008, 1018\})}{\text{supp}(\{1018\})} \\ &= \frac{0.107143}{0.178571} = 0.6\end{aligned}$$

$$\begin{aligned}\text{lift}(\{1018\} \rightarrow \{1008\}) &= \frac{\text{conf}(\{1018\} \rightarrow \{1008\})}{\text{supp}(\{1008\})} = \frac{\text{supp}(\{1008, 1018\})}{\text{supp}(\{1018\}) \cdot \text{supp}(\{1008\})} \\ &= \frac{0.107143}{0.17857 \cdot 0.285714} = 2.1\end{aligned}$$

$$\begin{aligned}\text{lvr}(\{1018\} \rightarrow \{1008\}) &= \text{supp}(\{1008, 1018\}) - \text{supp}(\{1018\}) \cdot \text{supp}(\{1008\}) \\ &= 0.107143 - 0.17857 \cdot 0.285714 \approx 0.056122\end{aligned}$$

$$\begin{aligned}\text{conv}(\{1018\} \rightarrow \{1008\}) &= \frac{1 - \text{supp}(\{1008\})}{1 - \text{conf}(\{1018\} \rightarrow \{1008\})} \\ &= \frac{1 - 0.285714}{1 - 0.6} \approx 1.785714\end{aligned}$$

Solutions for the above rule metrics, calculated for a series of arbitrarily chosen association rules over the 28 transaction sample data set, are shown in Table 3.5

3.2.2 Frequent Itemset Mining Approaches

Generally, the Association Rule Mining (ARM) process is divided into two steps. First, the data set is analyzed to discover all frequent itemsets. These frequent itemsets are used to create association rules, which can be scored and filtered based on rule metrics, as discussed above, to identify interesting rules. Frequent itemsets are defined as all those itemsets whose support exceeds some minimum threshold value.

The process of discovering the frequent itemsets is computationally intensive, and has been the subject of significant research and development. It is easy to imagine a trivial algorithm for discovering the frequent itemsets - exhaustively generating every non-empty subset s of the itemset I , i.e. $S = \{s \in \mathcal{P}(I) : |s| > 0\}$, and traversing the database once for each such subset to count its support, would be sufficient to arrive at a correct answer. However, for an itemset I of size $K = |I|$, there are $|S| = 2^K - 1$ non-empty subsets of I for which support values would need to be calculated. For itemsets and data sets of any realistic or interesting size, better algorithms are clearly needed to discover these frequent itemsets.

Many algorithms have been proposed to efficiently solve the frequent itemset mining problem. Two classes of frequent itemset mining algorithms are of particular interest: a priori-like candidate generation and prefix growth.

3.2.2.1 Apriori and Candidate Generation

Candidate generation algorithms solve the frequent itemset discovery problem by implementing a procedure to limit the candidate itemsets being generated much more than the trivial algorithm imagined above. The most famous example of such an algorithm is the Apriori algorithm [48]. Apriori is so well known that similar candidate generation and

Table 3.5: The association rule metrics defined in this Section 3.2.1, calculated for an arbitrary selection of rules ($X \rightarrow Y$) over the 28 sample transactions shown in Table 3.3.

X	Y	supp(X)	supp(Y)	supp($X \cup Y$)	conf($X \rightarrow Y$)	lift(.)	lvrg(.)	conv(.)
{1003}	{1008}	0.143	0.286	0.071	0.500	1.750	0.031	1.429
{1025}	{1026}	0.071	0.107	0.071	1.000	9.333	0.064	Infinity
{1026}	{1025}	0.107	0.071	0.071	0.667	9.333	0.064	2.786
{1008}	{1017}	0.286	0.286	0.071	0.250	0.875	-0.010	0.952
{1035}	{1017}	0.107	0.286	0.071	0.667	2.333	0.041	2.143
{1018}	{1008}	0.179	0.286	0.107	0.600	2.100	0.056	1.786
{1003}	{1008, 1018, 1035}	0.143	0.071	0.071	0.500	7.000	0.061	1.857
{1035}	{1003, 1008, 1018}	0.107	0.071	0.071	0.667	9.333	0.064	2.786
{1003, 1008}	{1018, 1035}	0.071	0.071	0.071	1.000	14.000	0.066	Infinity
{1035}	{1003, 1018}	0.107	0.107	0.071	0.667	6.222	0.060	2.679
{1003, 1018}	{1035}	0.107	0.107	0.071	0.667	6.222	0.060	2.679
{1003, 1035}	{1018}	0.071	0.179	0.071	1.000	5.600	0.059	Infinity
{1030}	{1004, 1017}	0.107	0.071	0.071	0.667	9.333	0.064	2.786
{1017, 1030}	{1004}	0.107	0.214	0.071	0.667	3.111	0.048	2.357
{1034}	{1008, 1017}	0.143	0.071	0.071	0.500	7.000	0.061	1.857
{1035}	{1008, 1018}	0.107	0.107	0.071	0.667	6.222	0.060	2.679
{1008, 1018}	{1035}	0.107	0.107	0.071	0.667	6.222	0.060	2.679
{1003}	{1008, 1035}	0.143	0.071	0.071	0.500	7.000	0.061	1.857
{1035}	{1003, 1008}	0.107	0.071	0.071	0.667	9.333	0.064	2.786
{1003}	{1008, 1018}	0.143	0.107	0.071	0.500	4.667	0.056	1.786
{1003, 1008}	{1018}	0.071	0.179	0.071	1.000	5.600	0.059	Infinity
{1003, 1018}	{1008}	0.107	0.286	0.071	0.667	2.333	0.041	2.143
{1008, 1018}	{1003}	0.107	0.143	0.071	0.667	4.667	0.056	2.571

pruning algorithms are referred to generally in the literature as “a priori-like” [49, 50, 35, 51], each being named for the a priori support counts from previous iterations used to prune the candidate search space.

The Apriori algorithm is a breadth-first search that makes use of the fact that every subset of a frequent itemset must itself be a frequent itemset (the “downward closure property” of support) to prune the itemset search space. Since it is a breadth-first algorithm, it begins with finding frequent itemsets of size 1, i.e. 1-itemsets, then finds frequent 2-itemsets, and 3-itemsets, and so on. As such, when candidate k -itemsets are generated they can immediately be compared to the list of frequent $(k - 1)$ -itemsets. For each candidate k -itemset, every subset of size $(k - 1)$ must itself be a frequent $(k - 1)$ -itemset in order to satisfy the downward closure property. If any $(k - 1)$ subsets are not frequent $(k - 1)$ -itemsets, then that candidate k -itemset can be pruned, and there is no need to traverse the database and count its support. This has the effect of drastically reducing the size of the frequent itemset search space, significantly improving processing speed.

The Apriori algorithm for finding frequent itemsets is outlined in Algorithm 3.1.

3.2.2.2 *Pattern Growth, FP-growth*

Other approaches to the frequent itemset mining problem seek to avoid the candidate generation and maintenance process entirely. Typical of these approaches is the FP-growth [50] algorithm.

FP-growth avoids the need to generate candidate sequences by storing support counts in a data structure designed such that the frequent patterns “grow” in-place. This is accomplished by inserting the support counts into a trie, also known as a prefix tree, a tree-like data structure useful for key-value storage in applications where the key is itself a sequence or collection. Each element of the key corresponds to one node on the trie, and it is the position of the node within the tree, rather than the contents of the node itself, that encodes the complete key value. To optimize the compression of the data structure and

Algorithm 3.1 Apriori Algorithm [48]

```
1: procedure APRIORI( $T, I, m$ )
2:    $L_1 \leftarrow \{\{i\} : i \in I \wedge |\{t \in T : i \subseteq t\}|/|T| > m\}$  ▷ support threshold  $m$ 
3:    $k \leftarrow 2$ 
4:   while  $|L_{k-1}| > 0$  do
5:      $C_k \leftarrow \text{APRIORIGEN}(L_{k-1})$ 
6:     for all  $t \in T$  do
7:        $c_t \leftarrow \{c \in C_k : c \subseteq t\}$ 
8:       for all  $c \in c_t$  do
9:          $\text{count}[c]++$ 
10:     $L_k \leftarrow \{c \in C_k : \text{count}[c]/|T| > m\}$ 
11:     $k++$ 
12:   return  $\bigcup_k L_k$ 
13: procedure APRIORIGEN( $L_{k-1}$ )
14:   for all  $(p, q) \in L_{k-1} \times L_{k-1}$  do ▷ “join” step
15:     if  $p_1 = q_1 \wedge p_2 = q_2 \wedge \dots \wedge p_{k-2} = q_{k-2} \wedge p_{k-1} < q_{k-1}$  then
16:       append  $\{p_1, p_2, \dots, p_{k-1}, q_{k-1}\}$  to  $C_k$ 
17:   for all  $c \in C_k$  do ▷ “prune” step
18:     for all  $\{s \in \mathcal{P}(c) : |s| = k - 1\}$  do
19:       if  $s \notin L_{k-1}$  then
20:         delete  $s$  from  $C_k$ 
21:   return  $C_k$ 
```

avoid equivalent itemsets, the items within transactions are first be sorted in descending order of frequency. Subsets of resulting trie are then recursively searched in a bottom-up manner, yielding a complete list of frequent itemsets.

3.2.3 Generating Association Rules

Association rules can now be generated from the set of frequent itemsets \mathcal{F} . Each frequent itemset is divided into all possible pairs of mutually exclusive, non-empty subsets. Each pair of such subsets is then treated as a candidate rule.

$$\text{rules}(\mathcal{F}) = \bigcup_{f \in \mathcal{F}} \left\{ (X \rightarrow f \setminus X) \mid X \in \mathcal{P}(f) \text{ and } |X| > 0 \text{ and } |f \setminus X| > 0 \right\} \quad (3.6)$$

These candidate rules can then be scored, filtered, and sorted according to the rule metrics outlined in Section 3.2.1.

3.2.4 Case Study: ARM with Apriori Algorithm

The Apriori algorithm is one of the most common and important algorithms used in association rule mining [23]. In this section, the Apriori algorithm is implemented in the Scala programming language. Scala is a JVM-based language supporting both object-oriented and functional programming paradigms, which has found use in the field of big data analytics. This code additionally supports the the scoring of association rules derived from Apriori's mined frequent itemsets. Rule mining results are shown after running on the training fold of the MSWeb [39] data, and the application of the rules to the testing data will be discussed. A simplified and truncated version of the source code is shown in Figure 3.1; the complete working source code for this implementation, including the functionality to download and parse the MSWeb data set, is published along with this thesis.

Table 3.6: The Apriori code was used to sweep a range of minimum support thresholds, and the number of frequent n -itemsets found for each threshold are reported, up to a maximum size of $n = 7$.

Support	Frequent n -Itemsets for $n =$						
	1	2	3	4	5	6	7
10^{-1}	7	2	0	0	0	0	0
10^{-2}	47	80	64	6	0	0	0
10^{-3}	162	924	1612	1278	433	97	17

3.2.4.1 Implementation of the Apriori Algorithm

The Apriori algorithm is implemented in Scala, and shown in Figure 3.1. Here, a class called Apriori defines an object which is initialized with the transaction data set. This object presents public methods for the mining of frequent itemsets and association rules. Note that while the ARM problem formulation does not require transactions to be ordered, or that they even be sortable, in this implementation sorted sets are used to facilitate the candidate generation process. If the items in the itemset I did not have comparison operations defined, simply mapping the transaction items to the integer indices of a list representation of I would allow compatibility.

3.2.4.2 Rule Mining Results

The frequent itemset and association rule mining code was run on the training fold of the MSWeb data, and the sizes of the result sets are shown in Table 3.6 and Table 3.7. Thresholds for rules of meeting a minimum support of 10^{-2} and a minimum confidence of 0.5 are chosen based on visual inspection, and the fifteen resulting rules with highest lift are shown in Table 3.8.

3.3 Sequential Rule Mining

Sequential Rule Mining (SRM) is an extension of ARM’s frequent itemset mining that incorporates additional ordering information, such as a time-stamped transactions, into the


```

import scala.collection.immutable.SortedSet
class Apriori[I](data: Seq[SortedSet[I]], itemset: Set[I])(implicit ordering: Ordering[I]) {
  import ordering.mkOrderingOps
  type Itemsets = Seq[SortedSet[I]]
  case class Rule(X: Set[I], Y: Set[I], score: Map[String, Double])
  private val N = data.size.toLong
  private def suppN(s: Set[I]): Long = data.map(t => s.subsetOf(t)).count(identity)

  def find_frequent_itemsets(minsupport: Double, maxlength: Int = Int.MaxValue): Itemsets = {
    val m = Math.ceil(minsupport * data.size).toLong
    val R = collection.mutable.Map.empty[Int, Itemsets]
    R += 1 -> initialCandidates(m)
    var k = 2
    while (R(k-1).nonEmpty && k<=maxlength) {
      val C = generate_candidates(R(k-1), k)
      R += k -> C.map(c=>c->suppN(c)).filter(_._2 > m).map(_._1).seq
      k += 1
    }
    R.values.flatten.toSeq
  }
  private def initialCandidates(m: Long): Itemsets = {
    data.flatten.groupBy(identity).mapValues(_._2.size).filter(_._2>=m).keys.toSeq.map(k=>SortedSet(k))
  }
  private def generate_candidates(L: Itemsets, k: Int): Itemsets = {
    val C = collection.mutable.HashSet.empty[SortedSet[I]]
    for (p <- L) {
      for (q <- L if p.last < q.last && q.dropRight(1).subsetOf(p)) {
        val c = p + q.last
        if (c.subsets(k-1).map(L.contains).reduce(_&_)) {
          C += c
        }
      }
    }
    C.toSeq
  }

  def find_association_rules(minsupport: Double, minconfidence: Double): Seq[Rule] = {
    val L = find_frequent_itemsets(minsupport)
    val rules_all = for ((a,b) <- L.flatMap(rule_splitter).distinct) yield Rule(a,b,score_rule(a,b))
    val rules = rules_all.filter(_._3.score("confidence")>=minconfidence)
    rules.seq
  }
  private def score_rule(x: Set[I], y: Set[I]): Map[String, Double] = {
    val suppx = 1.0 * suppN(x) / N
    val suppy = 1.0 * suppN(y) / N
    val suppxy = 1.0 * suppN(x union y) / N
    val confxy = suppxy / suppx
    Map(
      "support_antecedent" -> suppx,
      "support_consequent" -> suppy,
      "support" -> suppxy,
      "confidence" -> confxy,
      "lift" -> (suppxy / (suppx * suppy)),
      "leverage" -> (suppxy - (suppx * suppy)),
      "conviction" -> (1.0 - suppy)/(1.0 - confxy)
    )
  }
  private def rule_splitter(l: Set[I]): Iterator[(Set[I], Set[I])] =
    l.subsets.filter(_._1.nonEmpty).map(s=>s->l.diff(s)).filter(_._2.nonEmpty)
}

```

Figure 3.1: Simplified implementation of the Apriori frequent itemset mining algorithm in Scala, with association rule mining and scoring capability.

Table 3.7: The number of association rules found satisfying the minimum support threshold, as well as meeting or exceeding the additional metric constraints (e.g. minimum confidence) shown in the top row.

Support	Conf \geq		Lift \geq	
	0	0.5	5	10
10^{-1}	2	1	0	0
10^{-2}	628	91	154	24
10^{-3}	49992	5893	24282	13002

KDD process. For datasets containing ordered sequences of items, or time-stamped events, Sequential Rule Mining (SRM) can provide even more valuable insights into the structure and trends of the data. These insights are expressed as frequent sequential patterns S , or sequential rules relating an antecedent A to a consequent B of the form $(A \Rightarrow B)$.

3.3.1 Comparison To ARM

While much of the problem definition and approach for SRM is similar to that in ARM, there are a few key differences. In SRM sequences, transactions, and the antecedents / consequents of rules are composed of ordered sequences of elements or events, with each element or event being a set of items from the itemset I .

$$S = (s_p) = (s_1, s_2, \dots, s_P), \text{ where } s_p \subseteq I$$

$$p \in \mathbb{N}_{\leq |S|}^*, \quad |S| = P$$

Similarly:

$$A = (a_n), a_n \subseteq I, \quad B = (b_n), b_n \subseteq I$$

Table 3.8: Rules of meeting a minimum support of 10^{-2} and a minimum confidence of 0.5 are chosen based on visual inspection, and the fifteen resulting rules with highest lift are shown.

X	Y	$\text{supp}(X)$	$\text{supp}(Y)$	$\text{supp}(X \cup Y)$	$\text{conf}(X \rightarrow Y)$	$\text{lift}(.)$	$\text{lvr}(.)$	$\text{conv}(.)$
{1014}	{1000}	0.022	0.028	0.011	0.512	18.377	0.011	1.994
{1008,1035}	{1009,1018}	0.027	0.045	0.020	0.748	16.629	0.019	3.794
{1001,1009,1018}	{1035}	0.016	0.055	0.011	0.682	12.463	0.010	2.976
{1008,1009,1018}	{1035}	0.030	0.055	0.020	0.673	12.287	0.019	2.888
{1035}	{1009,1018}	0.055	0.045	0.029	0.525	11.663	0.026	2.010
{1009,1018}	{1035}	0.045	0.055	0.029	0.639	11.663	0.026	2.615
{1014}	{1040}	0.022	0.046	0.012	0.521	11.308	0.011	1.990
{1001,1035}	{1003,1018}	0.030	0.047	0.015	0.519	11.059	0.014	1.980
{1003,1035}	{1001,1018}	0.024	0.059	0.015	0.635	10.692	0.014	2.580
{1003,1009}	{1035}	0.020	0.055	0.010	0.508	9.272	0.009	1.920
{1038,1041}	{1026}	0.012	0.098	0.011	0.896	9.103	0.010	8.672
{1009,1035}	{1008,1018}	0.033	0.073	0.020	0.620	8.487	0.018	2.440
{1008,1038}	{1026}	0.015	0.098	0.012	0.825	8.385	0.011	5.163
{1038}	{1026}	0.034	0.098	0.027	0.805	8.173	0.024	4.612
{1001,1018,1035}	{1003}	0.024	0.091	0.015	0.633	6.977	0.013	2.478

$$T = (t_n) = \left(\begin{array}{c} t_1 = (t_{1,m}) = (t_{1,1}, t_{1,2}, t_{1,3}, \dots, t_{1,M_1}), \\ t_2 = (t_{2,m}) = (t_{2,1}, t_{2,2}, t_{2,3}, \dots, t_{2,M_2}), \\ \dots \\ t_N = (t_{N,m}) = (t_{N,1}, t_{N,2}, t_{N,3}, \dots, t_{N,M_N}) \end{array} \right), \quad t_{n,m} \subseteq I$$

This illustrates two key differences between frequent itemset mining and Sequential Pattern Mining (SPM) - that the elements s_p of the sequence S are ordered, and that each element is not necessarily a single item, but is itself a set of items. The same holds true in transaction databases $T = (t_n)$ suited to SPM/SRM. This is of particular importance in temporally-ordered SPM applications, where such elements are the representation of multiple simultaneous items in one event.

Some other elements of the problem definition must be redefined, augmenting set-based concepts applicable in ARM with sequence concepts, reflecting the ordered nature of the transactions. A sequence $S = (s_p)$ is called a k -sequence when it contains k total items, i.e. $k = \sum_p |s_p|$. To accomodate the ordered structure, support is redefined in Equation 3.8 such that for the sequence S , a transaction t_n supports S only if S is equal to or is a subsequence of t_n . The definition of a subsequence is defined here in an application-specific manner, reflecting that these sequences are sequences of sets, using the operator $A \triangleleft B$ to mean that A is a subsequence of B, and $A \trianglelefteq B$ to mean that A is a subsequence of or equal to B. A subsequence is defined such that for every a_p in A, the sequence B can be divided at an element b_q such that $a_p \subseteq b_q$, all elements preceding a_p are subsets of elements preceding b_q , and all elements following a_p are subsets of elements following b_q .

$$A \trianglelefteq B \triangleq \forall p \in \mathbb{N}_{\leq |A|}^*, \exists q \in \mathbb{N}_{\leq |B|}^* \left\{ \begin{array}{l} a_p \subseteq b_q \\ \wedge \forall x \in \mathbb{N}_{< p}^*, \exists y \in \mathbb{N}_{< q}^* (a_x \subseteq b_y) \\ \wedge \forall x \in \mathbb{N}_{> p}^*, \exists y \in \mathbb{N}_{> q}^* (a_x \subseteq b_y) \end{array} \right. \quad (3.7)$$

Applying Equation 3.7 to the concept of support:

$$\text{supp}_{SPM}(S) = \frac{1}{N} \sum_{n=1}^N \begin{cases} 1, & S \leq t_n \\ 0, & \text{otherwise} \end{cases} \quad (3.8)$$

In ARM, the mining of frequent itemsets is only the first half of the problem - it is most often followed by the processing of these itemsets into association rules, as shown in Equation 3.6. And this can be the case in SRM as well, as shown in Section 3.3.3. However, some SPM applications omit this step, using the frequent sequences themselves, thresholded to some minimum level of support, as the key output.

3.3.2 Sequential Pattern Mining Approaches

Frequent sequences then need to be found, analogous to the mining of frequent itemsets in ARM. Similarly to ARM frequent itemset mining, SPM frequent sequence mining can be divided into approaches that do and do not use candidate generation and maintenance procedures.

3.3.2.1 A Priori-like Methods

The Generalized Sequential Pattern (GSP) algorithm [52] is an a priori-like algorithm adapted to the mining of sequential patterns in ordered transaction data. The main procedure is very similar to that of Apriori, with the largest differences being in the candidate generation logic. While Apriori's itemset "join" method was able to simply append additional items to the candidate itemsets, a SPM solution must take the ordering into account, and also be able to generate separate candidates both by adding items to the final element of a sequence and by adding a new single-item element to the sequence.

To achieve this, the authors introduce the concept of contiguous subsequences. Given a sequence of sets of items, contiguous subsequences can be generated by (1) dropping items

from the first or last elements; (2) dropping items from interior elements when those elements have two or more item; and (3) recursively generating the contiguous subsequences of the contiguous subsequences of the original sequence. This definition has the property of making Apriori's downward closure property of support applicable in SPM context; the contiguous subsequences of a frequent sequence will all be frequent sequences themselves. This enables a priori-like candidate generation.

Other a priori-like approaches, such as AprioriAll [53], employ a variety of technical details to more efficiently generate and prune their search spaces. However, a priori-like methods have been broadly overshadowed by much more performant candidate-less approaches [54].

3.3.2.2 *Pattern Growth*

As with Apriori in ARM, a priori-like methods in SPM suffer from low performance [55]. To overcome this, techniques for frequent sequence mining with candidate generation are a subject of significant research interest [54, 56, 52, 55, 57].

Pattern growth is an offshoot of the concepts introduced by the FP-Growth algorithm [50], adapted for sequential pattern applications. Rather than repeatedly generating large candidate sets and scanning the entire transaction database to count their support, the database is projected onto patterns already found, creating smaller sub-databases which can be searched for locally frequent sequences and then recursively projected onto those sequences. In addition to the efficient design philosophy that this technique shares with ARM algorithms like FP-Growth, its strategy of recursively partitioning the the search space is readily parallelizable, and thus well-suited to the distributed computing environments characteristic of big data analysis [58].

FreeSpan [59] is a notable algorithm for mining frequent sequences by pattern growth. For a given sequence $S = (s_p)$, the projected itemset of the sequence is defined as $\bigcup_p s_p$. First, the set of all frequent items in the database L_1 is found. The set of all frequent patterns

in the database can then be divided into $|L_1|$ disjoint subsets: those containing only the first element of L_1 , those containing only the first and second elements of L_1 , and so on. For each of these cases, a projected database, removing infrequent items and containing only the specified elements of L_1 , is created. Each of these projected sub-databases can then be scanned for frequent sequential patterns of length-2, and a new sub-database containing only the transactions matching each found frequent 2-pattern can be recursively projected, analyzed, and so on, until the complete set of all frequent sequences has been discovered.

PrefixSpan [57], an extension of FreeSpan, is one of the most prevalent algorithms for mining frequent sequential patterns by way of pattern growth. PrefixSpan improves on FreeSpan in several ways. First, an ordering is imputed to the itemset I such that the items within each element of a transaction can be sorted into a consistent order. This allows the definition of the sequence prefix:

Sequence $A = (a_1, a_2, \dots, a_M)$ is a prefix of sequence $B = (b_1, b_2, \dots, b_N)$ only if all of the following are true:

- (a) $M \leq N$
- (b) $\forall i \in \mathbb{N}_{<M}^* a_i = b_i$
- (c) $a_M \subseteq b_M$, and
- (d) all items in $b_M \setminus a_M$ follow in order all items in a_M

Sequence C is then the suffix of the same sequence B with respect to prefix A :

$$C = (b_M \setminus a_M, b_{M+1}, \dots, b_N)$$

The projection procedure of FreeSpan can then be modified such that only the sequences of which a frequent pattern is a prefix are projected into that pattern's projected sub-database. This ensures that the projected databases shrink at every iteration of the

recursive pattern search, reducing the computational intensiveness of the SPM task.

3.3.2.3 Other Useful Algorithms and Concepts

Many other algorithms and techniques for SPM, having various advantages and disadvantages. The popular SPADE [55] algorithm has the useful property of being able to find all frequent sequences with only three scans of the database, with execution time that grows linearly with the database size. BIDE [60, 61], another popular candidate-less SPM procedure, has the useful property of finding only closed frequent sequences \mathcal{C} - those frequent sequences that are not subsequences of another frequent sequence with the same level of support. This reduces the size of the frequent sequence list returned, without any loss of specificity. If, depending on the application, it is more desirable to have the complete list of frequent sequences \mathcal{G} , these can be easily generated from the closed frequent sequence list by including the subsequences of the closed sequences.

$$\mathcal{G} = \text{open}(\mathcal{C}) = \bigcup_{c \in \mathcal{C}} \{s : s \trianglelefteq c\} \quad (3.9)$$

3.3.3 Generating and Scoring Sequential Rules

Analogous to the ARM problem, sequential rules must now be mined from the frequent sequential patterns \mathcal{G} found using the algorithms from Section 3.3.2 using Equation 3.10.

$$\text{rules}(\mathcal{G}) = \bigcup_{g \in \mathcal{G}} \left\{ \left((g_1, \dots, g_n) \Rightarrow (g_{n+1}, \dots, g_{|g|}) \right) : n \in \mathbb{N} \wedge 1 \leq n < |g| \right\} \quad (3.10)$$

While the problem formulations of ARM and SPM are well settled and accepted, there are several slightly different approaches for how frequent sequences can be processed into sequential rules, and how the resulting rules should be assessed and measured. For instance, some algorithms require that for the sequential rule $(A \Rightarrow B)$ to be supported by transaction t , all of the elements of the antecedent A and consequent B must occur in the t in the same

order as in A and B [56]. Other approaches apply looser constraints, requiring only that all items in A must occur in t before the items of B [54]. In some other constructions, sequential rules are represented in the form $(\alpha \Rightarrow \beta)$, where $\alpha \triangleleft \beta$, $\alpha = A$ and $\beta = A \# B$ [55]. To enable this definition, the operator $\#$ is used to denote the concatenation of sequences.

$$\begin{aligned} A \# B &\triangleq (a_1, a_2, \dots, a_{|A|}, b_1, b_2, \dots, b_{|B|}) \\ A &= (a_n), \ a_n \subseteq I, \quad B = (b_n), \ b_n \subseteq I \\ |A \# B| &= |A| + |B| \end{aligned}$$

For the purposes of this section and later work, the stricter of these possible definitions of support is adopted, with A and B in rules $(A \Rightarrow B)$ representing distinct antecedents and consequents.

$$\text{supp}_{SRM}(A \Rightarrow B) = \text{supp}_{SPM}(A \# B) \quad (3.11)$$

With these foundational definitions established, it is now possible to generalize association rule metrics to the domain of sequential rule analysis. A selection of the metrics introduced in Section 3.2.1 are shown here, adapted to the SRM problem.

$$\text{conf}_{SRM}(A \Rightarrow B) = \frac{\text{supp}_{SRM}(A \Rightarrow B)}{\text{supp}_{SPM}(A)} = \frac{\text{supp}_{SPM}(A \# B)}{\text{supp}_{SPM}(A)} \quad (3.12)$$

$$\text{lift}_{SRM}(A \Rightarrow B) = \frac{\text{conf}_{SRM}(A \Rightarrow B)}{\text{supp}_{SPM}(Y)} = \frac{\text{supp}_{SPM}(A \# B)}{\text{supp}_{SPM}(X) \cdot \text{supp}_{SPM}(Y)} \quad (3.13)$$

$$\text{lvr}_{SRM}(A \Rightarrow B) = \text{supp}_{SPM}(A \# B) - \text{supp}_{SPM}(X) \cdot \text{supp}_{SPM}(Y) \quad (3.14)$$

3.4 Application to Biomedical Informatics

Association Rule Mining (ARM), Sequential Pattern Mining (SPM), and Sequential Rule Mining (SRM) each have many important applications in the biomedical informatics domain. Transactional data sets, of the kind these algorithms are suited to analyze, occur in throughout the healthcare domain. With diagnoses as the items, association rules can be used to predict the likelihood of encountering common co-morbidities. ARM also finds use in the operational, business-oriented areas of healthcare. When mined on itemsets of healthcare resources utilized, association rules and sequential patterns can be used to track and control spending and to examine census patterns (the movement of patients between units of a hospital). More broadly, association rules can provide benefits as recommender systems for expanding on common relationships.

Given that many important bioinformatics problems occur in the form of sequences, such as DNA and protein sequences, the applications of SPM are clear. In the healthcare domain, even the earliest medical record systems recognized the importance of being able to track a patient's condition over time [8, 9], due to the inherently temporally ordered nature of health events.

In this work, sequence mining techniques are applied in the domain of public health informatics, to discover patterns in the reported sequences of causes of death gathered in mortality reporting [62].

CHAPTER 4

IMPROVING VALIDITY OF CAUSE OF DEATH ON DEATH CERTIFICATES

4.1 Preface

Work in this chapter is adapted from the publication "Improving Validity of Cause of Death on Death Certificates" [63], of which Ryan Hoffman is the first and leading author. This publication is copyrighted by the Association for Computing Machinery (ACM), and reused with permission. For detailed copyright disclosure, please see Appendix F. It has been modified throughout to comply with formatting requirements and to better integrate with the rest of this work.

R. A. Hoffman, J. Venugopalan, L. Qu, H. Wu, and M. D. Wang, "Improving validity of cause of death on death certificates," in *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, ACM, 2018, pp. 178–183

Proceedings, Association for Computing Machinery by Association for Computing Machinery. Reproduced with permission of Association for Computing Machinery in the format Republish in a thesis/dissertation via Copyright Clearance Center.

4.2 Abstract

Accurate reporting of causes of death on death certificates is essential to formulate appropriate disease control, prevention and emergency response by national health-protection institutions such as Center for disease prevention and control (CDC). In this study, we utilize knowledge from publicly available expert-formulated rules for the cause of death to determine the extent of discordance in the death certificates in national mortality data with the expert knowledge base. We also report the most commonly occurring invalid causal

pairs which physicians put in the death certificates. We use sequence rule mining to find patterns that are most frequent on death certificates and compare them with the rules from the expert knowledge based. Based on our results, 20.1% of the common patterns derived from entries into death certificates were discordant. The most probable causes of these discordance or invalid rules are missing steps and non-specific ICD-10 codes on the death certificates.

4.3 Introduction

Approximately 2.6 million deaths occur each year in the United States (US) and 56 million deaths occur per year worldwide [1, 64]. Accurate death statistics are imperative to help the national health protection institutions such as the National Center for Health Statistics prevent epidemics and disease outbreak, formulate a response to communicable diseases and evaluate statistics such as the birth and death trends. The data from death certificates is also used for the estimation of the trends in chronic conditions such as the prevalence of diabetes and cardiovascular conditions. Causes of death from death records are often used by the reporting agencies to carry out the tasks mentioned above. To ensure a timely and accurate response to disease threats, high-quality death information on death certificates is essential.

The World Health Organization (WHO) has classified the Cause(s) of Death (COD) using the International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10) which contains 22 chapters covering 2,046 categories of diseases [65, 66]. Despite the pressing need for high quality cause of death information, challenges such as lack of adequate knowledge and practice still exist for the accurate filling of death certificates. These challenges lead to death certificates of uncertain quality. Studies have found disagreements between the death certificate COD and the sequence of events reported in the medical record [67, 68]. The reporting of COD with a higher accuracy is difficult to implement in some situations, i.e. when the certifying physician is not the primary care

provider, when the primary care provider may be part of a separate healthcare organization or located in a different area, when certificate completion is significantly delayed after the death, or when the certifying physician does not have ready access to relevant medical records. These situations are part of the reality of medical practice, where the clinicians providing care the time of a patient's death may not have a comprehensive knowledge of the patient's medical history. In fact, the errors in COD reports results from improperly maintained and improved data quality within both systems in state and local hospitals. One study found that only 56.9% of attending physicians, 56.0% of resident physicians, and 55.7% of medical students matched experts for the correct cause of death in clinical case studies [69]. Another study found that 45% of resident respondents incorrectly identified a cardiovascular event as the primary cause of death [70]. The Framingham Heart Study and other studies have indicated that coronary artery disease is overestimated on death certificates as a cause of death in the general population by 24% and two times more in older patients than in younger ones [71]. The extent of such overestimation affects the quality of death certificate information and national mortality statistics is not known.

The information on COD constitutes the basis of health systems in hospitals and is essential for studying the relationships among diseases. The inaccurate reporting of these data could lead to inappropriate public health interventions [72]. Moreover, discrepancies in vital statistics documents pose a challenge in implementing effective public health interventions and accurate reporting of morbidity and mortality information [73, 74]. For example, Lauren E. Johns et. al. have shown that inaccurate COD reports on health disparity have a large impact on New York City premature cardiovascular mortality [74]. Accurate scientific results have a strong need for improving the accuracy and validity of mortality statistics [75, 76]. In addition, accurate COD reports are also important in clinical trials and studies. Besides, COD reports are the basis for the National Center for Health Statistics (NCHS) to help with surveillance of disease and proper allocation of funds for public health programs and research, and to help prioritize governmental decisions and actions

regarding health care. Because health statistics, national mortality and morbidity statistics, and data on disease prevalence in society are largely derived from COD reports, the accuracy of COD reports is essential.

Although ICD provides detailed and specific rules, COD reports still have many issues. Some studies have focused on the coding problems that are related to particular diseases in specific countries, such as hypertension [77], myocardial infarction [78]. Some studies have come up with the incompleteness and inaccuracy in cause-of-death statement [79]. Several researchers have studied patient medical charts for errors in cause-of-death reporting, and several have demonstrated inaccurate COD reporting among residents [70]. A few studies have discussed the types of coding errors and the reasons for them [80]. No study has provided effective solutions to the errors in the COD reports. Methods to remedy errors in COD reports are urgently needed. To increase the accuracy of COD reports, physician training in death certificate completion is found effective [79]. The NYC Office of Vital Statistics is solving this issue, especially on heart disease death reporting. An intervention was implemented within eight NYC hospitals from 2009 to 2010. The proportion of heart disease deaths reported at the intervention hospitals decreased 50% to the level of those at the nonintervention hospitals after the intervention [81, 82]. Many low-cost or free methods are also used, including an e-learning module created by New York City DOHMH in cooperation with the National Association for Public Health Statistics and Information Systems, and cause-of-death documentation handbooks provided by the CDC [83, 84]. However, these methods can only reduce the possibility of errors when recording COD reports, but not avoid or find the errors in the COD reports after recording the COD reports.

There is a gap in research about the efficient determination of the true match among several potential matches for COD reports [85]. This process is extremely complex and time intensive. Luciana K.T. et.al. identified accurate deaths focusing on only anaphylaxis-caused deaths in Brazil [86]. Perviz A. M. et.al. revealed that only 10–15% of assigned underlying causes in data from England are not true pathophysiological causes of death.

Although this research showed that the COD reports have some inaccurate records, they focus on only one disease in a specific country or area. Another research study focused on accurate cause classification for only stillbirths and neonatal deaths [87]. There are also studies providing new methods for reclassification of the underlying cause of death using all the COD codes including the errors [88].

In order to document acceptable causal relationships to be used in automated and manual mortality coding, experts NCHS have published a comprehensive list of acceptable sequence codes for the cause of death [89]. The list consists of valid causes of death and the list of ICD-10 codes which can cause the COD code. Accuracies on the death certificates can be improved by multiple ways, one of which is to capture concordance with this expert pool of knowledge. However, studies have not actively used this resource to identify the discordances in codes put on death certificates, their sources and help improve the current clinical practice regarding the filling of death certificates.

In this paper, we utilized this publicly available resource of expert knowledge to determine the discordance in death certificates in national mortality repositories (National Vital Statistics System (NVSS) death certificate data). To determine frequently filled patterns from NVSS data, we used sequential rule mining. Then we compared the rules obtained from the sequential rule mining with those obtained from the expert knowledge base. The goal is to develop processes for eliminating the wrong and inaccurate COD records. In addition, our methods will help researchers better understand the relationships among diseases as well as the right intervention of public health.

We structure the rest of the paper as follows: we first describe the data sources and modeling in Section 4.4, followed by the results and discussion in Section 4.5. Finally, we conclude with the conclusions, limitations, and potential for future work in Section 4.6.

4.4 Experimental and Computational Details

As mentioned above, this study seeks to find the extent of discordances in the death certificates in national mortality database with an expert pool of knowledge, the most common sources of differences, and the sources of these discordances. We used the death certificates from those made public by the NVSS.

4.4.1 Data Sources: Expert Knowledge Base

Medical experts working on the Cause(s) of Death (COD) have published a comprehensive list of acceptable sequence codes for the cause of death, which integrates NCHS's guidance as well as other international sources of data [90]. This data consists of a valid cause of death relationships between ICD-10 codes. Each chain given consists of an address (COD) ($F3$) followed by one or two sub-addresses ($F2 : F1$). The relationships are such that each address is caused by the following sub-address, i.e. $F2 \rightarrow F3$. In case of more than one sub-address, the address is caused by the following sub-address and all the sub-addresses which fall in between the two ($F1 : F2 \rightarrow F3$).

In this dataset, some of the relationships that are marked ambivalent with ambivalent codes are defined. Relationships which are marked ambivalent indicate that further clarifications may be needed in the future. In this analysis, we utilized all the relationships, including the ones marked as ambivalent.

4.4.2 Data Sources: Death Certificate Data

National Vital Statistics System (NVSS), coordinated by the National Center for Health Statistics aggregates the causes of death for all deaths occurring within the United States from 1959 to 2014 [64]. For this analysis, we used the mortality data from 2012, which contains 2,547,864 deaths. Each death certificate format in vital statistics offices of each state, the District of Columbia, and other special jurisdictions vary but generally consists

of the underlying cause of death as recorded by physicians and other details such as the demographics, comorbid conditions, race and ethnicity. The cause of death on the death certificates was recorded as entity access codes and record access codes (up to 20 conditions). The entity axis codes refer to raw data put on death certificates and the record access axis codes refer to the codes cleaned by the NVSS. The entity access codes for this data contains two parts, with part one containing the ordered set of COD (sequences) codes, and part two containing additional related COD codes, which are unordered. Since the goal of this study is to find the discrepancies in the sequence of COD on the actual death certificates (not cleaned or filtered codes), we used the part one of the entity access codes for this analysis. Using the COD information, we extracted rules indicative of the most frequently used sequences on COD using sequential rule mining.

4.4.3 Deriving Frequent COD Patterns from Death Certificates

In this analysis, we derive the most frequently used patterns from death certificates using sequence rule mining (SRM). Sequence rule mining, sequence pattern mining [91, 92, 93, 94, 49] or association rule mining [95, 96] are the most commonly used temporal models in literature for finding temporal relationships among sequences. SRM has diverse pattern mining applications in finance and market analysis [97, 98], travel analysis [99], mobile learning [100] and database projections. (PrefixSpan [57], MEMISP [101]). In healthcare, SRM has applications in multi-dimensional EEG analysis [102], administrative data analysis [103], heart disease prediction [95, 96], healthcare auditing [104], and neurological diagnosis [105]. It was first introduced by Agrawal et al. to extract regularities between products in large-scale warehouse databases [106]. Ordonez et al. adopted SRM in medical data and proposed an improved algorithm to constrain rules so as to speed up the mining process [96]. In this analysis, SRM was chosen as the method for analysis as opposed to direct comparison of the sequences on the death records, since SRM discovers the sequences which are more commonly used on the death certificates. This helps the clinicians and the

national health institutions find top sequences where discordances occur with the expert knowledge base described above. This allows for targeted interventions for clinician training and clinical decision support systems. SRM is used to discover all temporal sequences frequently found in the dataset. The rules are determined useful if there exists a minimum presence in the dataset. The rules are included for analysis if they have a minimum support. Support of a rule is defined as the proportion of sequences in the data that exhibit the pattern [52]. In COD data mining, a rule of the form “ $X \rightarrow Y$, with a support B ” can be interpreted as follows. If the sequence has COD X , there is $B\%$ possibility of it being followed by COD Y . In the current NVSS dataset, which does not have temporal relationships, the pattern mining is performed on the sequences of COD in the death certificates.

The training dataset in the death certificates consists of a list of causes of death $C = [C_1, C_2, \dots, C_K]$. Using this training dataset, we discover a set of N rules, $S = [R_1, R_2, \dots, R_N]$. Each rule R in the set of rule S is given by $R = \langle r_1, r_2, \dots, r_T \rangle$, such that r_1, r_2, \dots, r_N is the sequence of COD in the rule R . They are sequentially ordered to reflect the relationship of $r_1 \rightarrow r_2 \rightarrow r_3 \dots \rightarrow r_T$ (T is the number of COD in the sequence). The support of a rule R in the set of sequences S is defined as the number of sequences that contain this rule. The support value in this analysis is used as a metric to pick the valid rules, which have a value larger than a minimum support. In our experiments, we use the BIDE algorithm, short for BI-Directional-Extension-based frequent closed sequence mining, proposed by Wang et al. [60]. The BIDE algorithm was selected because it is an implementation of frequent closed sequence mining which emphasizes scalability and real-world performance.

4.4.4 Deriving Rules from Expert Knowledge

In the expert-derived relationship data we mentioned above, a total of 10,849 unique ICD-10 codes were found. In addition, we also have a list of ICD-10 codes, which were marked as not valid in the US for mortality reporting (1,301 out of 10,849). In this analysis, after

consultation with experts from the Center for Disease Control and Prevention, we used only the codes which were valid in the US. Following that consultation, we excluded from analysis death records containing external (V-Y) or nature of injury (S-T) codes, as being outside of the scope for physician mortality certifiers.

4.5 Results and Discussion

4.5.1 Results from SRM Analysis

We applied BIDE to the cause of death data to part 1 of the entity access codes from 2012 death certificates.

After performing the filtering procedure in Section 4.4.4, 224,608 death records were omitted from the analysis. From the remaining records, using a minimum support of 50 occurrences, we extracted 11,815 sequential rules, together accounting for 4,010,150 causal relationships between codes. Of these rules, 61 had length 4 and 3,386 had length 3, with the remaining 8,368 being length 2, seen in Table 4.1.

4.5.2 Results from Comparing Rules from SRM with Expert Knowledge

For each rule of length greater than two found using SRM, we checked for the validity using the relationship dataset mentioned above. Rules were mapped as invalid if the address and sub-address from SRM were not found in the relationship data set. For multi-part rules, if one part of the SRM rule did not conform, the entire rule was marked invalid, as seen in Figure 4.1.

Of the total 11,815 rules, 2,378 (20.1%) of the rules were marked as invalid. Based on

Table 4.1: Size of Frequent Patterns from SRM Analysis

Length	Count
2	8368
3	3386
4	61
≥ 5	0

the counts, there the cumulative count of relationships from SRM was 4,010,150. Of these relationships, a cumulative count of 491,955 (12.3%) were marked invalid. Rules can be invalid if any individual step is invalid or if the ICD10 code not allowable or is not specific enough. Table 4.2 and Table 4.3 show the top rules, of lengths two and three respectively, which were marked invalid based on the frequency of occurrence. Support is expressed in the number of records which contained the rule. For comparison, valid rules are shown in Table 4.4. The valid rules show relationships that are accepted by medical consensus as being plausibly causally linked.

In summary, 20.1% of the frequently occurring patterns from the death certificate data showed discordance with the expert knowledge data. Major causes of this could include the use of non-specific codes and missing entities in the sequences entered. Finding the codes which are top areas of discordance on the COD section of death certificates can help the national health institutions such the CDC, NCHS, and NVSS provide the certifying clinicians with the requisite training to avoid potential inaccuracies. They could also help these institutions to assist the physicians to record a more detailed information. Also, finding out the missing links in the COD sequence for death analysis has the potential to help in clinical decision support.

4.6 Conclusions

The lack of methods and evaluation of the accuracy of current death records in national mortality databases is a challenge and can result in inappropriate public health interven-

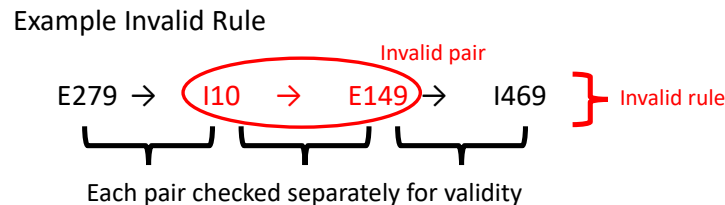


Figure 4.1: Interpretation of a multi-part rule, where any invalid link identifies the rule as invalid.

Table 4.2: Top 10 Invalid Rules of Length 2

Support	Rule ($c_1 \rightarrow c_2$)
7278	Essential (primary) hypertension \rightarrow Unspecified diabetes mellitus without complications
7175	Chronic obstructive pulmonary disease, unspecified \rightarrow Malignant neoplasm of bronchus or lung, unspecified
6598	Congestive heart failure \rightarrow Chronic obstructive pulmonary disease, unspecified
6248	Congestive heart failure \rightarrow Atherosclerotic heart disease
5998	Essential (primary) hypertension \rightarrow Chronic obstructive pulmonary disease, unspecified
5992	Chronic obstructive pulmonary disease, unspecified \rightarrow Atherosclerotic heart disease
5670	Atherosclerotic heart disease \rightarrow Chronic obstructive pulmonary disease, unspecified
5665	Essential (primary) hypertension \rightarrow Unspecified dementia
3550	Atrial fibrillation and flutter \rightarrow Atherosclerotic heart disease
3407	Essential (primary) hypertension \rightarrow Non-insulin-dependent diabetes mellitus without complications

Table 4.3: Top 10 Invalid Rules of Length 3

Support	Rule ($c_1 \rightarrow c_2 \rightarrow c_3$)
1329	Essential (primary) hypertension \rightarrow Unspecified diabetes mellitus without complications \rightarrow Cardiac arrest, unspecified
1284	Congestive heart failure \rightarrow Atherosclerotic heart disease \rightarrow Cardiac arrest, unspecified
1186	Chronic obstructive pulmonary disease, unspecified \rightarrow Atherosclerotic heart disease \rightarrow Cardiac arrest, unspecified
1079	Essential (primary) hypertension \rightarrow Unspecified diabetes mellitus without complications \rightarrow Atherosclerotic heart disease
879	Essential (primary) hypertension \rightarrow Unspecified diabetes mellitus without complications \rightarrow Acute myocardial infarction, unspecified
861	Congestive heart failure \rightarrow Chronic obstructive pulmonary disease, unspecified \rightarrow Respiratory failure, unspecified
798	Essential (primary) hypertension \rightarrow Chronic obstructive pulmonary disease, unspecified \rightarrow Cardiac arrest, unspecified
733	Atherosclerotic heart disease \rightarrow Chronic obstructive pulmonary disease, unspecified \rightarrow Cardiac arrest, unspecified
607	Congestive heart failure \rightarrow Chronic obstructive pulmonary disease, unspecified \rightarrow Cardiac arrest, unspecified
600	Chronic obstructive pulmonary disease, unspecified \rightarrow Malignant neoplasm of bronchus or lung, unspecified \rightarrow Respiratory failure, unspecified

Table 4.4: Top 10 Valid Rules of Length 2

Support	Rule ($c_1 \rightarrow c_2$)
68064	Atherosclerotic heart disease \rightarrow Cardiac arrest, unspecified
47089	Atherosclerotic heart disease \rightarrow Acute myocardial infarction, unspecified
38386	Atherosclerotic heart disease \rightarrow Congestive heart failure
30814	Congestive heart failure \rightarrow Cardiac arrest, unspecified
30787	Essential (primary) hypertension \rightarrow Cardiac arrest, unspecified
28401	Chronic obstructive pulmonary disease, unspecified \rightarrow Respiratory failure, unspecified
27733	Pneumonia, unspecified \rightarrow Septicemia, unspecified
26349	Pneumonia, unspecified \rightarrow Respiratory failure, unspecified
24728	Essential (primary) hypertension \rightarrow Atherosclerotic heart disease
24535	Acute myocardial infarction, unspecified \rightarrow Cardiac arrest, unspecified

tions, loss of life, or increased expense. In this study, we develop a framework to showcase the discordances in COD coding of death certificates in mortality databases with an expert knowledge base. We also identified the rules with the most frequent discrepancies. This provides us with the knowledge to improve the training of physicians to improve the filling of death certificates. It also gives us an insight into the common discordances and the systemic changes required for improving the accuracy of death certificates. These systematic differences may highlight potential opportunities for updating and revising either clinician training or the causal classification procedures themselves. In addition, this can be incorporated into intelligent analytics with future potential for improving the accuracy of death reporting.

There are limitations to the methods and results described by this work. Though potential root causes for the discordance are proposed and discussed throughout, the lack of ground truth hampers direct evaluation of the causes of the discordance. Additionally, only one year of historical data was used for the rule mining step. In the future, we will extend our analysis to data from spanning multiple years. We will also use graph based ontology analysis which can potentially provide the missing links to the clinicians at the time of filling of the death certificates. We believe that this work demonstrates the applicability and value of decision support systems to mortality reporting, and hope that future work in this area may definitively identify the sources of the sources of these errors. We will also investigate the combination of our methods with decision support systems.

4.7 Acknowledgements

This work was carried out in collaboration with the Centers for Disease Control and Prevention (CDC). This work was supported in part by grants from the National Center for Advancing Translational Sciences of the National Institutes of Health (NIH) under Award UL1TR000454 to Dr. May D. Wang, National Science Foundation (NSF) Award NSF-1651360, and the US Department of Health and Human Services (HHS) Centers for Dis-

ease Control and Prevention (CDC) HHS-D2002015F62550B to Dr. May D. Wang, and Microsoft Research and Hewlett Packard. This article does not reflect the official policy or opinions of the CDC, NSF, or the US Department of HHS and does not constitute an endorsement of the individuals or their programs.

For this work, the authors thank Paula Braun (CDC), and Charles Sirc (CDC) for their invaluable assistance and support in shaping this project. We also thank Donna Hoyert and Robert Anderson at the National Center for Health Statistics for their invaluable feedback and support.

Proceedings, Association for Computing Machinery by Association for Computing Machinery. Reproduced with permission of Association for Computing Machinery in the format Republish in a thesis/dissertation via Copyright Clearance Center. See Appendix F for license details. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.

CHAPTER 5

ADVANCED RULE MINING FOR MORTALITY REPORTING DECISION SUPPORT

5.1 Introduction

There is a need in mortality reporting for decision support systems which can aid clinicians in providing complete and accurate causal relationships, using conditions extracted from a patient's electronic medical records. Better methods of generating, filtering, and ranking mortality association rules can be used to create effective decision support systems to achieve this goal. Both hierarchical relationships between codes in the ICD-10 system, and published cause of death ontologies specifying what are valid causal relationships between codes, provide potential sources for novel mortality-specific rule metrics for use in rule mining computation.

5.2 Data

5.2.1 Mortality Data

The multiple causes of death data published by the NVSS, a division of the NCHS of the CDC was used as the primary data set for rule mining. Data for 16 years, from 2003-2018 was downloaded directly from https://www.cdc.gov/nchs/data_access/vitalstatsonline.htm. In total, 40,990,408 deaths are included in the rule mining system, an average of 2.561 million records per year. A summary of the available data is shown in Table 5.1.

Table 5.1: Size and Summary Statistics for the NVSS Multiple Causes of Death Data Set

Year	Records	Distinct Codes	Average Elements / Seq.	Average Items / Element
2003	2 452 154	5541	2.2868	1.2407
2004	2 401 400	5433	2.3005	1.2458
2005	2 452 506	5433	2.3196	1.2488
2006	2 430 725	5439	2.3263	1.2566
2007	2 428 343	5426	2.3265	1.2623
2008	2 476 811	5600	2.3317	1.2637
2009	2 441 219	5529	2.3290	1.2686
2010	2 472 542	5602	2.3414	1.2738
2011	2 519 842	5483	2.3524	1.2840
2012	2 547 864	5485	2.3534	1.2883
2013	2 601 452	5559	2.3613	1.2930
2014	2 631 167	5581	2.3608	1.2964
2015	2 718 198	5672	2.3754	1.3044
2016	2 749 860	5822	2.3772	1.3228
2017	2 820 028	5792	2.3959	1.3350
2018	2 846 297	5907	2.4074	1.3463
Total	40 990 408	9012	2.3483	1.2852

5.3 Methods

5.3.1 Computing Environment and Tools

All data was stored in, and processed using, cloud services obtained through Amazon Web Services (AWS). AWS’s Elastic MapReduce (AWS EMR) was used as the primary computing environment. Elastic MapReduce (AWS EMR) is a service that automates the provisioning of Apache Hadoop clusters and the deployment of applications based on the Hadoop ecosystem. Apache Spark was used for distributed machine learning code, and Apache Hive for the database system. Spark code was written in Scala. This work uses the AWS EMR software distribution version 5.31.0, and code was written primarily using Scala 2.11.

5.3.2 PrefixSpan SPM

PrefixSpan [57] is a Sequential Pattern Mining (SPM) algorithm based on the principle of projecting the database into smaller subsets based on frequent sequence prefixes. The foundations of SPM are outlined in Chapter 3, and more details about the PrefixSpan algorithm itself are presented in Section 3.3.2.

5.3.2.1 Parameter Selection

The maximum sequence length was set to 5. This parameter value was set by domain knowledge, as the NVSS multiple causes of death data set breaks the sequence of causes of death into five “lines” and, as can be seen in Table 5.1, most of these lines, or elements of the sequences in SPM parlance, have an average of around one item. Applying the strict definitions adopted in Section 3.3.3, sequences longer than five elements should not occur, and would not have any cognizable meaning if they did. Since elements of the source sequences have on average only

In previous work, we have used a minimum support threshold of 50 occurrences in one year of mortality reporting data [62, 107, 63]. This value was chosen based on domain expert knowledge, based on discussions with collaborators at the CDC and NCHS, based on what they would find interesting in their own experience. With the NVSS multiple causes of death data set containing approximately 2.5 million records per year, this equates to a minimum support threshold of 0.00002, or $2 \cdot 10^{-5}$. For this analysis, a range of potential minimum support values $\{10^{-3}, 10^{-4}, 2 \cdot 10^{-5}, 10^{-5}\}$ are explored. This range is chosen so as to encompass the domain expert-recommended level of support, while also exploring the parameter space to allow characterization of the parameter response.

5.3.3 Sequential Rule Mining

The mining of frequent sequential patterns and their supports is orders of magnitude more computationally intensive than the scoring, filtering, and sorting of sequential rules. As

such, the rule generation algorithm was run on the most expansive of the frequent sequence lists obtained by Section 5.3.2, with minimum support of 10^{-5} .

The rule generation algorithm is presented in detail in Section 3.3.3; briefly, each frequent sequential pattern is divided at every possible index that would leave non-empty sequences on both sides. Because PrefixSpan does not limit its output to the mining of closed frequent sequences, all frequent sequences are included in the database and there is no need for a procedure to generate additional subsequences.

$$\text{rules}(\mathcal{G}) = \bigcup_{g \in \mathcal{G}} \left\{ \left((g_1, \dots, g_n) \Rightarrow (g_{n+1}, \dots, g_{|g|}) \right) : n \in \mathbb{N} \wedge 1 \leq n < |g| \right\}$$

These candidate rules will then be scored according to the definitions of support, confidence, lift, and leverage, as derived in Section 3.3.3.

$$\begin{aligned} \text{supp}_{SRM}(A \Rightarrow B) &= \text{supp}_{SPM}(A \# B) \\ \text{conf}_{SRM}(A \Rightarrow B) &= \frac{\text{supp}_{SRM}(A \Rightarrow B)}{\text{supp}_{SPM}(A)} = \frac{\text{supp}_{SPM}(A \# B)}{\text{supp}_{SPM}(A)} \\ \text{lift}_{SRM}(A \Rightarrow B) &= \frac{\text{conf}_{SRM}(A \Rightarrow B)}{\text{supp}_{SPM}(Y)} = \frac{\text{supp}_{SPM}(A \# B)}{\text{supp}_{SPM}(X) \cdot \text{supp}_{SPM}(Y)} \\ \text{lvr}_{SRM}(A \Rightarrow B) &= \text{supp}_{SPM}(A \# B) - \text{supp}_{SPM}(X) \cdot \text{supp}_{SPM}(Y) \end{aligned}$$

The notation used in listing sequences and sequential rules is similar to that introduced in Chapter 3, with some deviations to more clearly support the long string representations of ICD-10 codes. In representing sequences, the plus symbol (+) is used to indicate that two items are part of the same element or event. The sequential flow between the elements of the sequence is shown by connecting them with the single arrow (\rightarrow) symbol. The divide between the antecedent and consequent of a sequential rule is called out with a double arrow symbol (\Rightarrow).

The design methodology is summarized in Figure 5.1

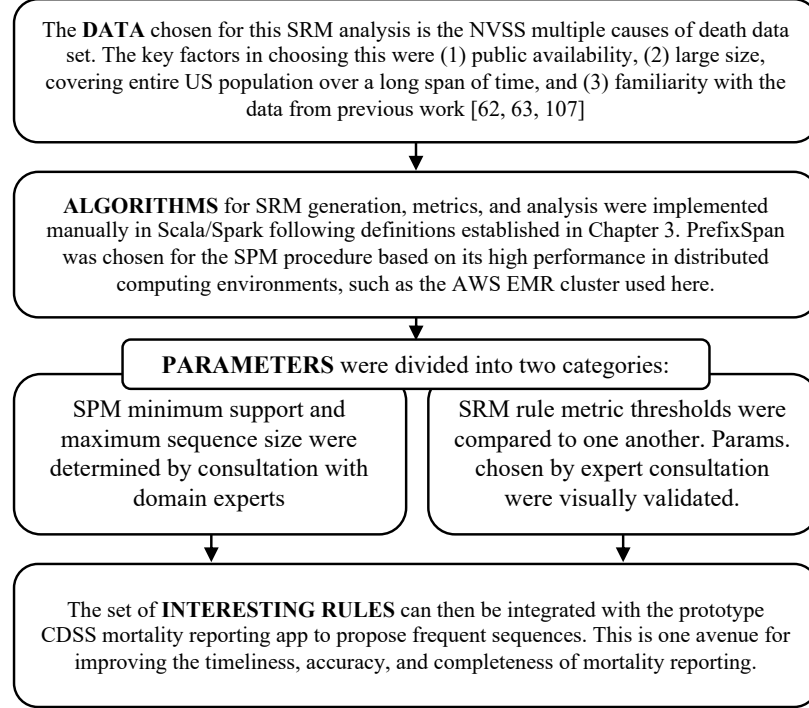


Figure 5.1: Summary of key design decisions and considerations for the SPM/SRM procedure.

5.4 Results and Discussion

5.4.1 Sequential Pattern Mining

First, the implementation of PrefixSpan included in Apache Spark’s MLlib package was run with a minimum support threshold of 10^{-3} , i.e. only sequences supported by at least one in one thousand records are retained. Table 5.2 shows the number of frequent sequences found, broken down by the sequence length. Table 5.3, Table 5.4, and Table 5.5 illustrate the most highly supported sequences. Due to the very high minimum support threshold, no frequent 4-sequences or 5-sequences were found in the data, and a total of 905 frequent sequences were found.

This analysis was then repeated with progressively lower support thresholds, to produce larger sets of frequent sequential patterns. The minimum support was set to 10^{-4} , i.e. only sequences supported by at least one in ten thousand records are retained. The maximum

sequence size was set to 5. Table 5.6 shows the number of sequences found, broken down by the sequence length. For brevity, additional tables are presented in Appendix A. Table A.1, Table A.2, Table A.3, illustrate the most highly supported sequences included in this frequent sequence set that were not included in the set for minimum support of 10^{-3} . Sequences with length four and five items were found in the data, shown in Table A.4 and Table A.5, and a total of 10 560 frequent sequences of any length were found.

The precise expert recommended minimum support value, was then applied. The minimum support was set to $2 \cdot 10^{-5}$ Table 5.7 shows the number of sequences found, broken down by the sequence length. Table A.6, Table A.7, Table A.8, Table A.9, and Table A.10 illustrate the most highly supported sequences found. A total of 57 570 frequent patterns of any length.

The minimum support was set to 10^{-5} , i.e. only sequences supported by at least one in one hundred thousand records are retained. The maximum sequence size was set to 5. Table 5.8 shows the number of sequences found, broken down by the sequence length. Table A.11, Table A.12, Table A.13, Table A.14, and Table A.15 illustrate the most highly supported sequences found. Due to the low minimum support threshold many sequences as long as five elements were found in the data, with a total of 117 510 frequent patterns of any length.

The relationship between the minimum support threshold, sequence size, and the number of frequent sequences mined is visualized in Figure 5.2. Visual analysis of this relationship suggests a smooth, power law-like relationship between minimum support and the number of sequences found, within the range 10^{-5} to 10^{-3} . A notable exception to the prevailing trend is that for frequent items, i.e. frequent 1-sequences, the slope of the relationship is significantly more flat, showing some nonlinearity. This is likely due to the saturation of the possible space of frequent items as the minimum support is reduced. Visual analysis of the parameter response also validates the design decision to set a maximum rule length of 5. It is clear that only very few additional frequent sequences could possibly

Table 5.2: Length of Frequent Sequences, PrefixSpan with minsupport= 10^{-3}

k	Number of k -Sequences
1	288
2	535
3	82
Total	905

have been found, at the cost of a potentially significant increase to computation time.

5.4.2 Sequential Rule Mining

Following the SPM process, the largest database of frequent sequences was post-processed to generate candidate rules. In all, 130 349 candidate rules were identified. The rules were then scored as outlined in Section 5.3.3. The support, confidence, leverage, and sequence sizes of all identified rules are shown in Figure 5.3.

Based on a visual inspection of Figure 5.3, the rule metric values appear to be evenly distributed, as one would expect from noisy or random data. However, there is a clear discontinuity in this pattern in the lift metric around $\text{lift}(A \Rightarrow B) = 10$. This distribution suggests these rules may be very valuable for decision support and recommendation systems. Rules are thresholded to a minimum support value of 10, and visualized again in Figure 5.4. Compared to the set of all candidate rules, the distribution of confidence has shifted to the right, indicating increased confidence. This is an expected result, as the definition of lift in Equation 3.3 is dependent on confidence. However, the notable rightward shift in the distribution of sizes, and the significant reduction in 1-sequences as rule consequents, suggests that thresholding by lift will yield more complex, interesting and useful sequential rules. Selections of these rules are shown in Table 5.10, Table 5.11, and Table 5.9

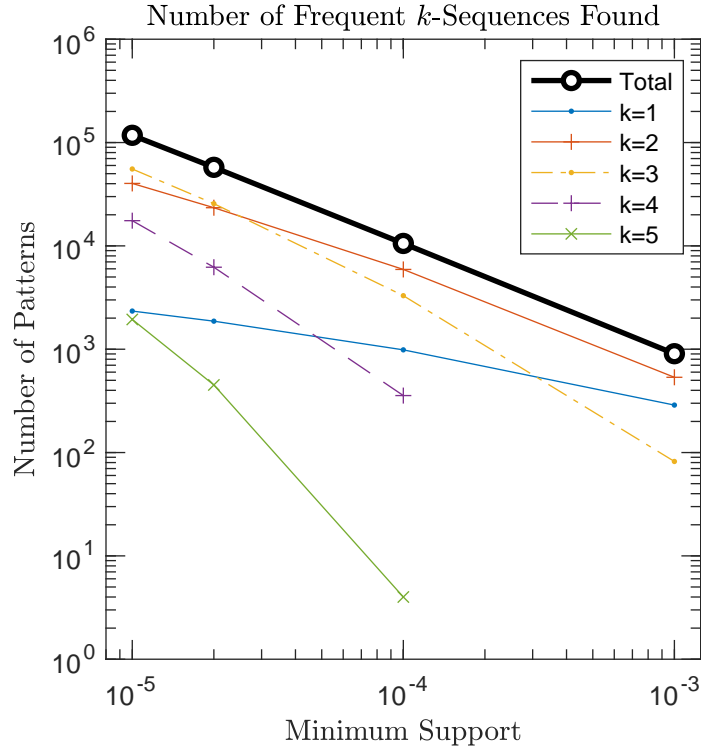


Figure 5.2: The relationship between the minimum support threshold and the number of frequent k -sequences mind is shown for a variety of sequence sizes k , on a log-log axis. Trends between minimum support and the number of rules found are approximately linear in this log-log space, suggesting a power law-like response. The only discontinuities apparent are the lack of larger sequences when the minimum support value becomes prohibitively high. Based on this analysis, the use of the expert-recommended minimum support threshold of $2 \cdot 10^{-5}$ should not have any unexpected impact on the analysis.

Table 5.3: Top 15 Frequent 1-Sequences, PrefixSpan with minsupport= 10^{-3}

Length	Support	Count
Atherosclerotic heart disease of native coronary artery	1.3133E-01	5383225
Cardiac arrest, cause unspecified	1.2969E-01	5315907
Essential (primary) hypertension	1.2466E-01	5109662
Heart failure	1.0529E-01	4316082
Chronic obstructive pulmonary disease, unspecified	9.7797E-02	4008742
Nicotine dependence	9.1477E-02	3749675
Unspecified dementia	7.2078E-02	2954502
Pneumonia, unspecified organism	6.8466E-02	2806463
Malignant neoplasm of unspecified part of bronchus or lung	6.3963E-02	2621865
Acute myocardial infarction, unspecified	6.3844E-02	2617004
Respiratory failure, unspecified	6.3824E-02	2616173
Sepsis, unspecified organism	6.3719E-02	2611863
E149	6.2777E-02	2573243
I64	5.0310E-02	2062247
Atrial fibrillation and flutter	4.6208E-02	1894099

Table 5.4: Top 15 Frequent 2-Sequences, PrefixSpan with minsupport= 10^{-3}

Length	Support	Count
Atherosclerotic heart disease of native coronary artery \rightarrow Cardiac arrest, cause unspecified	2.9056E-02	1191014
Nicotine dependence \rightarrow Malignant neoplasm of unspecified part of bronchus or lung	2.2184E-02	909342
Nicotine dependence \rightarrow Chronic obstructive pulmonary disease, unspecified	2.2038E-02	903352
Essential (primary) hypertension \rightarrow Cardiac arrest, cause unspecified	2.1379E-02	876334
Atherosclerotic heart disease of native coronary artery \rightarrow Acute myocardial infarction, unspecified	1.9295E-02	790903
Essential (primary) hypertension \rightarrow Atherosclerotic heart disease of native coronary artery	1.8681E-02	765753
Atherosclerotic heart disease of native coronary artery \rightarrow Heart failure	1.6926E-02	693784
Heart failure \rightarrow Cardiac arrest, cause unspecified	1.4435E-02	591692
Essential (primary) hypertension + E149	1.4378E-02	589365
Essential (primary) hypertension \rightarrow Heart failure	1.3370E-02	548055
Essential (primary) hypertension \rightarrow Acute myocardial infarction, unspecified	1.2210E-02	500482
Chronic obstructive pulmonary disease, unspecified \rightarrow Respiratory failure, unspecified	1.2100E-02	496002
Chronic obstructive pulmonary disease, unspecified \rightarrow Cardiac arrest, cause unspecified	1.1736E-02	481071
Chronic obstructive pulmonary disease, unspecified + Nicotine dependence	1.1567E-02	474152
E149 \rightarrow Atherosclerotic heart disease of native coronary artery	1.0877E-02	445834

Table 5.5: Top 10 Frequent 3-Sequences, PrefixSpan with minsupport=10⁻³

Length	Support	Count
Essential (primary) hypertension → Atherosclerotic heart disease of native coronary artery → Cardiac arrest, cause unspecified	4.6974E-03	192549
Essential (primary) hypertension → Atherosclerotic heart disease of native coronary artery → Acute myocardial infarction, unspecified	3.9240E-03	160846
Nicotine dependence → Chronic obstructive pulmonary disease, unspecified → Respiratory failure, unspecified	3.5381E-03	145028
Asphyxiation → Asphyxiation + X70	3.2844E-03	134629
Atherosclerotic heart disease of native coronary artery → Acute myocardial infarction, unspecified → Cardiac arrest, cause unspecified	3.1072E-03	127367
Chronic obstructive pulmonary disease, unspecified + Nicotine dependence → Malignant neoplasm of unspecified part of bronchus or lung	2.8036E-03	114919
E149 → Atherosclerotic heart disease of native coronary artery → Cardiac arrest, cause unspecified	2.6729E-03	109562
Essential (primary) hypertension + E149 → Cardiac arrest, cause unspecified	2.5221E-03	103383
Essential (primary) hypertension → Atherosclerotic heart disease of native coronary artery → Heart failure	2.4578E-03	100745
Essential (primary) hypertension + E149 → Atherosclerotic heart disease of native coronary artery	2.3972E-03	98262

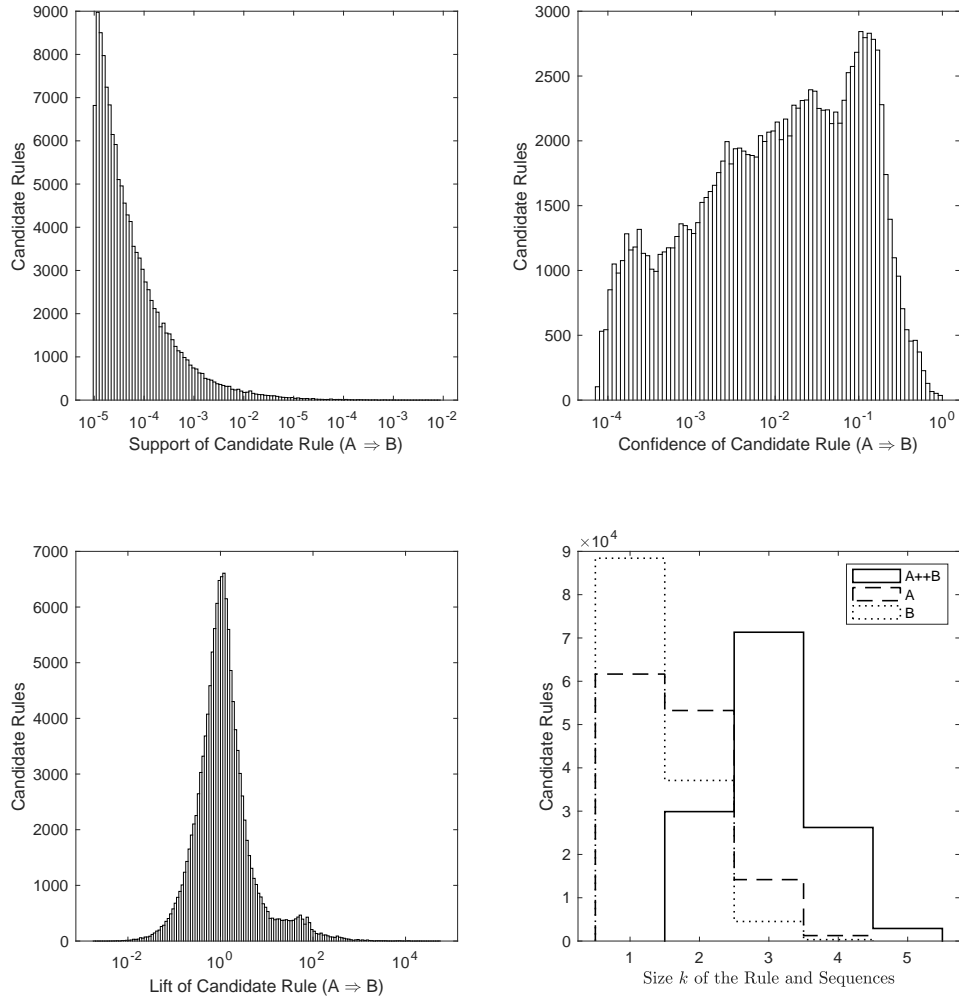


Figure 5.3: The distribution of various rule scoring metrics and sequence sizes are shown. This visualization was generated using the data from all 130 349 candidate rules generated from the frequent sequences with minimum support of 10^{-5} .

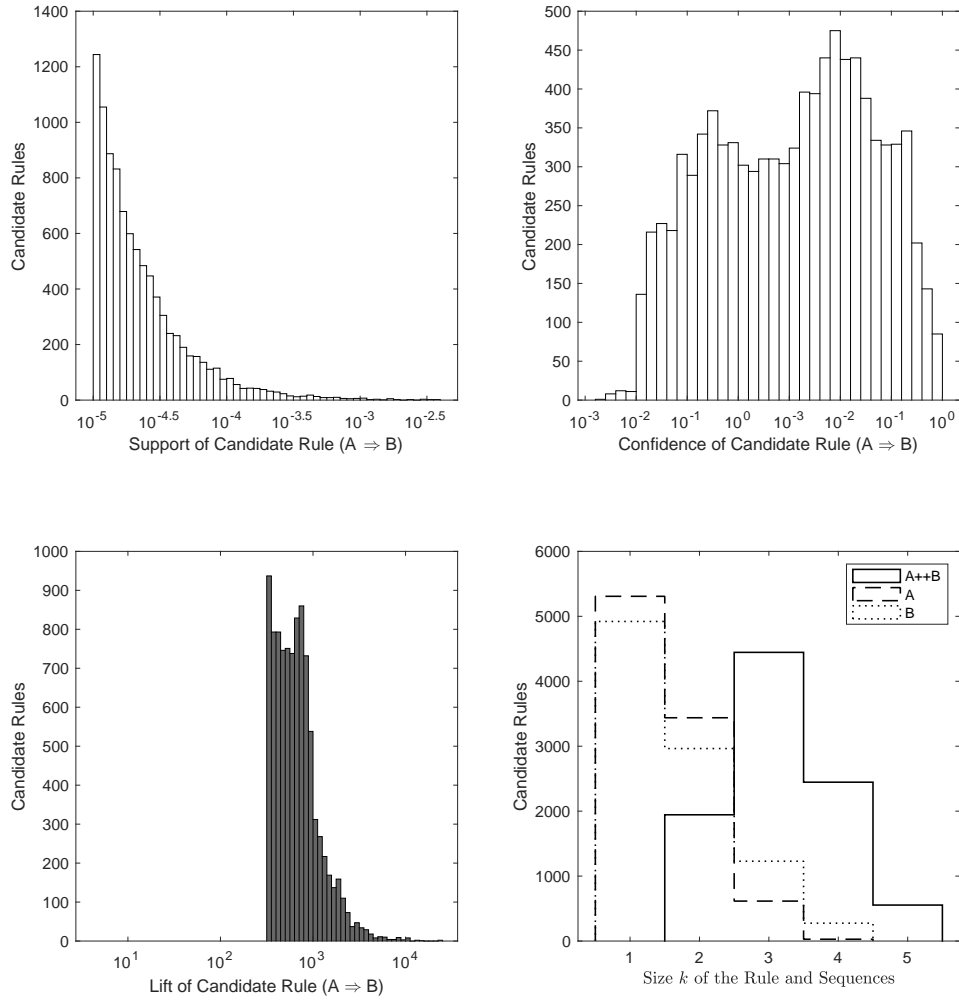


Figure 5.4: The distribution of various rule scoring metrics and sequence sizes are shown, after thresholding to a minimum lift threshold of 10. This thresholding reduced the size of the rule set from 130 349 candidate rules to 9 389 interesting rules.

Table 5.6: Length of Frequent Sequences, PrefixSpan with minsupport= 10^{-4}

k	Number of k -Sequences
1	986
2	5917
3	3297
4	356
5	4
Total	10560

Table 5.7: Length of Frequent Sequences, PrefixSpan with minsupport= $2 \cdot 10^{-5}$

k	Number of k -Sequences
1	1868
2	23412
3	25623
4	6216
5	451
Total	57570

5.5 Conclusions and Future Work

Sequential rules are well suited to the problem of mortality reporting, due to the strictly ordered and causally linked nature of the reported events in the NVSS mortality data. The sequential rules mined in this chapter have multiple applications in mortality reporting, including decision support and augmenting the validity analysis published in Chapter 4. The next steps would be to include integration of this data with the next iteration of the mortality reporting application. Future work could also look for supervised data sources that could support further validation and optimization of the thresholds and parameters chosen here.

Table 5.8: Length of Frequent Sequences, PrefixSpan with minsupport= 10^{-5}

k	Number of k -Sequences
1	2340
2	40223
3	55466
4	17540
5	1941
Total	117510

Table 5.9: Top Sequential Rules, Ordered by Lift

Rule	Support	Condifence	Lift
T590 \Rightarrow X67 + T590	1.1881e-05	0.7550	5.5564e+04
T590 \Rightarrow T590	1.2808e-05	0.8140	5.1728e+04
T634 \Rightarrow T634	2.1102e-05	0.6633	2.0852e+04
T634 \Rightarrow T634 + X23	1.4052e-05	0.4417	1.8551e+04
T634 \Rightarrow X23	1.5077e-05	0.4739	1.7423e+04
T487 \Rightarrow X64 + T487	4.4742e-05	0.6921	1.2459e+04
O96 \Rightarrow O96 \rightarrow O96	1.1905e-05	0.2642	1.2406e+04
O96 \rightarrow O96 \Rightarrow O96	1.1905e-05	0.5590	1.2406e+04
T487 \Rightarrow T487	4.9792e-05	0.7702	1.1913e+04
O961 \Rightarrow O960	1.1954e-05	0.2579	1.1515e+04
T383 \Rightarrow X64 + T383	1.8760e-05	0.3610	1.0802e+04
O96 \Rightarrow O96	2.1298e-05	0.4727	1.0490e+04
O961 \Rightarrow O961	2.2054e-05	0.4758	1.0265e+04
T390 \Rightarrow X60 + T390	1.1295e-05	0.1985	9.5930e+03
T383 \Rightarrow T383	2.3518e-05	0.4526	8.7096e+03

Table 5.10: Top Sequential Rules, Ordered by Support

Rule	Support	Condifence	Lift
T71 \Rightarrow T71	3.9405e-03	0.7398	1.3889e+02
T141 \Rightarrow S019	3.9377e-03	0.4493	5.9450e+01
T71 \Rightarrow X70	3.2896e-03	0.6176	1.6123e+02
T71 \Rightarrow T71 + X70	3.2844e-03	0.6166	1.6194e+02
R13 \Rightarrow J690	3.0111e-03	0.2871	1.2338e+01
T141 \Rightarrow X95	2.9504e-03	0.3367	7.9791e+01
T509 \Rightarrow X44	2.9171e-03	0.3282	4.8374e+01
T509 \Rightarrow T509	2.4394e-03	0.2745	3.0882e+01
T141 \Rightarrow X74	2.2940e-03	0.2618	5.4960e+01
K746 \Rightarrow K729	1.9328e-03	0.1188	1.1929e+01
S019 \Rightarrow S019	1.8978e-03	0.2511	3.3221e+01
T509 \Rightarrow T509 + X44	1.7654e-03	0.1986	5.5336e+01
T141 \Rightarrow S019 + X74	1.7265e-03	0.1970	5.5369e+01
B182 \Rightarrow K746	1.6588e-03	0.2774	1.7051e+01
T509 \Rightarrow X42	1.6204e-03	0.1823	2.6910e+01

Table 5.11: Top Sequential Rules, Ordered by Confidence

Rule	Support	Condifence	Lift
T71 + F319 \Rightarrow T71	1.8151e-05	0.9960	1.8699e+02
F329 + T71 \Rightarrow T71	1.3820e-04	0.9925	1.8633e+02
T71 \rightarrow F329 \Rightarrow T71	2.9373e-05	0.9821	1.8438e+02
T71 + T519 + X65 \Rightarrow T71	1.3662e-05	0.9773	1.8349e+02
T71 + F419 \Rightarrow T71	1.0076e-05	0.9764	1.8331e+02
T71 + T519 + X65 \Rightarrow X70	1.3589e-05	0.9721	2.5376e+02
T71 + T519 + X65 \Rightarrow T71 + X70	1.3564e-05	0.9703	2.5483e+02
F329 + T71 \Rightarrow X70	1.3491e-04	0.9688	2.5291e+02
T71 + X65 \Rightarrow T71	2.9934e-05	0.9669	1.8153e+02
I679 \rightarrow R630 \rightarrow R090 \Rightarrow J969	1.1173e-05	0.9662	1.5139e+01
F329 + T71 \Rightarrow T71 + X70	1.3454e-04	0.9662	2.5375e+02
T71 + F319 \Rightarrow X70	1.7516e-05	0.9612	2.5092e+02
T71 + F319 \Rightarrow T71 + X70	1.7516e-05	0.9612	2.5243e+02
T71 + T510 + X65 \Rightarrow T71	1.6589e-05	0.9605	1.8032e+02
T71 \rightarrow F329 \Rightarrow X70	2.8616e-05	0.9568	2.4977e+02

PART 2

USING FHIR TO ENABLE PUBLIC HEALTH INFORMATICS

CHAPTER 6

FHIR PROFILING FOR MORTALITY REPORTING

6.1 Preface

Some parts of sections of this chapter are adapted from the publication "Intelligent Mortality Reporting with FHIR" [62], of which Ryan Hoffman is the first and leading author. This publication is copyrighted by IEEE, and reused with permission. For detailed copyright disclosure, please see Appendix F. It has been modified throughout to comply with formatting requirements and to better integrate with the rest of this work.

©2018 IEEE. Reprinted, with permission, from R. A. Hoffman, H. Wu, J. Venugopalan, P. Braun, and M. D. Wang, "Intelligent mortality reporting with FHIR," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 5, pp. 1583–1588, 2018. DOI: 10.1109/JBHI.2017.2780891.

6.2 Problem Statement

In seeking to develop widely interoperable data systems, there is a constant struggle to maintain balance between standardization and flexibility. All other things being equal, the more structured a store of data is, the easier it will be to achieve rich interoperability with other systems, and the more powerful it can become for secondary research use. However, significant variation between medical practitioners, EMR and Electronic Health Record (EHR) software vendors, information technology systems, and even political policy can drive users toward ad-hoc, incompatible infrastructure. Without significant effort towards data and policy harmonization, there is little chance of developing the deep interoperability that is needed for next-generation healthcare and biomedical informatics research. However, the more tightly constrained and specialized an informatics solution becomes, the

narrower its potential user base becomes, while at the same time the financial and technological obstacles to its adoption continue to mount. This fundamental tension between useful structure and broad appeal is one of the core challenges at the heart of the health informatics profession.

6.3 Background

6.3.1 FHIR

Fast Healthcare Interoperability Resources (FHIR) is a next-generation healthcare data interoperability system, consisting of two major components: a set of powerful and flexible common data models [108]; and a modern, web-like Application Programming Interface (API) for interacting with the data [109]. The FHIR Resources, as of 2019's FHIR Release 4 (R4), are 145 data models of varying complexity and maturity, each representing a component of the overall FHIR common data model ecosystem [108]. A Resource consists of a list of elements, each of which may include restrictions to its cardinality (i.e. how many of that element can appear in a Resource, or whether the element is in fact required for a valid Resource), coding systems, and other constraints. Elements can be arranged hierarchically and may consist of complex data types defined as part of the FHIR platform.

Consider as an example FHIR's Patient resource [110]. This is one of the foundational resources of the FHIR data ecosystem, and is one of just a few resources in the developing standard to have reached the "Normative" level of maturity and standardization. The Patient resource is a natural nexus between many other data resources, and many other resources refer to Patients, either directly or indirectly, as one of the organizing principles of the overall data structure. However, the resource is not over-constrained or over-defined. In an earlier version of the FHIR specification [111], Draft Standard for Trial Use (DSTU) 2, the Patient resource even had a field for specifying the species of the Patient, ensuring that the standard remained flexible enough for veterinary and animal research use.

However, as can be seen in Figure 6.1 and in the full specification at [110], this field


Name	Flags	Card.	Type	Description & Constraints
Patient	 N		DomainResource	Information about an individual or animal receiving health care services
identifier	 Σ	0..*	Identifier	Elements defined in Ancestors: id , meta , implicitRules , language , text , contained , extension , modifierExtension An identifier for this patient
active	 Σ	0..1	boolean	Whether this patient's record is in active use
name	 Σ	0..*	HumanName	A name associated with the patient
telecom	 Σ	0..*	ContactPoint	A contact detail for the individual
gender	 Σ	0..1	code	male female other unknown AdministrativeGender (Required)
birthDate	 Σ	0..1	date	The date of birth for the individual
deceased[x]	 Σ	0..1		Indicates if the individual is deceased or not
deceasedBoolean	 Σ		boolean	
deceasedDateTime	 Σ		dateTime	
address	 Σ	0..*	Address	An address for the individual
maritalStatus	 Σ	0..1	CodeableConcept	Marital (civil) status of a patient MaritalStatus (Extensible)
multipleBirth[x]	 Σ	0..1		Whether patient is part of a multiple birth
multipleBirthBoolean	 Σ		boolean	
multipleBirthInteger	 Σ		integer	
photo	 Σ	0..*	Attachment	Image of the patient
contact	 I	0..*	BackboneElement	A contact party (e.g. guardian, partner, friend) for the patient + Rule: SHALL at least contain a contact's details or a reference to an organization
relationship	 Σ	0..*	CodeableConcept	The kind of relationship Patient Contact Relationship (Extensible)
name	 Σ	0..1	HumanName	A name associated with the contact person
telecom	 Σ	0..*	ContactPoint	A contact detail for the person
address	 Σ	0..1	Address	Address for the contact person
gender	 Σ	0..1	code	male female other unknown AdministrativeGender (Required)
organization	 I	0..1	Reference(Organization)	Organization that is associated with the contact
period	 Σ	0..1	Period	The period during which this contact person or organization is valid to be contacted relating to this patient

Figure 6.1: This excerpt from the FHIR Patient resource specification at HL7.org [110] illustrates the hierarchical structure, strict data typing enforcement, and cardinality constraints typical of the FHIR Resource specifications.

is no longer present in the newer FHIR Release 4 (R4) version of the specification. This might seem at first like a regression, or a design compromise, but is in fact an illustration of a key aspect of FHIR development: FHIR Profiling.

6.3.2 FHIR Profiling

With Fast Healthcare Interoperability Resources (FHIR), its developers aim to resolve the fundamental tension between useful structure and burdensome specificity gap through the fusion of broadly defined core data structures, combined with programmatically interpretable mechanisms for extending or constraining the underlying FHIR Resources and/or Application Programming Interface (API) for a particular application. This facility is referred to as FHIR "profiling", and is an essential part of developing useful FHIR-based medical informatics systems.

Profiling provides a number of mechanisms by which the FHIR data models and/or APIs can be extended or constrained for a given specific application. Critically, this is accomplished in a way that forbids directly contradicting the base specification, as this would break the promise of data interoperability that had driven FHIR's development and adoption. These techniques include:

- Constraining the cardinality of an element (i.e. making an optional element from the specification mandatory).
- Requiring that certain data elements must have certain values, or values within some range, etc.
- Dividing lists of elements into subsets, each of which may have additional constraints put upon it.
- Defining what terminology systems or coding schemes can be used, and where.

FHIR also provides a mechanism to define extensions to the base resources. Extending a data system is naturally more hazardous to interoperability than constraining it, however

there are mechanisms in place to prevent problems from emerging. FHIR Resources have object-oriented programming-like patterns of inheritance, where Resources extend and are derived from one another. Most FHIR resources are derived from a few common abstract types such as the `DomainResource`, and it is through this inheritance that all `DomainResources` (including, for example, `Patient`, `Practitioner`, `Observation`, `Condition`) inherit an optional list (cardinality `0..*`) of Extension data elements. Extensions are implemented as a data type, which must define a URL (cardinality `1..1`) that links to a file defining: (1) in human-readable terms, what the extension represents; and (2) in machine-readable terms, exactly what data is involved in the extension. This removes a great deal of friction in the process of enabling a FHIR systems to store and retrieve a new extended resource.

As a hypothetical example, consider a hospital that was contracted to provide medical services to a local boarding school. It is easy to imagine the legal and practical challenges that could arise from taking on the business of treating children whose guardians might be remote or unreachable, and the hospital would naturally reach an agreement with the school to share student identity and emergency contact information to facilitate the relationship. Allowing the school nurse's EHR system to interoperate with the hospital's own systems would be a very natural and potentially a very valuable means for facilitating this exchange of information.

The power of FHIR in this hypothetical is that it can take these restrictions out of the realm of policy and communication, and enforce them at a technical level, as a condition of interoperability. The hospital could develop a `ResidentialStudentPatient` profile, to be used "on top of" the base resource, to enforce the data sharing requirements. For example, FHIR `Patient` resources may have between zero and infinity identifiers associated with it. In the FHIR specification, this is expressed as the identifier element having a cardinality of `0..*`. To accept a boarding school patient into their system, a FHIR profile might enforce that the student must have at least one identifier (`1..*`), and furthermore that the identifier's type, system, or issuer must be some agreed upon value to identify the student patients.

Because this constraint is more restrictive than the base specification, there is no penalty to interoperability - a compliant resource using this ResidentialStudentPatient profile is still a completely valid Patient in the eyes of any FHIR-compliant data system.

6.3.3 FHIR Package Registries

There are dozens of profiles and extensions defined as part of the FHIR standard, such as the patient-animal extension discussed above. These are first-party parts of the FHIR ecosystem, controlled by the same HL7 community that governs the base FHIR standard. However, anyone may make and distribute profiles for any purpose they wish. HL7 [112] and other organizations such as Simplifier [113] provide free registries where such profiles and extensions can be shared and worked on collaboratively, to strengthen the open-source FHIR platform.

6.4 Mortality Reporting Profiling

The core FHIR data model includes a generic Bundle resource, allows multiple other resources to be packaged together for purposes ranging from searching and storage to attribution and reporting [114]. For this work, we propose developing prototype FHIR profiles defining a death certificate, building off of the FHIR idiom of a Bundle with the Composition resource summarizing its contents to serve as an analogue to a real-world document. Additional Resources are then added to the bundle to support the information contained in the complex document, which is illustrated in Figure 6.2.

Unlike the older HL7 CDA document standard, FHIR documents are modular compositions of full EHR Resources, which can be readily split apart and incorporated into another interoperable system. A significant milestone in developing a FHIR-based electronic death record (EDR) is the mapping of all elements in a death certificate to FHIR Resources and common profiles. Our mapping is: (i) easy to use so that application developers can get what various Resource elements are used for; (ii) scalable and modular so that the rich vari-

U.S. STANDARD CERTIFICATE OF DEATH									
LOCAL FILE NO.					STATE FILE NO.				
1. DECEDENT'S LEGAL NAME (Include AKA's if any) (First, Middle, Last)					2. SEX		3. SOCIAL SECURITY NUMBER		
4a. AGE-Last Birthday (Years)		4b. UNDER 1 YEAR		4c. UNDER 1 DAY		5. DATE OF BIRTH (Mo/Day/Yr)		6. BIRTHPLACE (City and State or Foreign Country)	
7a. RESIDENCE-STATE		7b. COUNTY		7c. CITY OR TOWN					
7d. STREET AND NUMBER				7e. APT. NO.		7f. ZIP CODE		7g. INSIDE CITY LIMITS? <input type="checkbox"/> Yes <input type="checkbox"/> No	
8. EVER IN US ARMED FORCES? <input type="checkbox"/> Yes <input type="checkbox"/> No		9. MARITAL STATUS AT TIME OF DEATH <input type="checkbox"/> Married <input type="checkbox"/> Married, but separated <input type="checkbox"/> Widowed <input type="checkbox"/> Divorced <input type="checkbox"/> Never Married <input type="checkbox"/> Unknown		10. SURVIVING SPOUSE'S NAME (If wife, give name prior to first marriage)					
11. FATHER'S NAME (First, Middle, Last)					12. MOTHER'S NAME PRIOR TO FIRST MARRIAGE (First, Middle, Last)				
13a. INFORMANT'S NAME			13b. RELATIONSHIP TO DECEDENT			13c. MAILING ADDRESS (Street and Number, City, State, Zip Code)			
14. PLACE OF DEATH (Check only one: see instructions)									
IF DEATH OCCURRED IN A HOSPITAL:					IF DEATH OCCURRED SOMEWHERE OTHER THAN A HOSPITAL:				
15. FACILITY NAME (If not institution, give street & number)					16. CITY OR TOWN, STATE, AND ZIP CODE				
17. COUNTY OF DEATH									
18. METHOD OF DISPOSITION: <input type="checkbox"/> Burial <input type="checkbox"/> Cremation <input type="checkbox"/> Donation <input type="checkbox"/> Entombment <input type="checkbox"/> Removal from State <input type="checkbox"/> Other (Specify):					19. PLACE OF DISPOSITION (Name of cemetery, crematory, other place)				
20. LOCATION-CITY, TOWN, AND STATE					21. NAME AND COMPLETE ADDRESS OF FUNERAL FACILITY				
22. SIGNATURE OF FUNERAL SERVICE LICENSEE OR OTHER AGENT							23. LICENSE NUMBER (Of licensee)		
ITEMS 24-28 MUST BE COMPLETED BY PERSON WHO PRONOUNCES OR CERTIFIES DEATH									
24. DATE PRONOUNCED DEAD (Mo/Day/Yr)					25. TIME PRONOUNCED DEAD				
26. SIGNATURE OF PERSON PRONOUNCING DEATH (Only when applicable)					27. LICENSE NUMBER			28. DATE SIGNED (Mo/Day/Yr)	
29. ACTUAL OR PRESUMED DATE OF DEATH (Mo/Day/Yr) (Spell Month)			30. ACTUAL OR PRESUMED TIME OF DEATH			31. WAS MEDICAL EXAMINER OR CORONER CONTACTED? <input type="checkbox"/> Yes <input type="checkbox"/> No			
CAUSE OF DEATH (See instructions and examples)									
32. PART I. Enter the <u>chain of events</u> —diseases, injuries, or complications—that directly caused the death. DO NOT enter terminal events such as cardiac arrest, respiratory arrest, or ventricular fibrillation without showing the etiology. DO NOT ABBREVIATE. Enter only one cause on a line. Add additional lines if necessary.									
IMMEDIATE CAUSE (Final disease or condition resulting in death) a. _____ Due to (or as a consequence of): _____									
Sequentially list conditions, if any, leading to the cause listed on line a. Enter the UNDERLYING CAUSE (disease or injury that initiated the events resulting in death) LAST c. _____ Due to (or as a consequence of): _____									
d. _____									
33. PART II. Enter other significant conditions contributing to death but not resulting in the underlying cause given in PART I									
34. WAS AN AUTOPSY PERFORMED? <input type="checkbox"/> Yes <input type="checkbox"/> No									
35. DID TOBACCO USE CONTRIBUTE TO DEATH? <input type="checkbox"/> Yes <input type="checkbox"/> Probably <input type="checkbox"/> No <input type="checkbox"/> Unknown									
36. IF FEMALE: <input type="checkbox"/> Not pregnant within past year <input type="checkbox"/> Pregnant at time of death <input type="checkbox"/> Not pregnant, but pregnant within 42 days of death <input type="checkbox"/> Not pregnant, but pregnant 43 days to 1 year before death <input type="checkbox"/> Unknown if pregnant within the past year									
37. MANNER OF DEATH <input type="checkbox"/> Natural <input type="checkbox"/> Homicide <input type="checkbox"/> Accident <input type="checkbox"/> Pending investigation <input type="checkbox"/> Suicide <input type="checkbox"/> Could not be determined									
38. DATE OF INJURY (Mo/Day/Yr) (Spell Month)			39. TIME OF INJURY		40. PLACE OF INJURY (e.g., Decedent's home, construction site, restaurant, wooded area)			41. INJURY AT WORK? <input type="checkbox"/> Yes <input type="checkbox"/> No	
42. LOCATION OF INJURY: State: _____ City or Town: _____									
43. DESCRIBE HOW INJURY OCCURRED: _____									
44. IF TRANSPORTATION INJURY, SPECIFY: <input type="checkbox"/> Driver/Operator <input type="checkbox"/> Passenger <input type="checkbox"/> Pedestrian <input type="checkbox"/> Other (Specify)									
45. CERTIFIER (Check only one): <input type="checkbox"/> Certifying physician-To the best of my knowledge, death occurred due to the cause(s) and manner stated. <input type="checkbox"/> Pronouncing & Certifying physician-To the best of my knowledge, death occurred at the time, date, and place, and due to the cause(s) and manner stated. <input type="checkbox"/> Medical Examiner/Coroner-On the basis of examination, and/or investigation, in my opinion, death occurred at the time, date, and place, and due to the cause(s) and manner stated.									
Signature of certifier: _____									
46. NAME, ADDRESS, AND ZIP CODE OF PERSON COMPLETING CAUSE OF DEATH (Item 32)									
47. TITLE OF CERTIFIER		48. LICENSE NUMBER		49. DATE CERTIFIED (Mo/Day/Yr)			50. FOR REGISTRAR ONLY- DATE FILED (Mo/Day/Yr)		
51. DECEDENT'S EDUCATION Check the box that best describes the highest degree or level of school completed at the time of death. <input type="checkbox"/> 8th grade or less <input type="checkbox"/> 9th - 12th grade; no diploma <input type="checkbox"/> High school graduate or GED completed <input type="checkbox"/> Some college credit, but no degree <input type="checkbox"/> Associate degree (e.g., AA, AS) <input type="checkbox"/> Bachelor's degree (e.g., BA, AB, BS) <input type="checkbox"/> Master's degree (e.g., MA, MS, MEng, MEd, MSW, MBA) <input type="checkbox"/> Doctorate (e.g., PhD, EdD) or Professional degree (e.g., MD, DDS, DVM, LLB, JD)					52. DECEDENT OF HISPANIC ORIGIN? Check the box that best describes whether the decedent is Spanish/Hispanic/Latino. Check the "No" box if decedent is not Spanish/Hispanic/Latino. <input type="checkbox"/> No, not Spanish/Hispanic/Latino <input type="checkbox"/> Yes, Mexican, Mexican American, Chicano <input type="checkbox"/> Yes, Puerto Rican <input type="checkbox"/> Yes, Cuban <input type="checkbox"/> Yes, other Spanish/Hispanic/Latino (Specify) _____				
53. DECEDENT'S RACE (Check one or more races to indicate what the decedent considered himself or herself to be) <input type="checkbox"/> White <input type="checkbox"/> Black or African American <input type="checkbox"/> American Indian or Alaska Native (Name of the enrolled or principal tribe) _____ <input type="checkbox"/> Asian Indian <input type="checkbox"/> Chinese <input type="checkbox"/> Filipino <input type="checkbox"/> Japanese <input type="checkbox"/> Korean <input type="checkbox"/> Vietnamese <input type="checkbox"/> Other Asian (Specify) _____ <input type="checkbox"/> Native Hawaiian <input type="checkbox"/> Guamanian or Chamorro <input type="checkbox"/> Samoan <input type="checkbox"/> Other Pacific Islander (Specify) _____ <input type="checkbox"/> Other (Specify) _____									
54. DECEDENT'S USUAL OCCUPATION (Indicate type of work done during most of working life. DO NOT USE RETIRED).									
55. KIND OF BUSINESS/INDUSTRY									

REV. 11/2003

Figure 6.2: U.S Standard Certificate of Death, 2003 Revision. This complex document requires input from multiple stakeholders to certify a death. Item 32 contains the causes of death. CDC 2003 - Work of the US Federal Government, Public Domain [32]

ation in data elements in different USA State’s health agencies can be represented by an interoperable set of Resources, contained in a Bundle; (iii) idiomatically correct so that potential users do not misuse the standard fields; and (iv) designed with stakeholders’ current data practices in mind to mirror existing processes wherever feasible so that the integration friction is minimized and more adoption is accomplished. In this application profiling, we mapped death certificate data to FHIR Draft Standard for Trial Use (DSTU) 2 resources and data elements because it has much wider acceptance over the newer Standard for Trial Use (STU) 3 version at the time of development.

We used the standard FHIR metaphor of a “document” to represent a death certificate object. As illustrated in Figure 6.3, we profiled a FHIR document with a defined minimum set of resources to represent death certificate data elements. To maximize its usefulness and to enable interoperability before electronic death reporting systems are truly or fully FHIR-enabled, we have completed a mapping of this profile FHIR resource data elements to known standard HL7 Vital Records Domain Analysis Model (VR DAM) section on mortality reporting.

Some death certificate data elements map very naturally and directly to FHIR Resources (e.g. the decedent name and address map directly to a FHIR Patient). Others require more careful thought and design. For example, to choose between FHIR Patient.contacts or RelatedPersons in representing the decedent’s family members, assuming both representations are complete, are able to represent all of the data elements from the U.S. Standard Certificate of Death and the HL7 VR DAM without extension, and are easy to understand, the deciding factor becomes “idiomatic FHIR correctness” – i.e., which choice most clearly follows from the originally intended uses. The contacts field of a FHIR Patient is intended primarily to enable contacting the patient or his/her decision-makers in a clinical setting, whereas RelatedPerson resources are sources of patient information with non-healthcare relationships to the patient.

Another informative example is the choice between a collection of Observation re-

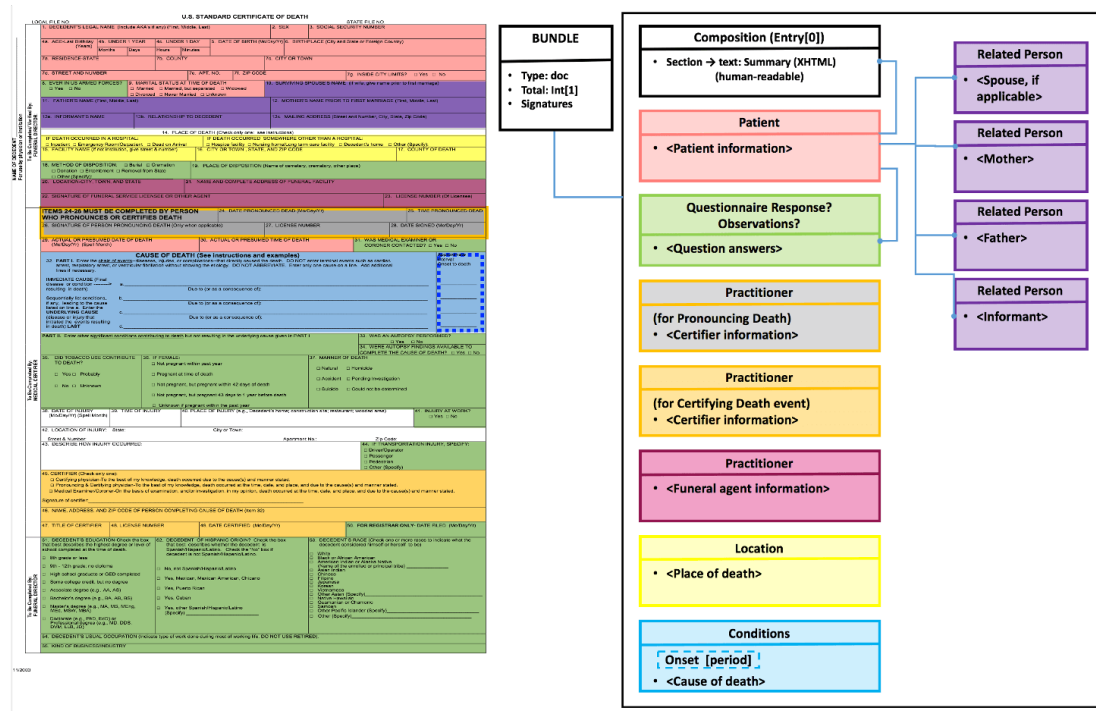


Figure 6.3: The U.S. Standard Certificate of Death is overlaid with colored blocks representing mapped FHIR resources (left), to be included into a death certificate FHIR Bundle (right).

sources or Questionnaire / QuestionnaireResponse resources to represent the demographics and medical history sections of the death certificate (shown in green in Figure 6.3). Using questionnaire to represent a list of questions on a physical death certificate form is attractive. However, we used Observations instead because the variations in lists of medical history questions and valid responses among different US State reporting jurisdictions makes it infeasible to develop a single comprehensive Questionnaire for the entire US. If each jurisdiction were using a unique, “flat” Questionnaire to represent their set of questions and answers, it would complicate the processes of designing decision support systems in US interoperable across 57 State jurisdictional boundaries.

6.5 FHIR - VR DAM Mappings

The mapping tables developed though the VR DAM Working Group efforts are presented in [62] and in Appendix B.

CHAPTER 7

PROTOTYPING MORTALITY REPORTING WITH FHIR

7.1 Preface

Work in this chapter is adapted from the publication "Intelligent Mortality Reporting with FHIR" [107], of which Ryan Hoffman is the first and leading author. This publication is copyrighted by IEEE, and reused with permission. For detailed copyright disclosure, please see Appendix F. It has been modified throughout to comply with formatting requirements and to better integrate with the rest of this work.

©2017 IEEE. Reprinted, with permission, from R. A. Hoffman, H. Wu, J. Venugopalan, P. Braun, and M. D. Wang, "Intelligent mortality reporting with FHIR," in *2017 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, 2017, pp. 181–184. DOI: 10.1109/BHI.2017.7897235. Later published in an expanded and revised form as R. A. Hoffman, H. Wu, J. Venugopalan, P. Braun, and M. D. Wang, "Intelligent mortality reporting with FHIR," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 5, pp. 1583–1588, 2018. DOI: 10.1109/JBHI.2017.2780891. This expanded work is presented in Chapter 8.

7.2 Abstract

One pressing need in the area of public health is timely, accurate, and complete reporting of deaths and the conditions leading up to them. Fast Healthcare Interoperability Resources (FHIR) is a new Health Level Seven (HL7) interoperability standard for electronic health record (EHR), while Sustainable Medical Applications and Reusable Technologies (SMART)-on-FHIR enables third-party app development that can work "out of the box". This research demonstrates the feasibility of developing SMART-on-FHIR applications to

enable medical professionals to perform timely and accurate death reporting within multiple different jurisdictions of US. We explored how the information on a standard certificate of death can be mapped to resources defined in the FHIR standard (DSTU2). We also demonstrated analytics for potentially improving the accuracy and completeness of mortality reporting data.

7.3 Introduction

Mortality is one of the most reliable sources of health-related data that is comparable across different geographical locations and is a large source of population-level health data, with approximately 56 million deaths per year world-wide [1]. In the United States of America (US) alone 2.6 million people die each year [2], with 60-80

Accurate collection and aggregation of high-quality mortality data remains an ongoing challenge primarily due to issues such as the lack of practice for physicians to perform death certification (on the order of 1-2 times a year), non-standard methods to determine the cause of death information, complex data flow between the funeral home, the certifying physician and the registrar, non-standard practices of data acquisition and transmission [3, 4]. The issue is further compounded by the fact that each location has different laws regarding the format of the death certificates and the type of information collected. For example, in the US, the National Center for Health Statistics aggregates mortality data from the 57 reporting jurisdictions around the country. However, the precise regulations and local laws of each reporting jurisdiction differ [115].

A decision support system that can assist the physician to fill the appropriate cause of death and put on the death certificate in the requisite format can largely assist to mitigate these challenges of data accuracy. Additionally, current efforts towards mortality reporting standardization using technologies such as HL7 V2 [116, 117] and CDA [118] have some shortcomings, such as challenges in integrating with large-scale web services. As a result, the current flow of information between the various providers and registrars is not

optimal. Also, under the new meaningful use of EHR, the government requires healthcare institutions to show at least partial support of APIs to show potential for sharing data, interoperability and clinical decision support [119]. The current systems of HL7 and CDA do not have the capability to support APIs and incorporate decision support systems that facilitate the use of data from the decedent's electronic health records to notify the certifying medical professional of events that may be associated with the death.

To overcome these challenges, we propose a framework that utilizes HL7's Fast Healthcare Interoperability Resources (FHIR) as both an application platform and a means of accessing EHR data. FHIR is a new emerging health standard that is aimed at streamlining and standardizing healthcare communication using a resource-centric approach (as opposed to document-centric) for specification of data elements. It is designed to allow simple implementation using existing technologies such as restful APIs, OAuth security, and XML/JSON data [21]. FHIR was chosen for the application of death certificates, because it is vendor-neutral, scalable, and is positioned to emerge as a global standard. FHIR is designed to work within current EHR systems using APIs and can be used to pre-populate death certificates to aid physicians in determining the cause of death information. FHIR is also currently the only standard which supports the addition of analytics into the EHR systems [120, 121]. The ultimate goals of this project are to generate information that will aid in more complete physician reporting of the causes of death, and to provide valuable mortality information to registrars, public health department, and other authorized parties in a timelier manner.

7.4 Web Application Design

A web application was implemented in HTML and JavaScript, using the SMART-on-FHIR [5] JavaScript client library (<https://github.com/smart-on-fhir/client-js>). The application runs in the browser, securely accessing the FHIR server using OAuth2 authentication. The application was developed and tested using a virtual FHIR server. The develop-

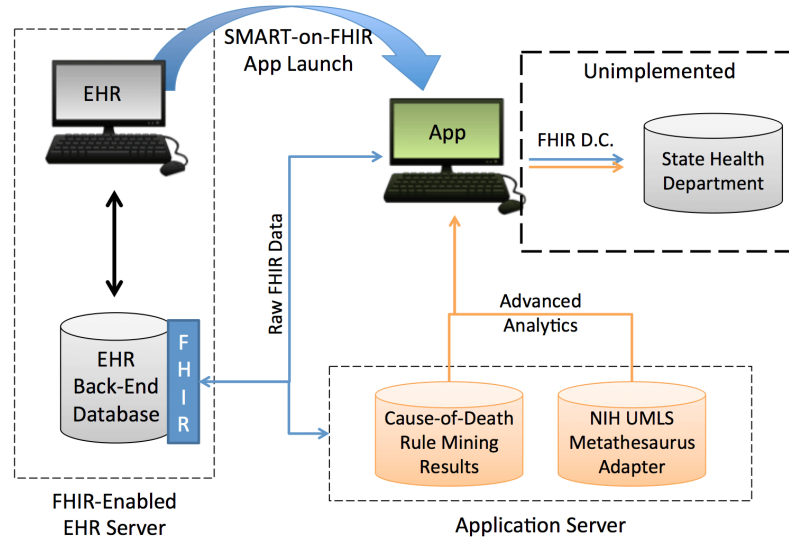


Figure 7.1: Proposed Infrastructure for Death Reporting Application. A user’s existing EHR system with compatible FHIR interface can directly launch the application with patient context. The application can contact an internal or external analytics server for decision support and other tools, before packaging the death certificate object and sending it.

ment server is based on a Vagrant VM configuration created by the SMART project team (<https://github.com/smart-on-fhir/installer>). In addition to the SMART-on-FHIR compliant EHR server, an application server hosts CGI interfaces to UMLS and data mining functionality. An outline of the proposed infrastructure for a SMART-on-FHIR-based mortality is shown in Figure 7.1.

7.4.1 Application Features

This application was designed with the ultimate goal of enabling not just more timely, but also more accurate and complete data about the chain of events ultimately leading to death. As such, it is designed to allow the simultaneous visualization of a large portion of the patient history. As mortality reporting in the United States has adopted the ICD-10 standard since 1999, integration with the Unified Medical Language System (UMLS) Metathesaurus is necessary to enable crosswalk between medical event coding systems.

7.4.2 Illustrative Synthetic Data

To aid in interface prototyping and illustration of the application interface in Figure 7.2, synthetic patient data was created for patients with a variety of conditions surrounding a hypothetical recorded death.

7.4.3 Interface Design

The application main interface is illustrated in Figure 7.2. The application's interface is broken into horizontal panes. The topmost panes recapitulate the patient's information, displaying basic identification data as well as recent noted entered into the patient's record. This information is pulled from Patient, Condition, and other FHIR resources.

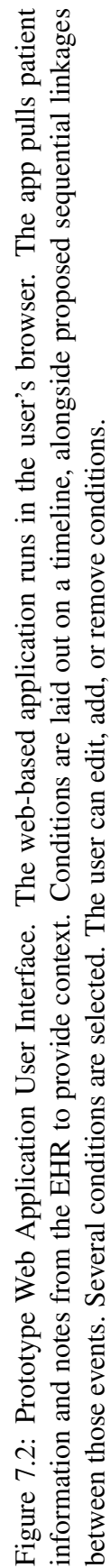
The largest, central pane displays an interactive patient history timeline generated using the popular D3 and D3-tip visualization libraries (<https://d3js.org>, <https://github.com/Caged/d3-tip>). Events displayed on the timeline are spaced logarithmically, with the axis anchored at the time of death. This allows simultaneous visualization of events occurring around the time of death alongside relevant context from the patient's more distant history. Scrolling adjusts the scaling to allow focus on past and recent events. The events shown on this timeline are generated using Condition resources accessed from the FHIR server.

The bottommost panes are designed so as to recreate the familiar-to-users appearance of the US Standard Certificate of Death's cause-of-death field layout, with a chain of one or more events occurring as consequences of one another.

Buttons are provided to access important features, such as closing the application, submitting a death certificate package, or interfacing with UMLS.

7.5 Sequential Pattern Mining Analytics

This application illustrates how next-generation web services can be developed to aid in timely, accurate mortality reporting. To better understand the availability of this data and



demonstrate this capability, we use sequential pattern mining on one year of NVSS data to mine a list of rules that can be used directly in the application to propose common pathways of events that may have lead to death.

7.5.1 Data

National Vital Statistics System, coordinated by the National Center for Health Statistics aggregates the causes of death for all deaths occurring within the United States from 1959 to 2014 [64]. Each death certificate format in vital statistics offices of each state, the District of Columbia, and other special jurisdictions varies, but generally consists of the underlying cause of death as recorded by physicians and other details such as the demographics, comorbid conditions, race and ethnicity.

7.5.2 SPM Background and Related Work

The temporal models commonly seen in the literature include models such as sequence analysis [91, 92, 93, 94, 49] and association rule mining [94, 122, 123]. Sequential Pattern Mining (SPM) is a data mining technique that seeks temporal relationships among events (in this case the underlying causes of death) [124] and has been extensively examined in the literature with applications in pattern mining [48] (AprioriAll [53], SPADE [55]) and database projections. (PrefixSpan [57], MEMISP [101]). Recently, privacy preserving pattern mining [125, 126, 127] and distributed mining [125] have attracted considerable interests. In health care, SPM has applications in heart disease prediction [128], healthcare auditing [123], and neurological diagnosis [129], violent death reporting [130] etc. The input data to an SPM is a set of sequences, which comprises a list of events ordered by temporal relations.

The goal of SPM is to discover all valid sequential patterns with pre-specified minimum support, where support of a candidate pattern is the proportion of sequences in the data that exhibit the pattern [52]. For example, in the Multiple Cause-of-Death data, each record

contains a list of ordered conditions (up to 20 conditions) that could lead to a person's death. While SPM's output an ordered list of sequences which correlates with the target outcome and is able to find rules such as "*Condition1* \rightarrow *Condition2*". This means that if we observe Condition 1, we can assert that Condition 2 will possibly follow Condition 1. This was introduced as an improvement over Association Rule Mining (ARM) [48], which doesn't take the temporal relations into consideration, so will only output rules like "[Condition 1, Condition 2]", meaning if we observe Condition 1, we are also likely to observe Condition 2 for some confidence. Hence SPMs form an ideal algorithm for the current task of discovering the most probable sequence of events that led to the cause of death to help certifying physicians fill the death certificates.

7.5.3 SPM Problem Formulation

As discussed above, the NCHS database consists of up to 20 underlying conditions $C = [C_1, C_2, \dots, C_K]$ which lead to death, where $C = \{C_1, C_2, \dots, C_K\}$ is the list of unique events / conditions. Using this data as the training, set our goal is find the list of most frequent sequence conditions $S = \langle s_1, s_2, \dots, s_T \rangle$ which can occur given the outcome and comorbidities. A sequence S is an ordered set of items, denoted as $S = \langle s_1, s_2, \dots, s_T \rangle$, where each $e_i \in C$ is an item and the sequence is of length T . A set of sequences D is a collection of sequences, $D = \{S^{(1)}, S^{(2)}, \dots, S^{(N)}\}$, where the superscript denotes the index of an individual sequence and N is the number of total number sequences.

We define the contain relationship \subset between two sequences A, B as follows: For $A = \langle a_1, a_2, \dots, a_{T1} \rangle$, $B = \langle b_1, b_2, \dots, b_{T2} \rangle$, if $\forall a_i \in A, \exists m(i), \text{ s.t. } b_{m(i)} \in B$, then $A \subset B$. A sequential pattern, or a rule, R of a sequence S is an ordered list that satisfies $R \subset S$. The relative support of a rule R in the set of sequences D is defined as the percentage of sequences that contain this rule, i.e.,

$$rel_support(R) = \frac{|\{S | S \in D \ \& \ R \subset S\}|}{|D|}$$

Where $|\cdot|$ is the cardinality of a set. And we could also use the number of sequences that contain the rule as an evaluation metric, i.e., $support(R) = |\{S | S \in D \ \& \ R \subset S\}|$. Oftentimes, we want to find rules of length maximum possible, which leads to the definition of frequent closed rule. A rule R is a frequent closed rule if there exists no P , such that $R \subset P$ and $support(R) = support(P)$.

SPM aims to discover sequential patterns that have support larger than a pre-specified minimum support. For example, we have the following set of sequences in Table 7.2 (the arrow “ \rightarrow ” stands for temporal orders) and want to identify rules with a minimum support of 0.8.

7.5.4 SPM Methodology

In our experiments, we use the BIDE algorithm, short for BI-Directional-Extension-based frequent closed sequence mining, proposed in Wang et al. [60]. Conventional sequence mining algorithms adopt a candidate maintenance- and-test paradigm, in which they maintain a list of discovered closed rules and use the rules to prune the search space and determine whether new rules are promising to be closed. Such paradigm is accurate but lacks scalability with respect to the number of frequent closed rules, both in time and storage. On the other hand, the BIDE algorithm aims to find all the frequent closed rules, without candidate maintenance.

7.5.5 Pattern Mining Results

We apply the aforementioned algorithm BIDE to Multiple Cause-of-Death Mortality Data from NCHS to find most frequent sequences of conditions before people’s deaths. We picked Year 2012’s Mortality Data, which contains 2, 547, 864 deaths. We set the minimum support to be 50 and identified a total of 65,915 frequent closed rules. We present the distribution of rules of different lengths in Table 7.2.

Due to the space limits, we present the top 5 rules of length-2 in Table 7.1 for illus-

Table 7.1: Top 5 Prototype Rules of Length-2 from 2012 NCHS Mortality Data

	Rule	Count	Percent
Mental and behavioral disorders due to use of tobacco	→ Other chronic obstructive pulmonary disease	100,920	3.96%
	Chronic ischemic heart disease → Cardiac arrest	85,952	3.37%
Essential (primary) hypertension	→ Chronic ischemic heart disease	77,249	3.03%
	Mental and behavioral disorders due to use of tobacco → Malignant neoplasm of bronchus and lung	73,212	2.87%
	Chronic ischemic heart disease → Heart failure	63,283	2.48%

Table 7.2: Count of Prototype Frequent Rules by Length

Length	Count
1	961
2	19,160
3	33,179
4	11,508
5	1,081
6	26
7	0

tration. The full set of rules was deployed as a lookup table service using a CGI script, integrating it into the death reporting prototype application.

7.6 Conclusions and Future Work

This work demonstrates the feasibility of using the SMART-on-FHIR application framework to develop public health applications for mortality reporting, improving the timeliness and accessibility of such reports. Intelligent analytics have been show integrated with the prototype application, demonstrating future potential for improving the accuracy of death reporting. Future work may focus on using alternative data sets, as well as the more complete patient information exposed though the interoperability of FHIR, to construct more personalized and precise analytics systems. Further development is ongoing to develop precise FHIR resource profiles to concisely, completely, and flexibly represent death certificate data.

7.7 Acknowledgements

The authors thank Dr. Mark Braunstein (Georgia Tech), Dr. Myung Choi (Georgia Tech Research Institute), and Charles Sirc (CDC) for their invaluable assistance and support in shaping this project.

This work was supported in part by grants from the National Center for Advancing Translational Sciences of the National Institutes of Health (NIH) under Award UL1TR-

000454 to Dr. May D. Wang, T32 GM105490 Traineeship to Dr. Greg Gibson of Georgia Tech for trainee R. A. Hoffman, National Science Foundation Award NSF1651360, and the US Department of Health and Human Services (HHS) Centers for Disease Control and Prevention (CDC) HHSD2002015F62550B to Dr. May D. Wang, and Microsoft Research and Hewlett Packard. This article does not reflect the official policy or opinions of the CDC, NSF, or the US Department of HHS and does not constitute an endorsement of the individuals or their programs.

R. A. Hoffman and J. Venugopalan: Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA 30332 USA.

H. Wu: School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA.

P. Braun: Centers for Disease Control and Prevention (CDC), Atlanta, GA 30329 USA (co-corresponding author, e-mail: pabraun@cdc.gov).

M. D. Wang is with the Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA 30332 USA (co-corresponding author, phone: 404-385-2954; e-mail: maywang@bgatech.edu).

©2017 IEEE. Reprinted, with permission, from R. A. Hoffman, H. Wu, J. Venugopalan, P. Braun, and M. D. Wang, “Intelligent mortality reporting with FHIR,” in *2017 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, 2017, pp. 181–184. DOI: 10.1109/BHI.2017.7897235.

PART 3

APP IMPROVEMENT AND FUTURE DIRECTIONS

CHAPTER 8

INTELLIGENT MORTALITY REPORTING WITH FHIR

8.1 Preface

Work in this chapter is adapted from the publication "Intelligent Mortality Reporting with FHIR" [62], of which Ryan Hoffman is the first and leading author. This publication is copyrighted by IEEE, and reused with permission. For detailed copyright disclosure, please see Appendix F. It has been modified throughout to comply with formatting requirements and to better integrate with the rest of this work, and is presented here in full.

©2018 IEEE. Reprinted, with permission, from R. A. Hoffman, H. Wu, J. Venugopalan, P. Braun, and M. D. Wang, "Intelligent mortality reporting with FHIR," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 5, pp. 1583–1588, 2018. DOI: 10.1109/JBHI.2017.2780891.

8.2 Abstract

One pressing need in the area of public health is timely, accurate, and complete reporting of deaths and the diseases or conditions leading up to them. Fast Healthcare Interoperability Resources (FHIR) is a new HL7 interoperability standard for electronic health record (EHR), while Sustainable Medical Applications and Reusable Technologies (SMART)-on-FHIR enables third-party app development that can work "out of the box". This research demonstrates the feasibility of developing SMART-on-FHIR applications that enables medical professionals to perform timely and accurate death reporting within multiple different USA State jurisdictions. We explored how the information on a standard certificate of death can be mapped to resources defined in the FHIR standard Draft Standard for Trial Use Version 2 (DSTU 2) and common profiles. We also demonstrated analytics for potentially

improving the accuracy and completeness of mortality reporting data.

8.3 Introduction

There are approximately 56 million deaths per year world-wide [1], with 2.6 million happening in the United States of America (USA) [2]. Accurate and timely mortality reporting is essential for gathering this important public health data in order to formulate emergency response to epidemics and new disease threats, to prevent communicable diseases such as flu, and to determine vital statistics such as life expectancy, mortality trends, etc. However, accurate collection and aggregation of high-quality mortality data remains an ongoing challenge primarily due to issues such as the average low frequency with which physicians perform death certification (on the order of 1-2 times a year), inconsistent training in determining the causes of death, complex data flow among the funeral home, the certifying physician and the registrar, and non-standard practices of data acquisition and transmission [3, 4]. In addition, the US National Center for Health Statistics (NCHS) aggregates mortality data from the 57 reporting jurisdictions around the country, where the precise regulations and local laws of each reporting jurisdiction differ [115]. Current efforts towards mortality reporting standardization using technologies such as Clinical Document Architecture (CDA) [116, 117] and HL7 V2 [118] have some shortcomings in integrating with large-scale web services, and Integrating the Healthcare Enterprise (IHE) have limited adoptions that result in the non-optimal flow of information among various providers and registrars. Thus, the emerging solutions are (1) to use the new meaningful use of electronic health records (EHR) as the government requires healthcare institutions to show at least partial support of patient-facing application programming interfaces (APIs) for sharing data, interoperability and clinical decision support [119]; and (2) to develop a death decision support system capable of assisting the physician in determining the appropriate cause of death to put on the death certificate in a standardized format.

In this article, we designed and developed a platform that utilizes the new HL7 health

standard Fast Healthcare Interoperability Resources (FHIR) to access EHR data. FHIR aims to streamline and standardize healthcare communication in electronic death reporting using a resource-centric approach (as opposed to document-centric) to specify data elements. It adopts existing technologies such as RESTful (REpresentational State Transfer) APIs, OAuth security, and XML (Extensible Markup Notation) / JSON (JavaScript Object Notation) data to form an application platform [21]; it uses APIs to work with current EHR systems and supports the addition of data analytics [5, 120, 131]; and it is vendor-neutral and scalable. Specifically, FHIR can pre-populate sections of the death certificates to provide information from the decedent’s health history. Also, it can incorporate data-driven analytics to provide mortality decision support for physicians in reporting the causes of death. Ultimately, the accurate and timely reporting will provide valuable mortality information for registrars, public health departments, and other authorized parties in deciding public policies. We have presented a preliminary version of this work at the IEEE Biomedical and Health Informatics (BHI) conference in February 2017 [107]. In this article, we report the complete system.

8.4 Web Application Design

To enable timely, accurate, and complete capture of the chain of diseases or conditions leading to death, we developed SMART-on-FHIR-based mortality reporting system consisting of two parts as shown in Figure 8.1 architecture diagram: the SMART-on-FHIR compliant EHR server, and the application server. First, we developed a web app using the SMART-on-FHIR JavaScript client library (<https://github.com/smart-on-fhir/client-js>) and tested it using a virtual FHIR server. Using this library and HTML (Hypertext Markup Language) JavaScript, we implemented a web application that securely accesses the FHIR server using OAuth2 authentication. Then we developed interactive visualization of a large portion of the decedent’s health history by using simple RESTful interfaces to UMLS (Unified Medical Language System) and data mining functionality. The application’s graphical

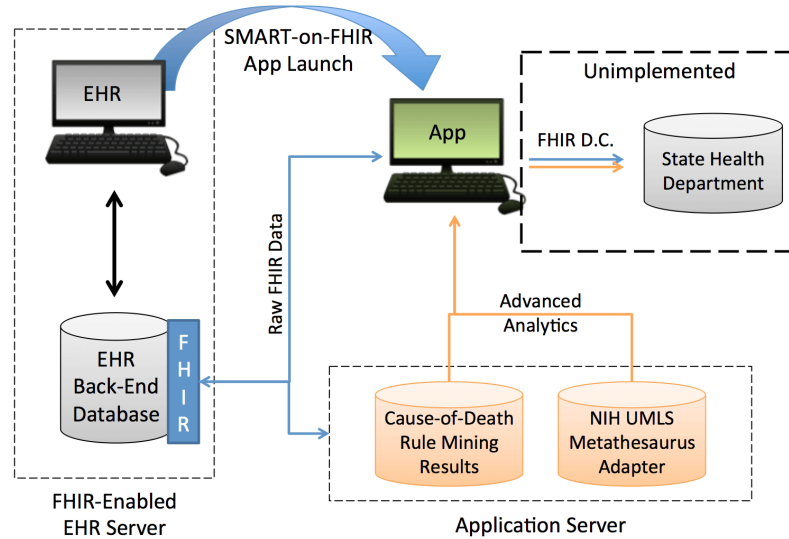


Figure 8.1: Proposed Infrastructure for Death Reporting Application. A user’s existing EHR system with compatible FHIR interface can directly launch the application with patient context. The application can contact an internal or external analytics server for decision support and other tools, before packaging the death certificate object and sending it.

user interface (GUI) is broken into sequential pages, with each page addressing a section of the death certificate to assist in determining the causes of death. Because US mortality reporting has adopted the ICD-10 standard since 1999, by using UMLS Metathesaurus, our application enables crosswalk among medical event coding systems, and allows interoperability among records retrieved from a variety of coding systems and downstream analytics.

To aid in GUI prototyping, we used synthetic patient data available in the Cerner SMART-on-FHIR app development sandbox (<https://code.cerner.com>). As shown in Figure 8.2, diseases and conditions displayed on the timeline are spaced logarithmically, with the axis anchoring at the time of death. The interactive patient history timeline is generated using the D3 and D3-tip visualization libraries (<https://d3js.org>, <https://github.com/Caged/d3-tip>). This enables simultaneous visualization of events occurring around the time of death, along with relevant context from the patient’s more distant history. The events shown on this timeline are generated using Condition resources accessible through

Figure 8.2: Prototype Web Application User Interface. The web-based application runs in the user’s browser. The app pulls patient information and notes from the EHR to provide context. Conditions are laid out on a timeline, alongside proposed sequential linkages between those events. Several conditions are selected. The user can edit, add, or remove conditions.

the FHIR server. In addition, scrolling adjusts the scaling to allow users to focus on distant past or recent events. The bottommost section is designed to recreate the familiar-to-users appearance of the cause-of-death field layout in the US Standard Certificate of Death, with a chain of one or more causes occurring as consequences of one another. Buttons are provided to access additional pages with fields such as injury information, the provider’s information, and submission / download controls.

The application can be downloaded from <http://miblab.bme.gatech.edu/software>.

8.5 Representing Death Certificate Data in FHIR

Unlike the older HL7 CDA document standard, FHIR documents are modular compositions of full EHR Resources, which can be readily split apart and incorporated into another interoperable system. FHIR documents are a Bundle of resources, where the first entry in the Bundle is a Composition that contains a human readable summary of the Bundle’s

contents. Additional Resources are then added to the bundle to support the information contained in the document. A significant milestone in developing a FHIR-based electronic death record (EDR) is the mapping of all elements in a death certificate to FHIR Resources and common profiles. Our mapping is: (i) easy to use so that application developers can get what various Resource elements are used for; (ii) scalable and modular so that the rich variation in data elements in different USA State's health agencies can be represented by an interoperable set of Resources, contained in a Bundle; (iii) idiomatically correct so that potential users do not misuse the standard fields; and (iv) designed with stakeholders' current data practices in mind to mirror existing processes wherever feasible so that the integration friction is minimized and more adoption is accomplished. In this application profiling, we mapped death certificate data to FHIR Draft Standard for Trial Use (DSTU) 2 resources and data elements because it has much wider acceptance over the newer Standard for Trial Use (STU) 3 version at the time of development.

We used the standard FHIR metaphor of a “document” to represent a death certificate object. As illustrated in Figure 8.3, we profiled a FHIR document with a defined minimum set of resources to represent death certificate data elements. To maximize its usefulness and to enable interoperability before electronic death reporting systems are truly or fully FHIR-enabled, we have completed a mapping of this profile FHIR resource data elements to known standard HL7 Vital Records Domain Analysis Model (VR DAM) section on mortality reporting.

Some death certificate data elements map very naturally and directly to FHIR Resources (e.g. the decedent name and address map directly to a FHIR Patient). Others require more careful thought and design. For example, to choose between FHIR Patient.contacts or RelatedPersons in representing the decedent's family members, assuming both representations are complete, are able to represent all of the data elements from the U.S. Standard Certificate of Death and the HL7 VR DAM without extension, and are easy to understand, the deciding factor becomes “idiomatic FHIR correctness” – i.e., which choice most clearly

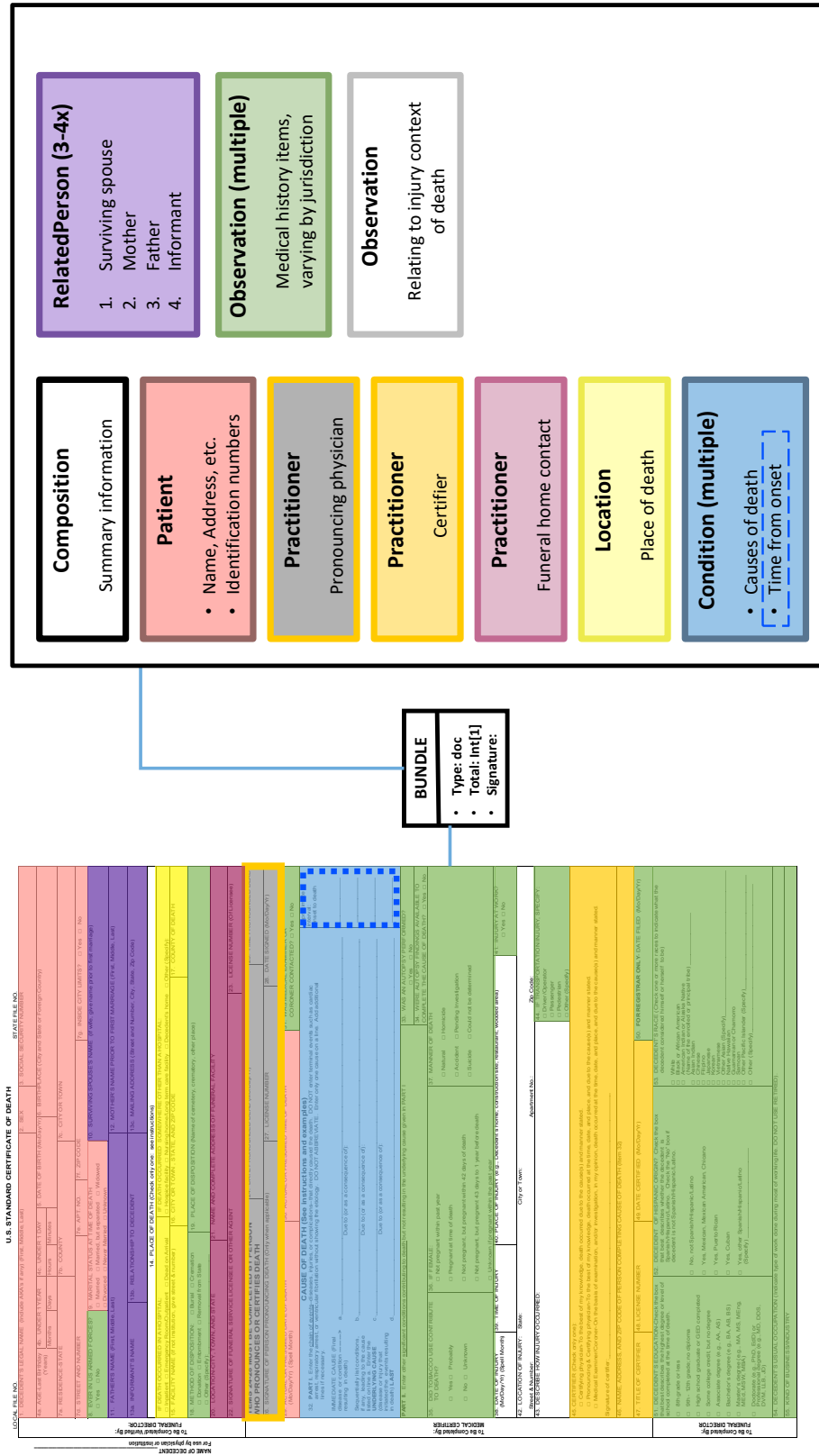


Figure 8.3: Resource-Level Mapping Overview. (Left) The U.S. Standard Certificate of Death is overlaid with colored blocks representing mapped FHIR resources, to be bundled into a death certificate document (Right). The exact number of Conditions will vary with the circumstances of death, and specific observations may be specified by public health reporting jurisdictions.

follows from the originally intended uses. The contacts field of a FHIR Patient is intended primarily to enable contacting the patient or his/her decision-makers in a clinical setting, whereas RelatedPerson resources are sources of patient information with non-healthcare relationships to the patient.

Another informative example is the choice between a collection of Observation resources or Questionnaire / QuestionnaireResponse resources to represent the demographics and medical history sections of the death certificate (shown in green in Figure 8.3). Using questionnaire to represent a list of questions on a physical death certificate form is attractive. However, we used Observations instead because the variations in lists of medical history questions and valid responses among different US State reporting jurisdictions makes it infeasible to develop a single comprehensive Questionnaire for the entire US. If each jurisdiction were using a unique, “flat” Questionnaire to represent their set of questions and answers, it would complicate the processes of designing decision support systems in US interoperable across 57 State jurisdictional boundaries.

8.6 Sequential Pattern Mining Analytics

As illustrated in Figure 8.1, with FHIR-based platform in place, using data-driven analytics to extract possible causes of death can help accomplish timely, accurate mortality reporting. Thus, we have tried sequential pattern mining on one year of US public causes of death data from the National Center for Health Statistics in US National Vital Statistics System (NVSS) to derive a list of frequent sequences of events (i.e. diseases and conditions) that may have led to death.

The temporal models commonly seen in the literature are sequence analysis [91, 92, 93, 94] and association rule mining [94, 122, 123]. Sequential Pattern Mining (SPM) seeks temporal relationships among a sequence of events [124, 48] with multiple algorithmic approaches such as AprioriAll [53], SPADE [55], PrefixSpan [57], and MEMISP [101]. In diseases modeling with multiple patient conditions, Association Rule Mining (ARM)

Table 8.1: Death Record Layout in Multiple Cause-of-Death Mortality Data from NCHS

Type	Information Included
Demographics	Age, Gender, Residence, Death Time, etc.
Underlying Cause of Death	Cause of death coded according to ICD and several other coding systems.
Conditions	A maximum of 20 conditions that correlate with the death
Race and Ethnicity	The reported race for States that are reporting single race or the bridged race for States that are reporting multiple race.

[48] only outputs rules like $[Condition1, Condition2]$, meaning that Condition 1 is likely to coexist with Condition 2 for some confidence. It doesn't take temporal relations into consideration. On the other hand, SPM outputs an ordered list of sequences relate to the target outcome with rules such as $Condition1 \rightarrow Condition2$, which means that if we observe Condition 1, we can assert that Condition 2 will possibly follow Condition 1. SPM was applied successfully to clinical data [132, 133] ranging from heart disease prediction [128], healthcare auditing [104], neurological diagnosis [129], and violent death reporting [130]. Thus, we chose SPMs to discover the most probable sequence of events that lead to a death to help certifying physicians in filling out death certificates.

Through the National Vital Statistics System (NVSS), the National Center for Health Statistics aggregates data from death certificates reported by 57 vital records jurisdictions across the United States [64]. Each record in this data set contains a list of ordered conditions (up to 20 conditions) that could lead to a person's death as the causes of death recorded by physicians, medical examiners, and coroners, and other details such as the demographics, race and ethnicity. The available fields are shown in Table 8.1.

In SPM, support of a candidate pattern is the proportion of sequences in the data that exhibit the pattern [52]. The relative support of a rule R in the set of sequences D is defined as the percentage of sequences that contain this rule, i.e.,

$$rel_support(R) = \frac{|\{S | S \in D \ \& \ R \subset S\}|}{|D|}$$

where $|\cdot|$ is the cardinality of a set. SPM aims to discover sequential patterns that have support larger than a pre-specified minimum support (e.g. the support is a number between 0 and 1 and we can specify the minimum support of a pattern of interest). If we use $C = [C_1, C_2, \dots, C_K]$, $K = 20$ to represent 20 unique underlying conditions/diseases as the causes leading to death, under a pre-specified minimum support requirement, SPM will discover all valid sequential patterns $S = \langle s_1, s_2, \dots, s_T \rangle$ that can occur. More specifically, we used one SPM approach BI-Directional-Extension-based frequent closed sequence mining (BIDE) proposed by Wang et al. [60]. Closed frequent sequences are those sequences that are not subsets of other frequent sequences having equal support. Because conventional sequence mining algorithms must maintain a list of discovered closed sequences, while using the patterns to determine whether new sequences are promising to be closed, it cannot scale up in both time and storage for the number of frequent closed sequences. Thus, the BIDE approach was developed by Wong et al. to find all the frequent closed sequences without candidate maintenance.

In our project, under given outcomes (i.e. mortality) and comorbidities, we applied BIDE to find the list of most frequent sequence conditions $S = \langle s_1, s_2, \dots, s_T \rangle$ that can occur. Using 2012's Multiple Cause-of-Death Mortality Data from the NVSS public that contains 2,547,864 deaths, we set the minimum support to be approximately 2×10^{-5} (equivalent to 50 occurrences for this data set). We identified a total of 59,864 frequent closed sequences of length-2 or greater for the most frequent sequences of diseases or conditions before people's deaths as shown in Table 8.2. We present the top 20 sequences of length-2 in Table 8.3 for illustration.

To use these analytics for FHIR-enabled mortality reporting, a lookup table service with simple CGI-based API is developed to access the full results, where a user can review the most frequent patterns present in a given patient's history as illustrated in orange in Figure 8.2.

Table 8.2: Frequent Sequence Count of Different Lengths

Length	Count
2	22450
3	29344
4	7661
5	405
6	4
≥ 7	0

8.7 Conclusions and Future Work

This work demonstrates the feasibility of using the SMART-on-FHIR application framework in order to improve the timeliness and accessibility of public health mortality reports. This platform is able to incorporate intelligent analytics to further improve the accuracy of death reporting. Future work includes: (i) using alternative data sets, and more complete patient information made accessible through the interoperability of FHIR, to construct more personalized and precise analytics systems; and (ii) developing precise FHIR resource profiles to concisely, completely, and flexibly represent death certificate data.

8.8 Acknowledgements

The authors thank Dr. Mark Braunstein (Georgia Tech), Dr. Myung Choi (Georgia Tech Research Institute), and Charles Sirc (CDC) for their invaluable assistance and support in shaping this project.

This work has been supported in part by National Institutes of Health (NIH) T32 GM105490 Traineeship to Professor Dr. Greg Gibson of Georgia Tech for trainee R. A. Hoffman, grants from US Department of Health and Human Services Centers for Disease Control and Prevention (DHHS CDC) Award HHSD2002015F62550B, NIH National Center for Advancing Translational Sciences Award UL1TR000454, and National Science Foundation Award NSF1651360 to Professor Dr. May D. Wang, and Microsoft Research and Hewlett Packard. This article does not reflect the official policy or opinions of the

Table 8.3: Rules from 2012 NCHS Mortality Data

Rule	Count	Percent
Unspecified mental and behavioral disorder due to use of tobacco → Chronic obstructive pulmonary disease, unspecified	99587	3.91%
Unspecified mental and behavioral disorder due to use of tobacco → Malignant neoplasm of bronchus or lung, unspecified	72211	2.83%
Atherosclerotic heart disease → Cardiac arrest, unspecified	70490	2.77%
Rheumatic heart disease, unspecified → Atherosclerotic heart disease	62927	2.47%
Rheumatic heart disease, unspecified → Cardiac arrest, unspecified	56497	2.22%
Atherosclerotic heart disease → Acute myocardial infarction, unspecified	45916	1.8%
Atherosclerotic heart disease → Congestive heart failure	45721	1.79%
Unspecified mental and behavioral disorder due to use of tobacco → Atherosclerotic heart disease	44285	1.74%
Rheumatic heart disease, unspecified → Congestive heart failure	42919	1.68%
Unspecified mental and behavioral disorder due to use of tobacco → Rheumatic heart disease, unspecified	38518	1.51%
Congestive heart failure → Cardiac arrest, unspecified	35406	1.39%
Rheumatic heart disease, unspecified → Unspecified diabetes mellitus without complications	34966	1.37%
Unspecified diabetes mellitus without complications → Atherosclerotic heart disease	33793	1.33%
Chronic obstructive pulmonary disease, unspecified → Respiratory failure, unspecified	32287	1.27%
Rheumatic heart disease, unspecified → Acute myocardial infarction, unspecified	32042	1.26%
Hyperlipidemia, unspecified → Rheumatic heart disease, unspecified	31846	1.25%
Chronic obstructive pulmonary disease, unspecified → Cardiac arrest, unspecified	30581	1.2%
Rheumatic heart disease, unspecified → Chronic obstructive pulmonary disease, unspecified	29440	1.16%
Unspecified diabetes mellitus without complications → Rheumatic heart disease, unspecified	29211	1.15%
Rheumatic heart disease, unspecified → Stroke, not specified as hemorrhage or infarction	28401	1.11%

CDC, NSF, NIH, and DHHS and does not constitute an endorsement of the individuals or their programs.

R. A. Hoffman and J. Venugopalan: Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA 30332 USA.

H. Wu: School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA.

P. Braun: Centers for Disease Control and Prevention (CDC), Atlanta, GA 30329 USA (co-corresponding author, e-mail: pabraun@cdc.gov).

M. D. Wang: Corresponding Author, Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA 30332 USA (phone: 404-385-2954; e-mail: maywang@gatech.edu).

©2018 IEEE. Reprinted, with permission, from R. A. Hoffman, H. Wu, J. Venugopalan, P. Braun, and M. D. Wang, “Intelligent mortality reporting with FHIR,” *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 5, pp. 1583–1588, 2018. DOI: 10.1109/JBHI.2017.2780891.

CHAPTER 9

CONCLUSIONS

Based on the work presented in this thesis, there are several avenues for future work building on this foundation. Finally, in closing, we will review some specific insights into how future investigators might maximize usability and adoption of decision support technology in the mortality reporting field.

9.1 Future Work

9.1.1 Application Platforms

Incremental revisions of the application have continued since the published research presented herein. Since publication, revisions have been made towards partial STU3 / R4 compatibility. Though this application has been enormously valuable as a prototype, future development on this platform may not make the most sense, given industry trends towards mobile and embedded applications. Future efforts could explore implementing mobile applications along the lines of the work presented here, in the hopes of improving the timeliness, accuracy, and completeness of mortality reporting Using FHIR apps and machine learning.

9.1.2 Mortality Pattern Mining

Sequential rules are well suited to the problem of mortality reporting, due to the strictly ordered and causally linked nature of the reported events in the NVSS mortality data. The sequential rules mined in Chapter 5 have multiple applications in mortality reporting, including decision support and augmenting the validity analysis published in Chapter 4. The next steps would be to include integration of this data with the next iteration of the mor-

tality reporting application. Future work could also look for supervised data sources that could support further validation and optimization of the thresholds and parameters chosen in this work.

9.2 Author’s Perspectives

In this section, the author presents his personal philosophy, insights, advice, and opinions on the broader future of the field.

9.2.1 Data Harmonization and FHIR

In this work, much has been made of the need for improved healthcare data harmonization and for standards (such as FHIR) to enable interoperability. It is important to consider the history of how the current state of affairs arose, and what insights this provides for the future.

9.2.1.1 Message- and Document-Based Standards

Historically, the siloing of biomedical data in obscure or proprietary formats has represented a competitive advantage for EMR vendors. This creates vendor lock-in, a situation where there are significant costs, both in time and resources, required to switch between vendors offering seemingly interchangeable solutions. This environment also creates the opportunity to market paid support services for the proprietary, esoteric software. These advantages align to disincentivize vendor support for strong interoperability.

Even in light of these incentives, valuable standards for limited healthcare data exchange have still arisen, such as HL7 v3 messaging or the Clinical Document Architecture (CDA) [117]. However, none of these standards on its own could full the role of FHIR. Messaging standards, designed to support real-time medical communication, enable a more limited, one-way form of interoperability, where “messages” from disparate sources can be pooled into one EMR database. A practical example of where messaging standards might

be used is to allow data from a device like a bedside vital signs monitor to be ingested and included in the medical record. This limited “interoperability of sources” serves the primary purpose of EMR systems — communication between medical practitioners enabling patient care — but is of very limited value for secondary research reuse of EMR data. Document-based standards, such as Clinical Document Architecture (CDA), are designed to allow the export of EMR data in a format emphasizing human-readability, but lacking in computer-readability and defined structure. Again, this reflects the primary use of EMR systems — enabling medical care — but is of limited use for secondary research re-use.

9.2.1.2 Meaningful Use and FHIR

With the primary uses of EMR data well supported, and competitive advantages in proprietary vendor lock-in, there may not have been any drive to support interoperability at the deep data level needed for research re-use. This changed in 2009 with the HITECH Act, a part of the stimulus plan enacted by the US Federal Government in response to the financial crisis of 2008 [134]. Under this policy, US Department of Health and Human Services (HHS) and Center for Medicare and Medicaid Services (CMS) invested in and incentivized the adoption of EMR systems. However, it also set standards for “meaningful use” of EMR systems, capabilities which must be achieved to meet program requirements. Meaningful Use included requirements for interoperability between EMR systems, and thus created a strong financial incentive for deep interoperability. This is the environment in which FHIR came about, with the first draft version published in 2011 [135].

9.2.1.3 Implications for the Future

Viewed in this light, it is not likely that FHIR would have been embraced by EMR vendors, especially to the degree that it has been, without the significant regulatory push for Meaningful Use. However, specific political efforts to reconsider Obama-era healthcare regulation, on top of the future uncertainty inherent to any government regulatory effort,

AllergyIntolerance

The AllergyIntolerance data models describe a patient's intolerance to a foreign substance and an associated reaction that occurs from exposure.

Technical Specifications:

- AllergyIntolerance.Read (DSTU2)
- AllergyIntolerance.Search (DSTU2)
- AllergyIntolerance.Create (STU3)
- AllergyIntolerance.Read (STU3)
- AllergyIntolerance.Search (STU3)
- AllergyIntolerance.Create (R4)
- AllergyIntolerance.Read (R4)
- AllergyIntolerance.Search (R4)

Figure 9.1: Fragmentation of support for FHIR versions is illustrated by this excerpt from Epic's FHIR interface documentation. Three different versions of the FHIR standard are implemented for various APIs of one resource, AllergyIntolerance, highlighted in gray. Retrieved from <https://open.epic.com/Interface/FHIR> on April 28, 2021.

mean that we cannot count on such pressure existing in the future. As such, we must make the most out of those standards which already have achieved significant support, and cannot depend on the adoption of new standards.

This advice is not just relevant to the future, but today as well. Efforts to improve and optimize FHIR have resulted in multiple version of several major version of FHIR [136]. Already, there is some significant fragmentation among vendors regarding what versions of FHIR are supported for which resources, with support sometimes varying even between APIs for a single resource, as shown in Figure 9.1.

Without significant external incentives to drive adoption, it is reasonable to assume that this fragmentation will continue with each additional future release version. In light of this, and keeping in mind FHIR's significant capacity for standards-friendly extension through the FHIR Profiling mechanism, it is possible that future releases of FHIR may begin to do more harm than good. If fragmentation of support continues, FHIR's usefulness as a true, deep interoperability standard may be degraded. This could be addressed by stopping, or at least significantly slowing, the pace of FHIR standard development. Development of the FHIR platform could continue, using a set of standard profiles as its primary tool for

driving change.

9.2.2 Maximizing Adoption and Value

With any decision support system, maximizing stakeholder buy-in and adoption is a prerequisite for realizing any value whatsoever from the system. No machine learning insight is so powerful as to impact patient care if no practicing clinicians ever see it.

This is a design problem as much as it is an engineering or business problem, and as with any design task the most important factor is keeping the wants and needs of the end-user foremost in the decision-making process. As a specific example related to this mortality informatics work, in a meeting with representatives of the CDC to hand over some parts of this project, there was some hesitancy about predictive analytics from the CDC representatives. While they appreciated the overall value of the work, they had concerns about programmatically proposing causes of death to certifiers, for fear of biasing their responses. This is a valuable example of the perspectives of end users versus the perspectives of data scientists, that there was such a bright-line distinction drawn between machine learning prediction and decision support, with some more “advanced” features actually undesirable to critical stakeholders.

This insight from our collaborators in the CDC highlights one of the most important points for adoption and buy-in: thoughtful design that considers human factors. It is not sufficient for a data scientist to be an engineer, developing a machine learning system that produces some insight about a biomedical system — a data scientist must also be a designer, understanding how that insight can be communicated to the end-user and how that could change the user’s behavior. Without meaningful insights and value to add, design is not enough to make a tool impactful; and without thoughtful and considered design, the potential value of a tool will never become practical, real-world impact.

Future data scientists, both in the specific field of mortality reporting and throughout the entire data science field, must constantly keep their specific end-users and their needs in

mind to ensure that they create tools that can effectively translate into real-world practice.

9.3 Closing

In this work, we have presented a mortality reporting application prototype, machine learning work to provide decision support insights to reporting clinicians, and thoughts on this work's place in the larger and evolving public health informatics field. It is the author's hope that this work will serve as an effective guide and a strong foundation for future research efforts in this area. With this work, we can realize the great potential value of accurate, timely, and complete mortality reporting data.

APPENDICES

APPENDIX A

ADDITIONAL TABLES AND RESULTS

A.1 Preface

This appendix contains supporting tables and additional results, expanding on the results presented earlier in this work where the tables were too large or disrupted the flow of the discussion.

Table A.1: Top 15 Frequent 1-Sequences, PrefixSpan with minsupport= 10^{-4} , not included in minsupport= 10^{-3}

Length	Support	Count
Vascular dementia	9.9916E-04	40956
Disorders of calcium metabolism	9.9782E-04	40901
Nicotine dependence	9.9653E-04	40848
Bipolar disorder, unspecified	9.9106E-04	40624
Hypertensive chronic kidney disease with stage 1 through stage 4 chronic kidney disease, or unspecified chronic kidney disease	9.8760E-04	40482
Cerebral infarction due to embolism of cerebral arteries	9.8101E-04	40212
Other specified intracranial injuries	9.7008E-04	39764
Alzheimer's disease with late onset	9.6457E-04	39538
Fracture of neck, unspecified	9.6452E-04	39536
Infectious gastroenteritis and colitis, unspecified	9.6071E-04	39380
Malignant neoplasm of tongue, unspecified	9.6032E-04	39364
Opioid use, unspecified	9.4405E-04	38697
Exposure to uncontrolled fire in building or structure	9.3385E-04	38279
Spinal stenosis	9.3134E-04	38176
Diffuse large B-cell lymphoma	9.2527E-04	37927

Table A.2: Top 15 Frequent 2-Sequences, PrefixSpan with minsupport= 10^{-4} , not included in minsupport= 10^{-3}

Length	Support	Count
Unspecified dementia → Cardiac arrhythmia, unspecified	9.9777E-04	40899
I64 → Lack of expected normal physiological development in childhood and adults	9.9762E-04	40893
Unspecified injury of face and head + Unspecified fall	9.9706E-04	40870
E149 → Chronic kidney disease (CKD)	9.9152E-04	40643
Chronic kidney disease, stage 5 → Atherosclerotic heart disease of native coronary artery	9.9145E-04	40640
Alzheimer's disease, unspecified → Acute myocardial infarction, unspecified	9.9013E-04	40586
Nicotine dependence → Malignant neoplasm of esophagus, unspecified	9.8977E-04	40571
Chronic kidney disease, unspecified + Anemia, unspecified	9.8928E-04	40551
Perforation of intestine (nontraumatic) → Sepsis, unspecified organism	9.8860E-04	40523
Atrial fibrillation and flutter → Cardiac arrhythmia, unspecified	9.8799E-04	40498
Unspecified dementia → Acute respiratory failure	9.8708E-04	40461
Atherosclerotic heart disease of native coronary artery + Unspecified kidney failure	9.8638E-04	40432
E149 → Heart failure, unspecified	9.8074E-04	40201
Unspecified dementia → Unspecified kidney failure	9.8035E-04	40185
Acute myocardial infarction, unspecified → Atherosclerotic heart disease of native coronary artery	9.7960E-04	40154

Table A.3: Top 10 Frequent 3-Sequences, PrefixSpan with minsupport= 10^{-4} , not included in minsupport= 10^{-3}

Length	Support	Count
E149 \rightarrow Heart failure \rightarrow Cardiac arrest, cause unspecified	9.8435E-04	40349
Heart failure \rightarrow Chronic obstructive pulmonary disease, unspecified \rightarrow Respiratory failure, unspecified	9.7889E-04	40125
Chronic obstructive pulmonary disease, unspecified \rightarrow Respiratory failure, unspecified \rightarrow Cardiac arrest, cause unspecified	9.7616E-04	40013
Nicotine dependence \rightarrow Pneumonia, unspecified organism \rightarrow Respiratory failure, unspecified	9.6859E-04	39703
E149 \rightarrow Essential (primary) hypertension \rightarrow Cardiac arrest, cause unspecified	9.6696E-04	39636
Essential (primary) hypertension + Nicotine dependence \rightarrow Cardiac arrest, cause unspecified	9.6637E-04	39612
E149 \rightarrow Chronic ischemic heart disease + Hypertensive heart disease without heart failure	9.5625E-04	39197
Chronic obstructive pulmonary disease, unspecified \rightarrow Acute respiratory failure + Chronic respiratory failure	9.4695E-04	38816
Pneumonia, unspecified organism \rightarrow Respiratory failure, unspecified \rightarrow Cardiac arrest, cause unspecified	9.4515E-04	38742
Asphyxiation \rightarrow Asphyxiation \rightarrow X70	9.4398E-04	38694

Table A.4: Top 10 Frequent 4-Sequences, PrefixSpan with minsupport= 10^{-4} , not included in minsupport= 10^{-3}

Length	Support	Count
Asphyxiation \rightarrow Asphyxiation \rightarrow Asphyxiation + X70	9.4256E-04	38636
Essential (primary) hypertension + E149 \rightarrow Atherosclerotic heart disease of native coronary artery \rightarrow Cardiac arrest, cause unspecified	6.6628E-04	27311
Essential (primary) hypertension + E149 \rightarrow Atherosclerotic heart disease of native coronary artery \rightarrow Acute myocardial infarction, unspecified	6.2737E-04	25716
Essential (primary) hypertension + Hyperlipidemia, unspecified \rightarrow Atherosclerotic heart disease of native coronary artery \rightarrow Acute myocardial infarction, unspecified	5.7106E-04	23408
Essential (primary) hypertension \rightarrow Atherosclerotic heart disease of native coronary artery \rightarrow Acute myocardial infarction, unspecified	5.5730E-04	22844
Essential (primary) hypertension + Hyperlipidemia, unspecified \rightarrow Atherosclerotic heart disease of native coronary artery \rightarrow Cardiac arrest, cause unspecified	4.8733E-04	19976
Chronic obstructive pulmonary disease, unspecified + Nicotine dependence \rightarrow Atherosclerotic heart disease of native coronary artery \rightarrow Acute myocardial infarction, unspecified	4.1739E-04	17109
Nicotine dependence \rightarrow Chronic obstructive pulmonary disease, unspecified \rightarrow Acute respiratory failure + Chronic respiratory failure	4.1002E-04	16807
Essential (primary) hypertension + Nicotine dependence \rightarrow Atherosclerotic heart disease of native coronary artery \rightarrow Acute myocardial infarction, unspecified	4.0290E-04	16515
Essential (primary) hypertension + Chronic obstructive pulmonary disease, unspecified + Nicotine dependence \rightarrow Malignant neoplasm of unspecified part of bronchus or lung	4.0053E-04	16418

Table A.5: Top 4 Frequent 5-Sequences, PrefixSpan with minsupport= 10^{-4} , not included in minsupport= 10^{-3}

Length	Support	Count
Essential (primary) hypertension + E149 + Hyperlipidemia, unspecified → Atherosclerotic heart disease of native coronary artery → Acute myocardial infarction, unspecified	1.2418E-04	5090
Essential (primary) hypertension + E149 → Atherosclerotic heart disease of native coronary artery → Acute myocardial infarction, unspecified	1.1764E-04	4822
Essential (primary) hypertension + Nicotine dependence + Hyperlipidemia, unspecified → Atherosclerotic heart disease of native coronary artery → Acute myocardial infarction, unspecified	1.1003E-04	4510
Essential (primary) hypertension + E149 + Hyperlipidemia, unspecified → Atherosclerotic heart disease of native coronary artery → Cardiac arrest, cause unspecified	1.0473E-04	4293

Table A.6: Top 15 Frequent 1-Sequences, PrefixSpan with minsupport= $2 \cdot 10^{-5}$, not included in minsupport= 10^{-4}

Length	Support	Count
Postprocedural (acute) (chronic) kidney failure	9.9609E-05	4083
Driver of special all-terrain or other off-road motor vehicle injured in traffic accident	9.9267E-05	4069
Open wound of lower back and pelvis	9.9218E-05	4067
Toxic encephalopathy	9.9072E-05	4061
Crushed chest	9.8779E-05	4049
Rheumatic diseases of endocardium, valve unspecified	9.8779E-05	4049
Inconclusive laboratory evidence of human immunodeficiency virus [HIV]	9.8584E-05	4041
Malignant neoplasm of accessory sinus, unspecified	9.8096E-05	4021
Sarcoidosis of lung	9.7828E-05	4010
Injury of unspecified body region	9.7608E-05	4001
Infection and inflammatory reaction due to prosthetic device, implant and graft in urinary system	9.7584E-05	4000
Myositis, unspecified	9.7486E-05	3996
Other specified disorders of the skin and subcutaneous tissue	9.7291E-05	3988
Incisional hernia with obstruction, without gangrene	9.7023E-05	3977
Irritable bowel syndrome without diarrhea	9.6998E-05	3976

Table A.7: Top 15 Frequent 2-Sequences, PrefixSpan with minsupport= $2 \cdot 10^{-5}$, not included in minsupport= 10^{-4}

Length	Support	Count
Other and unspecified atherosclerosis → Shock, unspecified	9.9999E-05	4099
Poisoning by, adverse effect of and underdosing of methadone + Y12	9.9999E-05	4099
Coagulation defect, unspecified → Other general symptoms and signs	9.9975E-05	4098
Atherosclerotic heart disease of native coronary artery + Rheumatic mitral valve disease, unspecified	9.9950E-05	4097
Hypertensive heart disease without heart failure → Bronchopneumonia, unspecified organism	9.9950E-05	4097
Malignant neoplasm of unspecified part of bronchus or lung → Anoxic brain damage, not elsewhere classified	9.9926E-05	4096
Unspecified kidney failure → Dilated cardiomyopathy	9.9901E-05	4095
Preterm [premature] newborn [other] → Neonatal cardiac dysrhythmia	9.9853E-05	4093
Unspecified dementia + Pulmonary embolism without acute cor pulmonale	9.9853E-05	4093
Influenza due to unidentified influenza virus with other respiratory manifestations → Respiratory failure, unspecified	9.9828E-05	4092
E149 + Other interstitial pulmonary diseases with fibrosis	9.9804E-05	4091
Pulmonary collapse → Respiratory failure, unspecified	9.9804E-05	4091
I64 → Chronic respiratory failure	9.9804E-05	4091
Malignant neoplasm of unspecified part of bronchus or lung → Cachexia	9.9779E-05	4090
Syncope and collapse → Cardiac arrest, cause unspecified	9.9731E-05	4088

Table A.8: Top 10 Frequent 3-Sequences, PrefixSpan with minsupport= $2 \cdot 10^{-5}$, not included in minsupport= 10^{-4}

Length	Support	Count
Pneumonia, unspecified organism \rightarrow Cardiac arrest, cause unspecified \rightarrow Anoxic brain damage, not elsewhere classified	9.9999E-05	4099
Hyperlipidemia, unspecified \rightarrow Chronic ischemic heart disease \rightarrow Acute myocardial infarction, unspecified	9.9999E-05	4099
Major depressive disorder, single episode, unspecified \rightarrow Atherosclerotic heart disease of native coronary artery \rightarrow Cardiac arrest, cause unspecified	9.9975E-05	4098
Essential (primary) hypertension + Hyperlipidemia, unspecified \rightarrow Cardiogenic shock	9.9950E-05	4097
Parkinson's disease \rightarrow Pneumonia, unspecified organism \rightarrow Cardiac arrest, cause unspecified	9.9926E-05	4096
Essential (primary) hypertension + Hypothyroidism, unspecified \rightarrow Malignant neoplasm of unspecified part of bronchus or lung	9.9926E-05	4096
Atrial fibrillation and flutter \rightarrow Alzheimer's disease, unspecified \rightarrow Lack of expected normal physiological development in childhood and adults	9.9926E-05	4096
Essential (primary) hypertension \rightarrow Cardiac arrest, cause unspecified + Respiratory arrest	9.9926E-05	4096
Type 2 diabetes mellitus without complications \rightarrow Other and unspecified atherosclerosis \rightarrow Cardiac arrest, cause unspecified	9.9901E-05	4095
Essential (primary) hypertension + Chronic obstructive pulmonary disease, unspecified \rightarrow Unspecified kidney failure	9.9901E-05	4095

Table A.9: Top 10 Frequent 4-Sequences, PrefixSpan with minsupport= $2 \cdot 10^{-5}$, not included in minsupport= 10^{-4}

Length	Support	Count
Atherosclerotic heart disease of native coronary artery + Essential (primary) hypertension + Atrial fibrillation and flutter \rightarrow Pneumonia, unspecified organism	9.9999E-05	4099
Nicotine dependence \rightarrow E149 \rightarrow Atherosclerotic heart disease of native coronary artery \rightarrow Cardiac arrest, cause unspecified	9.9779E-05	4090
Essential (primary) hypertension + Type 2 diabetes mellitus without complications + Hyperlipidemia, unspecified \rightarrow Heart failure	9.9731E-05	4088
E149 + Unspecified kidney failure \rightarrow Atherosclerotic heart disease of native coronary artery \rightarrow Heart failure	9.9609E-05	4083
Atherosclerotic heart disease of native coronary artery + Essential (primary) hypertension + E149 + Peripheral vascular disease, unspecified	9.9584E-05	4082
Essential (primary) hypertension + Unspecified dementia + E149 \rightarrow Atherosclerotic heart disease of native coronary artery	9.9438E-05	4076
Chronic obstructive pulmonary disease, unspecified \rightarrow Pneumonia, unspecified organism \rightarrow Respiratory failure, unspecified \rightarrow Cardiac arrest, cause unspecified	9.9316E-05	4071
Essential (primary) hypertension + Heart failure + Atrial fibrillation and flutter \rightarrow Pneumonia, unspecified organism	9.9267E-05	4069
Poisoning by, adverse effect of and underdosing of other and unspecified drugs, medicaments and biological substances \rightarrow X44 + Poisoning by and adverse effect of heroin + Poisoning by, adverse effect of and underdosing of benzodiazepines	9.9023E-05	4059
Nicotine dependence \rightarrow Pneumonia, unspecified organism \rightarrow Acute respiratory failure + Chronic respiratory failure	9.8877E-05	4053

Table A.10: Top 10 Frequent 5-Sequences, PrefixSpan with minsupport=2 · 10⁻⁵, not included in minsupport=10⁻⁴

Length	Support	Count
Essential (primary) hypertension + Hyperlipidemia, unspecified → Atherosclerotic heart disease of native coronary artery → Acute myocardial infarction, unspecified → Cardiac arrest, cause unspecified	8.9118E-05	3653
Essential (primary) hypertension + Chronic obstructive pulmonary disease, unspecified + Nicotine dependence → Atherosclerotic heart disease of native coronary artery → Acute myocardial infarction, unspecified	8.6581E-05	3549
Essential (primary) hypertension + Nicotine dependence + E149 → Atherosclerotic heart disease of native coronary artery → Acute myocardial infarction, unspecified	8.5142E-05	3490
Essential (primary) hypertension + Type 2 diabetes mellitus without complications + Hyperlipidemia, unspecified → Atherosclerotic heart disease of native coronary artery → Acute myocardial infarction, unspecified	7.8335E-05	3211
X44 + Poisoning by, adverse effect of and underdosing of other opioids + X45 + Poisoning by, adverse effect of and underdosing of benzodiazepines + Toxic effect of ethanol	7.6847E-05	3150
Atherosclerotic heart disease of native coronary artery + Essential (primary) hypertension + Chronic obstructive pulmonary disease, unspecified + Nicotine dependence → Malignant neoplasm of unspecified part of bronchus or lung	7.6189E-05	3123
Essential (primary) hypertension + Nicotine dependence + E149 + Hyperlipidemia, unspecified → Atherosclerotic heart disease of native coronary artery	6.8992E-05	2828
Essential (primary) hypertension + Nicotine dependence + Hyperlipidemia, unspecified → Atherosclerotic heart disease of native coronary artery → Cardiac arrest, cause unspecified	6.7967E-05	2786
Poisoning by, adverse effect of and underdosing of other and unspecified drugs, medicaments and biological substances → X44 + X45 + Poisoning by, adverse effect of and underdosing of benzodiazepines + Toxic effect of ethanol	6.5869E-05	2700
Alcohol use, unspecified → X42 + X45 + Poisoning by and adverse effect of heroin + Toxic effect of ethanol	6.5723E-05	2694

Table A.11: Top 15 Frequent 1-Sequences, PrefixSpan with minsupport= 10^{-5} , not included in minsupport= $2 \cdot 10^{-5}$

Length	Support	Count
Sepsis due to other specified staphylococcus	1.9980E-05	819
Cauda equina syndrome	1.9907E-05	816
Open wound of hip	1.9883E-05	815
Exposure to other inanimate mechanical forces	1.9883E-05	815
Malignant neoplasm of jejunum	1.9858E-05	814
Other disorders resulting from impaired renal tubular function	1.9834E-05	813
Disorders of fatty-acid metabolism	1.9834E-05	813
Edema of larynx	1.9834E-05	813
Congenital malformation of nervous system, unspecified	1.9834E-05	813
Unspecified maltreatment, confirmed	1.9834E-05	813
Neurosphilis, unspecified	1.9810E-05	812
Renovascular hypertension	1.9712E-05	808
Other specified chromosome abnormalities	1.9712E-05	808
Common arterial trunk	1.9712E-05	808
Adenovirus infection, unspecified	1.9712E-05	808

Table A.12: Top 15 Frequent 2-Sequences, PrefixSpan with minsupport=10⁻⁵, not included in minsupport=2·10⁻⁵

Length	Support	Count
Hypothyroidism, unspecified → Malignant neoplasm of stomach, unspecified	1.9980E-05	819
Endocarditis, valve unspecified → Hypotension, unspecified	1.9980E-05	819
Other and unspecified asthma → Secondary malignant neoplasm of liver and intrahepatic bile duct	1.9980E-05	819
Atrial fibrillation and flutter → Disease of pericardium, unspecified	1.9980E-05	819
Pneumonitis due to inhalation of food and vomit → Pleural effusion, not elsewhere classified	1.9980E-05	819
Poisoning by, adverse effect of and underdosing of other opioids + Poisoning by, adverse effect of and underdosing of other nonsteroidal anti-inflammatory drugs [NSAID]	1.9980E-05	819
Secondary malignant neoplasm of lung → Malignant neoplasm of uterus, part unspecified	1.9980E-05	819
Gastric ulcer, unspecified as acute or chronic, without hemorrhage or perforation → Acute myocardial infarction, unspecified	1.9980E-05	819
Exposure to excessive natural cold + Other specified effects of external causes	1.9980E-05	819
Lack of expected normal physiological development in childhood and adults → Anoxic brain damage, not elsewhere classified	1.9980E-05	819
Embolism and thrombosis of unspecified artery → Cerebral infarction, unspecified	1.9980E-05	819
Cerebral infarction due to embolism of cerebral arteries → Acute and subacute infective endocarditis	1.9980E-05	819
Essential (primary) hypertension → Malignant neoplasm of abdomen	1.9980E-05	819
Nicotine dependence → Hyperosmolality and hypernatremia	1.9980E-05	819
Cachexia → Foreign body in respiratory tract, part unspecified	1.9980E-05	819

Table A.13: Top 10 Frequent 3-Sequences, PrefixSpan with minsupport=10⁻⁵, not included in minsupport=2 · 10⁻⁵

Length	Support	Count
Nicotine dependence → Unspecified protein-calorie malnutrition → Malignant neoplasm of unspecified part of bronchus or lung	1.9980E-05	819
Essential (primary) hypertension → Heart disease, unspecified → Respiratory arrest	1.9980E-05	819
Nicotine dependence → Sequelae of cerebrovascular disease → Atherosclerotic heart disease of native coronary artery	1.9980E-05	819
Essential (primary) hypertension + Gastro-esophageal reflux disease without esophagitis → Other general symptoms and signs	1.9980E-05	819
Essential (primary) hypertension → Lack of expected normal physiological development in childhood and adults → Sepsis, unspecified organism	1.9980E-05	819
Unspecified dementia + Major depressive disorder, single episode, unspecified → Atrial fibrillation and flutter	1.9980E-05	819
Unspecified dementia + X590 → Respiratory failure, unspecified	1.9980E-05	819
Poisoning by, adverse effect of and underdosing of other and unspecified drugs, medicaments and biological substances + X44 + Poisoning by, adverse effect of and underdosing of other and unspecified narcotics	1.9980E-05	819
Nicotine dependence + Chronic kidney disease, unspecified → Acute kidney failure, unspecified	1.9980E-05	819
Nicotine dependence + Unspecified fall + Unspecified injury	1.9980E-05	819

Table A.14: Top 10 Frequent 4-Sequences, PrefixSpan with minsupport=10⁻⁵, not included in minsupport=2 · 10⁻⁵

Length	Support	Count
Essential (primary) hypertension + Atrial fibrillation and flutter → I64 → Aphagia and dysphagia	1.9980E-05	819
Atherosclerotic heart disease of native coronary artery + Essential (primary) hypertension + Nicotine dependence → Acute kidney failure, unspecified	1.9980E-05	819
Nicotine dependence + Type 2 diabetes mellitus without complications + Hyperlipidemia, unspecified → Essential (primary) hypertension	1.9980E-05	819
Essential (primary) hypertension + Type 2 diabetes mellitus without complications → Cardiomyopathy, unspecified → Cardiac arrest, cause unspecified	1.9980E-05	819
E149 + Hyperlipidemia, unspecified → Chronic ischemic heart disease + Hypertensive heart disease without heart failure	1.9980E-05	819
Fracture of head and neck of femur + X590 → Unspecified dementia → Lack of expected normal physiological development in childhood and adults	1.9956E-05	818
Poisoning by, adverse effect of and underdosing of other and unspecified drugs, medicaments and biological substances → Poisoning by, adverse effect of and underdosing of other and unspecified drugs, medicaments and biological substances + X44 → Pulmonary edema	1.9956E-05	818
Chronic ischemic heart disease + Hypertensive heart disease without heart failure → Unspecified fall + Unspecified injury	1.9956E-05	818
Nicotine dependence + Hyperlipidemia, unspecified → Atherosclerotic heart disease of native coronary artery → Chronic obstructive pulmonary disease, unspecified	1.9956E-05	818
Chronic obstructive pulmonary disease, unspecified + Nicotine dependence + Pneumonia, unspecified organism → Acute myocardial infarction, unspecified	1.9956E-05	818

Table A.15: Top 10 Frequent 5-Sequences, PrefixSpan with minsupport=10⁻⁵, not included in minsupport=2 · 10⁻⁵

Length	Support	Count
Essential (primary) hypertension + Heart failure + E149 → Chronic obstructive pulmonary disease, unspecified → Respiratory failure, unspecified	1.9980E-05	819
Essential (primary) hypertension + E149 + Hyperlipidemia, unspecified → Atherosclerotic heart disease of native coronary artery → Ischemic cardiomyopathy	1.9980E-05	819
Essential (primary) hypertension + Chronic obstructive pulmonary disease, unspecified + Unspecified dementia → Atherosclerotic heart disease of native coronary artery → Cardiac arrest, cause unspecified	1.9931E-05	817
Other psychoactive substance abuse → X42 + X45 + Poisoning by, adverse effect of and underdosing of cocaine + Toxic effect of unspecified alcohol	1.9931E-05	817
Essential (primary) hypertension + Unspecified dementia + Hyperlipidemia, unspecified → Atherosclerotic heart disease of native coronary artery → Cardiac arrest, cause unspecified	1.9907E-05	816
Other psychoactive substance abuse → X44 + Poisoning by, adverse effect of and underdosing of other opioids + Poisoning by, adverse effect of and underdosing of benzodiazepines + Poisoning by, adverse effect of and underdosing of other and unspecified antidepressants	1.9883E-05	815
X42 + Poisoning by, adverse effect of and underdosing of other opioids + X45 + Poisoning by and adverse effect of heroin + Toxic effect of ethanol	1.9883E-05	815
Essential (primary) hypertension + Hyperlipidemia, unspecified → E149 → Atherosclerotic heart disease of native coronary artery → Cardiac arrest, cause unspecified	1.9883E-05	815
Atherosclerotic heart disease of native coronary artery + Chronic obstructive pulmonary disease, unspecified + Nicotine dependence → Pneumonia, unspecified organism → Sepsis, unspecified organism	1.9883E-05	815
Essential (primary) hypertension + Chronic obstructive pulmonary disease, unspecified + E149 → Acute myocardial infarction, unspecified → Cardiac arrest, cause unspecified	1.9858E-05	814

APPENDIX B

DEATH CERTIFICATE DATA MAPPING

B.1 Preface

This appendix contains the mapping from the HL7 Vital Records Domain Analysis Model (VR DAM) to FHIR elements that we developed in collaboration with the VR DAM Working Group ca. 2017. It is presented in a pipe (|) delimited format, and can be easily converted to a table.

Work in this appendix is adapted from the publication "Intelligent Mortality Reporting with FHIR" [62], of which Ryan Hoffman is the first and leading author. This publication is copyrighted by IEEE, and reused with permission. For detailed copyright disclosure, please see Appendix F. It has been modified throughout to comply with formatting requirements and to better integrate with the rest of this work.

©2018 IEEE. Reprinted, with permission, from R. A. Hoffman, H. Wu, J. Venugopalan, P. Braun, and M. D. Wang, "Intelligent mortality reporting with FHIR," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 5, pp. 1583–1588, 2018. DOI: 10.1109/JBHI.2017.2780891.

B.2 Mapping

To comply with formatting constraints, the mapping table presented here is simplified. The full table, including details such as terminology set OIDs, can be found at [62].

Table B.1: VR DAM to FHIR Resource Mapping

DAM Class.Attribute	DAM Type	FHIR Resource.Attribute
CauseOfDeathDescription.causeofDeathDescription	char	Condition.code.text OR code.coding
CauseOfDeathDescription. causeOfDeathIntervalSequenceOrder	char	Condition.onset
CauseOfDeathInformation.autopsyFindingsIndicator	boolean	Observation.valueCodeableConcept
CauseOfDeathInformation.autopsyPerformedIndicator	boolean	Observation.valueCodeableConcept
CauseOfDeathInformation.mannerOfDeathCode	char:MannersOfDeath	Observation.valueCodeableConcept
CauseOfDeathInformation.pregnancyStatusCode	char:PregnancyStatuses	Observation.valueCodeableConcept
CauseOfDeathInformation. causeOfDeathSignificantConditionDescriptiveText	char	Observation.valueString
CauseOfDeathInformation.tobaccoUseCode	char:Contributory TobaccoUses	Observation.valueCodeableConcept
ClinicianCoronerMedicalExaminer.certifierAddress	AddressLocation	Practitioner.address
ClinicianCoronerMedicalExaminer.certifierName	PersonName	Practitioner.name
ClinicianCoronerMedicalExaminer.certifierRoleCode	char:CertifierRoles	Practitioner.practitionerRole
ClinicianCoronerMedicalExaminer.certifierSignature	char	Composition.attester

Table B.1 (continued)

DAM Class.Attribute	DAM Type	FHIR Resource.Attribute
ClinicianCoronerMedicalExaminer.certifierTitle	char	Practitioner.practitionerRole
ClinicianCoronerMedicalExaminer.certifierIdentifier	char	Practitioner.qualification
CodedCauseOfDeath.causeOfDeathUnderlyingCode	char	Observation.valueCodeableConcept
Death.bodyNotRecoveredIndicator	boolean	Observation.valueCodeableConcept
Death.certificationDate	char	Observation.valueDateTime
Death.decedentCertifierKnownName	PersonName	Patient.name
Death.decedentDateOfDeath	char	Patient.deceased
Death.decedentDeathDateEstimatedIndicator	boolean	Observation.valueCodeableConcept
Death.decedentTimeOfDeath	char	Patient.deceased
Death.decedentDeathTimeEstimatedIndicator	boolean	Observation.valueCodeableConcept
Death.facilityName	char	Location.name
Death.locationOfDeathDescription	char	Location.description
Death.placeOfDeathAddress	AddressLocation	Location.address
Death.placeOfDeathCode	char:PlacesOfDeath	Observation.valueCodeableConcept
Death.placeOfDeathDescription	char	Observation.valueString
Death.pronouncedDateOfDeath	char	Observation.valueDateTime

Table B.1 (continued)

DAM Class.Attribute	DAM Type	FHIR Resource.Attribute
Death.pronouncedTimeOfDeath	char	Observation.valueDateTime
Death.pronouncerIdentifier	char	Practitioner.qualification
Death.pronouncerSignature	char	Composition.attester
Death.pronouncerSignatureDate	char	Composition.attester.time
Death.pronouncerSignatureTime	char	Composition.attester.time
Death.referredToMedicalExaminerIndicator	boolean	Observation.valueCodeableConcept
Decedent.decedentAgeAtTimeOfDeath	char	Patient.birthDate
Decedent. decedentAgeAtTimeOfDeathUnitOfMeasure	char:TimeDurationUnits	Patient.birthDate
Decedent.decedentAliasName	PersonName (0..*)	Patient.name
Decedent.decedentDateOfBirth	char	Patient.birthDate
Decedent.decedentEducationLevelCompletedCode	char:EducationLevels	Observation.valueCodeableConcept
Decedent. decedentEducationLevelMissingValueReason	char:EducationLevelMVRs	Observation.dataAbsentReason
Decedent.decedentFatherName	PersonName	RelatedPerson.name

Table B.1 (continued)

DAM Class.Attribute	DAM Type	FHIR Resource.Attribute
Decedent.	char:EthnicityMVRs	Observation.dataAbsentReason
decedentHispanicOriginMissingValueReason		
Decedent.decedentBusinessOrIndustryDescription	char	Observation.valueString
Decedent.decedentLegalName	PersonName	Patient.name
Decedent.decedentMaritalStatusCode	char:MaritalStatuses	Patient.maritalStatus
Decedent.decedentMotherName	PersonName	RelatedPerson.name
Decedent.decedentNamePriorToFirstMarriage	PersonName	Patient.name
Decedent.decedentOccupationDescription	char	Observation.valueString
Decedent.decedentPlaceOfBirthCity	char	Patient.extension-birthplace
Decedent.decedentPlaceOfBirthCountry	char	Patient.extension-birthplace
Decedent.decedentPlaceOfBirthState	char	Patient.extension-birthplace
Decedent.decedentRaceMissingValueReason	char:RaceMVRs	Observation.dataAbsentReason
Decedent.decedentResidentialAddress	AddressLocation	Patient.address
Decedent.decedentSexCode	char:Genders	Patient.gender
Decedent.decedentSpouseName	PersonName	RelatedPerson.name
Decedent.decedentSSN	char	Observation.valueString

Table B.1 (continued)

DAM Class.Attribute	DAM Type	FHIR Resource.Attribute
Decedent.decedentSSNMissingValueReason	char:SocialSecurityNumber	Observation.dataAbsentReason
	MVRs	
Decedent.decedentUSArmedForcesIndicator	boolean	Observation.valueCodeableConcept
Disposition.dispositionMethodCode	char	Observation.valueCodeableConcept
Disposition.dispositionMethodDescription	char	Observation.valueString
Disposition.placeOfDispositionCityOrTown	char	Location.address
Disposition.placeOfDispositionName	char	Location.name
Disposition.placeOfDispositionState	char	Location.address
EntityAxisCode.causeOfDeathCausalSequence	int	Condition(
		CauseOfDeathDescription)
EntityAxisCode.causeOfDeathCode	char:DiseasesOrConditions	Condition.entityAxisCodes.coding
EntityAxisCode.causeOfDeathIntraSequenceOrder	int	Condition.entityAxisCodes.text
FuneralFacility.facilityLicenseIdentifier	char	Organization.identifier
FuneralFacility.funeralDirectorIdentifier	char	Practitioner.qualification
FuneralFacility.funeralDirectorSignature	char	Composition.attester
FuneralFacility.funeralFacilityAddress	AddressLocation	Organization.address

Table B.1 (continued)

DAM Class.Attribute	DAM Type	FHIR Resource.Attribute
FuneralFacility.funeralFacilityName	char	Organization.name
HispanicOrigin.hispanicOriginCode	char:HispanicOrigins	Observation.valueCodeableConcept
HispanicOrigin.hispanicOriginDescription	char	Observation.valueString
Informant.informantDecedentRelationshipDescription	char	RelatedPerson.relationship
Informant.informantMailingAddress	AddressLocation	RelatedPerson.address
Informant.informantName	PersonName	RelatedPerson.name
Injury.injuryAddress	AddressLocation	Condition.location (Reference(Location))
Injury.injuryAtWorkIndicator	boolean	Condition.workplaceInjury (boolean)
Injury.injuryDate	char	Condition.onsetDateTime
Injury.injuryDateEstimatedIndicator	boolean	Condition.dateEstimated (boolean)
Injury.injuryOccurrenceDescription	char	Condition.code
Injury.injuryPlaceDescription	char	Condition.location (Reference(Location))
Injury.injuryTime	char	Condition.onsetDateTime

Table B.1 (continued)

DAM Class.Attribute	DAM Type	FHIR Resource.Attribute
Injury.injuryTimeEstimatedIndicator	boolean	Condition.timeEstimated (boolean)
Injury.injuryTransportationRelationshipCode	char:TransportationRoles	Condition.transportationRelationship (CodeableConcept)
Injury.injuryTransportationRelationshipDescription	char	Condition.transportationRelationship (CodeableConcept)
Race.raceCode	char:Races	Observation.valueCodeableConcept
Race.raceDescription	char	Observation.valueString
Record.AxisCode.causeOfDeathCode	char:DiseasesOrConditions	Condition.recordAxisCodes.coding
VitalRecordAmendment.VRAmendedDate	char	Provenance.recorded
VitalRecordAmendment.VRAmendmentNew Value	char	Provenance.target[0]
VitalRecordAmendment.VRAmendmentOld Value	char	Provenance.target[1]
VitalRecordAmendment.VRAmendmentType	char	Provenance.activity
VitalRecordCertification.VRAmendedRecordIndicator	boolean	Composition.status
VitalRecordCertification.VRAuxiliaryFileNumber	char	Composition.auxiliaryIdentifier
VitalRecordCertification. VRCertificationOrReportFileNumber	char	Composition.auxiliaryIdentifier

Table B.1 (continued)

DAM Class.Attribute	DAM Type	FHIR Resource.Attribute
VitalRecordCertification.VRDateFiledByRegistrar	char	Composition.date
VitalRecordCertification. VRLinkedBirthCertificateNumber	char	Composition.section[0].entry
VitalRecordCertification. VRLinkedDeathCertificateNumber	char	Composition.section[0].entry
VitalRecordCertification.VRMannerOfFilingCode	char	Composition.event.code
VitalRecordCertification.VRRRegistrarSignature	char	Composition.attester
VitalRecordCertification.VRUniqueIdentifier	char	Composition.identifier.value
VitalRecordCertification. VRVoidCertificateOrReportIndicator	char	Composition.status

APPENDIX C

COMPARISON OF NORMALIZATION ALGORITHMS FOR CROSS-BATCH COLOR SEGMENTATION OF HISTOPATHOLOGICAL IMAGES

C.1 Preface

Work in this chapter is adapted from the publication "Comparison of normalization algorithms for cross-batch color segmentation of histopathological images" [137], of which Ryan Hoffman is the first and leading author. This publication is copyrighted by IEEE, and reused with permission. For detailed copyright disclosure, please see Appendix F. It has been modified throughout to comply with formatting requirements and to better integrate with the rest of this work.

©2017 IEEE. Reprinted, with permission, from R. A. Hoffman, H. Wu, J. Venugopalan, P. Braun and M. D. Wang, "Comparison of normalization algorithms for cross-batch color segmentation of histopathological images", in Proceedings of the 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Aug. 2014, DOI: 10.1109/BHI.2017.7897235.

C.2 Abstract

Automated processing of digital histopathology slides has the potential to streamline patient care and provide new tools for cancer classification and grading. Before automatic analysis is possible, quality control procedures are applied to ensure that each image can be read consistently. One important quality control step is color normalization of the slide image, which adjusts for color variances (batch-effects) caused by differences in stain preparation and image acquisition equipment. Color batch-effects affect color-based features and reduce the performance of supervised color segmentation algorithms on images acquired

separately. To identify an optimal normalization technique for histopathological color segmentation applications, five color normalization algorithms were compared in this study using 204 images from four image batches. Among the normalization methods, two global color normalization methods normalized colors from all stain simultaneously and three stain color normalization methods normalized colors from individual stains extracted using color deconvolution. Stain color normalization methods performed significantly better than global color normalization methods in 11 of 12 cross-batch experiments ($p < 0.05$). Specifically, the stain color normalization method using k-means clustering was found to be the best choice because of high stain segmentation accuracy and low computational complexity.

C.3 Introduction

Histopathology is an integral part of the detection, monitoring, and research of cancer. Digital histopathology slides, also known as whole-slide images (WSIs), are a modern, high-resolution tool to store the information from a tissue sample fixed on a glass slide for later analysis. WSIs have uses in training, healthcare record management, and telemedicine [138]. The availability of large, public banks of WSIs such as the Cancer Genome Atlas (TCGA) has created a growing area of research devoted to the automated analysis of these images [139]. Reliable, accurate, and automatic processing of WSIs has the potential to cut costs, improve patient outcomes, and take modern pathology into environments not previously possible [140].

Before useful automated processing, digital histopathology slides must undergo a number of quality control steps. These quality control steps ensure that no artifacts or technical variations, created during image acquisition, affect the biological data and the performance of image analysis and machine learning algorithms. Due to the great variability that exists between slides processed using different equipment or reagents, color normalization, which will normalize colors across batches, is a vital quality control step in the slide analysis pro-

cess [141].

Tissue samples are stained to highlight different cellular structures. For instance, in the most common slide staining for histopathology — H&E or hematoxylin and eosin — hematoxylin stains nuclear structures purple or blue, and eosin stains cytoplasmic structures pink. Analysis of WSIs often requires that the contributions from these two stains be extracted and considered separately. For example, nuclear segmentation algorithms may begin by identifying high concentrations of hematoxylin. The shape and texture features of the isolated stain channels have been shown to have diagnostic value in classification problems. Accurate normalization is thus a necessary first step for extracting any features based on color, texture, or stain segmentation. In this paper, the role of color normalization methods in a supervised stain segmentation pipeline is studied.

Researchers have previously studied color normalization methods for histopathological images [141, 142, 143]. Among the published research, there are two categories of methods: global color normalization that normalizes colors of all pixels irrespective of their stain and stain color normalization that separates stains and then normalizes each stain individually. The latter category would be ideal if the stains could be separated accurately. However, unsupervised stain segmentation of histopathological images is often not straightforward. Kothari et al. proposed two global color normalization methods that normalize images using quantile normalization of all pixels in the RGB color space and the quantile normalization of the unique color map [4]. Magee et al. proposed a stain color normalization method that roughly separates stains using color deconvolution and clustering and then normalizes each stain individually using Reinhard’s method [143, 144]. In their study, Magee et al. used a variational Bayesian Gaussian mixture model to cluster the areas where each stain is present in deconvolved images and compared original and normalized colors after normalization rather than comparing segmentation performance. However, variational Bayesian methods are computationally complex. Thus, in this study, two additional stain normalization procedures are developed that use the less complex k-

means clustering and expectation-maximization methods to identify stain classes, rather than variational Bayesian methods.

In summary, a quantitative comparison of the impact of five normalization algorithms, two global normalization and three stain color normalization methods, on color segmentation performance is presented in this paper.

C.4 Methods

C.4.1 Data

Manually curated portions of digital histopathology slides from four separately acquired image batches are used in this study. Two image batches/datasets, ovarian serous adenocarcinoma (OV) and glioblastoma multiforme (GBM), are from The Cancer Genome Atlas (TCGA). Images in these datasets are cropped sections of 1024x1024 pixels. The other two datasets, renal cell carcinoma (RCC1 & RCC2), were acquired at Emory University. Images in renal datasets are cropped sections of 1600x1200 pixels. In total, 204 images are considered, out of which 50 were derived from OV samples, 52 from GBM, 55 from RCC1, and 47 from RCC2.

Ground truth segmentation for all images is obtained using an interactive system, where an experienced user selected sample pixels belonging to one of the four classes: hematoxylin, eosin, erythrocyte, and stain-free regions. All the image pixels were grouped into one of the four classes based on their Euclidian distance to selected pixels. These ground truth labels are used for training segmentation classifiers and evaluating segmentation performance.

C.4.2 Color Normalization Algorithms

Color normalization methods affect the value of color features and performance of color segmentation algorithms. In this paper, performance of color segmentation using five candidate normalization algorithms (as outlined in Figure C.1) is studied. Previous work pub-

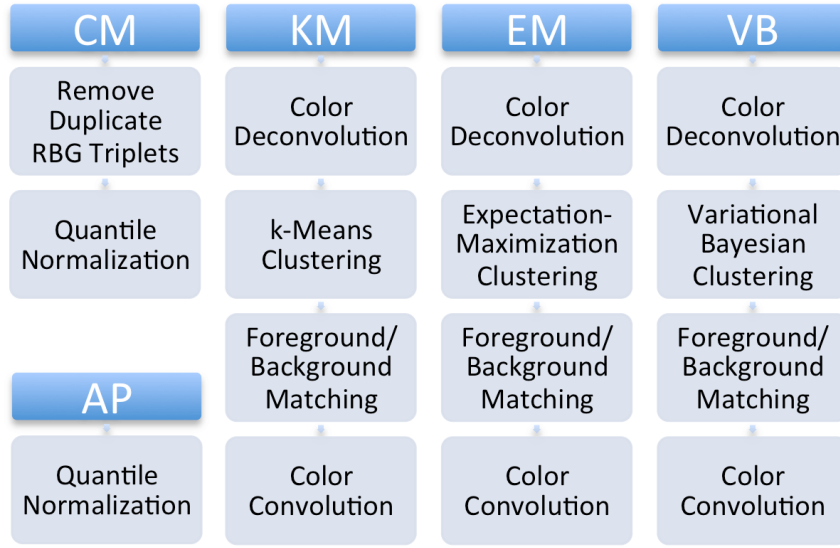


Figure C.1: Normalization Algorithm Candidates. All five candidate algorithms are compared.

lished color segmentation results using two global color normalization methods: all pixel and color map normalization, and as such, it is used here as a control [141]. The three other methods are derived from the color normalization methods published in [143, 144]. These methods use stain deconvolution as a first step, splitting the sample image into separate channels for hematoxylin and eosin staining. Three different clustering algorithms are then applied to segment those channels into stain is present / is not present regions. After normalizing different stains in a sample image to stains in a reference image, sample image stains are convolved to produce a normalized sample image.

C.4.2.1 Global Color Normalization

All-pixel quantile normalization performs simple quantile normalization of the red (R), green (G), and blue (B) color channel intensity distributions from the sample image to a reference image [141]. In quantile normalization, the largest value from the sample is replaced by the largest value from the reference, the second largest sample value by the second largest reference value, etc. The color distributions of the quantile normalized sam-

ple image will then share important statistical properties such as the mean and variance with the color distributions of the reference image.

In color map normalization, a color map is first constructed for the reference image by creating a list of every unique RGB triplet that occurs within the image [141]. This process is repeated with the target image to create its color map. Quantile normalization is then used to normalize individual color channel distributions for the sample color map to the color channel distributions of the reference color map.

C.4.2.2 Stain Color Normalization

Stain color normalization normalizes each stain separately using the following steps: (1) stain separation, (2) clustering, (3) multimodal color deconvolution (CVD-MM) normalization [142], and (4) stain combination.

(1) Stain Separation

First, the RGB image I produced over the background I_0 is broken down into channels representing the contribution from each stain A . This is accomplished using a fixed optical density matrix Q based on the nominal color of each stain: hematoxylin and eosin [142, 145].

$$Q = \begin{bmatrix} 0.65 & 0.704 & 0.285 \\ 0.072 & 0.990 & 0.105 \\ 0.6218 & 0 & 0.7831 \end{bmatrix}, A = \log_{10} \frac{I}{I_0} Q^{-1}$$

(2) Clustering

Color deconvolution returns grayscale images corresponding to each stain, where intensity at each pixel represents stain intensity. Pixels may have some intensity in each stain channel. Various clustering methods are employed to separate the foreground (strong staining) and background (weak staining) classes for each stain. The three clustering algorithms are employed and compared in this study are k-means, expectation-maximization

for a Gaussian mixture, and variational Bayesian inference for a Gaussian mixture. All three clustering techniques were run with the number of classes constrained at $k = 2$.

The k-means algorithm randomly chooses two cluster centers, adds each of the observations to the nearest of those clusters, then updates the cluster center and iterates until it converges to a final solution when the cluster assignments no longer change between iterations [30]. In this implementation, Euclidian distances to cluster centers are used.

The expectation-maximization algorithm used in this study works by estimating the mean and variance parameters of a mixture of two Gaussian distributions that fit the data. The expectation-maximization process consists of two steps. First, the probability that each observation falls into each distribution is determined and each observation is assigned a preliminary class based on the highest probability. The next step assumes that the labels assigned in the first are all true, and generates new parameters to best fit those classes. The EM algorithm used in this study is specifically fitting a Gaussian mixture model, rather than optimizing Euclidian distances to cluster centers as in k-means.

Rather than finding an approximation of the posterior distribution as in expectation-maximization algorithms, the variational Bayesian method attempts to estimate the posterior distribution for all unknown variables [26]. The main difference between variational Bayesian and expectation-maximization is that variational Bayesian calculates the probable distributions of the variables, rather than estimating the parameter values (such as Gaussian mixture means) directly.

Figure C.2 shows the color deconvolution and clustering processes for a sample image from RCC1, where the image is broken down into hematoxylin and eosin “channels” before foreground and background clustering.

(3) CVD-MM normalization

A similar deconvolution and clustering takes place for both sample and reference images. Once this is done, the clusters of the sample image are normalized to match the mean and variance of those clusters found in the reference image by the CVD-MM method de-

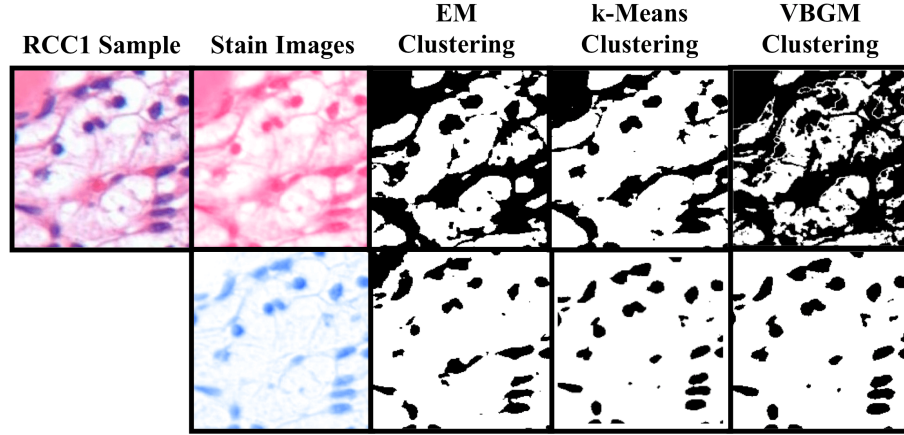


Figure C.2: Clustering comparison. A sample image from the RCC1 data set is segmented into hematoxylin and eosin channels. These channels are then separated into foreground (strong staining) background (weak staining) clusters by each of three clustering algorithms.

scribed by Magee et al. [142], which is conceptually similar to Reinhard’s method [144] implemented in a stain-specific color space.

Reference Gaussian distributions are generated using the means and standard deviations from the clustering step. Background and foreground weights are calculated at each pixel by linear interpolation of the reference Gaussian distributions. A separate saturated-pixel weight is defined such that near-white pixels will not be significantly changed. These weights and reference distributions are combined to yield a normalized stain component pixel [142].

(4) Stain Combination

The normalized stain-domain image is then converted back to the RGB color space using color convolution, in an inverse operation of the deconvolution performed in step (1).

C.4.3 Stain Segmentation

Images are segmented using a four-step, supervised color segmentation system [140]. First, a test image is normalized to a standard reference image using one of the five color normalization methods, discussed in the previous section. Second, every pixel in the test image

based on its RGB color values is classified as one the of four tissue classes using a supervised classifier. The system uses a 4-class linear discriminant (LDA) classifier, which is trained using ground truth labels and RGB colors values of the reference image. The four tissue classes refer to the hematoxylin, eosin, erythrocyte, and stain-free regions of the image. The first and second steps are repeated with ten different references resulting in ten slightly different segmentations. Ten top references are selected from the same batch using internal cross-validation. More details on cross-validation and validation are described in the next section. Third, the segmentation labels are combined for each pixel using max-voting. Because images are segmented in the normalized color space, decision planes for each segmented tissue class may be irregular when transformed into the original color space. Therefore, to refine the segmentation in the original color space, a classifier is trained using the segmentation labels from the third step and the image's original RGB color values [140].

C.4.4 Validation

The normalization methods are compared using the performance of the color segmentation system, when images are normalized with any method in the first step. The performance is assessed for each binary combination of four batches, where one batch is the train set while another is the test set. In total, 12 cross-batch combinations are assessed during the validation process.

The performance of normalization methods and classifier model depends on the selection of reference images. Therefore, multiple images are selected to avoid bias due to the selection of any single reference image. Cross-validation within a batch is used to select the top ten references for a batch. First, each image within the data set is used as a reference to normalize and segment all of the other images, after which the mean stain segmentation accuracy is recorded. This is repeated for all members of a data set, after which the 10 highest scoring images are saved as the reference set for that batch.

C.5 Results and Discussion

Table C.1 lists the mean and standard deviation of the segmentation accuracy using two global color normalization methods—all pixel (AP) and color map (CM)—and four stain color normalization methods—k-means (KM), expectation-maximization (EM), and variational Bayesian inference (VB)—for all cross-batch experiments. As reported in previous work as well, among global color normalization methods, CM performs better than AP [140]. However, in most cases stain color normalization methods outperform global color normalization methods. This was expected because stain color normalization normalized each stain separately and prevents color intermixing between stains. To more statistically compare these methods, Student’s t-test was performed between the performances using different normalization methods within each test case, i.e., a train and test batch combination. The following can be concluded based on t-test p-values: (1) There is no statistical difference between stain color normalization methods (KM, EM, and VB) using different clustering methods, (2) In all but one case (RCC2 train set and RCC1 test set), CM performs statistically better than or equivalent to AP, and (2) In all but one case (OV train set and RCC2 test set), KM performs statistically better than or equivalent to CM. Statistical significance was established using $p < 0.05$. Figure C.3 illustrates qualitative differences in the segmentation masks generated by the KM, EM, and VB algorithms.

Although there was no significant difference in the performance using either of the stain color normalization methods, there was a significant difference in computational complexity between the KM, EM, and VB clustering methods. To quantify the differences in performance between these three algorithms, a single standardized sample from the RCC1 data set was normalized against 10 randomly selected reference images, and the total time elapsed was recorded. The results are reported in Table C.2. KM was the fastest, with 10 normalizations taking only 29.28 seconds. It was found to be approximately 6.5x faster than the EM procedure and over 17x faster than VB. Thus, based on our experiments, KM

Table C.1: Average stain segmentation accuracy using color deconvolution normalization (KM, EM, and VB) and quantile normalization (AP and CM) methods. Performance of KM methods are highlighted in **bold** where either (1) performance is significantly better than all other methods for the particular test case, or (2) performance is not significantly different from any other methods. (Student's t-tests, $p < 0.05$).

Testing		OV		GBM		RCC1		RCC2	
Training	Method	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
OV	KM	-	-	83	7.2	91	6.4	80	12.6
	EM	-	-	83	7.1	91	6.4	80	12.6
	VB	-	-	83	7.2	90	6.5	80	12.6
GBM	KM	94	6.5	-	-	92	7.6	84	10
	EM	94	6.5	-	-	93	7.5	84	10
	VB	94	6.5	-	-	92	7.5	85	10
RCC1	KM	92	4.2	87	6.7	-	-	91	6.8
	EM	92	4.2	87	6.7	-	-	91	6.8
	VB	92	4.2	85	7	-	-	91	6.8
RCC2	KM	91	4.5	87	6.5	96	9.2	-	-
	EM	91	4.5	87	6.5	96	9.2	-	-
	VB	91	4.5	87	6.7	95	9.3	-	-

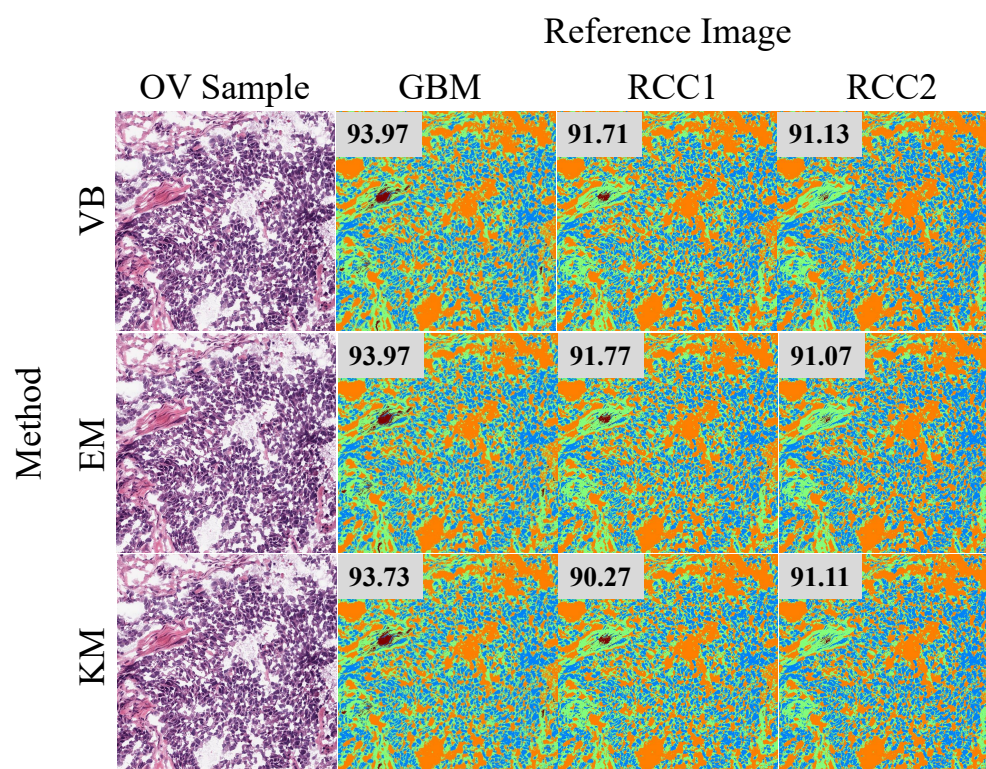


Figure C.3: Segmentation accuracy results for a single sample. A single OV sample (left column) is segmented after normalization using three different algorithms (rows) against three different references (columns). Color segmentation accuracy is shown in the top-right of each segmented color map.

Table C.2: Performance comparison for KM, EM, and VB

Method	N	Time (s)
KM	10	36.38
EM	10	191.1
VBGM	10	475.8

is clearly the ideal choice because it performs better or equivalent to global normalization methods and it is fastest among stain color normalization methods.

C.6 Conclusion

Color normalization is an important quality control step for histopathological images to insure accurate downstream processing of these images. In this work, based on the performance of color segmentation system, five color normalization methods were compared. Among these methods, three methods were previously published but two were novel extensions of an existing method. One of our novel extensions using k-means clustering was found to be the optimal normalization algorithm based on high segmentation accuracy and low computational time. This preliminary study used only four batches of manually curated images. In future work, this work would be extended by evaluating several other normalization methods on more image batches and complete whole-slide images.

C.7 Acknowledgements

The authors thank Sumit Joshi for his contributions towards implementation and assistance in collecting results.

Ryan A. Hoffman and Sonal Kothari, PhD are with the Department of Biomedical Engineering, Georgia Institute of Technology, Atlanta, GA 30332.

May D. Wang, PhD is with the Department of Biomedical Engineering, Winship Cancer Institute, Parker H. Petit Institute of Bioengineering and Biosciences, Institute of People and Technology, Georgia Institute of Technology and Emory University, Atlanta, GA (e-

mail: maywang@bme.gatech.edu).

APPENDIX D

A HIGH-RESOLUTION TILE-BASED APPROACH FOR CLASSIFYING BIOLOGICAL REGIONS IN WHOLE-SLIDE HISTOPATHOLOGICAL IMAGES

D.1 Preface

Work in this chapter is adapted from the publication "A High-Resolution Tile-Based Approach for Classifying Biological Regions in Whole-Slide Histopathological Images" [146], of which Ryan Hoffman is the first and leading author. This publication is copyrighted by Springer, and reused with permission. For detailed copyright disclosure, please see Appendix F. It has been modified throughout to comply with formatting requirements and to better integrate with the rest of this work.

Reprinted by permission from Springer: Proceedings of the International Federation for Medical and Biological Engineering (IFMBE) "A High-Resolution Tile-Based Approach for Classifying Biological Regions in Whole-Slide Histopathological Images", R.A. Hoffman, S. Kothari, J.H. Phan, and M.D. Wang, ©2014 Springer

D.2 Abstract

Computational analysis of histopathological whole slide images (WSIs) has emerged as a potential means for improving cancer diagnosis and prognosis. However, an open issue relating to the automated processing of WSIs is the identification of biological regions such as tumor, stroma, and ne-crotic tissue on the slide. We develop a method for classifying WSI portions (512x512-pixel tiles) into biological regions by (1) extracting a set of 461 image features from each WSI tile, (2) optimizing tile-level prediction models using nested cross-validation on a small (600 tile) manually annotated tile-level training set, and (3) validating the models against a much larger (1.7×10^6 tile) data set for which ground

truth was available on the whole-slide level. We calculated the predicted prevalence of each tissue region and compared this prevalence to the ground truth prevalence for each image in an independent validation set. Results show significant correlation between the predicted (using automated system) and reported biological region prevalences with $p < 0.001$ for eight of nine cases considered.

D.3 Introduction

Histopathology plays a vital support role in oncology. Whole-slide histopathological images (WSIs) are digital images of sectioned and stained tissue samples that are scanned and digitally recorded at high resolutions. In traditional use, WSIs have enabled (1) more streamlined record-keeping, (2) training of laboratory technicians, and (3) offsite consultation for diagnoses and prognoses [147, 138]. WSIs have emerged as a growing area of interest in image processing and imaging informatics. WSIs have been shown to contain significant diagnostic and prognostic data, which can be extracted in reproducible and quantitative ways to support fast, accurate decisions [148, 140]. The emergence of large, multi-modal cancer data repositories, such as The Cancer Genome Atlas (TCGA), has also stimulated research in automated informatics methods for WSIs [139].

Significant computational, experimental, and biological challenges exist when attempting to use WSIs for diagnostic or prognostic applications. The size and resolution of the images pose computational challenges, with many images having dimensions on the order of gigapixels. This has limited many previous studies to relatively small data sets, e.g., on the order of dozens of WSIs. Experimentally, WSIs are subject to many of the same image artifacts as traditional slides, including tissue folds and non-tissue markings in the image [149]. WSIs are very heterogeneous in terms of the types of cells and tissues captured. Thus, a major biological challenge is region-of-interest (ROI) selection and classification, which is essential for accurate diagnosis. There is a pressing need for automated WSI informatics methods for artifact correction, feature extraction, ROI selection, and image

classification [140].

This paper focuses on a specific biological challenge of ROI classification. Three distinct tissue types can appear in any cancer WSI: stroma, tumor tissue, and necrotic tissue. Tumor and necrotic tissue each contain distinct diagnostic and prognostic features [150], and it is important to exclude connective stroma from consideration with these regions. Here, we seek to develop and validate methods for the image-based classification of these tissue regions. We examine two types of carcinoma (ovarian serous cystadenocarcinoma [OV] and renal clear cell carcinoma [KIRC]); optimize disease-specific and pooled classification models; and illustrate informative imaging markers for classifying biological regions in WSIs. Finally, we validate our approach by comparing prediction results to manually annotated ground truth data as well as to annotation data reported by TCGA.

D.4 Methods

D.4.1 Data

WSIs for both OV and KIRC were taken from the TCGA database. The high-resolution WSIs are first broken into 512x512-pixel tiles. This serves both to subdivide the computationally intensive work of extracting features and to limit those features to describing small neighborhoods of cells. Figure D.1(b) shows the scope of a single WSI tile. We selected 300 high-resolution tiles from each cancer type for the ground truth data to be used in training. For each data set, we selected one hundred representative tiles for each tissue class considered: stroma, tumor tissue, and necrotic tissue.

D.4.2 Quality Control

Before extracting features from the tiled WSIs, quality control is performed to isolate the tissue from the slide background and remove artifacts such as pen markings and tissue folds (Figure D.2). Slide background and pen marks are defined by thresholding in the HSV color space [151]. Tissue folds are identified using a supervised, connectivity-based

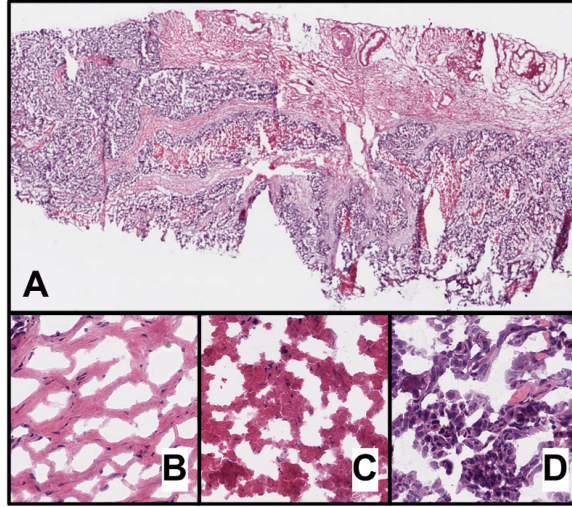


Figure D.1: (A) A typical whole-slide image. (B) A 512x512 high-resolution WSI tile showing stroma. (C) A typical necrotic tissue tile. (D) A tile containing typical tumor tissue.

soft thresholding method called ConnSoftT [149]. Finally, tiles within the tissue region of interest are defined as those tiles with less than 10% tissue fold and less than 80% background and pen markings combined. We use these tiles for the subsequent feature-extraction step.

The full set of 461 features used in this analysis has previously been shown to capture useful information about the image. These features fall into one of nine classes: global color, global texture, eosinophilic texture, eosinophilic object shape, basophilic texture, basophilic object shape, no-stain object shape, nuclear shape, or nuclear topology [152]. The tile is first color-normalized against a fixed set of reference images to mitigate batch effects due to variation in staining and equipment (Figure D.3). It is then divided into regions based on the presence of hematoxylin or eosin staining. Each of these regions is then processed separately, yielding its own shape and texture characteristics. Finally, the nuclear region mask is also processed for topological features describing the distribution of the nuclei in the tissue section [153].

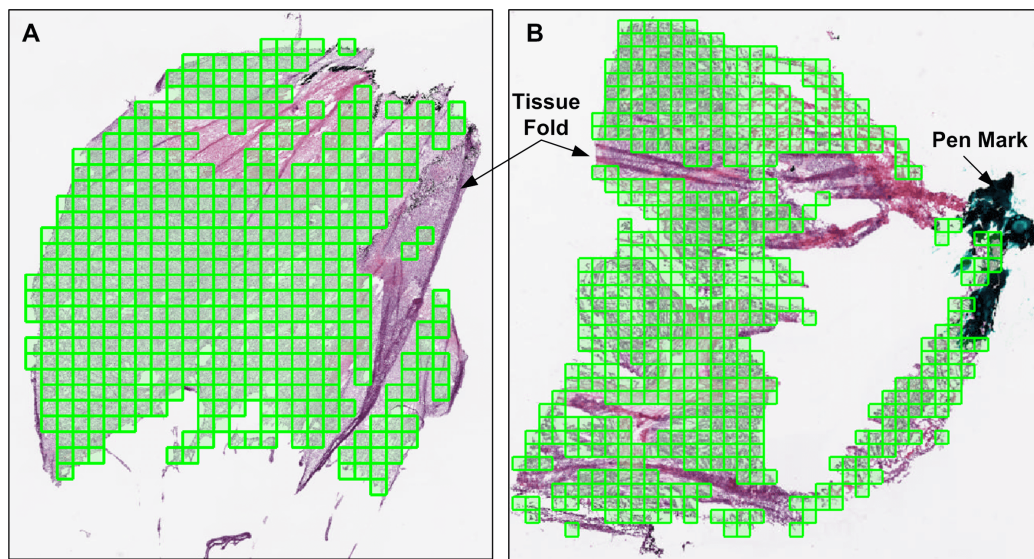


Figure D.2: Two typical WSIs with the ROI tiles after quality control outlined in green.

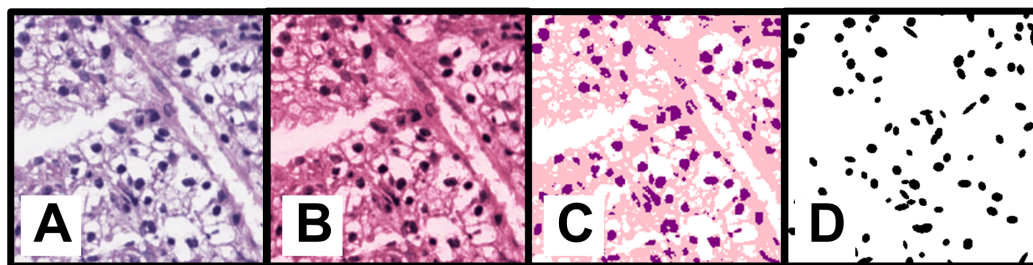


Figure D.3: (A) A portion of a high-resolution WSI tile. (B) The tile is color-normalized. (C) Thresholding is used to segment stain regions (light pink = eosinophilic, dark purple = basophilic). (D) A nuclear mask of the tile.

D.4.3 Feature Selection and Classification

We use the minimum redundancy, maximum relevance (mRMR) method to select features from among the 461 image features [154]. We apply the mRMR method in two ways: mutual information difference (MID) and mutual information quotient (MIQ). We use the linear support vector machine (SVM) [29] and various forms of discriminant analysis (DA) for classification. We consider four different DA classifiers: linear, diagonal linear, quadratic, and diagonal quadratic.

D.4.4 Model Optimization and Validation

In order to compare the various models of interest and estimate classifier accuracy, we used a nested cross-validation scheme. The outer CV loop (10 iterations, 3 fold) randomly splits the annotated training set into a test set and a training set. Each training set was fit to a model using every permutation of each parameter. SVM cost parameters considered were $2^{-5}, 2^{-4}, \dots, 2^{10}$. Feature sizes tested were 5, 10, \dots 50 for both the MID and MIQ mRMR methods. Each of these models were themselves analyzed using a 3-fold, 10 iteration internal CV loop. This nested CV structure is necessary to produce unbiased estimates of model performance. By subdividing the training set before optimizing the model, a portion of unused training tiles remains to test the best model (Figure D.4).

The best model was defined as the simplest model whose mean internal CV accuracy was within one standard deviation of the accuracy of the most accurate model. Smaller feature sets were preferred and small costs were preferred, in that order. In DA classification, linear classifiers were considered simpler than quadratic, and diagonal methods were considered simpler than non-diagonal.

As part of the metadata available for histopathology slides, TCGA provides expert annotation of the proportion of each tissue sample that falls into various biological regions. The OV data set consisted of 1087 slides with complete metadata. After applying tile-level quality control, 861,430 high-resolution tiles remained. For KIRC, 922 slides containing

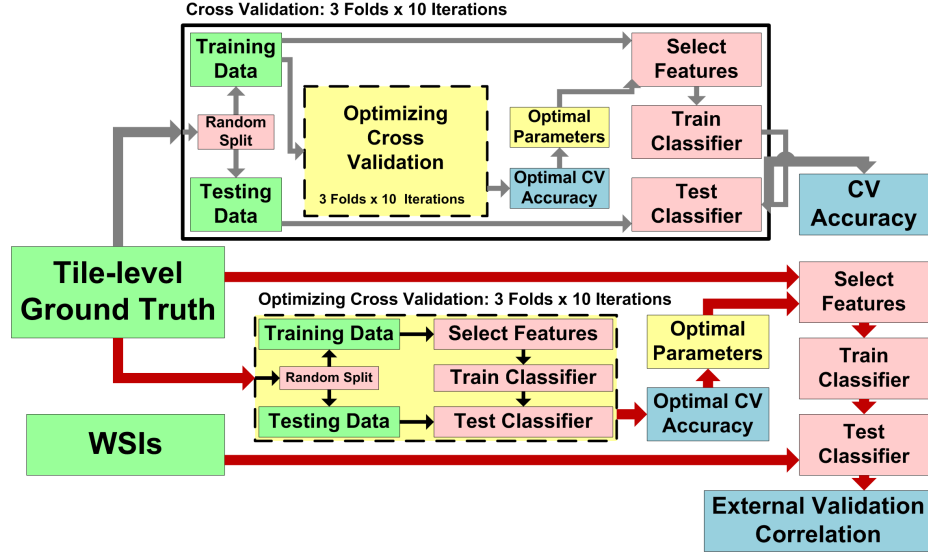


Figure D.4: The “tile-level ground truth” data set consisted of 600 expert-annotated high-resolution tile. The “WSIs” data set consisted of 1.7×10^6 tiles. These tiles were grouped into their constituent WSIs and analyzed together for normalized region sizes, as slide-level metadata is what is available from TCGA.

802,833 tissue tiles were used. The “WSIs” block shown in Figure D.4 refers to this data set.

Tiles for each slide were processed with the trained classifier and estimates of the prevalence of each bi-ological region were generated for that slide. We compared these results to TCGA reported values and computed the Pearson correlation coefficient.

D.5 Results and Discussion

D.5.1 Informative Features

Using the mRMR method and number of features found to be most robust by the inner cross-validation loop, a Fisher’s exact test was performed to determine if any class of features was significantly over-represented in the optimized feature spaces. Results are shown in Table D.1.

Values marked with a (*) are significant with $p < 0.05$. Color, texture, and shape features were each found to be significantly over-represented in at least one classifier. This

Table D.1: Contribution of Each Feature Class to Each Classifier Model

Feature Class	OV SVM	OV DA	KIRC SVM	KIRC DA	Pooled SVM	Pooled DA
Color	0.20	0.22*	0.23*	0.23*	0.21	0.22*
Global texture	0.22	0.37*	0.32	0.34	0.27	0.33
Eosinophilic shape	0.08	0.08	0.04	0.05	0.09	0.10
Eosinophilic texture	0.06	0.04	0.09*	0.12*	0.03	0.07
No-stain shape	0.08	0.04	0.04	0	0.07	0.04
Basophilic shape	0.20*	0.16	0.18*	0.17*	0.18*	0.16
Basophilic texture	0.04	0	0.04	0.03	0.04	0.02
Nuclear shape	0.08	0.04	0.07	0.05	0.08	0.05
Nuclear topology	0.04	0.04	0	0	0.02	0

Table D.2: Cross-Validation Accuracy of Each Classifier Model

	SVM	DA
OV	97.6 \pm 1.59 *	97.5 \pm 1.46
KIRC	98.7 \pm 0.76 †	98.4 \pm 1.22 ‡
Pooled	95.9 \pm 1.14 * † °	96.8 \pm 1.29 ‡ °

validates the use of a large, varied set of image features over the use of texture-only or color-only feature spaces, as are common in prior literature [155, 156].

D.5.2 Classification Cross-Validation

Table D.2 shows the outer CV accuracy of SVM classification for each disease and classifier model considered. The most likely sources of bias in these results are selection bias in the training data set and the possibility of multiple tiles from one WSI spanning the training and test sets. Models marked with a symbol are significantly different from models marked with the same symbol by a two-tailed Student’s t-test with $p < 0.05$. Values are given as mean \pm one standard deviation. While Table D.2 may not be an unbiased estimate of external validation accuracy, the relative accuracies show which models consistently outperform the others.

Two-tailed t-tests were used to compare models across disease cases and classifier models. SVM models were found to perform slightly better for KIRC and OV, however this difference was not significant. Pooled DA significantly outperformed the corresponding SVM. Pooled models significantly underperformed compared to disease-specific models in all cases except OV-DA.

D.5.3 External Validation

SVM based predictions of the relative size of biological regions in TCGA WSIs had higher correlations than DA counterparts in 8 of 9 cases, as shown in Table D.3. Correlation was higher for pooled samples than disease-specific classifiers in 5 of 6 cases, and statistically significant for all classes. Neither the OV nor KIRC classifiers yielded significant corre-

Table D.3: Correlation Coefficients for Slide-Level Classification Models

Class	SVM			DA		
	Necrosis	Stroma	Tumor	Necrosis	Stroma	Tumor
OV	0.031	0.425 *	0.371 *	-0.002	0.439 *	0.253 *
KIRC	0.130 *	0.299 *	0.351 *	0.075	0.346 *	0.276 *
Pooled	0.098 *	0.491 *	0.479 *	0.109 *	0.469 *	0.389 *

lation in all cases shown in Table D.3, where values marked with a (*) have significant correlation with $p < 0.001$. Figure D.5 shows larger and more intense regions of correlation between predicted and reported regions sizes in Pooled results.

The selected SVM classification model for each disease case was trained using the entire 300 or 600 tile ground truth data set, and then used to predict the class of the corresponding large TCGA data set.

In all disease cases, when validating on ground-truth tiles, pooled models underperform as compared to disease-specific models. This difference was significant in three of four cases but the performance of pooled models was still high. Moreover, pooled models tended to perform better on larger datasets in external validation. Hence, pooled models are more generalized. This indicates the possibility of developing a generalized tissue-classification model for different cancer, subject to further investigation.

D.6 Conclusion

In this paper, we developed and validated an automated system for classifying biological regions in WSIs. The proposed system uses a tile-based approach, where each tile is represented by a comprehensive set of features and classified into one of the regions using pre-optimized classifiers. The overall trend of the external validation shows significant correlation between the classifier-predicted values and expert-reported values suggesting successful classification of biological regions. In future work, these methods could be expanded to include information about a tile’s neighborhood to arrive at more accurate and better predictions of biological regions. These methods may also be applicable to other

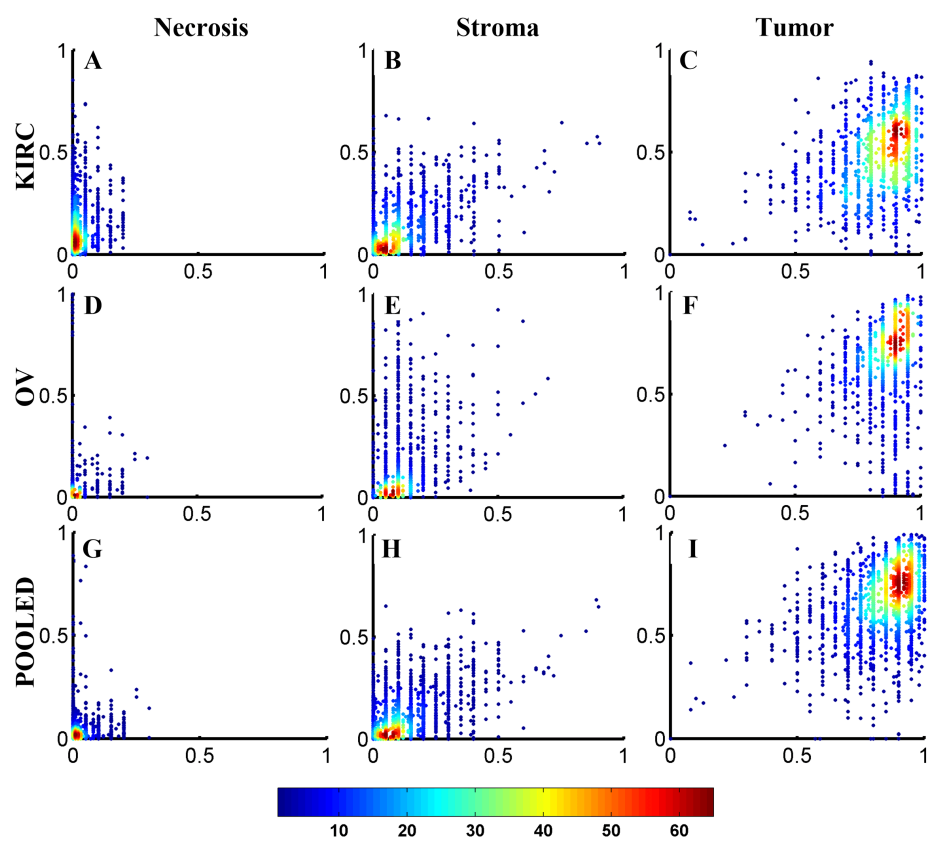


Figure D.5: Each scatterplot shows the classifier predicted biological region size versus the TCGA reported biological region size.

cancer types and biological regions.

The authors thank the NCI, the NHGRI, and the patients who contributed to the TCGA data set. We thank Michael Glover for his work in assembling the tile-level training set. This research has been supported by grants from NIH (U54CA119338, 1RC2CA148265, and R01CA163256).

APPENDIX E

PUBLICATIONS LIST

The following are selected publications of which Ryan Hoffman is an author.

R. A. Hoffman, H. Wu, J. Venugopalan, P. Braun, and M. D. Wang, “Intelligent mortality reporting with FHIR,” *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 5, pp. 1583–1588, 2018. DOI: 10.1109/JBHI.2017.2780891

R. A. Hoffman, H. Wu, J. Venugopalan, P. Braun, and M. D. Wang, “Intelligent mortality reporting with FHIR,” in *2017 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, 2017, pp. 181–184. DOI: 10.1109/BHI.2017.7897235

R. A. Hoffman, J. Venugopalan, L. Qu, H. Wu, and M. D. Wang, “Improving validity of cause of death on death certificates,” in *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, ACM, 2018, pp. 178–183

R. A. Hoffman, S. Kothari, and M. D. Wang, “Comparison of normalization algorithms for cross-batch color segmentation of histopathological images,” in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, IEEE, 2014, pp. 194–197. DOI: 10.1109/EMBC.2014.6943562

R. A. Hoffman, S. Kothari, J. H. Phan, and M. D. Wang, “A high-resolution tile-based approach for classifying biological regions in whole-slide histopathological images,” *Proceedings of the International Federation for Medical and Biological Engineering (IFMBE)*, vol. 42, pp. 280–283, 2014, ISSN: 1680-0737. DOI: 10.1007/978-3-319-03005-0_71. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4983443/>

C. Tanade, N. Pate, E. Paljug, R. A. Hoffman, and M. D. Wang, “Hybrid modeling of ebola propagation,” in *2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE)*, IEEE, 2019, pp. 204–210

- J. Dunn, H. Qiu, S. Kim, D. Jjingo, R. Hoffman, C. W. Kim, I. Jang, D. J. Son, D. Kim, C. Pan, *et al.*, “Flow-dependent epigenetic DNA methylation regulates endothelial gene expression and atherosclerosis,” *The Journal of clinical investigation*, vol. 124, no. 7, pp. 3187–3199, 2014
- E. R. Cosman Jr, J. R. Dolensky, and R. A. Hoffman, “Factors that affect radiofrequency heat lesion size,” *Pain Medicine*, vol. 15, no. 12, pp. 2020–2036, 2014
- J. H. Phan, R. Hoffman, S. Kothari, P.-Y. Wu, and M. D. Wang, “Integration of multi-modal biomedical data to predict cancer grade and patient survival,” in *2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, IEEE, 2016, pp. 577–580
- P.-Y. Wu, C.-W. Cheng, C. D. Kaddi, J. Venugopalan, R. Hoffman, and M. D. Wang, “-omic and electronic health record big data analytics for precision medicine,” *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 2, pp. 263–273, 2017
- L. Tong, R. Hoffman, S. R. Deshpande, and M. D. Wang, “Predicting heart rejection using histopathological whole-slide imaging and deep neural network with dropout,” in *2017 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, IEEE, 2017, pp. 1–4
- A. Bhatia, L. Tong, R. Hoffman, P. Wu, H. Hassanzadeh, M. Wang, and S. Deshpande, “Refinement of automated whole slide image analysis in pediatric heart transplants,” *The Journal of Heart and Lung Transplantation*, vol. 36, no. 4, S103–S104, 2017
- S. K. Phan, R. Hoffman, and M. D. Wang, “Biomedical imaging informatics for diagnostic imaging marker selection,” in *Health Informatics Data Analysis*, Springer, 2017, pp. 115–127
- Y. Zhu, M. Saqib, E. Ham, S. Belhareth, R. Hoffman, and M. D. Wang, “Mitigating patient-to-patient variation in eeg seizure detection using meta transfer learning,” in *2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)*, 2020, pp. 548–555. DOI: 10.1109/BIBE50027.2020.00095

Y. Zhu, Y. Sha, H. Wu, M. Li, R. A. Hoffman, and M. D. Wang, “Public health informatics: Proposing causal sequence of death using neural machine translation,” Pre-print, 2020

APPENDIX F

COPYRIGHT STATEMENTS

Various chapters and sections of this work have been previously published in peer-reviewed journals or refereed conference proceedings. That material is reused in this work with permission. The following sections contain the written permission and copyright licenses obtained from the copyright holders and publishers, included as an Appendix to this work as required by Georgia Institute of Technology policy.

The epigraph is attributed to Nobel laureate Ronald H. Coase, and referred to in his 1981 lecture "How Should Economists Choose?" [167].

F.1 IEEE

The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:

Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:

1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line ©2011 IEEE. 2) In the case of illustrations or tabular material, we require that the copyright line ©[Year of original publication] IEEE appear prominently with each reprinted figure and/or table. 3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:

1) The following IEEE copyright/ credit notice should be placed prominently in the references: ©[year of original publication] IEEE. Reprinted, with permission, from [author

names, paper title, IEEE publication title, and month/year of publication] 2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line. 3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

©2018 IEEE. Reprinted, with permission, from R. A. Hoffman, H. Wu, J. Venugopalan, P. Braun and M. D. Wang, "Intelligent Mortality Reporting With FHIR", in IEEE Journal of Biomedical and Health Informatics, vol. 22, no. 5, pp. 1583-1588, Sept. 2018, DOI: 10.1109/JBHI.2017.2780891.

©2017 IEEE. Reprinted, with permission, from R. A. Hoffman, H. Wu, J. Venugopalan, P. Braun and M. D. Wang, "Intelligent Mortality Reporting With FHIR", in Proceedings of the 2017 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI), pp. 181-184, Feb. 2017, DOI: 10.1109/BHI.2017.7897235.

©2017 IEEE. Reprinted, with permission, from R. A. Hoffman, H. Wu, J. Venugopalan, P. Braun and M. D. Wang, "Comparison of normalization algorithms for cross-batch color segmentation of histopathological images", in Proceedings of the 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Aug. 2014, DOI: 10.1109/BHI.2017.7897235.

F.2 ACM

Proceedings, Association for Computing Machinery by Association for Computing Machinery. Reproduced with permission of Association for Computing Machinery in the format Republish in a thesis/dissertation via Copyright Clearance Center.

ACM (Association for Computing Machinery) LICENSE TERMS AND CONDITIONS

Jun 20, 2020

This is a License Agreement between Georgia Institute of Technology – Ryan Hoffman ("You") and ACM (Association for Computing Machinery) ("ACM (Association for Computing Machinery)") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by ACM (Association for Computing Machinery), and the payment terms and conditions. All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

License Number: 4657071364632

License date: Aug 22, 2019

Licensed content publisher: ACM (Association for Computing Machinery)

Licensed content title: Proceedings, Association for Computing Machinery

Licensed content date: Jan 1, 1900

Type of Use: Thesis/Dissertation

Requestor type: Academic institution

Format: Print, Electronic

Portion: chapter/article

The requesting person/organization is: Ryan Hoffman / Graduate Student / Georgia Institute of Technology

Title or numeric reference of the portion(s): Entire conference proceeding article, in-

cluding text and figures, with right to minor modifications, including adapting format and integration with other material for purpose of inclusion in doctoral thesis.

Title of the article or chapter the portion is from: Improving Validity of Cause of Death on Death Certificates

Author of portion(s): Ryan Hoffman (Self)

2018 Proceedings

Duration of use: Life of current and all future editions

TERMS AND CONDITIONS

The following terms are individual to this publisher:

None

Other Terms and Conditions:

STANDARD TERMS AND CONDITIONS

1. Description of Service; Defined Terms. This Republication License enables the User to obtain licenses for republication of one or more copyrighted works as described in detail on the relevant Order Confirmation (the “Work(s)”). Copyright Clearance Center, Inc. (“CCC”) grants licenses through the Service on behalf of the rightsholder identified on the Order Confirmation (the “Rightsholder”). “Republication”, as used herein, generally means the inclusion of a Work, in whole or in part, in a new work or works, also as described on the Order Confirmation. “User”, as used herein, means the person or entity making such republication.

2. The terms set forth in the relevant Order Confirmation, and any terms set by the Rightsholder with respect to a particular Work, govern the terms of use of Works in connection with the Service. By using the Service, the person transacting for a republication license on behalf of the User represents and warrants that he/she/it (a) has been duly authorized by the User to accept, and hereby does accept, all such terms and conditions on behalf of User, and (b) shall inform User of all such terms and conditions. In the event such person is a “freelancer” or other third party independent of User and CCC, such party

shall be deemed jointly a “User” for purposes of these terms and conditions. In any event, User shall be deemed to have accepted and agreed to all such terms and conditions if User republishes the Work in any fashion.

3. Scope of License; Limitations and Obligations.

3.1 All Works and all rights therein, including copyright rights, remain the sole and exclusive property of the Rightsholder. The license created by the exchange of an Order Confirmation (and/or any invoice) and payment by User of the full amount set forth on that document includes only those rights expressly set forth in the Order Confirmation and in these terms and conditions, and conveys no other rights in the Work(s) to User. All rights not expressly granted are hereby reserved.

3.2 General Payment Terms: You may pay by credit card or through an account with us payable at the end of the month. If you and we agree that you may establish a standing account with CCC, then the following terms apply: Remit Payment to: Copyright Clearance Center, 29118 Network Place, Chicago, IL 60673-1291. Payments Due: Invoices are payable upon their delivery to you (or upon our notice to you that they are available to you for downloading). After 30 days, outstanding amounts will be subject to a service charge of 1-1/2

3.3 Unless otherwise provided in the Order Confirmation, any grant of rights to User (i) is “one-time” (including the editions and product family specified in the license), (ii) is non-exclusive and non-transferable and (iii) is subject to any and all limitations and restrictions (such as, but not limited to, limitations on duration of use or circulation) included in the Order Confirmation or invoice and/or in these terms and conditions. Upon completion of the licensed use, User shall either secure a new permission for further use of the Work(s) or immediately cease any new use of the Work(s) and shall render inaccessible (such as by deleting or by removing or severing links or other locators) any further copies of the Work (except for copies printed on paper in accordance with this license and still in User’s stock at the end of such period).

3.4 In the event that the material for which a republication license is sought includes third party materials (such as photographs, illustrations, graphs, inserts and similar materials) which are identified in such material as having been used by permission, User is responsible for identifying, and seeking separate licenses (under this Service or otherwise) for, any of such third party materials; without a separate license, such third party materials may not be used.

3.5 Use of proper copyright notice for a Work is required as a condition of any license granted under the Service. Unless otherwise provided in the Order Confirmation, a proper copyright notice will read substantially as follows: “Republished with permission of [Rightsholder’s name], from [Work’s title, author, volume, edition number and year of copyright]; permission conveyed through Copyright Clearance Center, Inc. ” Such notice must be provided in a reasonably legible font size and must be placed either immediately adjacent to the Work as used (for example, as part of a by-line or footnote but not as a separate electronic link) or in the place where substantially all other credits or notices for the new work containing the republished Work are located. Failure to include the required notice results in loss to the Rightsholder and CCC, and the User shall be liable to pay liquidated damages for each such failure equal to twice the use fee specified in the Order Confirmation, in addition to the use fee itself and any other fees and charges specified.

3.6 User may only make alterations to the Work if and as expressly set forth in the Order Confirmation. No Work may be used in any way that is defamatory, violates the rights of third parties (including such third parties’ rights of copyright, privacy, publicity, or other tangible or intangible property), or is otherwise illegal, sexually explicit or obscene. In addition, User may not conjoin a Work with any other material that may result in damage to the reputation of the Rightsholder. User agrees to inform CCC if it becomes aware of any infringement of any rights in a Work and to cooperate with any reasonable request of CCC or the Rightsholder in connection therewith.

4. Indemnity. User hereby indemnifies and agrees to defend the Rightsholder and CCC,

and their respective employees and directors, against all claims, liability, damages, costs and expenses, including legal fees and expenses, arising out of any use of a Work beyond the scope of the rights granted herein, or any use of a Work which has been altered in any unauthorized way by User, including claims of defamation or infringement of rights of copyright, publicity, privacy or other tangible or intangible property.

5. Limitation of Liability. UNDER NO CIRCUMSTANCES WILL CCC OR THE RIGHTSHOLDER BE LIABLE FOR ANY DIRECT, INDIRECT, CONSEQUENTIAL OR INCIDENTAL DAMAGES (INCLUDING WITHOUT LIMITATION DAMAGES FOR LOSS OF BUSINESS PROFITS OR INFORMATION, OR FOR BUSINESS INTERRUPTION) ARISING OUT OF THE USE OR INABILITY TO USE A WORK, EVEN IF ONE OF THEM HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. In any event, the total liability of the Rightsholder and CCC (including their respective employees and directors) shall not exceed the total amount actually paid by User for this license. User assumes full liability for the actions and omissions of its principals, employees, agents, affiliates, successors and assigns.

6. Limited Warranties. THE WORK(S) AND RIGHT(S) ARE PROVIDED "AS IS". CCC HAS THE RIGHT TO GRANT TO USER THE RIGHTS GRANTED IN THE ORDER CONFIRMATION DOCUMENT. CCC AND THE RIGHTSHOLDER DISCLAIM ALL OTHER WARRANTIES RELATING TO THE WORK(S) AND RIGHT(S), EITHER EXPRESS OR IMPLIED, INCLUDING WITHOUT LIMITATION IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. ADDITIONAL RIGHTS MAY BE REQUIRED TO USE ILLUSTRATIONS, GRAPHS, PHOTOGRAPHS, ABSTRACTS, INSERTS OR OTHER PORTIONS OF THE WORK (AS OPPOSED TO THE ENTIRE WORK) IN A MANNER CONTEMPLATED BY USER; USER UNDERSTANDS AND AGREES THAT NEITHER CCC NOR THE RIGHTSHOLDER MAY HAVE SUCH ADDITIONAL RIGHTS TO GRANT.

7. Effect of Breach. Any failure by User to pay any amount when due, or any use by

User of a Work beyond the scope of the license set forth in the Order Confirmation and/or these terms and conditions, shall be a material breach of the license created by the Order Confirmation and these terms and conditions. Any breach not cured within 30 days of written notice thereof shall result in immediate termination of such license without further notice. Any unauthorized (but licensable) use of a Work that is terminated immediately upon notice thereof may be liquidated by payment of the Rightsholder's ordinary license price therefor; any unauthorized (and unlicensable) use that is not terminated immediately for any reason (including, for example, because materials containing the Work cannot reasonably be recalled) will be subject to all remedies available at law or in equity, but in no event to a payment of less than three times the Rightsholder's ordinary license price for the most closely analogous licensable use plus Rightsholder's and/or CCC's costs and expenses incurred in collecting such payment.

8. Miscellaneous.

8.1 User acknowledges that CCC may, from time to time, make changes or additions to the Service or to these terms and conditions, and CCC reserves the right to send notice to the User by electronic mail or otherwise for the purposes of notifying User of such changes or additions; provided that any such changes or additions shall not apply to permissions already secured and paid for.

8.2 Use of User-related information collected through the Service is governed by CCC's privacy policy, available online here: <http://www.copyright.com/content/cc3/en/tools/footer/privacypolicy.html>.

8.3 The licensing transaction described in the Order Confirmation is personal to User. Therefore, User may not assign or transfer to any other person (whether a natural person or an organization of any kind) the license created by the Order Confirmation and these terms and conditions or any rights granted hereunder; provided, however, that User may assign such license in its entirety on written notice to CCC in the event of a transfer of all or substantially all of User's rights in the new material which includes the Work(s) licensed

under this Service.

8.4 No amendment or waiver of any terms is binding unless set forth in writing and signed by the parties. The Rightsholder and CCC hereby object to any terms contained in any writing prepared by the User or its principals, employees, agents or affiliates and purporting to govern or otherwise relate to the licensing transaction described in the Order Confirmation, which terms are in any way inconsistent with any terms set forth in the Order Confirmation and/or in these terms and conditions or CCC's standard operating procedures, whether such writing is prepared prior to, simultaneously with or subsequent to the Order Confirmation, and whether such writing appears on a copy of the Order Confirmation or in a separate instrument.

8.5 The licensing transaction described in the Order Confirmation document shall be governed by and construed under the law of the State of New York, USA, without regard to the principles thereof of conflicts of law. Any case, controversy, suit, action, or proceeding arising out of, in connection with, or related to such licensing transaction shall be brought, at CCC's sole discretion, in any federal or state court located in the County of New York, State of New York, USA, or in any federal or state court whose geographical jurisdiction covers the location of the Rightsholder set forth in the Order Confirmation. The parties expressly submit to the personal jurisdiction and venue of each such federal or state court. If you have any comments or questions about the Service or Copyright Clearance Center, please contact us at 978-750-8400 or send an e-mail to urlinfo@copyright.com.

v 1.1

Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.

F.3 Springer

SPRINGER NATURE LICENSE TERMS AND CONDITIONS

Jun 20, 2020

This Agreement between Georgia Institute of Technology – Ryan Hoffman ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

License Number: 4654491358113

License date: Aug 22, 2019

Licensed Content Publisher: Springer Nature

Licensed Content Publication: Springer eBook

Licensed Content Title: A High-Resolution Tile-Based Approach for Classifying Biological Regions in Whole-Slide Histopathological Images

Licensed Content Author: R. A. Hoffman, S. Kothari, J. H. Phan et al

Licensed Content Date: Jan 1, 2014

Type of Use: Thesis/Dissertation

Author of this Springer Nature content: yes

Title: Graduate Student

Institution name: Georgia Institute of Technology

Requestor Location

Georgia Institute of Technology 313 Ferst Drive Dept. of Biomedical Engineering

ATLANTA, GA 30332 United States Attn: Georgia Institute of Technology

Terms and Conditions

Springer Nature Customer Service Centre GmbH Terms and Conditions

This agreement sets out the terms and conditions of the licence (the Licence) between you and Springer Nature Customer Service Centre GmbH (the Licensor). By clicking 'accept' and completing the transaction for the material (Licensed Material), you also confirm your acceptance of these terms and conditions.

Grant of License

The Licensor grants you a personal, non-exclusive, non-transferable, world-wide licence to reproduce the Licensed Material for the purpose specified in your order only.

Licences are granted for the specific use requested in the order and for no other use, subject to the conditions below.

The Licensor warrants that it has, to the best of its knowledge, the rights to license reuse of the Licensed Material. However, you should ensure that the material you are requesting is original to the Licensor and does not carry the copyright of another entity (as credited in the published version).

If the credit line on any part of the material you have requested indicates that it was reprinted or adapted with permission from another source, then you should also seek permission from that source to reuse the material.

Scope of Licence

You may only use the Licensed Content in the manner and to the extent permitted by these Ts&Cs and any applicable laws.

A separate licence may be required for any additional use of the Licensed Material, e.g. where a licence has been purchased for print only use, separate permission must be obtained for electronic re-use. Similarly, a licence is only valid in the language selected and does not apply for editions in other languages unless additional translation rights have been granted separately in the licence. Any content owned by third parties are expressly excluded from the licence.

Similarly, rights for additional components such as custom editions and derivatives require additional permission and may be subject to an additional fee. Please apply to Journalpermissions@springernature.com/bookpermissions@springernature.com for these rights.

Where permission has been granted free of charge for material in print, permission may also be granted for any electronic version of that work, provided that the material is incidental to your work as a whole and that the electronic version is essentially equivalent to, or substitutes for, the print version.

An alternative scope of licence may apply to signatories of the STM Permissions Guide-

lines, as amended from time to time.

Duration of Licence

A licence for is valid from the date of purchase ('Licence Date') at the end of the relevant period in the below table:

Scope of Licence	Duration of Licence	Post on a website	12 months	Presentations	12 months	Books and journals	Lifetime of the edition in the language purchased	Acknowledgement
------------------	---------------------	-------------------	-----------	---------------	-----------	--------------------	---	-----------------

The Licensor's permission must be acknowledged next to the Licenced Material in print. In electronic form, this acknowledgement must be visible at the same time as the figures/tables/illustrations or abstract, and must be hyperlinked to the journal/book's homepage. Our required acknowledgement format is in the Appendix below.

Restrictions on use

Use of the Licensed Material may be permitted for incidental promotional use and minor editing privileges e.g. minor adaptations of single figures, changes of format, colour and/or style where the adaptation is credited as set out in Appendix 1 below. Any other changes including but not limited to, cropping, adapting, omitting material that affect the meaning, intention or moral rights of the author are strictly prohibited.

You must not use any Licensed Material as part of any design or trademark.

Licensed Material may be used in Open Access Publications (OAP) before publication by Springer Nature, but any Licensed Material must be removed from OAP sites prior to final publication.

Ownership of Rights

Licensed Material remains the property of either Licensor or the relevant third party and any rights not explicitly granted herein are expressly reserved.

Warranty

IN NO EVENT SHALL LICENSOR BE LIABLE TO YOU OR ANY OTHER PARTY OR ANY OTHER PERSON OR FOR ANY SPECIAL, CONSEQUENTIAL, INCIDENT-

TAL OR INDIRECT DAMAGES, HOWEVER CAUSED, ARISING OUT OF OR IN CONNECTION WITH THE DOWNLOADING, VIEWING OR USE OF THE MATERIALS REGARDLESS OF THE FORM OF ACTION, WHETHER FOR BREACH OF CONTRACT, BREACH OF WARRANTY, TORT, NEGLIGENCE, INFRINGEMENT OR OTHERWISE (INCLUDING, WITHOUT LIMITATION, DAMAGES BASED ON LOSS OF PROFITS, DATA, FILES, USE, BUSINESS OPPORTUNITY OR CLAIMS OF THIRD PARTIES), AND WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION SHALL APPLY NOTWITHSTANDING ANY FAILURE OF ESSENTIAL PURPOSE OF ANY LIMITED REMEDY PROVIDED HEREIN.

Limitations

BOOKS ONLY:Where 'reuse in a dissertation/thesis' has been selected the following terms apply: Print rights of the final author's accepted manuscript (for clarity, NOT the published version) for up to 100 copies, electronic rights for use only on a personal website or institutional repository as defined by the Sherpa guideline (www.sherpa.ac.uk/romeo/).

Termination and Cancellation

Licences will expire after the period shown in Clause 3 (above).

Licensee reserves the right to terminate the Licence in the event that payment is not received in full or if there has been a breach of this agreement by you.

Appendix 1 — Acknowledgements:

For Journal Content: Reprinted by permission from [the Licensor]: [Journal Publisher (e.g. Nature / Springer / Palgrave)] [JOURNAL NAME] [REFERENCE CITATION (Article name, Author(s) Name), [COPYRIGHT] (year of publication)] For Advance Online Publication papers: Reprinted by permission from [the Licensor]: [Journal Publisher (e.g. Nature/Springer/Palgrave)] [JOURNAL NAME] [REFERENCE CITATION (Article name, Author(s) Name), [COPYRIGHT] (year of publication), advance online publication, day month year (doi: 10.1038/sj.[JOURNAL ACRONYM].)] For Adapta-

tions/Translations: Adapted/Translated by permission from [the Licensor]: [Journal Publisher (e.g. Nature / Springer / Palgrave)] [JOURNAL NAME] [REFERENCE CITATION (Article name, Author(s) Name), [COPYRIGHT] (year of publication) Note: For any republication from the British Journal of Cancer, the following credit line style applies: Reprinted/adapted/translated by permission from [the Licensor]: on behalf of Cancer Research UK: : [Journal Publisher (e.g. Nature / Springer / Palgrave)] [JOURNAL NAME] [REFERENCE CITATION (Article name, Author(s) Name), [COPYRIGHT] (year of publication) For Advance Online Publication papers: Reprinted by permission from The [the Licensor]: on behalf of Cancer Research UK: [Journal Publisher (e.g. Nature / Springer / Palgrave)] [JOURNAL NAME] [REFERENCE CITATION (Article name, Author(s) Name), [COPYRIGHT] (year of publication), advance online publication, day month year (doi: 10.1038/sj.[JOURNAL ACRONYM]) For Book content: Reprinted/adapted by permission from [the Licensor]: [Book Publisher (e.g. Palgrave Macmillan, Springer etc) [Book Title] by [Book author(s)] [COPYRIGHT] (year of publication) Other Conditions:

Version 1.2

Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.

REFERENCES

- [1] World Health Organization, *The top 10 causes of death*, <http://www.who.int/mediacentre/factsheets/fs310/en/index2.html>, Web Page, 2013.
- [2] K. D. Kochanek, S. L. Murphy, J. Xu, and B. Tejada-Vera, “Deaths: Final data for 2014,” *National vital statistics reports : from the Centers for Disease Control and Prevention, National Center for Health Statistics, National Vital Statistics System*, vol. 65, no. 4, pp. 1–122, 2016.
- [3] B. Randall, “Death certification: A primer part I - an introduction to the death certificate,” *South Dakota Medicine*, vol. 67, no. 5, 2014.
- [4] N. C. for Health Statistics, *Possible solutions to common problems in death certification*, http://www.cdc.gov/nchs/nvss/death_certification_problems.htm#references, Government Document, 1997.
- [5] J. C. Mandel, D. A. Kreda, K. D. Mandl, I. S. Kohane, and R. B. Ramoni, “SMART on FHIR: A standards-based, interoperable apps platform for electronic health records,” *Journal of the American Medical Informatics Association*, vol. 23, no. 5, pp. 899–908, 2016.
- [6] World Health Organization, *Global Health Expenditure Database*, <http://apps.who.int/nha/database/Home/Index/en>, Database.
- [7] The World Bank, *Gross Domestic Product 2010*, <https://siteresources.worldbank.org/DATASTATISTICS/Resources/GDP.pdf>, Report.
- [8] L. L. Weed, “Can the PROMIS be kept?” *Möbius: A Journal for Continuing Education Professionals in Health Sciences*, vol. 3, no. 2, pp. 17–24, Apr. 1983.
- [9] K. Häyrinen, K. Saranto, and P. Nykänen, “Definition, structure, content, use and impacts of electronic health records: A review of the research literature,” *International Journal of Medical Informatics*, vol. 77, no. 5, pp. 291–304, May 2008.
- [10] L. Jacobs, “Interview with Lawrence Weed, MD— the father of the problem-oriented medical record looks ahead,”
- [11] Institute of Medicine, *The Computer-Based Patient Record: An Essential Technology for Health Care, Revised Edition*, ser. An Essential Technology for Health Care, Revised Edition. Washington, D.C.: National Academies Press, Nov. 1997, ISBN: 978-0-309-08684-4.

- [12] C. S. Kruse, K. Bolton, and G. Freriks, "The effect of patient portals on quality outcomes and its implications to meaningful use: A systematic review," *Journal of Medical Internet Research*, vol. 17, no. 2, e44, 2015.
- [13] E. W. Ford, B. W. Hesse, and T. R. Huerta, "Personal health record use in the United States: Forecasting future adoption levels," *Journal of Medical Internet Research*, vol. 18, no. 3, e73, Mar. 2016.
- [14] M. Lester, S. Boateng, J. Studeny, and A. Coustasse, "Personal health records: Beneficial or burdensome for patients and healthcare providers?" *Perspectives in health information management*, vol. 13, 1h, 2016.
- [15] M. J. Bietz, C. S. Bloss, S. Calvert, J. G. Godino, J. Gregory, M. P. Claffey, J. Sheehan, and K. Patrick, "Opportunities and challenges in the use of personal health data for health research," *Journal of the American Medical Informatics Association*, vol. 23, no. e1, e42–8, Apr. 2016.
- [16] S. S. Jones, R. S. Rudin, T. Perry, and P. G. Shekelle, "Health information technology: An updated systematic review with a focus on meaningful use," *Annals of Internal Medicine*, vol. 160, no. 1, pp. 48–54–54, Jan. 2014.
- [17] B. M. C. Silva, J. J. P. C. Rodrigues, I. de la Torre Diez, M. Lopez-Coronado, and K. Saleem, "Mobile-health: A review of current state in 2015," *Journal of Biomedical Informatics*, vol. 56, no. C, pp. 265–272, Aug. 2015.
- [18] C. L. Ventola, "Mobile devices and apps for health care professionals: Uses and benefits," *P & T : a peer-reviewed journal for formulary management*, vol. 39, no. 5, pp. 356–364, May 2014.
- [19] U. Varshney, "Mobile health: Four emerging themes of research," *Decision Support Systems*, vol. 66, no. C, pp. 20–35, Oct. 2014.
- [20] W. Raghupathi and V. Raghupathi, "Big data analytics in healthcare: Promise and potential," *Health Information Science and ...*, 2014.
- [21] D. Bender and K. Sartipi, "HL7 FHIR: An agile and RESTful approach to healthcare information exchange," in *Computer-Based Medical Systems (CBMS), 2013 IEEE 26th International Symposium on*, IEEE, pp. 326–331.
- [22] J. H. Phan, A. N. Young, and M. D. Wang, "omniBiomarker: A web-based application for knowledge-driven biomarker identification," *IEEE transactions on biomedical engineering*, vol. 60, no. 12, pp. 3364–3367, Dec. 2013.
- [23] C.-W. Cheng, N. Chanani, J. Venugopalan, K. Maher, and M. D. Wang, "icuARM - an ICU clinical decision support system using association rule mining," *Transla-*

tional Engineering in Health and Medicine, IEEE Journal of, vol. 1, pp. 4 400 110–4 400 110, 2013.

- [24] C. Kaddi, R. M. Parry, and M. D. Wang, “Multivariate hypergeometric similarity measure,” in *the ACM Conference*, New York, New York, USA: ACM Press, 2012, pp. 234–241.
- [25] P. Y. Wu, R. Chandramohan, J. H. Phan, W. T. Mahle, J. W. Gaynor, K. O. Maher, and M. D. Wang, “Cardiovascular transcriptomics and epigenomics using next-generation sequencing: Challenges, progress, and opportunities,”
- [26] C. Bishop, *Pattern Recognition and Machine Learning*, ser. Information Science and Statistics. Springer, 2006, ISBN: 9780387310732.
- [27] R. A. Fisher, “The use of multiple measurements in taxonomic problems,” *Annals of Eugenics*, vol. 7, no. 2, pp. 179–188, 1936. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-1809.1936.tb02137.x>.
- [28] A. Dobson and A. Barnett, *An Introduction to Generalized Linear Models*. CRC Press, 2018, ISBN: 9781351726221.
- [29] C.-C. Chang and C.-J. Lin, “LIBSVM: A library for support vector machines,” *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.
- [30] G. Seber, *Multivariate Observations*, ser. Wiley Series in Probability and Statistics. Wiley, 2009, ISBN: 9780470317310.
- [31] Wikimedia Commons, *File:Eddie August Schneider (1911-1940) death certificate.gif* — *Wikimedia Commons, the free media repository*, [https://commons.wikimedia.org/w/index.php?title=File:Eddie_August_Schneider_\(1911-1940\)_death_certificate.gif&oldid=511665719](https://commons.wikimedia.org/w/index.php?title=File:Eddie_August_Schneider_(1911-1940)_death_certificate.gif&oldid=511665719), Online, Public Domain; Accessed 5/17/2021, 2020.
- [32] National Center for Health Statistics (NCHS), *U.S. Standard Certificate of Death REV. 11/2003*, <https://www.cdc.gov/nchs/data/dvs/death11-03final-acc.pdf>, Accessed 9/8/2017, 2003.
- [33] E. Herrett, A. D. Shah, R. Boggon, S. Denaxas, L. Smeeth, T. van Staa, A. Timmis, and H. Hemingway, “Completeness and diagnostic validity of recording acute myocardial infarction events in primary care, hospital care, disease registry, and national mortality records: Cohort study,” *BMJ*, vol. 346, 2013. eprint: <https://www.bmj.com/content/346/bmj.f2350.full.pdf>.

- [34] M. K. Najafabadi, M. N. Mahrin, S. Chuprat, and H. M. Sarkan, “Improving the accuracy of collaborative filtering recommendations using clustering and association rules mining on implicit data,” *Computers in Human Behavior*, vol. 67, pp. 113–128, 2017.
- [35] F. Zhan, X. Zhu, L. Zhang, X. Wang, L. Wang, and C. Liu, “Summary of association rules,” *IOP Conference Series: Earth and Environmental Science*, vol. 252, no. 3, p. 032 219, 2019.
- [36] P.-N. Tan, V. Kumar, and J. Srivastava, “Selecting the right objective measure for association analysis,” *Information Systems*, vol. 29, no. 4, pp. 293–313, 2004.
- [37] Georgia Institute of Technology, *Data Access Policy*, <https://policylibrary.gatech.edu/print/book/export/html/994>, 2015.
- [38] U.S. Food and Drug Administration, *Clinical Decision Support Software: Draft Guidance for Industry and Food and Drug Administration Staff*, <https://www.fda.gov/media/109618/download>, FDA Draft Guidance, 2019.
- [39] J. S. Breese, D. Heckerman, and C. Kadie, *Anonymous microsoft web data data set*, <http://archive.ics.uci.edu/ml/datasets/Anonymous+Microsoft+Web+Data>, 1998.
- [40] D. Dua and C. Graff, *UCI machine learning repository*, <http://archive.ics.uci.edu/ml>, 2017.
- [41] J. S. Breese, D. Heckerman, and C. Kadie, “Empirical analysis of predictive algorithms for collaborative filtering,” Tech. Rep. MSR-TR-98-12, 1998, p. 18.
- [42] —, “Empirical analysis of predictive algorithms for collaborative filtering,” in *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence (UAI1998)*, 1998, pp. 43–52.
- [43] S. Brin, R. Motwani, J. D. Ullman, and S. Tsur, “Dynamic itemset counting and implication rules for market basket data,” in *Proceedings of the 1997 ACM SIGMOD international conference on Management of data*, 1997, pp. 255–264.
- [44] G. Piatetsky-Shapiro, “Discovery, analysis, and presentation of strong rules,” *Knowledge discovery in databases*, pp. 229–238, 1991.
- [45] J. Paul *et al.*, “The distribution of the flora in the alpine zone. 1,” *New phytologist*, vol. 11, no. 2, pp. 37–50, 1912.
- [46] P. Clark and T. Niblett, “The cn2 induction algorithm,” *Machine learning*, vol. 3, no. 4, pp. 261–283, 1989.

- [47] P. J. Azevedo and A. M. Jorge, “Comparing rule measures for predictive association rules,” in *European Conference on Machine Learning*, Springer, 2007, pp. 510–517.
- [48] R. Agrawal and R. Srikant, “Fast algorithms for mining association rules,” in *Proc. 20th int. conf. very large data bases, VLDB*, vol. 1215, 1994, pp. 487–499.
- [49] I. Batal, H. Valizadegan, G. F. Cooper, and M. Hauskrecht, “A pattern mining approach for classifying multivariate temporal data,” in *Bioinformatics and Biomedicine (BIBM), 2011 IEEE International Conference on*, IEEE, 2011.
- [50] J. Han, J. Pei, and Y. Yin, “Mining frequent patterns without candidate generation,” *ACM sigmod record*, vol. 29, no. 2, pp. 1–12, 2000.
- [51] L. M. Aouad, N.-A. Le-Khac, and T. M. Kechadi, “Performance study of distributed apriori-like frequent itemsets mining,” *Knowledge and information systems*, vol. 23, no. 1, pp. 55–72, 2010.
- [52] R. Srikant and R. Agrawal, *Mining sequential patterns: Generalizations and performance improvements*. Springer, 1996, ISBN: 354061057X.
- [53] R. Agrawal and R. Srikant, “Mining sequential patterns,” in *Data Engineering, 1995. Proceedings of the Eleventh International Conference on*, IEEE, pp. 3–14, ISBN: 0818669101.
- [54] P. Fournier-Viger, R. Nkambou, and V. S.-M. Tseng, “Rulegrowth: Mining sequential rules common to several sequences by pattern-growth,” in *Proceedings of the 2011 ACM symposium on applied computing*, 2011, pp. 956–961.
- [55] M. J. Zaki, “SPADE: An efficient algorithm for mining frequent sequences,” *Machine learning*, vol. 42, no. 1-2, pp. 31–60, 2001.
- [56] D. Lo, S.-C. Khoo, and L. Wong, “Non-redundant sequential rules—theory and algorithm,” *Information Systems*, vol. 34, no. 4-5, pp. 438–453, 2009.
- [57] J. Pei, J. Han, B. Mortazavi-Asl, J. Wang, H. Pinto, Q. Chen, U. Dayal, and M.-C. Hsu, “Mining sequential patterns by pattern-growth: The prefixspan approach,” *Knowledge and Data Engineering, IEEE Transactions on*, vol. 16, no. 11, pp. 1424–1440, 2004.
- [58] J. Deng, Z. Qu, Y. Zhu, G.-M. Muntean, and X. Wang, “Towards efficient and scalable data mining using spark,” 2014.
- [59] J. Han, J. Pei, B. Mortazavi-Asl, Q. Chen, U. Dayal, and M.-C. Hsu, “Freespan: Frequent pattern-projected sequential pattern mining,” in *Proceedings of the sixth*

ACM SIGKDD international conference on Knowledge discovery and data mining, 2000, pp. 355–359.

- [60] J. Wang, J. Han, and C. Li, “Frequent closed sequence mining without candidate maintenance,” *Knowledge and Data Engineering, IEEE Transactions on*, vol. 19, no. 8, pp. 1042–1056, 2007.
- [61] J. Wang and J. Han, “Bide: Efficient mining of frequent closed sequences,” in *Proceedings. 20th international conference on data engineering*, IEEE, 2004, pp. 79–90.
- [62] R. A. Hoffman, H. Wu, J. Venugopalan, P. Braun, and M. D. Wang, “Intelligent mortality reporting with FHIR,” *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 5, pp. 1583–1588, 2018.
- [63] R. A. Hoffman, J. Venugopalan, L. Qu, H. Wu, and M. D. Wang, “Improving validity of cause of death on death certificates,” in *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, ACM, 2018, pp. 178–183.
- [64] National Center for Health Statistics, “Health, United States, 2014: With special feature on adults aged 55–64,” 2015.
- [65] World Health Organization, *International statistical classification of diseases and related health problems*. World Health Organization, 2004, vol. 1, ISBN: 9241546492.
- [66] P. K. Mony and C. Nagaraj, “Health information management: An introduction to disease classification and coding,” *National Medical Journal of India*, vol. 20, no. 6, p. 307, 2007.
- [67] J. Sington and B. Cottrell, “Analysis of the sensitivity of death certificates in 440 hospital deaths: A comparison with necropsy findings,” *Journal of clinical pathology*, vol. 55, no. 7, pp. 499–502, 2002.
- [68] C. Hoff and R. Ratard, “Louisiana death certificate accuracy: A concern for the public’s health,” *J La State Med Soc*, vol. 162, no. 6, pp. 350–352, 2010.
- [69] J. Messite and S. D. Stellman, “Accuracy of death certificate completion: The need for formalized physician training,” *JAMA*, vol. 275, no. 10, pp. 794–796, 1996.
- [70] D. R. Lakkireddy, M. S. Gowda, C. W. Murray, K. R. Basarakodu, and J. L. Vacek, “Death certificate completion: How well are physicians trained and are cardiovascular causes overstated?” *The American Journal of Medicine*, vol. 117, no. 7, pp. 492–498, 2004.

- [71] R. Agarwal, J. M. Norton, K. Konty, R. Zimmerman, M. Glover, A. Lekachvili, H. McGruder, A. Malarcher, M. Casper, and G. A. Mensah, "Overreporting of deaths from coronary heart disease in new york city hospitals, 2003," *Preventing Chronic Disease*, vol. 7, no. 3, 2010.
- [72] L. M. Seske, L. J. Muglia, E. S. Hall, K. E. Bove, and J. M. Greenberg, "Infant mortality, cause of death, and vital records reporting in Ohio, United States," *Maternal and child health journal*, vol. 21, no. 4, pp. 727–733, 2017.
- [73] L. A. Johansson and R. Westerling, "Comparing Swedish hospital discharge records with death certificates: Implications for mortality statistics," *International journal of epidemiology*, vol. 29, no. 3, pp. 495–502, 2000.
- [74] L. E. Johns, A. M. Madsen, G. Maduro, R. Zimmerman, K. Konty, and E. Begier, "A case study of the impact of inaccurate cause-of-death reporting on health disparity tracking: New York City premature cardiovascular mortality," *American Journal of Public Health*, vol. 103, no. 4, pp. 733–739, 2013.
- [75] A. Madsen, S. Thihalolipavan, G. Maduro, R. Zimmerman, R. Koppaka, W. Li, V. Foster, and E. Begier, "An intervention to improve cause-of-death reporting in New York City hospitals, 2009–2010," *Preventing Chronic Disease*, vol. 9, 2012.
- [76] M. Gissler, R. Kauppila, J. Merilainen, H. Toukomaa, and E. Hemminki, "Pregnancy-associated deaths in Finland 1987-1994 - definition problems and benefits of record linkage," *Acta obstetricia et gynecologica Scandinavica*, vol. 76, no. 7, pp. 651–657, 1997.
- [77] J. Curb, C. Babcock, S. Pressel, B. Tung, R. Remington, and C. Hawkins, "Nosological coding of cause of death," *American journal of epidemiology*, vol. 118, no. 1, pp. 122–128, 1983.
- [78] R. L. Guibert, D. T. Wigle, and J. I. Williams, "Decline of acute myocardial infarction death rates not due to cause of death coding," *Canadian journal of public health= Revue canadienne de sante publique*, vol. 80, no. 6, pp. 418–422, 1989.
- [79] K. A. Myers and D. R. Farquhar, "Improving the accuracy of death certification," *Canadian Medical Association Journal*, vol. 158, no. 10, pp. 1317–1323, 1998.
- [80] C. Percy and C. Muir, "The international comparability of cancer mortality data: Results of an international death certificate study," *American Journal of Epidemiology*, vol. 129, no. 5, pp. 934–946, 1989.
- [81] A. Madsen and E. Begier, "Improving quality of cause-of-death reporting in New York City," *Preventing chronic disease*, vol. 10, 2013.

- [82] A. Madsen, S. Thihalolipavan, and M. Maduro, “A successful intervention to improve the quality of cause of death reporting in New York City hospitals. (in press),” *Preventing Chronic Disease*, 2017.
- [83] National Association for Public Health Statistics and Information Systems, *NAPHSIS training resources*, <http://www.naphsis.org/index.asp?bid=1156>, Web Page.
- [84] Centers for Disease Control and Prevention (CDC), *Physician’s handbook on medical certification of death, 2003 revision*, http://www.cdc.gov/nchs/data/misc/hb_cod.pdf, 2003.
- [85] N. A. Skopp, D. J. Smolenski, D. A. Schwesinger, C. J. Johnson, M. J. Metzger-Abamukong, and M. A. Reger, “Evaluation of a methodology to validate National Death Index retrieval results among a cohort of U.S. service members,” *Ann Epidemiol*, vol. 27, no. 6, pp. 397–400, 2017.
- [86] L. K. Tanno, A. L. Bierrenbach, M. A. Calderon, A. Sheikh, F. Estelle R Simons, and P. Demoly, “Increasing the accuracy of notification of anaphylaxis deaths in Brazil through the International Classification of Diseases (ICD)-11 revision,” *Journal of Allergy and Clinical Immunology*, vol. 139, no. 2, AB226,
- [87] V. Flenady, A. M. Wojcieszek, D. Ellwood, S. H. Leisher, J. J. H. Erwich, E. S. Draper, E. M. McClure, H. E. Reinebrant, J. Oats, and L. McCowan, “Classification of causes and associated conditions for stillbirths and neonatal deaths,” in *Seminars in Fetal and Neonatal Medicine*, Elsevier.
- [88] K. J. Foreman, M. Naghavi, and M. Ezzati, “Improving the usefulness of US mortality data: New methods for reclassification of underlying cause of death,” *Population health metrics*, vol. 14, no. 1, p. 14, 2016.
- [89] T. H. Lu, “Using ACME (Automatic Classification of Medical Entry) software to monitor and improve the quality of cause of death statistics,” *Journal of Epidemiology and Community Health*, vol. 57, pp. 470–471, 2003.
- [90] *DIMDI - Iris Institute*, <https://www.dimdi.de/static/en/klassi/irisinstitute/about-iris/index.htm>, Web Page.
- [91] C. Tao, K. Wongsuphasawat, K. Clark, C. Plaisant, B. Shneiderman, and C. G. Chute, “Towards event sequence representation, reasoning and visualization for EHR data,” in *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, ser. IHI ’12, Miami, Florida, USA: Association for Computing Machinery, 2012, pp. 801–806, ISBN: 9781450307819.
- [92] T. D. Wang, C. Plaisant, A. J. Quinn, R. Stanchak, S. Murphy, and B. Shneiderman, “Aligning temporal data by sentinel events: Discovering patterns in electronic

- health records,” in *Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM, pp. 457–466, ISBN: 1605580112.
- [93] H. Syed and A. K. Das, “Identifying chemotherapy regimens in electronic health record data using interval-encoded sequence alignment,” in *Artificial Intelligence in Medicine*. Springer, 2015, pp. 143–147, ISBN: 3319195506.
 - [94] I. J. Casanova, M. Campos, J. M. Juarez, A. Fernandez-Fernandez-Arroyo, and J. A. Lorente, “Using multivariate sequential patterns to improve survival prediction in intensive care burn unit,” in *Artificial Intelligence in Medicine: 15th Conference on Artificial Intelligence in Medicine, AIME 2015, Pavia, Italy, June 17-20, 2015. Proceedings*, H. J. Holmes, R. Bellazzi, L. Sacchi, and N. Peek, Eds. Cham: Springer International Publishing, 2015, pp. 277–286, ISBN: 978-3-319-19551-3.
 - [95] S. Konias, G. D. Giaglis, G. Gogou, P. D. Bamidis, and N. Maglaveras, “Uncertainty rule generation on a home care database of heart failure patients,” in *Computers in Cardiology, 2003*, pp. 765–768, ISBN: 0-7803-8170-X.
 - [96] C. Ordonez, E. Omiecinski, L. De Braal, C. A. Santana, N. Ezquerra, J. A. Taboada, D. Cooke, E. Krawczynska, and E. V. Garcia, “Mining constrained association rules to predict heart disease,” in *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*, pp. 433–440.
 - [97] H. Liu and H. Du, “Stock sequence pattern mining method based on SWI-GSP algorithm,” in *Proceedings of the 2017 International Conference on Data Mining, Communications and Information Technology*, ACM, p. 3, ISBN: 1450352189.
 - [98] D. Zhang and L. Zhou, “Discovering golden nuggets: Data mining in financial application,” *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 34, no. 4, pp. 513–522, 2004.
 - [99] H. Q. Vu, G. Li, R. Law, and Y. Zhang, “Travel diaries analysis by sequential rule mining,” *Journal of Travel Research*, pp. 399–413, 2018.
 - [100] S. Tiwari and L. K. Tiwari, “Sequential rule mining in M-learning domain,” *International Journal of Computer Applications*, vol. 134, no. 3, pp. 23–29, 2016.
 - [101] M.-Y. Lin and S.-Y. Lee, “Fast discovery of sequential patterns by memory indexing,” in *Data Warehousing and Knowledge Discovery*. Springer, 2002, pp. 150–160, ISBN: 3540441239.
 - [102] G. N. Pradhan and B. Prabhakaran, “Association rule mining in multiple, multidimensional time series medical data,” *Journal of Healthcare Informatics Research*, vol. 1, no. 1, pp. 92–118, 2017.

- [103] M. Vandromme, J. Jacques, J. Taillard, A. Hansske, L. Jourdan, and C. Dhaenens, “Extraction and optimization of classification rules for temporal sequences: Application to hospital data,” *Knowledge-Based Systems*, vol. 122, pp. 148–158, 2017.
- [104] S. Concaro, L. Sacchi, C. Cerra, P. Fratino, and R. Bellazzi, “Mining health care administrative data with temporal association rules on hybrid events,” *Methods Inf Med*, vol. 50, no. 2, pp. 166–179, 2011.
- [105] R. Chaves, J. M. Górriz, J. Ramírez, I. A. Illán, D. Salas-Gonzalez, and M. Gómez-Río, “Efficient mining of association rules for the early diagnosis of Alzheimer’s disease,” *Physics in Medicine and Biology*, vol. 56, no. 18, p. 6047, 2011.
- [106] R. Agrawal, T. Imieliński, and A. Swami, “Mining association rules between sets of items in large databases,” in *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD ’93, Washington, D.C., USA: Association for Computing Machinery, 1993, pp. 207–216, ISBN: 0897915925.
- [107] R. A. Hoffman, H. Wu, J. Venugopalan, P. Braun, and M. D. Wang, “Intelligent mortality reporting with FHIR,” in *2017 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, 2017, pp. 181–184.
- [108] HL7, *FHIR Resource List*, <https://www.hl7.org/fhir/resourcelist.html>, Web Page, Accessed 3/5/2021, 2019.
- [109] —, *FHIR RESTful API*, <https://www.hl7.org/fhir/http.html>, Web Page, Accessed 3/5/2021, 2019.
- [110] —, *Resource Patient - Content (v4.0.1)*, <https://www.hl7.org/fhir/patient.html>, Web Page, Accessed 3/5/2021, 2019.
- [111] —, *Resource Patient - Content (v1.0.2)*, <http://hl7.org/fhir/DSTU2/patient.html>, Web Page, Accessed 4/28/2021, 2015.
- [112] —, *FHIR Package Registry*, <http://registry.fhir.org>, Web Page, Accessed 4/28/2021, 2021.
- [113] SIMPLIFIER.NET, *The FHIR Collaboration Platform - SIMPLIFIER.NET*, <https://simplifier.net>, Web Page, Accessed 4/28/2021, 2021.
- [114] HL7, *Resource Bundle - Content (v4.0.1)*, <https://www.hl7.org/fhir/bundle.html>, Web Page, Accessed 3/5/2021, 2019.

- [115] D. C. Cowper, J. D. Kubal, C. Maynard, and D. M. Hynes, "A primer and comparative review of major US mortality databases," *Annals of epidemiology*, vol. 12, no. 7, pp. 462–468, 2002.
- [116] R. H. Dolin, L. Alschuler, S. Boyer, C. Beebe, F. M. Behlen, P. V. Biron, and A. Shabo, "HL7 clinical document architecture, release 2," *Journal of the American Medical Informatics Association*, vol. 13, no. 1, pp. 30–39, 2006.
- [117] R. H. Dolin, L. Alschuler, C. Beebe, P. V. Biron, S. L. Boyer, D. Essin, E. Kimber, T. Lincoln, and J. E. Mattison, "The HL7 clinical document architecture," *Journal of the American Medical Informatics Association*, vol. 8, no. 6, pp. 552–569, 2001.
- [118] M. Eichelberg, T. Aden, J. Riesmeier, A. Dogac, and G. B. Laleci, "Electronic health record standards-a brief overview," in *Proceedings of the 4th IEEE International Conference on Information and Communications Technology (ICICT 2006)*, Citeseer, 2006.
- [119] D. Blumenthal and M. Tavenner, "The "meaningful use" regulation for electronic health records," *New England Journal of Medicine*, vol. 363, no. 6, pp. 501–504, 2010.
- [120] G. Alterovitz, J. Warner, P. Zhang, Y. Chen, M. Ullman-Cullere, D. Kreda, and I. S. Kohane, "SMART on FHIR genomics: Facilitating standardized clinico-genomic apps," *Journal of the American Medical Informatics Association*, vol. 22, no. 6, pp. 1173–1178, 2015.
- [121] M. Smits, E. Kramer, M. Harthoorn, and R. Cornet, "A comparison of two detailed clinical model representations: Fhir and cda," *European Journal for Biomedical Informatics*, vol. 11, no. 2, 2015.
- [122] H. Yang and C. C. Yang, "Using health-consumer-contributed data to detect adverse drug reactions by association mining with temporal analysis," *ACM Trans. Intell. Syst. Technol.*, vol. 6, no. 4, pp. 1–27, 2015.
- [123] R. Bellazzi, F. Ferrazzi, and L. Sacchi, "Predictive data mining in clinical medicine: A focus on selected methods and applications," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, no. 5, pp. 416–430, 2011.
- [124] Q. Zhao and S. S. Bhowmick, "Sequential pattern mining: A survey," *ITechnical Report CAIS Nanyang Technological University Singapore*, pp. 1–26, 2003.
- [125] B. R. Mistry and A. Desai, "Privacy preserving heuristic approach for association rule mining in distributed database," in *Innovations in Information, Embedded and Communication Systems (ICIIECS), 2015 International Conference on*, IEEE, pp. 1–7, ISBN: 147996817X.

- [126] A. W.-C. Fu, R. C.-W. Wong, and K. Wang, "Privacy-preserving frequent pattern mining across private databases," in *Data Mining, Fifth IEEE International Conference on*, IEEE, 4 pp. ISBN: 0769522785.
- [127] K. Sathiyapriya and G. S. Sadasivam, "A survey on privacy preserving association rule mining," *International Journal of Data Mining and Knowledge Management Process*, vol. 3, no. 2, p. 119, 2013.
- [128] J. Soni, U. Ansari, D. Sharma, and S. Soni, "Predictive data mining for medical diagnosis: An overview of heart disease prediction," *International Journal of Computer Applications*, vol. 17, no. 8, pp. 43–48, 2011.
- [129] G. Dolce, M. Quintieri, S. Serra, V. Lagani, and L. Pignolo, "Clinical signs and early prognosis in vegetative state: A decisional tree, data-mining study," *Brain Injury*, vol. 22, no. 7-8, pp. 617–623, 2008.
- [130] M. R. McNally, C. L. Patton, and W. J. Fremouw, "Mining for murder-suicide: An approach to identifying cases of murder-suicide in the National Violent Death Reporting System Restricted Access Database," *Journal of forensic sciences*, 2015.
- [131] S. N. Kasthurirathne, B. Mamlin, H. Kumara, G. Grieve, and P. Biondich, "Enabling better interoperability for healthcare: Lessons in developing a standards based application programming interface for electronic medical record systems," *Journal of Medical Systems*, vol. 39, no. 11, p. 182, 2015.
- [132] S. Boytcheva, G. Angelova, D. Tcharaktchiev, and Z. Angelov, "Big data analytics in healthcare—pattern mining of temporal clinical events," 2011.
- [133] A. P. Wright, A. T. Wright, A. B. McCoy, and D. F. Sittig, "The use of sequential pattern mining to predict next prescribed medications," *Journal of biomedical informatics*, vol. 53, pp. 73–80, 2015.
- [134] The 111th United States Congress, *Public Law 111-5 (American Recovery and Reinvestment Act of 2009)*, <https://www.govinfo.gov/content/pkg/PLAW-111publ5/html/PLAW-111publ5.htm>, 2009.
- [135] G. Grieve, *Resources for healthcare*, <http://hl7.org/fhir/2011Aug/>, Web Page, 2011.
- [136] HL7, *FHIR Publication (Version) History*, <http://hl7.org/fhir/directory.html>, Web Page, Accessed 4/28/2021, 2021.
- [137] R. A. Hoffman, S. Kothari, and M. D. Wang, "Comparison of normalization algorithms for cross-batch color segmentation of histopathological images," in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, IEEE, 2014, pp. 194–197.

- [138] L. Pantanowitz, A. J. Evans, J. D. Pfeifer, L. C. Collins, P. N. Valenstein, K. J. Kaplan, D. C. Wilbur, and T. J. Colgan, “Review of the current state of whole slide imaging in pathology,” *Journal of Pathology Informatics*, vol. 2, no. 1, p. 36, 2011.
- [139] NIH, *The Cancer Genome Atlas*, <http://cancergenome.nih.gov>, Catalog.
- [140] S. Kothari, J. H. Phan, T. H. Stokes, and M. D. Wang, “Pathology imaging informatics for quantitative analysis of whole-slide images,” *Journal of the American Medical Informatics Association*, 2013.
- [141] S. Kothari, J. H. Phan, R. A. Moffitt, T. H. Stokes, S. E. Hassberger, Q. Chaudry, A. N. Young, and M. D. Wang, “Automatic batch-invariant color segmentation of histological cancer images,” *IEEE International Symposium on Biomedical Imaging: From Nano to Macro. Proceedings*, pp. 657–660, 2011.
- [142] D. Magee, D. Treanor, D. Crellin, M. Shires, K. Smith, K. Mohee, and P. Quirke, “Colour normalisation in digital histopathology images,” in *Proc Optical Tissue Image analysis in Microscopy, Histopathology and Endoscopy (MICCAI Workshop)*, pp. 100–111.
- [143] D. Magee, D. Treanor, D. Crellin, M. Shires, K. Smith, K. Mohee, and P. Quirke, “Colour normalisation in digital histopathology images,” pp. 100–111, 2009.
- [144] E. Reinhard, M. Adhikhmin, B. Gooch, and P. Shirley, “Color transfer between images,” *Computer Graphics and Applications, IEEE*, vol. 21, no. 5, pp. 34–41, 2001.
- [145] A. C. Ruifrok and D. A. Johnston, “Quantification of histochemical staining by color deconvolution,” *Anal Quant Cytol Histol*, vol. 23, no. 4, pp. 291–299, 2001.
- [146] R. A. Hoffman, S. Kothari, J. H. Phan, and M. D. Wang, “A high-resolution tile-based approach for classifying biological regions in whole-slide histopathological images,” *Proceedings of the International Federation for Medical and Biological Engineering (IFMBE)*, vol. 42, pp. 280–283, 2014.
- [147] M. M. Isaacs, J. K. J. K. Lennerz, S. S. Yates, W. W. Clermont, J. J. Rossi, and J. D. J. D. Pfeifer, “Implementation of whole slide imaging in surgical pathology: A value added approach,” *Journal of Pathology Informatics*, vol. 2, pp. 39–39, 2011.
- [148] S. Kothari, J. H. Phan, A. N. Young, and M. D. Wang, “Histological image feature mining reveals emergent diagnostic properties for renal cancer,” *BIBM*, pp. 422–425, 2011.

- [149] S. Kothari, J. H. Phan, and M. D. Wang, “Eliminating tissue-fold artifacts in histopathological whole-slide images for improved image-based prediction of cancer grade,” *Journal of pathology informatics*, 2013.
- [150] J. Costa, R. A. Wesley, E. Glatstein, and S. A. Rosenberg, “The grading of soft tissue sarcomas. results of a clinicohistopathologic correlation in a series of 163 cases,” *Cancer*, vol. 53, no. 3, pp. 530–541, 1984.
- [151] A. R. Smith, “Color gamut transform pairs,” *ACM Siggraph Computer Graphics*, vol. 12, no. 3, pp. 12–19, 1978.
- [152] S. Kothari, J. H. Phan, A. O. Osunkoya, and M. D. Wang, “Biological interpretation of morphological patterns in histopathological whole-slide images,” in *BCB '12: Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine*, ACM Request Permissions.
- [153] S. Kothari, Q. Chaudry, and M. D. Wang, “Automated cell counting and cluster segmentation using concavity detection and ellipse fitting techniques,” *IEEE International Symposium on Biomedical Imaging: From Nano to Macro. Proceedings*, pp. 795–798, 2009.
- [154] H. Peng, F. Long, and C. Ding, “Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.
- [155] B. Karaçali and A. Tözeren, “Automated detection of regions of interest for tissue microarray experiments: An image texture analysis,” *BMC Medical Imaging*, vol. 7, no. 1, p. 2, 2007.
- [156] K. Lesack and C. Naugler, “Performance of a simple chromatin-rich segmentation algorithm in quantifying basal cell carcinoma from histology images,” *BMC research notes*, vol. 5, no. 1, p. 35, 2012.
- [157] C. Tanade, N. Pate, E. Paljug, R. A. Hoffman, and M. D. Wang, “Hybrid modeling of ebola propagation,” in *2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE)*, IEEE, 2019, pp. 204–210.
- [158] J. Dunn, H. Qiu, S. Kim, D. Jjingo, R. Hoffman, C. W. Kim, I. Jang, D. J. Son, D. Kim, C. Pan, *et al.*, “Flow-dependent epigenetic DNA methylation regulates endothelial gene expression and atherosclerosis,” *The Journal of clinical investigation*, vol. 124, no. 7, pp. 3187–3199, 2014.
- [159] E. R. Cosman Jr, J. R. Dolensky, and R. A. Hoffman, “Factors that affect radiofrequency heat lesion size,” *Pain Medicine*, vol. 15, no. 12, pp. 2020–2036, 2014.

- [160] J. H. Phan, R. Hoffman, S. Kothari, P.-Y. Wu, and M. D. Wang, "Integration of multi-modal biomedical data to predict cancer grade and patient survival," in *2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, IEEE, 2016, pp. 577–580.
- [161] P.-Y. Wu, C.-W. Cheng, C. D. Kaddi, J. Venugopalan, R. Hoffman, and M. D. Wang, "--omic and electronic health record big data analytics for precision medicine," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 2, pp. 263–273, 2017.
- [162] L. Tong, R. Hoffman, S. R. Deshpande, and M. D. Wang, "Predicting heart rejection using histopathological whole-slide imaging and deep neural network with dropout," in *2017 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, IEEE, 2017, pp. 1–4.
- [163] A. Bhatia, L. Tong, R. Hoffman, P. Wu, H. Hassanzadeh, M. Wang, and S. Deshpande, "Refinement of automated whole slide image analysis in pediatric heart transplants," *The Journal of Heart and Lung Transplantation*, vol. 36, no. 4, S103–S104, 2017.
- [164] S. K. Phan, R. Hoffman, and M. D. Wang, "Biomedical imaging informatics for diagnostic imaging marker selection," in *Health Informatics Data Analysis*, Springer, 2017, pp. 115–127.
- [165] Y. Zhu, M. Saqib, E. Ham, S. Belhareth, R. Hoffman, and M. D. Wang, "Mitigating patient-to-patient variation in eeg seizure detection using meta transfer learning," in *2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)*, 2020, pp. 548–555.
- [166] Y. Zhu, Y. Sha, H. Wu, M. Li, R. A. Hoffman, and M. D. Wang, "Public health informatics: Proposing causal sequence of death using neural machine translation," Pre-print, 2020.
- [167] R. H. Coase, *How should economists choose?* Lecture, 1981.

VITA

Ryan Hoffman received a B.S. degree in Biomedical Engineering in 2012 from Georgia Tech, Atlanta, GA, USA. He has pursued his Ph.D. research working under Prof. May Wang in the Biomedical Informatics and Bio-Imaging Laboratory at Georgia Tech and Emory University. His graduate research has included wide-ranging work in the areas of histopathological image processing, critical care informatics, and FHIR application development.

In addition to his own graduate studies, Ryan places a heavy emphasis on teaching and mentorship. In his time both as a graduate and undergraduate student at Georgia Tech he has served as a teaching assistant for over 18 semester-courses, including experience as a TA for a study abroad course conducted in collaboration with Peking University. Multiple student groups that he has mentored have gone on to have their class projects peer-reviewed and published.