AUTOMATED DETECTION OF SURGICAL ADVERSE EVENTS FROM RETROSPECTIVE CLINICAL DATA

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA

BY

ZHEN HU

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

ADVISOR: GENEVIEVE B. MELTON-MEAUX, MD, PHD
CO-ADVISOR: GYORGY J. SIMON, PHD

AUGUST 2017

Acknowledgements

I would like to express the deepest appreciation to my advisor, Dr. Genevieve Melton-Meaux, and my co-advisor, Dr. Gyorgy Simon. It has been an honor and pleasure to work with them.

I would like to thank Genevieve, for providing me the opportunity to work on this challenging and exciting project. Her exceptional guidance and continuous support has kept me on track, and I have always been inspired by her numerous ideas and insights into the research in health informatics.

I would like to thank Gyorgy, for his great guidance in data analytics. Whenever I have a problem in project, he could always help me figure it out by his versatile knowledge and skills in machine learning and statistics. I have benefitted a lot from working closely with him.

Thanks are also due to Dr. David Pieczkiewicz, Dr. Terrence Adam, and Dr. Weihua Guan for agreeing to serve on my dissertation committee and for sparing their invaluable time reviewing this dissertation. Also, I would like to express my gratitude to Dr. Chih-Lin Chi, Dr. Lael Gatewood, Dr. Stuart Speedie, and Dr. Rui Zhang, at the Institute for Health Informatics for their help and support.

My collaborators and colleagues at the Institute of Health Informatics and the Department of Surgery have always been very supportive. Many thanks to:

- IHI: Dr. Yan Wang, Dr. Zohara Cohen, Reed McEwan, Jessica Whitcomb-Trance, and Lindsay Bork.
- Surgery: Dr. Elliot Arsoniadis, Dr. Steven Skube, Dr. Mary Kwaan, Dr. Eric Jensen, Alexandra Broek, and Carol Bigalke.

I would like to thank other students at the Institute for Health Informatics for their support and friendship. They are Era Kim, Jin Wang, Wonsuk Oh, and Ranyah Aldekhyyel.

Finally, I would like to acknowledge Fairview Health Services and the University of Minnesota Academic Health Center for support of the research.

Dedication

This dissertation is dedicated to the memory of my beloved mother, Guo Xiu-Rong.

Her love and encouragement have sustained me throughout my life.

Abstract

The Detection of surgical adverse events has become increasingly important with the growing demand for quality improvement and public health surveillance with surgery. Event reporting is one of the key steps in determining the impact of postoperative complications from a variety of perspectives and is an integral component of improving transparency around surgical care and ultimately around addressing complications. Manual chart review is the most commonly used method in identification of adverse events. Though the manual chart review is the most commonly used method that is considered the "gold-standard" for detecting adverse events for many patient safety studies (research setting), it could be very labor-intensive and time-consuming and thus many hospitals have found it too expensive to routinely use.

In this dissertation, aiming to accelerate the process of extracting postoperative outcomes from medical charts, an automated postoperative adverse events detection application has been developed by using structured electronic health record (EHR) data and unstructured clinical notes. First, pilot studies are conducted to test the feasibility by using only completed EHR data and focusing on three types of surgical site infection (SSI). The built models have high specificity as well as very high negative predictive values, reliably eliminating the vast majority of patients without SSI, thereby significantly reducing the chart reviewers' burden. Practical missing data treatments have also been explored and compared. To address modeling challenges, such as high-dimensional dataset, and imbalanced distribution, several machine learning methods haven been applied. Particularly, one single-task and five multi-task learning methods are developed and

compared for their detection performance. The models demonstrated high detection performance, which ensures the feasibility of accelerating the manual process of extracting postoperative outcomes from medical chart. Finally, the use of structured EHR data, clinical notes and the combination of these data types have been separately investigated. Models using different types of data were compared on their detection performance. Models developed with very high AUC score have demonstrated that supervised machine learning methods can be effective for automated detection of surgical adverse events.

Table of Contents

List of Tables

List of Figures

Chapter 1 Introduction

1.1 Problem and Significance

An adverse event is defined as "*an unintended injury or complication resulting in prolonged length of hospital stay, disability at the time of discharge or death caused by healthcare management and not by the patients' underlying disease*." Adverse events are costly (1-2), can result in significant patient harm (3-4), and overall 50% to 75% of all adverse events are associated with surgery.

According to recent reports, adverse events affect nearly one out of ten surgical patients during hospital admission, contributing significantly to postoperative morbidity and mortality. Surgical adverse events are very expensive to deal with and have been identified as an important cause of increased health care costs (5-6). In addition, about 40% of surgical adverse events, are potentially preventable (7-8).

Historically, despite their importance, there were few sources that healthcare institutions could use for understanding their surgical adverse event rates. Adverse event detection traditionally mainly relies on voluntary reporting systems and manual retrospective chart review. Voluntary reporting is the most often used method, but misses the majority of events, which means most adverse events go underreported, and approximately 63% of such events are undetected. The most reliable adverse event detection sources have historically relied on costly manual chart abstraction and many hospitals recognize the importance of having accurate data and utilize surgical clinical registries (e.g., NSQIP, STS, Trauma Registry) for quality improvement, outcome

surveillance, and determining the impact of postoperative complications on downstream clinical outcomes. Conducting research and quality improvement using manual chart review remains widely used in traditional observational clinical studies aimed at assessing detailed information on patients to understand disease course or outcomes and is also a primary modality used for quality improvement, epidemiologic assessments, and for graduate and ongoing professional education and assessment. Sources for surgical adverse event detection include administrative billing and coding data, discharge summaries, and other clinical data in the electronic health record (EHR). While manual chart review is considered the "gold-standard" for identifying adverse events for many patient safety studies (research setting), many hospitals have found it too expensive to routinely use.

The objective of this body of research is to explore, validate, and expand upon approaches for automatically detecting adverse events using structured electronic health record (EHR) data and to include information extraction from unstructured clinical notes leveraging adverse event detection conducted using laborious chart abstraction.

There are two main advantages of using electronic methods for surgical adverse event detection. They are based on routinely collected, readily available EHR data and clinical notes, and are therefore less expensive and less time-consuming than usual manual way. In addition, automated methods are based on objective criteria (e.g., diagnostic codes, value of lab test, etc.) that may be able to lead to standardized detection processes and should eliminate reviewer subjectivity and error.

1.2 Related work

A large body of research work has been conducted with the automated detection of adverse events caused by medication related harm. Using computerized algorithms to screen electronic healthcare databases for events has proven to be an effective method (9-11). There are, however, only a few automated applications for surgical adverse event detection developed to date (12-13).

Some rule-based approaches include "trigger tools" which work by detecting "triggers" of interest, like an abnormal laboratory value or a low blood pressure, which serve as clues to a possible adverse event. It is, however, difficult for these methods to deal with adverse events that are not based upon intuitive and obvious sets of rules.

Machine learning is a promising set of methods which could identify latent patterns associated with surgical adverse events and help with discovering the underlying cause and nature of adverse events that injure patients. These methods have been applied to a variety of use cases including the early detection of disease or mining health records for a disease's risks (14-15). These techniques can be used to exploit numeric or coded clinical EHR data (i.e., structured data) and also can be used on data from unstructured sources using text mining or natural language processing (NLP) from clinical notes, such as discharge summaries, operative reports, and nursing notes (16).

1.3 Aims and Hypotheses

The overall objective of this dissertation is to develop an automated detection platform for surgical adverse events based on local EHR data and clinical notes at the

University of Minnesota Medical Center. This work leverages data provided by surgical clinical reviewers who manually abstract surgical adverse events for surgical patients for a national surgical quality improvement registry.

The proposed task of automatically detecting surgical adverse events from clinical data is complicated due to a number of challenges with the associated dataset, which require specific considerations to address appropriately. Those main challenges include:

(1) how to preprocess EHR data and generate features for further modeling

(2) understanding reason(s) for missingness of EHR data and utilizing appropriate techniques to handle missing data (e.g., imputation methods)

(3) methods to address the high-dimensional nature of the associated dataset

(4) issues associated with processing an imbalanced dataset

Machine learning represents a complex field in itself and offers a wide range of solutions that can be potentially employed to address these challenges. In conducting the research, careful balance and analysis of the strengths and weaknesses of these approaches must be performed.

Overall, the hypotheses of the dissertation are:

(1) with reliably labeled adverse events as a gold standard, supervised machine learning methods should be effective for automatic detection of surgical adverse events

(2) the combination of structured EHR data and unstructured clinical notes would improve detection performance compared to solely using either structured EHR data, NLP or text mining from unstructured clinical notes.

1.4 Outline of thesis

Chapter 2 introduces the background of the clinical registry utilized in this work: American College of Surgeons (ACS) National Surgical Quality Improvement Program (NSQIP). NSQIP used well-defined and a validated process for documenting and detecting surgical adverse events. Additionally, a general primer around EHR data and clinical notes used is presented, as well, in Chapter 2.

Chapter 3 describes an initial study conducted that tests the feasibility of the overall project with surgical site infection as a use case. Three subtypes of surgical site infections are detected based on patients with complete EHR datasets. Methods for summarizing variables based on EHR data are presented. The study's results including model evaluation and significant features selected are discussed, as well.

Chapter 4 provides a specific analysis for exploring situations of missing data which occurs not uncommonly with EHR data. Several practical imputation methods are described, compared and discussed.

Chapter 5 expands upon previous methodologies and examines the application of multi-task learning for surgical adverse event detection. Though different adverse events have their own diagnosis method and treatment, a number of events (e.g., infectious events such as surgical site infection, sepsis, pneumonia) share some similar features in either diagnosis or treatment. Multi-task methods are designed and developed not only to find similar features but also solve the challenges caused by high-dimensional and imbalanced dataset.

As a culmination of the associated work, Chapter 6 examines a range of surgical adverse events and demonstrates the value of adding NLP and information extraction for clinical notes to this problem. In this study, automated detection based on both EHR data alone, NLP with clinical notes, or both sources are compared.

Finally, Chapter 7 summarizes and discusses the contribution and significance of the overall dissertation. Future directions and work for potential future investigations are discussed.

Chapter 2 Background

2.1 NSQIP

As stated in Chapter 1, our research leverages outcomes from the surgical registry, ACS NSQIP. NSQIP is widely recognized as "the best in the nation" surgical quality improvement resource in the United States (17-19). With the guidance of NSQIP, participating hospitals track outcomes and other patient variables using manual abstraction. Variables collected include preoperative, intraoperative, and postoperative clinical data elements and morbidity/complication occurrences. The preoperative and intraoperative clinical data elements include patient demographics, co-morbidities and disease history, functional status, laboratory results, operation duration, and wound classification scores. Postoperative morbidity outcomes include 21 well-defined surgical adverse events (i.e., complications) within the 30-day postoperative window. These adverse event occurrences include surgical site infections (SSIs), urinary tract infections (UTI), sepsis, and acute renal failure (ARF). These events each have detailed definitions with specific inclusion and exclusion criteria. For example, **Table 2-1** includes a standard definition used in NSIQP for sepsis. Surgical patients are selected at each hospital for inclusion into NSQIP (based on a cyclical schedule and a certain target number with stratified sampling to preferentially select major surgical cases). For each of selected patient cases, all preoperative, intraoperative, and postoperative data are collected, entered, and included in NSQIP which then has a number of statistical analysis performed on the data to provide feedback.

Table 2-1. The standard definition of sepsis used by the NSQIP registry

Report this event if the patient has two of the following clinical signs and symptoms of SIRS:

- Temp > 38°C (100.4 °F) or < 36 °C (96.8°F)
- HR > 90 bpm
- RR > 20 breaths/min or Pa CO_2<32 mmHg (<4.3 kPa)
- WBC > 12,000 cell/mm^3, < 4,000 cells/mm3, or > 10% immature band forms.
- Anion gap acidosis: this is defined by either:
    - [Na + K] – [Cl + HCO_3 (or serum CO_2)]. If this number is greater than 16, then an anion gap acidosis is present.
    - Na – [Cl + HCO_3 (or serum CO_2)]. If this number is greater than 12, then an anion gap acidosis is present.

*If anion gap lab values are performed at your facilities lab, ascertain which formula is utilized and follow guideline criteria.

And either A or B below:

A. One of the following:

- Positive blood culture
- Clinical documentation of purulence or positive culture from any site for which there is documentation noting the site as the acute cause of sepsis.

B. One of the following findings during the Principal Operative Procedure**:**

- Confirmed infarcted bowel requiring resection
- Purulence in the operative site
- Enteric contents in the operative site, or
- Positive intra-operative cultures

Guidance: if the patient meets criteria to assign preoperative sepsis, assign the risk factor; if the patient meets the criteria to assign postop sepsis, assign the occurrence and then assess for PATOS and assign if appropriate.

NSQIP uses the collected data from all member hospitals to calculate the hospital's relative performance for different types of operations with respect to adjusted postoperative morbidity and mortality and compares each member hospital's performance with a benchmark for each postoperative adverse event. Specifically, a ratio of observed to expected number of events is provided to each hospital for each event adjusted by patient morbidity, case complexity, and a number of other factors. An O/E ratio of 1 means the performance is as expected for a particular outcome given the composite patient and case severity, whereas less or greater than one indicates better or worse performance, respectively (20-23). With this feedback, NSQIP member hospitals are able to focus on areas of improvement and have achieved measurable improvement in surgical care quality and in many cases have saved money by reducing length of stay and preventable readmissions. **Figure 2-1** demonstrates the mechanism NSQIP utilizes for surgical quality improvement.

The success of NSQIP in improving surgical quality for member hospitals is ensured by the high quality collection of clinical data elements performed with manual abstraction. To maintain the high reliability of this data collection, formally trained surgical clinical reviewers are employed by hospitals. These individuals select surgery cases strictly following NSQIP inclusion and exclusion criteria, manually extract preoperative data characteristics, and then recognize and record 21 postoperative surgical adverse events and mortality. Because data collection for NSQIP is very time and labor intensive, only a subset of surgical patients are selected, and despite the registry's value, its cost poses a significant burden for medical institutions. Unfortunately, mainly due to this costly manual manner of

clinical data collection and other costs like NSQIP's associated participation fee, less than 20% of hospitals in the United States currently are enrolled in NSQIP.

To make NSQIP more accessible, one proposed and promising solution is to accelerate the process of data extraction by automatically or semi-automatically detecting NSQIP elements from EHR systems. While EHR data and specific modules have been built (e.g., Epic NSQIP module) for automated abstraction of preoperative and intraoperative clinical data elements, these approaches have not been used for adverse event outcome data. Preoperative and intraoperative data also tend to be structured, and are much easier to extract without complicated algorithms. Since reviewers spent most of their time identifying post-operative adverse events and these outcomes are most relevant for performance of a healthcare institution, the goal of the associated research is to utilize automated techniques, specifically machine-learning, to classify surgical patients with or without particular postoperative adverse events.

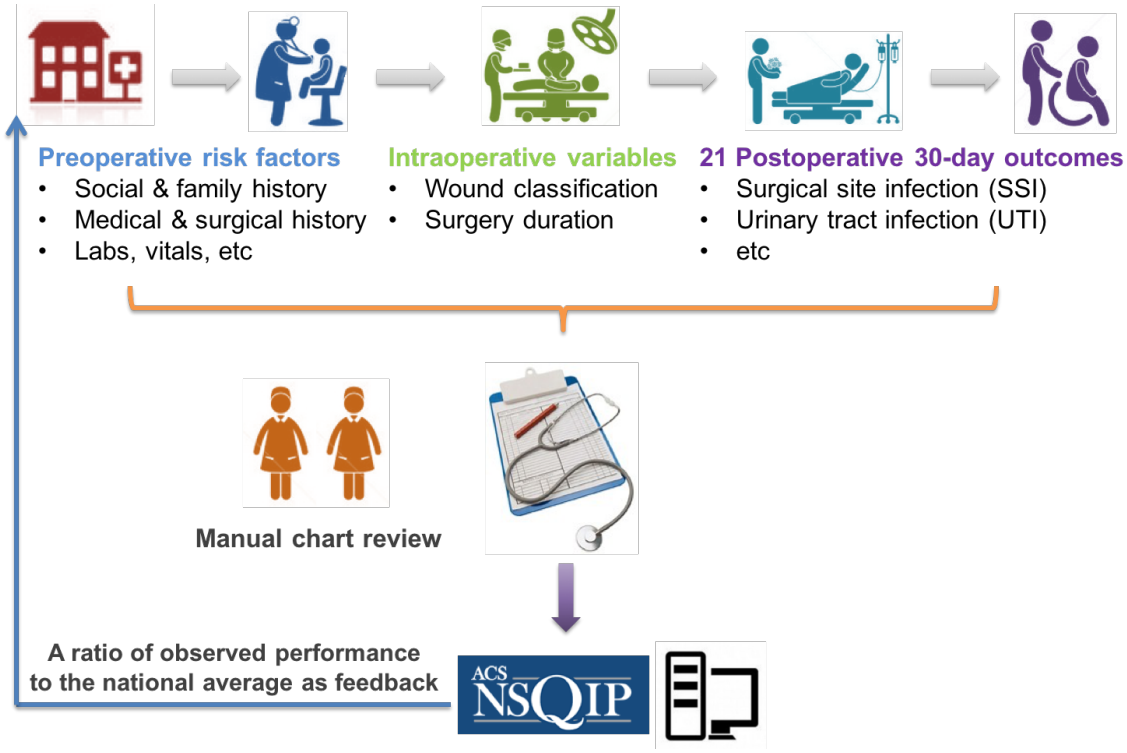

Figure 2-1. The overall review of NSQIP standard workflow

2.2 Hospital acquired infections

Hospital acquired infections (HAIs) are infections acquired in a hospital or other health care facility. **Table 2-2** list all 21 surgical adverse events defined and recorded in NSQIP. Among all events, infections in italic type are the HAIs.

HAIs account for nearly 60% of all complications (24-25). They are not only the most common surgical adverse events, but also very morbid. Severe SSIs and pneumonia could trigger sepsis and even septic shock, particularly in people who are already at risk. Sepsis and septic shock are common and deadly, and CDC has listed “septicemia” as the

11th leading cause of death nationwide (26-28). In addition, HAIs are expensive to treat. According to a recent study, 440,000 of these adverse events happen annually and cost overall up to 10 billion dollars per year in the United States (29-31). Given the significant influence on the quality and cost of healthcare, postoperative HAIs are increasingly and widely viewed as a quality benchmark and are a strong emphasis of national initiatives for infection prevention and control (32-35). Therefore, our research has a particular emphasis on HAI detection in surgical patients and SSI in particular in several of the studies.

Table 2-2. 21 defined postoperative surgical adverse events in NSQIP registry

Superficial surgical site infection	On ventilator > 48h
Deep surgical site infection	Pulmonary embolism
Organ space surgical site infection	Progressive renal insufficiency
Urinary tract infection	Acute renal failure
Pneumonia	Stroke/Cerebrovascular accident (CVA)
Sepsis	Bleeding requiring transfusion
Septic shock	Coma
Wound disruption	Perioperative nerve injuries
Cardiac arrest requiring CPR	Deep vein thrombosis (DVT)
Myocardial infarction	Postoperative death within 30 day
Unplanned reintubation	

2.3 Structured EHR data

Nationally, EHR systems have replaced paper-based systems in most healthcare organizations. EHR systems have resulted in a large amount of rich health data, which hold great value for reuse. As the American Medical Informatics Association (AMIA) states:

> *Secondary use of health data can enhance healthcare experiences for individuals, expand knowledge about disease and appropriate treatments, strengthen understanding about the effectiveness and efficiency of our healthcare systems, support public health and security goals, and aid businesses in meeting the needs of their customers*.

Retrospective analysis of health data holds promise to expedite scientific discovery in medicine and constitutes a significant part of clinical research. National initiatives have been created to facilitate greater use of EHR to support clinical research in the United States.

EHR systems include a wide range of data about patients (i.e., demographics, problem list, vital status, vaccines, surgical and medical histories), as well as specific data about their hospital and clinic visits (i.e., admission/discharge, diagnoses, procedures, medications, lab tests, orders, vitals and observations, location of the visit, specialty of the provider). In this work, the EHR data comes from our CDR at the University of Minnesota, which houses the EHRs of more than 2 million patients seen at 8 hospitals and more than 40 clinics. For this research, all relevant EHR data of surgical patients enrolled in NSQIP at the University of Minnesota were collected from CDR for the adverse event detection modeling.

2.4 Unstructured clinical notes

Narrative clinical notes are a valuable source of information for detection and characterization of outbreaks, decision support, recruiting patients for clinical trials, and translational research, because clinical notes contain information regarding signs, symptoms, treatments, and outcomes (36-38). For example, radiology, surgical pathology, molecular pathology, cytogenetic, and flow cytometry reports contain valuable information for translational cancer research that can be used for epidemiologic and descriptive studies and discovery of new relationships that impact diagnosis and prognosis or treatment. Most of the information contained in clinical documents, however, is locked in free-text format and must be encoded in a structured form to be useful for automated and computable applications. In many cases, the words or concepts indicating a specific surgical adverse event might be found in the clinical notes. Several examples are shown in **Table 2-3**, where the keywords and concepts are bolded.

It is worthy to note that "**no** UTI" in the sentence is an example of negation. Negation is an example of important contextual feature in clinical reports. As such, successful application of NLP needs to address negation and other contextual information (e.g., uncertainty, past/previous) contained in clinical notes. In this study, an NLP tool for clinical research developed by NLP/IE group at the University of Minnesota was used to analyze unstructured clinical notes (39).

Table 2-3. Relevant keywords and concepts in clinical notes

"…...Pt had recent hernia repair. Pt came from rehab center with a concern for **wound infection**….."
"…... Concern for **sepsis. Fever** 101.8, chills & hypoxia. Has wound **vac**, central line, PICC & oxygen. **Sepsis protocol** initiated.…..."
"…... Quick Note: **Has pneumonia**, on **antibiotics**….."
"……Let her know UC so far negative (so no **UTI**)….."

Chapter 3 Preliminary studies and pilot testing

3.1 Introduction

To test the feasibility of automated retrospective detection of surgical adverse events, pilot studies were conducted on detection of the family of SSI events. An SSI is an infection occurring after surgery in the part of body where surgery took place. While most surgical patients do not experience an SSI (40), SSIs are very expensive and morbid. According to the depth and severity of infection, SSIs are categorized into superficial, deep, and organ/space. Definitions for SSIs have been standardized by the CDC and are used by NSQIP SCR to identify and document each SSI category (41).

Previous work has explored risk factors associated with SSI, but few studies have focused on the detection of SSI. Most papers examining detection have relied heavily on administrative data or claims data bases (such as age, gender, principal diagnosis, and billing information about medications and procedures) (42-44). Since EHR data contains more detailed and richer clinical data (e.g. vital signs, lab results, and social history), compared with claims data it would provide additional significant indicators and signals to SSI and thus enhance the detection performance. In addition, most studies are procedure-specific, only processing SSIs following certain types of operation, such as hip and knee arthroplasty (44-46), instead of the current approach which is broadly inclusive of different types of surgery. To help reduce the labor and cost in reviewing patient records for postoperative surgical occurrences, we hypothesized that we could leverage both EHR data and historic NSQIP registry data to develop and validate an automated approach with

supervised machine learning algorithms to detect NSQIP occurrence outcomes. In particular, we focused on the postoperative SSI occurrences to develop a classifier of three SSI categories (superficial, deep, and organ/space) and the overall SSI, and to reduce the SCR's burden by eliminating the vast majority patients if the surgeries that did not result in SSI.

3.2 Methods and Materials

Our overall methodological approach for this study included four steps as outlined in **Figure 3-1**: (1) identification of the patient cohort and associated patient EHR data, (2) data preprocessing, (3) iterative supervised learning model development, and (4) evaluation of the final models using gold standard outcome data from the NSQIP registry. Institutional review board approval was obtained and informed consent waived for this minimal risk study.

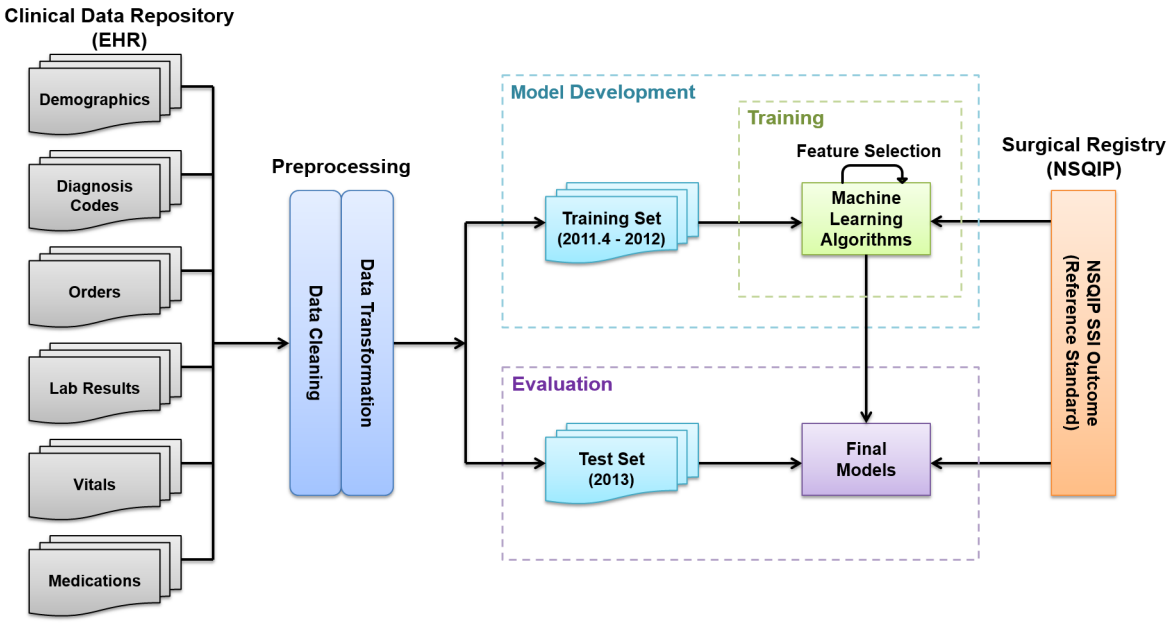

Figure 3-1. Overall materials and methods

3.2.1 Data collection and Patient cohort identification

The CDR at the University of Minnesota Medical Center (UMMC) is a database that makes EHR data accumulated from a larger, tertiary care medical center available for researchers. We extracted clinical EHR data from CDR for surgical patients included in the NSQIP registry 2011 through 2013 and retrieved their NSQIP postoperative SSI outcome from the registry. The patient's medical record number and date of surgery were used to link CDR data to the NSQIP registry. Though UMMC has been a member of NSQIP since 2007, the CDR only has consistent clinical data since 2011 when the institution implemented its current Enterprise EHR system (Epic systems). Patients without matching records in the CDR (22 total, from incorrectly entered medical record numbers) were removed. Our goal was to assess the models' robustness in the face of changes that take place over time, since the purpose of our model will be to ultimately detect future SSIs. Thus, our dataset was divided into a *training* set of patients with surgery dates between 2011 to the end of 2012 and a *test* set of patients with surgery dates in 2013. The training dataset was used for model development, while the test set was used solely for evaluation of the models we developed.

The standard definition of SSI by CDC has been used by NSQIP reviewers to determine if a patient experienced an SSI. However, some clinically important indicators mentioned in the standard definition, such as imaging orders and cultures, are not included in the NSQIP elements. We collected relevant data elements from six types of data: demographics, medications, orders, diagnosis codes, lab results and vital signs, based on

the opinion of three content experts (all surgeons familiar with NSQIP definitions and the EHR). Demographics contained each patient's basic information (e.g., gender, race, age). Among medications, we focused only on the use of antibiotics after surgery. Orders known to be associated with the diagnosis and treatment of an SSI were also gathered from CDR, including orders of imaging studies, infectious disease consultation, and interventional radiology drainage procedures for abscess drainage. Diagnosis codes consisted of relevant ICD-9 codes created during the encounter and hospital stay at the time of surgery from coding, as well as diagnoses from the past medical history and problem list. Lab values (e.g., WBC, hemoglobin, lactate, etc.) and vital signs (e.g., temperature, pain scale, etc.) before surgery and those generated during the postoperative window after surgery were extracted, as well, from the CDR. Microbiology cultures, such as wound culture, abscess culture, were collected. We also included surgical wound classification and American Society of Anesthesiologists (ASA) physical status classification that was recorded prior to surgery (47-48). The surgical wound classification is used to grade intra-operative wound contamination, which is highly correlated with the chance of developing a postoperative SSI, and is part of intra-operative case documentation. ASA classification reflects a patient's overall status with respect to surgical risk from normal healthy patient to a brain-dead patient (49-50).

3.2.2 Data preprocessing

EHR data of interest were collected, cleaned, and analyzed next. Identifying and removing outliers, and correcting inconsistent data were the very first tasks of data preprocessing. How to transform clinical data into meaningful features was our main

interest. Most clinical data, such as lab test results and vitals, tended to be longitudinal with repeated measures. Traditional methods to summarize those variables by calculating the moments (mean and standard deviation) or extremes tended not to be sufficient to describe the temporal behavior of such variables. To better summarize individual tests, we explored other features like the change of values during an "elevating period". An elevating period is a time period during which the measurement in question is near-monotonously increasing from a low level (trough point) to a high level (peak point). For patients with SSI, some lab results, like serum glucose (GLC), platelet count (PLT), and white blood cells (WBC), have significant increases in the measurement from the third day after operation.

As shown an example in **Figure 3-2**, GLC increased in three time periods: (I) day 3~7, GLC increased from 116 to 128; (II) day 7~9, from 104 to 140; and (III) day 15~28, from 87 to 148. Such elevation may indicate the onset of SSI. To capture the elevating period, a feature defined as the postoperative increase from a trough to its nearest peak was included in our tentative model. In case with multiple elevating periods, the feature was computed by using the period with the highest peak. For measures, where low values could indicate SSI, a "descending period" can be defined analogously.

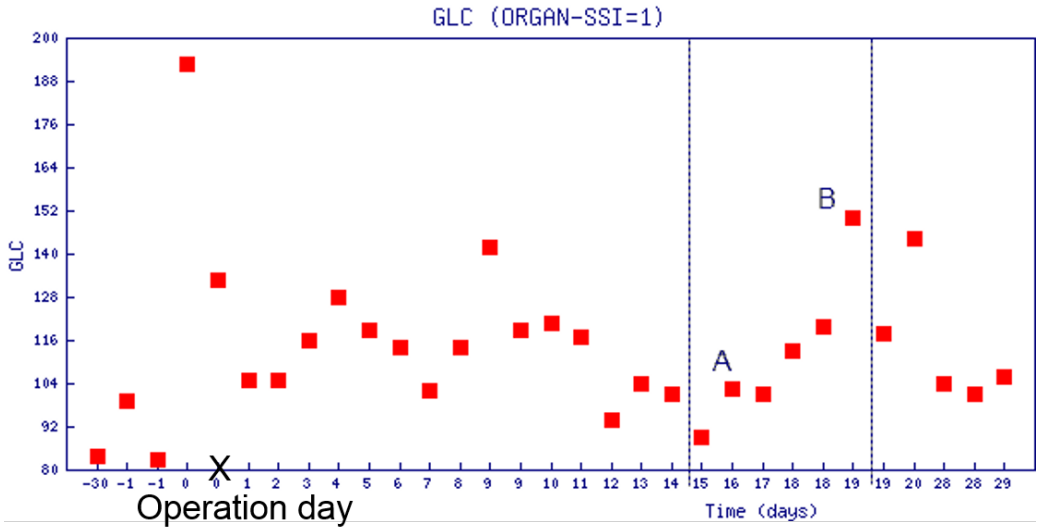

Figure 3-2. GLC values within 30 days before and after surgery

Figure 3-3 depicts the flow chart of the algorithm to compute this feature. The algorithm first searches for the maximum value (pm) from all results at least two days after the operation ({pi, i=0, …, n}), e.g., in **Figure 3-2**, point B is the maximum GLC value, which was measured nineteen days after the operation. Then the algorithm proceeds by searching for the trough point backward from point B. The algorithm is robust to filter out the abnormal point that temporarily breaks the rule of monotone. For example, in **Figure 3-2**, the elevating period is from day 15 to 19, however, there is an abnormal point A which breaks the monotone increasing trend between day 15 and 17; to overcome the problem and identify the real trough, the algorithm further compares day 15 and day 17 and see if they satisfy the criterion of monotone increasing.

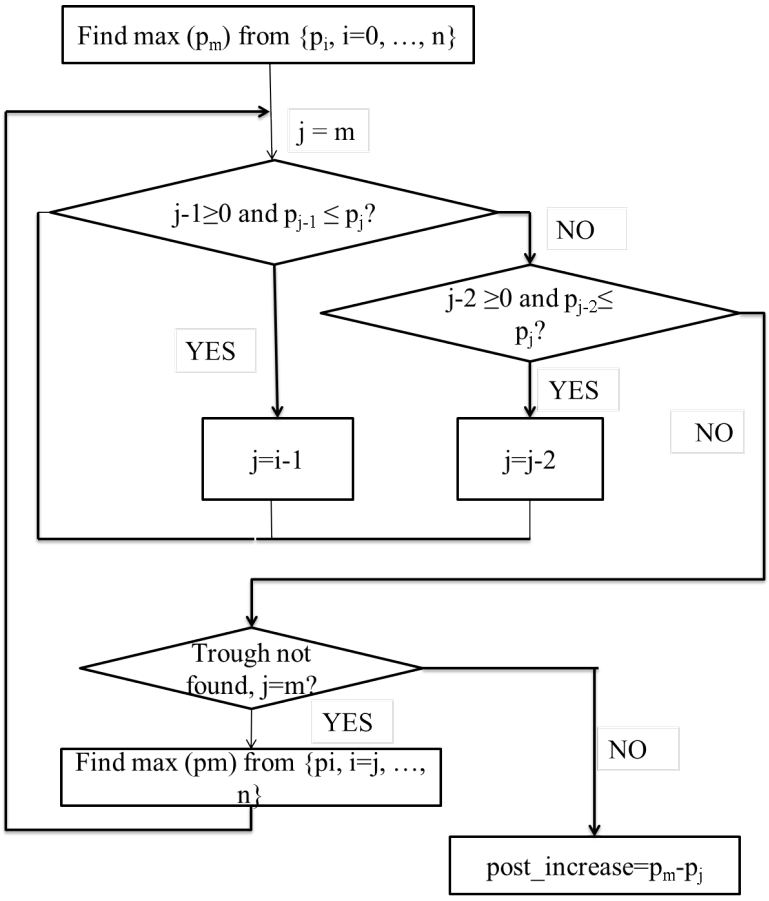

Figure 3-3. Finding the postoperative increase in GLC

For other data like antibiotic use and specific orders, we created binary variables to indicate whether a relevant element was observed. For example, a value of 1 for Interventional Radiology signifies that an abscess drainage order was placed for a patient; while a value of 0 signifies that no such test was ordered.

3.2.3 Model development

As our SSI detection model, we utilized multivariate logistic regression models. We constructed one model for overall SSI and one model for each of the three SSI subtypes.

Binary variables were entered as dummy indicator variables and continuous variables were entered unmodified. We used stepwise construction to select significant features and Akaike Information Criterion (AIC) for model selection.

3.2.4 Model evaluation

In assessing detection of surgical adverse event outcomes like SSI, since these events are relatively rare, overall detection accuracy percentage is not an optimal criterion for evaluating model validity. Instead, we report specificity, as well as the the area under the curve (AUC), in evaluation of our automated detection system. Our aim was to maximize the specificity under the constraint that the negative predictive value remains above 98%. This aim is reflective of our original expectation of actual use of the detection models: to assist a NSQIP chart extractor to eliminate patients who clearly did not suffer the adverse event and then accelerate the process of data abstraction from clinical charts.

3.3 Results

3.3.1 Significant variables selected

Tables 3-1 through **3-4** show the results for the multivariate detection models for the three kinds of SSI and the overall SSI, selected by AIC. The two most common variables included were diagnosis codes (the ICD-9 codes of SSI is 998.xx) and antibiotic use. Superficial SSI occurs just at the skin incision and thus relatively easily diagnosed. Therefore, imaging diagnostic orders tends to be unnecessary. Infection is sometimes diagnosed with microbiology cultures, however, frequently this diagnosis is based on the physical examination only. Actually only cultures ordered or not is a signal of SSI.

According to **Table 3-3** and **Table 3-4**, we can find that abscess culture, fluid culture and wound culture are significant factors for detecting deep and organ space SSI. Since this two kinds of SSI occurs deep within or under the wound, imaging orders for both diagnosis and treatment are frequently required.

We also found the postoperative elevating period of GLC for superficial and PLT for organ/space are indicative of clinical suspicion. Clinically these lab values can be altered in the setting of infection. For a unit increase in postoperative increase of GLC, we expect to see approximately 0.0112 increase in log-odds of superficial SSI. Similarly, for a unit postoperative increase of PLT, approximately 0.0115 increase in the log-odds of organ space SSI is expected.

Table 3-1. Significant indicators for detecting superficial SSI

Significant variables	**Estimate**	**P-value**
Diagnosis codes	2.1126	<0.0001
Wound culture ordered	2.1941	<0.0001
Antibiotic use	1.1321	<0.0001
Encounter type (inpatient)	1.6007	0.0010
ASA Classification (significant disturbance)	0.4342	0.0058
Abscess culture ordered	1.5020	0.0050
Postoperative increase of GLC	0.0112	0.0687

Table 3-2. Significant indicators for detecting deep SSI

Significant Variables	Estimate	p-value
Diagnosis codes	3.1959	<0.0001
Antibiotic Use	2.2276	<0.0001
Abscess culture ordered	1.2880	0.0868
Gram stain ordered	0.8040	0.0427
Imaging treatment ordered	1.5445	0.1107
Imaging diagnosis ordered	0.6254	0.0981
Tissue culture ordered	1.6516	0.1010

Table 3-3. Significant indicators for detecting organ space SSI

Significant Variables	Estimate	p-value
Imaging_treatment	1.3999	<0.0001
Imaging_diagnosis	1.2090	<0.0001
Antibiotic Use	1.1662	<0.0001
Abscess culture ordered	2.3041	<0.0001
Fluid culture ordered	1.4204	0.0003
Preoperative PLT	0.00332	0.0135
Drainage culture ordered	1.3760	0.0711
Diagnosis code	0.8259	0.0667
Postoperative increase of PLT	0.0115	0.0606

Table 3-4. Significant indicators for detecting overal SSI

Significant Variables	Estimate	p-value
Diagnosis codes	5.3940	<0.0001
Antibiotic use	1.3672	<0.0001
Abscess culture ordered	3.2565	<0.0001
Wound culture ordered	2.2926	<0.0001
Imaging diagnosis ordered	0.8741	<0.0001
Fluid culture ordered	1.2909	<0.0001
Encounter type (inpatient)	1.0185	0.0037
ASA Classification (significant disturbance)	0.4258	0.0031
Preoperative PLT	0.00214	0.0440
Post maximum pain	0.0775	0.0957

3.3.2 Model performance

Four detection models exhibited excellent specificity to eliminate the majority of non-SSI patients, which greatly accelerate the process of extracting postoperative SSI occurrences. **Table 3-5** presents the negative predictive value (NPV) for each of the SSI identification models. The highest specificity 0.988 was for detecting deep SSI at NPV equals to 0.99, and the lowest 0.787 was for detecting overall SSI at NPV equals to 0.99. AUC values for four models are 0.820, 0.898, 0.886 and 0.896.

Table 3-5. Negative predictive value and specificity for four SSI models

	NPV	Specificity
Superficial SSI	0.980	1.000
	0.985	0.987
	0.990	0.900
Deep SSI	0.980	1.000
	0.985	1.000
	0.990	0.988
Organ space SSI	0.980	1.000
	0.985	0.999
	0.990	0.974
Overall SSI	0.980	0.935
	0.985	0.888
	0.990	0.787

3.4 Discussion

The current research is a pilot study to examine the feasibility of automatically detecting postoperative SSI occurrences based on EHR data. The aim of this study is to assist a NSQIP SCR to eliminate patients who clearly did not suffer the adverse event. Therefore, very high negative predictive value is desired, which could assist in the reliable identification of patients without postoperative SSI. From the modeling results, we can see that all four models perform very well (with specificity ranging from 0.788 to 0.988) in eliminating the majority of patients without SSI based on the negative predictive value equals to 0.99. Considering the nature of NSQIP SCR's work, SCRs still need to review all clinical charts even the positive predictive value for a patient is 0.9 or even higher, since

they need to extract the clinical characteristics of patients with SSI. Therefore, achieving high negative predictive value, and thus allowing SCRs to eliminate patients, rather than achieving a high positive predictive value is the main focus of this research.

Among selected potential indicators, a few of them were found to be quite significant with very small p-values. Only the indicators that had p-value less than 0.0001 were also used to do logistic regression modeling, but this did not improve the detection performance. Other modeling methods, like Random Forest and Support Vector Machine, were employed; however, logistic regression models were found to outperform these methods for detection of all types of postoperative SSI events.

The current study was limited by the fact that it was conducted with only about complete cases in three years, which might have limited our ability to fully refine and optimize the automated detection model. In the future, more procedures will be included, and the treatment of missing data will be studied.

3.5 Conclusion

In this study, to accelerate the process of extracting postoperative SSI outcomes from medical charts and reduce the workload of NSIQP SCR, an automated postoperative SSI detection model based on supervised learning was proposed and validated. The models exhibited good performance, they reduced the SCR's burden by reliably eliminating the vast majority of patients with no SSI. The significant factors of detecting SSI identified by our models are in line with clinical knowledge. In addition, some useful patterns, e.g. postoperative increase of PLT and GLC, were extracted from the longitudinal lab results.

Chapter 4 Missing data in electronic health records

4.1 Capturing the context of "missing data"

Unfortunately, secondary use of EHR data can be challenging due to the inconsistent and incomplete nature of patient records within the EHR. The presence or absence of elements, the timing and sequence, and other characteristics of the collected data can vary greatly from patient to patient. Sometimes necessary or expected data elements might be missing in a patient's record. Missing data rates in the EHR have been previously reported from 20% to 80% (51-52). In this study, we were interested in clinical data between postoperative day 3 to 30 (which we refer to as the postoperative window henceforth) because the first two days after surgery often constitute a recovery period, where abnormal measurements are common and may simply be a result of healing from the trauma caused by surgery, rather than a sign of SSI. During the postoperative window, the problem of missingness commonly exists for many data elements. Researchers traditionally categorize missing data mechanisms into three types according to the characteristics of the missingness: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) (53).

- MCAR - Causes of missingness are not related with any characteristics of the dataset (e.g., whether a data point is missing is not related with any values in the dataset). For example, the urine culture test is usually ordered to help make diagnosis of urinary tract infection. However, a urine sample might be randomly broken and the test result is missing completely at random.

- MAR - Data are not missing at random, but the probability that a data element is missing depends on values of other observed variables in the dataset. As an example, suppose men are more likely to drop out of a clinical trial, but the chance of dropping out is the same for all men. We can say that male subjects are just MAR. Both MCAR and MAR are viewed as ignorable missingness.
- MNAR - When the likelihood of missingness is related to missing variables, a third type of non-ignorable non-response missingness, MNAR, arises. For example, consider a study aiming to evaluate treatments to reduce cocaine use. In this hypothetical study, the outcome drug level is measured from a urine drug test every Monday morning. Participants who use cocaine over the weekend and do not show up for their urine test would be expected to have higher cocaine metabolites. Therefore, the likelihood of the data being missing is directly related to the unobserved cocaine level, which is viewed as MNAR.

The traditional three missing data categories are not sufficient to capture the complexity of missing data in EHR-derived applications. Missing data in EHR-derived datasets could be caused by a lack of collection or a lack of documentation (54). Lack of collection, for example, refers to orders or other items that are not placed or measured. In this instance, the missing data element is typically a negative value, i.e. a normal state patient. Alternatively, the clinician may not be considering the measurement since the test or measure is thought to be low yield for the patient in question. Such missing values are MNAR. Lack of documentation refers to orders or other items that are placed or performed but the response values are not recorded or obtained during the process of data collection.

In this instance, data was lost during the extraction, transformation, and loading (ETL) of clinical data. Such missing values are MCAR or MAR. Furthermore, a good working knowledge of the specific research question is likely helpful for understanding missing data mechanisms and potentially for selecting the most suitable missing data imputation methods for a particular secondary use application of EHR data.

In our SSI detection use case with EHR data, possible missing data can potentially be caused by either lack of collection or lack of documentation and thus we are facing a mixture of MNAR and MCAR/MAR mechanisms. In the situation of lack of collection, for example, a WBC count is usually measured repeatedly to monitor a patient's status after surgery. However, WBC test is not necessarily ordered for all patients—patients doing well clinically are less likely to have the WBC test. Similarly, an image-guided order with interventional radiology related to SSI treatment or a microbiology culture test is less likely to be placed on patients for whom there is minimal to no suspicion of an SSI. There are also examples of lack of documentation. For instance, the microbiology gram stain specimen from a wound suspected of harboring an SSI may be sent to an outside laboratory and therefore not recorded in the system. It is difficult to tell to which category of data missingness a case of missingness belongs (e.g., lack of collection or lack of documentation). Additionally, performance of common missing value imputation methods in the context of MNAR is unknown. Therefore, we need to explore and compare different missing data imputation methods, and find the most suitable approach.

4.2 Related work on handling missing data

Though numerous missing data treatments have been developed, selecting the most appropriate one depends strongly on the problem at hand. Overall, most studies have not demonstrated one technique to be universally better than others. This section briefly summarizes traditional statistical and model-based methods. Our aim is to suggest ways that clinical research practitioners without extensive statistical backgrounds can handle missing data by exploring several of the most commonly used strategies to handle missing data for the real problem of postoperative SSI detection with EHR data.

The most common and easiest method is to exclude cases or single variables with missing data. Researchers either consciously or by default drop incomplete cases since many statistical and machine learning tools operate on complete cases and only rarely have built-in capabilities to handle missing data (55-56). However, discarding cases or variables with missing data not only decreases the number of available cases in a given dataset but may also result in significant bias (57-58). As an alternative to complete-case analysis, many researchers will impute missing values for variables with a small percentage of missing data (59-60), such as using the mean value of the observed cases on variables of interest. However, filling in the mean value usually causes standard errors to appear smaller than they actually are, since it ignores the uncertainty of missing data (60-61).

Compared with filling in the mean value, advanced methods, such as multivariate and maximum likelihood imputation, were developed several decades ago. In particular, multivariate imputation enables researchers to use existing data to generate or impute values approximating the “real” value and has been widely applied in clinical data analysis

(62-63). In addition, multivariate imputation by chained equations (MICE) approach generates a regression model for each variable with missing data, with other variables as predictors, to impute the missing data. This method, being a regression model, can handle different types of variables (continuous or discrete). More recently, imputation methods based on more sophisticated models have been developed. Some well-known approaches, such as multilayer perceptron, self-organizing maps, and K-nearest neighbors (KNN), have been employed as the predictive models to estimate values for the missing data in specific applications such as breast cancer diagnosis, detection of cardiovascular patients and intensive care unit monitoring (64-65). However, for clinical researchers, most complicated methods typically are not easy to implement. Also, to date they have failed to show a convincing and significant improvement over univariate imputation (e.g., filling in the mean value or MICE).

At the present time, most algorithms apply to MCAR or MAR. Imputation for MNAR is generally not recommended, and hence few algorithms exist. Algorithms like selection method and pattern mixture models could jointly model data and missingness. The former assigns weights to observations based on their propensity for missingness (66-67), while the latter constructs imputation models for each pattern of missingness (68-69). Both methods have the potential to reduce bias in the results. However, due to their untestable assumptions, they may perform worse than imputation methods developed for MCAR or MAR. Several researchers have previously applied the combination of Fourier transformation (70-71) and lagged KNN to impute biomedical time series data in which up to 50% of data are missing (72-73).

In our work, the potential for non-random missing data exists. Discarding patients with missing values would be a conservative choice. However, if we discarded all observations (patients) that contain missing values, we would discard close to 80% of our study population. This alone could fatally bias the results and hence imputation is imperative. In this work, we seek to explore several commonly used missing data imputation methods in our SSI dataset to increase our sample size and to avoid discarding a large portion of patients with missing values. In particular, eight imputation methods were used to fill in absent values for lab tests and vital signs in the postoperative SSI dataset. To compare different imputation methods, the performance of multiple detection models based on different missing data treatments were evaluated by using the reference standard SSI outcome from NSQIP.

4.3 Data collection and preprocessing

The EHR dataset used was the same one in pilot studies conducted previously. Data preprocessing consisted of transforming the data (if necessary), and correcting any inappropriate formatting in the data for further modeling (e.g., the erythrocyte sedimentation rate value could be entered as "56 H" in EHR system, however, to keep the value consistent in numerical format for further modeling, we needed to remove "H".) Lab results and vital signs, viewed as continuous variables, are measured periodically. The resulting longitudinal data were summarized into three features: two extreme values (highest and lowest values) as well as average value during the postoperative window. To establish a baseline, the preoperative extreme and average values were also extracted. Binary features (taking the values of 0 and 1) are created for medications, orders and

diagnosis codes, indicating the presence or absence of that data element during the postoperative window. For example, the value of 1 for a particular antibiotic (medication) signifies that the patient received the antibiotic during the postoperative window. Another two variables, ASA score and wound classification, are ordinal variables with multiple levels. A univariate logistic regression model was used to compare the effects of different levels, and levels regrouping might be necessary. In our dataset ASA classes I and II were grouped and classes III, IV, V and VI were grouped. For wound classification, classes I and II were grouped and classes of III and IV were grouped. We made age an ordinal variable—above the age of 65 and under 65. Other SSI risk factors, such as smoking, alcohol use, history of diabetes, anesthetic type, etc., however, were not selected as significant indicators to SSI by the detection model in our pilot study using the completed dataset, therefore, were not included in this study.

All the data were transformed into a data matrix amenable to statistical modeling. The rows of this matrix correspond to patients and the columns to features (predictors). It is worth noting that a patient may have multiple visits during the postoperative window. All the EHR data generated during the postoperative window were collected for modeling.

4.4 Missing data imputation of the incomplete dataset

We define a **missing value** as a specific lab result or vital sign that is missing entirely during the postoperative window. For example, a patient's WBC values could be completely missing during the postoperative window. In this instance, there is no way to summarize the WBC values. We need to impute the WBC related variables (i.e., maximum WBC, minimum WBC, and average of WBC). If a patient has several WBC measurements during the postoperative window, imputation is unnecessary. In this work, mean-imputation, 0-imputation, imputing normal values, and MICE methods were utilized. In the case of first three non-model-based methods, for each feature, a single value is imputed every time the value for that feature is missing. For example, in case of "mean" imputation, for each feature, the mean of the non-missing entries was calculated in the training dataset and imputed into both the training and test datasets every time the value was missing for that feature. In case of "0" imputation, we simply imputed the numeric value "0" for every missing entry; and in case of "normal" imputation, the average value of patients in the training set with no postoperative SSI was imputed into both the training and test datasets.

Non-model-based imputation ignores the concept that related features can be used to "predict" what the missing value could be. In *MICE method*, we utilized linear regression modeling to impute missing values based on the non-missing values of other features, essentially a multivariate imputation through chained equations (74-75).

In the course of imputation, bias can be introduced when values are not missing at random. To reduce some of this bias, indicator variables were used. An indicator variable, implemented as a dummy variable, takes the values of 1 and 0; 1 indicates that the

corresponding value is missing. For example, if a dummy variable for postoperative WBC is created and takes the value 1 for a patient (observation), then the patient in question does not have any postoperative WBC value during the postoperative window; the corresponding features (minimal, maximal postoperative WBC) contain imputed values. In total, for our dataset, this resulted in 15 original features, 33 transformed features and 22 dummy variables. **Table 4-1** summarized the different imputed datasets by different methods.

Table 4-1. Imputed datasets with eight imputation methods.

Imputed Datasets	Imputation Method
Mean	filling in the mean of all non-missing observations in training set; filling in the mean of all non-missing observations in test set, separately
Normal	filling in the mean of non-SSI patients in training set; filling in the mean of non-SSI patients in test set, separately
MICE	using multivariate regression model
0	filling in 0 for all missing values
Dummy+Mean	adding dummy variables to model "mean"
Dummy+Normal	adding dummy variables to model "normal"
Dummy+MICE	adding dummy variables to model "MICE"
Dummy+0	adding dummy variables to model "0"

4.5 Modeling method

LASSO regression estimates coefficients by minimizing the quantity of the regularization term plus the RSS, if fitting a linear regression (**Eq. 4-1**), or deviance, if using a logistic regression.

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda \sum_{j-1}^{p}\left|\beta_j\right| \qquad \text{(Eq. 4-1)}$$

The regularization term, or l_1 norm multiplied by a weight λ, favors sparse models that involve only a subset of predictors. Depending on the tuning parameter, λ, we can select arbitrary number of predictors by using LASSO (**Figure 4-1**).

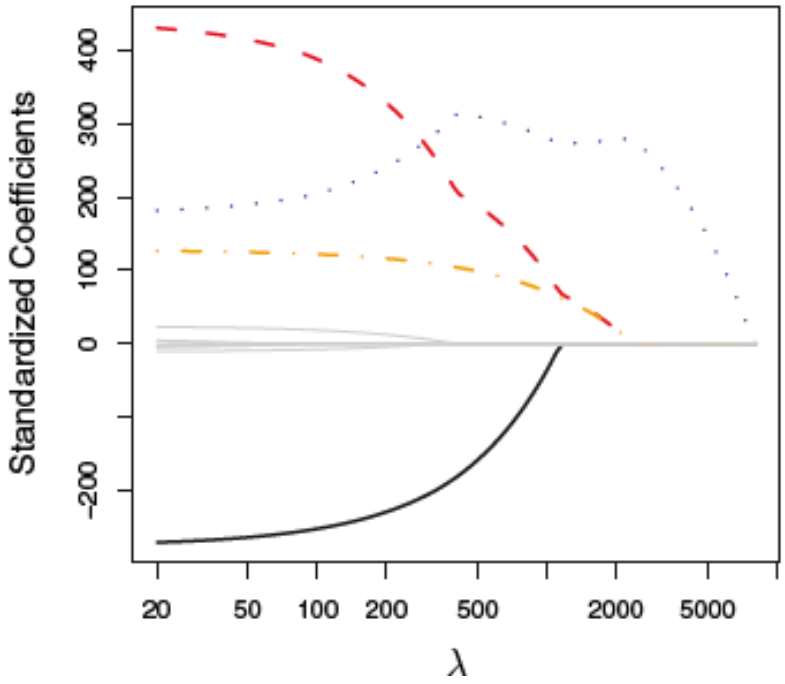

Figure 4-1. An example that indicates LASSO can produce a model with arbitrary number of predictors.

The curves of different styles show how the coefficients estimated change as λ

increases. As lambda increases, more coefficients become zero, in other words, only the more related predictors are selected in the LASSO regression model. Figure courtesy of The Elements of Statistical Learning (76)

By allowing the coefficient estimates to be zero, LASSO regression supports feature selection, which is the advantage compared with other regularization method. Compared with other variable selection methods such as best subset selection, LASSO regression has the advantage of computational feasibility.

Because of the additional regulation term (**Eq. 4-1**), when λ gets sufficiently large, the coefficients estimated by LASSO regression is shrunk from the coefficients estimated by linear regression or logistic regression, which significantly reduces their variance, at the expense of slight increase in bias. The best λ with the least error can be tuned, e.g., by cross validation. Thus, LASSO helps increase the prediction accuracy, especially when a relatively small subset of original predicators is related to the outcome. LASSO also improves model interpretability because irrelevant predictors are removed from the fitted model.

4.6 Model Evaluation

The performance of the SSI detection models was evaluated both on the training set and on the leave-out test set. In order to assess the detection performance of the model on the training set, 10-fold cross validation (CV) was employed and the values of area under the curve (AUC)43, as well as the bias, were calculated. AUC is an accepted performance metric and quantifies the ability of a model to discriminate between positive and negative

outcome. Bias is used to examine whether the estimation of outcome systematically differs from the true outcome18. The closer the absolute value of bias is to 0, the smaller the bias is for SSI detection. Positive or negative bias indicates the overestimate or underestimate of a model.

Evaluating the reference model raises important issues. We could evaluate its performance on the unimputed test dataset as a reference. However, the reference model would not be able to make predictions for the vast majority of patients, since many (around 50%) would be deleted due to missing values. For this reason, we applied the reference model (without imputation) to all imputed test sets and selected the one with best performance as the performance for the reference model. The reference model was evaluated on each imputed test dataset. Every model that was constructed on a specific imputed training set was evaluated by the imputed test set that used the same imputation method.

4.7 Results

We retrieved the clinical data from EHR for 4,491 patients in the NSQIP registry at UMMC between 2011 and 2013. Table 2 includes detailed demographic information. The training set covers years 2011 and 2012, and encompasses 2,840 patients with 132, 51, and 81 postoperative superficial SSI, deep SSI, and organ space SSI, respectively. The test data set covers the year 2013 and contains 1,651 patients with 41, 34, and 41 respective SSI types. Some patients may have multiple SSI types.

Table 4-2. Training and Test Set Patient and Surgical Site Infection (SSI) Characteristics

Characteristic	Training set (2011-2012)					Test set (2013)				
	ALL Procedure	Overall SSI	Superficial SSI	Deep SSI	Organ Space SSI	ALL Procedure	Overall SSI	Superficial SSI	Deep SSI	Organ Space SSI
Total	2840	252	132	51	81	1651	114	41	34	41
Encounter type										
Inpatient	2429	242	129	47	78	1052	104	35	32	38
Outpatient	411	10	3	4	3	599	10	6	2	3
Age group										
< 65	2259	210	109	41	67	1269	91	30	28	34
≧65	581	42	23	10	14	382	23	11	6	7
Gender										
Male	1246	119	63	22	39	774	51	19	15	18
Female	1594	133	69	29	42	877	63	22	19	23
Race										
White	2386	213	112	44	66	1481	99	34	30	38
African American	189	18	8	5	7	112	7	3	1	2
Other/ unknown	265	21	12	2	8	58	8	4	3	1

Model performance in detecting superficial, deep, organ space and overall SSI are shown in **Figure 4-2**. AUC scores obtained through a 10-fold cross validation on the training set and the final AUC scores on the test set are also reported. Generally, imputed models performed *substantially* better (statistically significant difference, with at least a 2nd digit difference in AUC) than the reference model, except for superficial SSI detection, where the reference model offered comparable performance.

The final AUC score and bias of each model are reported in **Table 4-3**. We observed that for superficial SSI, every model performed similarly except "Dummy+0" and the reference model, which had the largest bias. Among models of deep SSI, imputed models without dummy variables performed best and the reference model performed worst in terms of AUC. Also, "Dummy+MICE" had the smallest bias among the different models. For organ space SSI, models with dummy variables performed best except "Dummy+0", and all models had similar bias except "Dummy+0". For detecting any SSI, most imputed models had similar performance and were substantially better than the reference model. Biases were also similar except the reference model and "Dummy+0", which were more biased than the other models. Selected important variables, the estimated coefficients of variables and the 95% confidence intervals for the coefficients, are included in the supplemental appendix.

Pairwise t-tests on the 1000 replications of the Bootstrap procedure were conducted to test for statistical difference in AUC scores between the methods for each SSI event. The majority of imputation methods in each category of SSI showed statistical difference, with p-values less than 0.01, except "0" vs. "Dummy+Mean" for superficial SSI (p-

value=0.102), “MICE” vs. “Normal” for Deep SSI (p-value=0.129), “Dummy+Mean” vs. “Dummy+Normal” for organ space SSI (p-value=0.098), and “Reference” vs. “Mean” (p-value = 0.094) and “Normal” “vs. “0” (p-value=0.143) for overall SSI. The detailed results are included in the supplemental appendix.

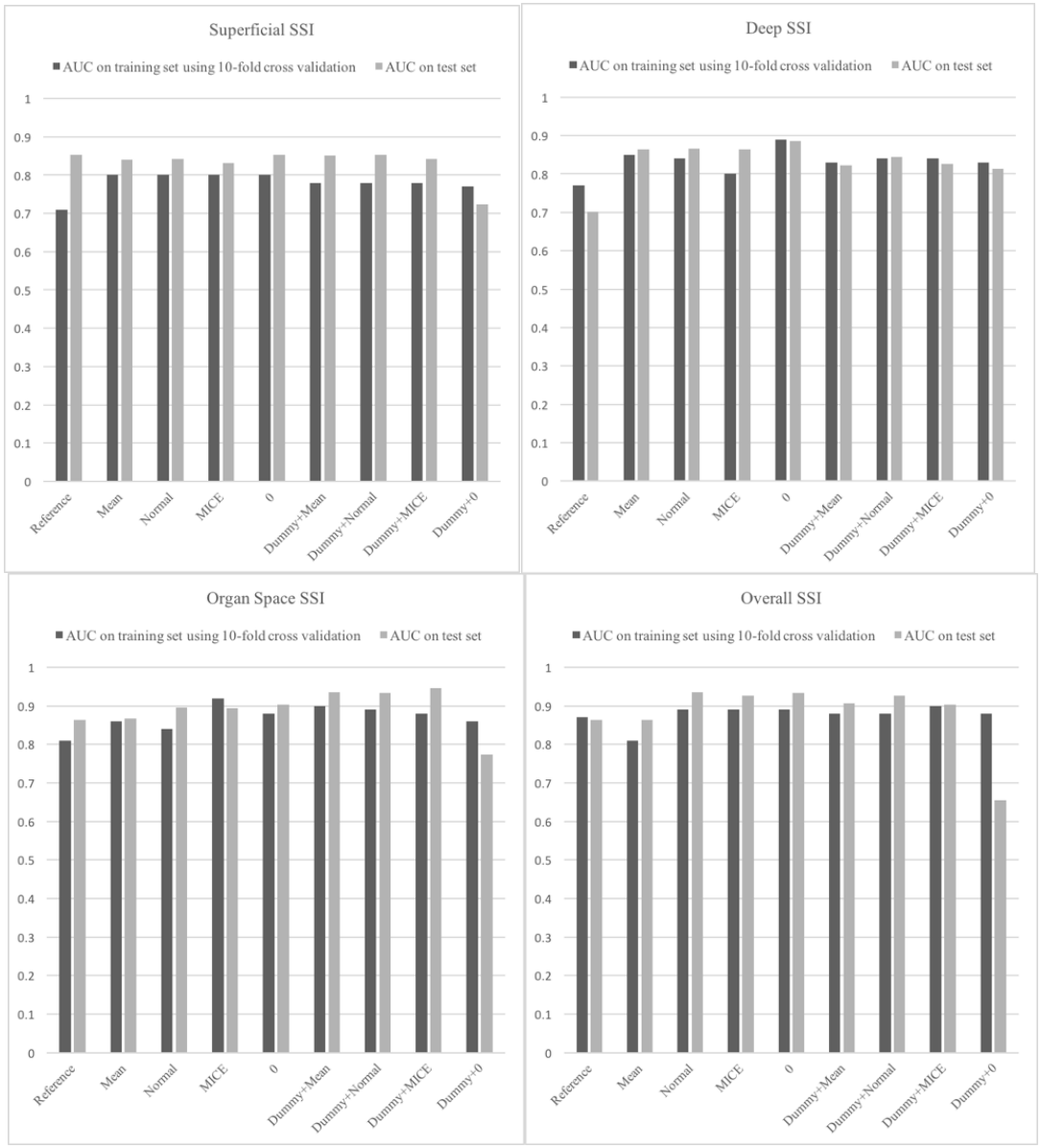

Figure 4-2. Detection Performance for each category of SSI with different imputation methods.

The AUC scores are calculated based on both the training set (using the 10-fold cross validation) and the test set. Generally, the results indicate that developed models have a better performance on the test sets.

Table 4-3. Bias analysis for eight imputed models as well as the reference model when detecting postoperative SSI. The average AUC and bias across the three sub-types of SSI and overall SSI are calculated as well.

	Superficial SSI		Deep SSI		Organ Space SSI		Overall SSI			
	AUC	Bias	AUC	Bias	AUC	Bias	AUC	Bias	Average AUC	Average Bias
Reference	0.855	0.0600	0.702	-0.0032	0.864	-0.0136	0.864	-0.0159	0.821	-0.0079
Mean	0.841	0.0131	0.864	-0.0085	0.867	-0.0086	0.863	-0.0055	0.858	-0.0024
Normal	0.845	0.0132	0.866	-0.0014	0.896	-0.0092	0.935	0.0038	0.884	0.0016
MICE	0.832	0.0191	0.865	-0.0010	0.894	-0.0087	0.927	0.0053	0.879	0.0037
0	0.852	0.0122	0.886	-0.0090	0.903	-0.0105	0.934	-0.0072	0.893	-0.0036
Dummy+Mean	0.851	0.0155	0.823	-0.0155	0.935	-0.0088	0.906	0.0052	0.878	-0.0009
Dummy+Normal	0.856	0.0155	0.844	-0.0009	0.934	-0.0082	0.926	0.0052	0.888	0.0029
Dummy+MICE	0.843	0.0195	0.826	0.0001	0.946	-0.0082	0.903	0.0089	0.879	0.0051
Dummy+0	0.724	0.1679	0.813	-0.0154	0.774	0.1349	0.655	0.0639	0.741	0.0879

4.8 Discussion

In this work, we explored the use of nine methods for treating missing data (one where records with missing values were completely discarded and eight methods using imputation for missing values) and evaluated the performance of the SSI-detection models constructed on nine training sets that utilized these imputation methods. Overall, we found imputation to be beneficial. Models built on imputed data outperformed the reference model for all SSI types except superficial SSI. In case of superficial SSI, essentially all models had very similar performance; only "Dummy+0", the model built on the 0-imputed dataset utilizing bias-correcting dummy variables, had lower performance. We will explore the reasons for its lower performance later.

The most surprising finding from this study is that the models with bias-correcting dummy variables did not perform as well as we expected. We expected that missing values signal that the patient is at a lower risk of SSI (the lab test is not necessary), giving rise to a "healthiness" bias. We originally thought that models without the dummy variables would have no ability to correct for this "healthiness" bias; hence, the addition of the bias-correcting dummy variables would allow the model to correct the bias, improving its performance. Instead, the performance did not improve. There are two possible reasons for this. First, potentially the rates of missing values in the cases (SSI patients) and the controls (patients without SSI) are significantly different between the training and test set. In 2013, non-SSI patients appear to have more results (e.g. WBC and vital signs like body temperature) than in 2011-2012; thus, the "healthiness" bias in the training set is different from that on the test set. Second, there are some variables in the dataset that can take on

the role of the dummy variables to some extent. For example, the variable “patient type”, which indicates whether the patient had an inpatient or outpatient surgery, captures the "healthiness" bias well: outpatient surgeries are traditionally less complicated and thus are less likely to have complications; or conversely, if a procedure is associated with higher risk and a higher complication rate, it is less likely to be performed in the outpatient setting. This is similar to dummy variables, which also indicate a lowered risk of complication when the corresponding lab tests are not ordered.

The “0” model, where the value 0 was imputed for missing elements, performed surprisingly well. It achieved AUC scores ranging from 0.852 to 0.934. In most cases, imputing 0 is not clinically meaningful. Typically, imputing 0 for temperature would be a disastrous choice, as it would create large biases. The model for superficial SSI is one example. Among its significant variables, two are related to temperature: the postoperative maximum temperature and the minimum temperature. Their coefficients have the opposite signs, with the maximal temperature having the positive coefficient and the minimum having the negative. This can be interpreted as the difference between the maximum and minimum postoperative temperatures, which automatically corrects for the bias.

MICE exploits the structure of the problem, namely the relationship between the variables with the missing value and other variables; compared with other non-model-based imputation methods that ignore such structure, MICE methods are expected to perform best. Surprisingly, we did not find that the performance of the MICE models is significantly better than that of other imputation models. This is a result of differences in the problem structure between unhealthy and healthy patients: variables of healthy non-

SSI patients are different in range from unhealthy SSI patients, and are more likely to have higher rate of missingness than unhealthy SSI patients, which affects the models in two ways: (1) the observations of unhealthy SSI patients contributes more in modeling imputation models because only complete observations are used to build imputation models; and (2) as a result, biases were likely introduced when applying the model to impute missing values for healthy non-SSI subjects. In spite of this, it is worth pointing out that "Dummy+MICE" achieved the best AUC on Organ Space SSI and the lowest bias on Superficial SSI. Overall, the performance of MICE method is good, but other simpler imputation techniques appear to be able to match their performance for the use case of SSI detection.

Another interesting fact worth noting is that in some cases, the performance of the models on test set was actually better than that in the training set. There are two possible reasons for this observation. First, this may be related to the SSI rates in the year of 2013 (test dataset) and 2011-2 (training dataset). For example, the rate of superficial SSI in the two sets was most different, 2.5% and 4.6%, respectively; consequently, the performance of models for superficial SSI between the two datasets differed and was higher for the test dataset. As for other types of SSI, rates were close between the two datasets, specifically 2.5% vs. 2.8% for organ SSI and 2.1% vs 2.1% for deep SSI. While the EHR system was not changed from the 2011 to 2013, other possible unseen factors which may influence the distribution of patients in 2013. Second, it is possible the constructed models underestimate the risk of SSI on the training set; therefore, the performance on the training dataset is relatively lower than for the test dataset; yet, the biases remain small. We also hypothesize

that the increased collection of lab results may have also biased the regression models we used to fill in missing values since they were constructed on the training set. This is an analogous effect to the inability of the dummy variables to "un-bias" the estimates.

4.9 Limitations

The NSQIP database can provide insight into the importance of adequately addressing the problem of missing data. Data in NSQIP is manually abstracted directly from the EHR by trained personnel. If WBC values are entirely missing in the NSQIP file, it is most likely that the cause of missingness is lack of collection (i.e. there was no need to measure it). However, in our experiment, we have not fully explored the characteristics of missingness in the NSQIP dataset. This will be addressed in future work.

The NSQIP population in this experiment had some patients with primary providers who utilized a different EHR from the one used for manual data extraction and entry into the NSQIP database. In general, the EHR for the institution enrolled in the NSQIP database includes all pre-, intra-, and post-operative data on included patients. However, there are examples where the surgeon's outpatient EHR or the patient's primary care provider's EHR differs from the EHR of NSQIP-enrolled institution. This is relevant to our present study of missing data, since preoperative data as well as post-operative complication data may be recorded in a database to which the trained manual abstractors do not have access. In our study, the EHR for the surgeon remained the same as the NSQIP-enrolled institutional EHR. However, the EHR for the primary care provider often differed (approximately 50% of cases). Patients with primary care providers who utilize a different outpatient EHR (compared to the NSQIP-enrolled institution's EHR) might have some

relevant data within the postoperative window missing after discharge from the hospital. We did not exclude/censor these patients. In addition, a subset of SSIs in our study were noted in the ICU or acute care (i.e., inpatient) setting. Some SSIs, most notably superficial SSI, can occur as wound infections in the outpatient and ambulatory settings after the index stay. Others, namely, deep and organ space SSI can be typically discovered during the index inpatient stay. However, some occur after discharge. These SSIs often require readmission and further inpatient treatment. Therefore, missing data after discharge could be an important potential limitation to applying our approach more widely. Actually, 5% of SSI cases in our cohort are those patients with SSIs who have no data collected. Presumably, these patients were both seen and treated at clinics that utilized a different EHR from the NSQIP-enrolled institutional EHR.

As introduced in section 2.1, our dataset was divided into a training set and a test set by calendar year rather than randomly sampling, since we are interested in investigating how robust the models are in face of institutional changes at a relatively short time horizon. It is inevitable that due to institutional changes the model performance will drift. With every passing year the model's performance can decrease. At some point in the future, the model will have to be recalibrated or outright reconstructed. It is undesirable to have to rebuild a model every year. Our method of dividing the data set by year allows us to assess how resilient the models are to such institutional changes.

The application of EHR data in surveillance continues to be an issue of importance in the informatics and quality literature. Though the main purpose of our work is to accelerate the manual process of NSQIP data collection, EHR data could be used to help

surveillance as well. At this point, we do not believe that we can entirely rely on the EHR since a number of challenges remain. These include the real-time availability of EHR data, the heterogeneity of EHR systems utilized by different providers treating the patients enrolled in the NSQIP database, and the variability of signals for event detection.

The relative infrequent nature of these events is part of the challenge with event detection. When events (e.g., myocardial infarction) are relatively rare, the imbalanced nature of the data could be a large part of the challenge. Possible solutions to deal with this challenge when investigating more adverse (and thankfully rarer) events will be explored in future work.

4.10 Conclusions

In summary, we found models with imputation perform almost always better than models that discarded patient records with missing values. However, the optimal choice of imputation method is not clear. Data characteristics and data collection variation all affect the performance of imputation methods. If the test and training datasets have similar characteristics in terms of missing values, the use of bias-correcting dummy variables can be advantageous; if the characteristics differ, the estimated bias will be incorrect and can be similar in magnitude to the bias caused by the missing value they try to correct for, which is what happened in the present study. Similarly, if variables present in the dataset, such as "patient type" can take on the role of correcting for the bias, then dummy variables may not be necessary.

If it is guaranteed that test datasets have the same missing value biases as training or evaluation datasets, then the use of bias-correcting dummy variables can be

advantageous. Similarly, MICE is advantageous only if the structure of the training dataset is similar to the structure of the test dataset. In our example, increased lab result collection created significant differences between the training and test datasets, rendering MICE only marginally useful. In our experiments, we found that imputing the mean of the non-SSI cases was successful in reducing the bias introduced by the fact that missing labs and vitals were suggestive of the lack of SSI event.

Chapter 5 Detection of postoperative complications using multi-task learning methods

5.1 Introduction

In this study, our aim is to build an automated platform for postoperative complications detection based on structured EHR data by using robust modeling techniques. Included in this analysis are the main postoperative complications of three subtypes of SSI (superficial, deep, and organ space), pneumonia, UTI, sepsis, and septic shock. We hypothesized that EHR data would include significant indicators and signals of postoperative complications and that sophisticated machine learning methods might be able to extract these signals, accelerating the Manual Chart Review (MCR) process of these adverse events. Compared with the gold standard MCR process, automated application has potential advantages. First, MCR lacks inter-rater reliability, while an automated abstraction system would provide an objective and consistent reporting protocol that can be applied across multiple medical institutions. Second, a successful automated abstraction system would allow for expansion to include other procedures where postoperative surveillance is not being performed currently.

Specifically, we explore several methods for developing postoperative complication detection models. The most straightforward way to detect each type of complications is to build an independent classifier for each of them, which could be viewed as single-task learning, where detecting each complication is a task. As shown in **Figure 5-1**, more than on surgical adverse event may happen on patients the same time. Among

all patients in population, 31 patients have organ space SSI and sepsis together. When tasks are known to share similar features, we expect the resultant models to be similar. Learning models for these tasks together allows us to introduce inductive bias to make the resultant models similar, providing us with more robust models. Learning models for related tasks together is referred to as multi-task learning. For example, when our task is to identify a particular SSI subtype, a related task can be to detect any SSI (overall SSI). With an overall SSI model in hand, identifying a particular SSI subtype is easier, because the classifier only needs to learn the difference between overall SSI and the particular SSI subtype, rather than the difference between the SSI subtype and any other complication.

In this work, we compare six methods for developing post-operative complications detection models. First we have single task learning, where predicting each complication is an independent task. We also explore five different methods for multi-task learning, assessing their value in improving the detection performance.

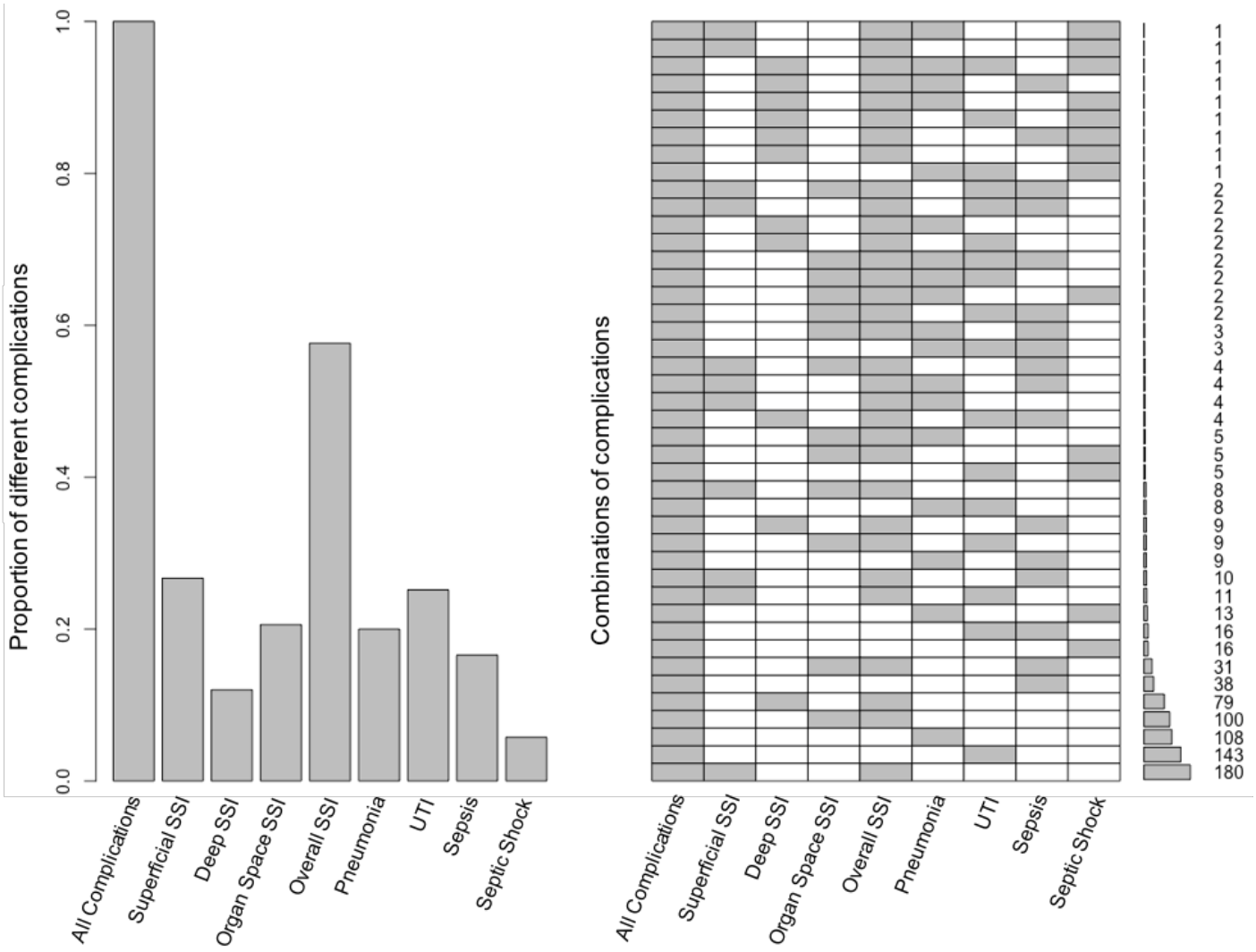

Figure 5-1. A hierarchical structure among postoperative complications

In our application, we have a hierarchy of tasks as shown in **Figure 5-2**. The first task is to distinguish patients with infection from those without. Next we distinguish among the various kinds of infections and finally, if the patient happens to have SSI, we distinguish among the three types of SSI. We assume that many infections share some characteristics that other diseases do not; and we further assume that many types of SSI share some characteristics that non-SSI infections do not. Our hypothesis is that by making a task-

similarity hierarchy available to the multi-task learning methods as domain knowledge, they can utilize these information towards building more robust and better performing detection models.

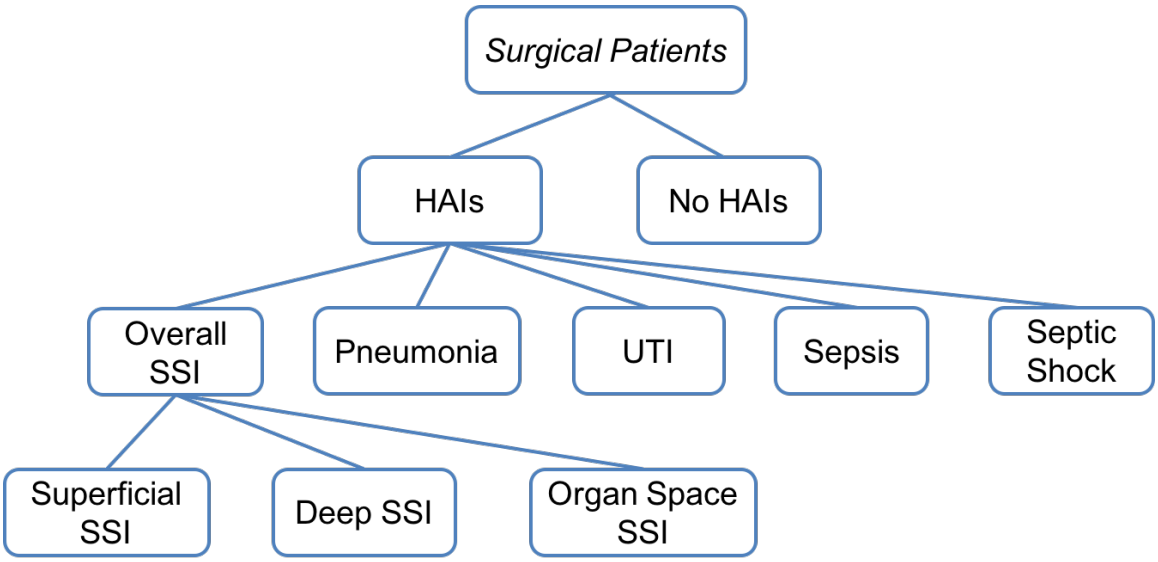

Figure 5-2. A hierarchical structure among postoperative adverse events

5.2 Data collection and preprocessing

We first identified surgical patients from 2011 to 2014 who had been selected for inclusion into the NSQIP. Clinical data for the identified patients were extracted from our CDR and their postoperative complication outcomes were retrieved from the NSQIP registry. The dataset was divided into training set (first 2.5 years) for model development and test set (last 1.5 years) for evaluation. The occurrences of postoperative complications in both training and test set are shown in **Table 5-1**. Overall SSI included any type of SSI.

Table 5-1. Postoperative complications distribution in training and test set

	All Complica-tions	Super-ficial SSI	Deep SSI	Organ Space SSI	Overall SSI	Pneu monia	UTI	Sepsis	Septic Shock	All Observa-tions
Training set	571	168	76	105	336	124	140	115	34	5280
Test set	279	59	26	73	157	48	76	30	17	3629

Data preprocessing generally included data cleaning and missing data imputation. Our previous work on missing data imputation methods suggests that filling in the average value of patients without complications (normal value) in the training set and the test set introduces the least bias17. Accordingly, in this research, we chose to follow this imputation method.

For longitudinal lab results and vital records, we used aggregated features. We included the most recent value before the operation as the baseline, and the extreme and mean values from day 3 to day 30 after the operation during follow-up. We assembled a list of relevant antibiotics for each specific outcome, and extracted different classes of antibiotics as features. For relevant orders, procedures, and diagnosis codes, binary variables were created to denote if they were assigned to a patient. Additionally, we not only considered whether a microbiology test was ordered or not, but also looked at the specific bacterial morphological types (e.g. gram positive rods, gram negative cocci). Two binary features were built for each test to represent a culture placed or not and a positive/negative result, separately.

5.3 Methods

After data collection and preprocessing, the six modeling techniques were applied.

When building our detection models, we face three key challenges. The first one is the skewed class distribution. A mere 10% of patients in the training set have any complications and some complications, like septic shock, occur in only .6% (half a percent) of the patients. The second challenge is the small sample size. Some complications, like septic shock, only have 34 observations in the training and 17 in the test set. While our problem does not appear particularly high dimensional, for these rare complications, the number of predictors (approx. 200) exceeds the number of samples. The third challenge is the heterogeneity of the outcomes. We have 9 outcomes, each having their own specific characteristics and there are also variations among patients who do not have any complications. A successful detection algorithm has to address some of these challenges.

Below, we explain each of the six methods and describe which of the above challenges they address.

5.3.1 Hierarchical Classification

Let us consider the hierarchical structure among surgical patients as shown in **Figure 5-2**. All tasks are divided into three levels and models are constructed in a top-down fashion. The top-most task identifies patients with any postoperative complication. Next, in patients who are predicted to have complications, a 2nd level task is carried out to distinguish between SSI, pneumonia, UTI, sepsis, and septic shock. If a patient is predicted to have SSI, a further 3rd level task is also carried out to identify the SSI type: superficial, deep, and organ space SSI. Each task utilizes Lasso-penalized logistic regression. When more than two classes are possible, the one-vs-all approach is used to break a multi-class classification into a set of binary classifications.

As the method progresses from the top towards the bottom of the hierarchy, it gradually focuses on subpopulations that are enriched in the outcome of interest. This addresses heterogeneity by explicitly ignoring patients without indication of the outcome and also addresses the skewed class distribution. The adoption of LASSO model can overcome the problem of small sample size.

5.3.2 Offset Method

Similar to the hierarchical method, the classifiers for different tasks are built in a top-down fashion. For the top level task (i.e. complication classifier), a LASSO logistic regression classifier is built directly. For the lower-level tasks, we essentially model the difference between the parent and the child task. For example, the deep SSI classifier models the difference between overall SSI and deep SSI. This is achieved through penalizing the child model against the parent model: the predictions from the parent model are included as an offset term (a term with fixed coefficient of 1) in the child model, which is a LASSO logistic regression classifier. Due to the Lasso penalty, variables that have the same effect in the parent and child model will have a coefficient of 0; and conversely, variables that have non-zero coefficient are the variables in which the parent and child tasks differ. The method addresses the challenge of small sample size in two ways. First, in contrast to method 2, it uses the entire population at each level, thus the problem does not become overly high dimensional. Also the offset biases the child classifier towards the parent model. This method only offers limited ability to address heterogeneity.

Figure 5-3 shows the mechanism of the offset method. Suppose we have hierarchical tasks r, s, and t, respectively; task r and s are the parent tasks of task s and t,

respectively; p is the number of features; N is the number of sample size; $y_{r,i}$, $y_{s,i}$, $y_{t,i}$ denote the gold standard of the task r, s, and t for a given subject *i*, respectively. From **Figure 5-3**, the estimation of the child model is dependent on the parent model.

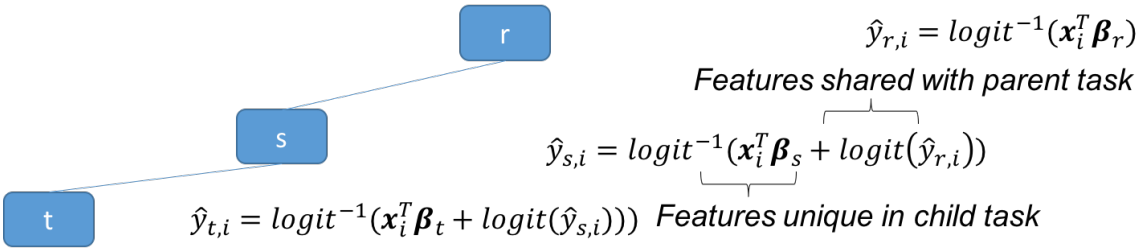

Figure 5-3. Mechanism of offset method.

For the top level task r, the model below estimates the probability of a given subject, *i*, having the event:

$$\hat{y}_{r,i} = logit^{-1}(\boldsymbol{x}_i^T\boldsymbol{\beta}_r) \quad \text{(Eq. 5-1)}$$

To estimate β_r, we need solve the optimization problem below:

$$\underset{\boldsymbol{\beta}_r \in \mathbb{R}^p}{\operatorname{argmin}}\, l(\boldsymbol{\beta}_r) + r(\boldsymbol{\beta}_r) \quad \text{(Eq. 5-2)}$$

where $l(\boldsymbol{\beta}_r)$ is the negative log likelihood function defined below:

$$l(\boldsymbol{\beta}_r) = -\frac{1}{N}\sum_{i=1}^{N}[y_{r,i}(\boldsymbol{x}_i^T\boldsymbol{\beta}_r) - \log\,(1 + e^{\boldsymbol{x}_i^T\boldsymbol{\beta}_r})] \quad \text{(Eq. 5-3)}$$

and $r(\boldsymbol{\beta}_r)$ is the lasso regulation term,

$$r(\boldsymbol{\beta}_r) = \lambda\|\boldsymbol{\beta}_r\|_1 = \lambda\sum_{j=1}^{p}|\beta_{r,j}| \quad \text{(Eq. 5-4)}$$

where λ is a tuning parameter which is estimated by cross validation.

For the secondary level task s, the model below estimates the probability of a given subject, *i*, having the event:

$$\hat{y}_{s,i} = logit^{-1}(\boldsymbol{x}_i^T \boldsymbol{\beta}_s + logit(\hat{y}_{r,i})) \quad \text{(Eq. 5-5)}$$

To estimate $\boldsymbol{\beta}_s$, we need solve the optimization problem below:

$$\underset{\boldsymbol{\beta}_s \in \mathbb{R}^p}{\operatorname{argmin}}\, l'(\boldsymbol{\beta}_s) + r(\boldsymbol{\beta}_s) \quad \text{(Eq. 5-6)}$$

where $l'(\boldsymbol{\beta}_s)$ is the negative log likelihood function defined below:

$$l'(\boldsymbol{\beta}_s) = -\frac{1}{N}\sum_{i=1}^{N}[y_{s,i}\left(\boldsymbol{x}_i^T \boldsymbol{\beta}_s + logit(\hat{y}_{r,i})\right) - \log\left(1 + e^{\boldsymbol{x}_i^T \boldsymbol{\beta}_s + logit(\hat{y}_{r,i})}\right)] \quad \text{(Eq. 5-7)}$$

$r(\boldsymbol{\beta}_s)$ is the same lasso regulation term as the parent model,

For the tertiary level task t, the model below estimates the probability of a given subject, *i*, having the event:

$$\hat{y}_{t,i} = logit^{-1}(\boldsymbol{x}_i^T \boldsymbol{\beta}_t + logit(\hat{y}_{s,i}))) \quad \text{(Eq. 5-8)}$$

To estimate $\boldsymbol{\beta}_t$, we need solve the optimization problem below:

$$\underset{\boldsymbol{\beta}_t \in \mathbb{R}^p}{\operatorname{argmin}}\, l"(\boldsymbol{\beta}_t) + r(\boldsymbol{\beta}_t) \quad \text{(Eq. 5-9)}$$

where $l''(\beta_t)$ is the negative log likelihood function defined below:

$$l"(\boldsymbol{\beta}_t) = -\frac{1}{N}\sum_{i=1}^{N}[y_{t,i}\left(\boldsymbol{x}_i^T \boldsymbol{\beta}_t + logit(\hat{y}_{s,i})\right) - \log\left(1 + e^{\boldsymbol{x}_i^T \boldsymbol{\beta}_t + logit(\hat{y}_{s,i})}\right)] \quad \text{(Eq. 5-10)}$$

$r(\boldsymbol{\beta}_t)$ is the same lasso regulation term as the parent model.

5.3.3 Propensity weighted observation

The propensity weighted observations method also builds classifiers from the top level to the bottom level. The classifier of the top-level task, the complication classifier, is LASSO-penalized logistic regression (same as all of the previous methods). The classifiers for the second level task are built on the entire population, however, the observations

(patients) are weighted by their propensity of having a complication. The propensity is obtained from the higher-level (complication) classifier. Patients, who are likely to have a complication receive a relatively large weight, while patients who are unlikely to have a complication receive a small weight. Therefore, patients with complication contribute more to the 2nd level classifiers than those who are unlikely to have complications. Similarly, the 3rd level classifiers, which distinguish between the three kinds of SSI, are also built on the entire population. The weights of the patients are their propensity of having SSI, thus the patients who likely have SSI contribute more to these classifiers than patients who are unlikely to have SSI. Similarly, to the offset method, the PWO method uses the entire population, but by applying weights, it reduces outcome heterogeneity (patients with unrelated complications receive small weights) and reduces the skew of the class distribution by enriching the training set with patients having the outcome of interest (these patients receive high weights).

Figure 5-4 shows the mechanism of the propensity weighted method. Suppose we have hierarchical tasks r, s, and t, respectively; task r and s are the parent tasks of task s and t, respectively; p is the number of features; N is the number of sample size; $y_{r,i}$, $y_{s,i}$, $y_{t,i}$ denote the gold standard of the task r, s, and t for a given subject *i*, respectively. The estimation of the child model in **Figure 5-4** is actually dependent on the parent model.

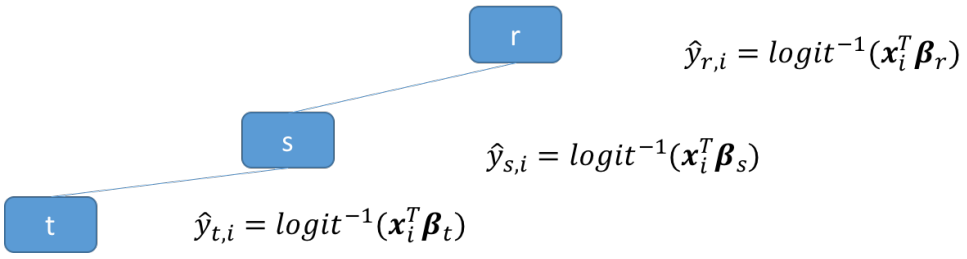

Figure 5-4. Mechanism of propensity weighted method.

For the top level task r, the model below estimates the probability of a given subject, *i*, having the event:

$$\hat{y}_{r,i} = logit^{-1}(\boldsymbol{x}_i^T \boldsymbol{\beta}_r) \quad \text{(Eq. 5-11)}$$

To estimate β_r, we need solve the optimization problem below:

$$\underset{\boldsymbol{\beta}_r \in \mathbb{R}^p}{\operatorname{argmin}}\, l(\boldsymbol{\beta}_r) + r(\boldsymbol{\beta}_r) \quad \text{(Eq. 5-12)}$$

where $l(\boldsymbol{\beta}_r)$ is the negative log likelihood function defined below:

$$l(\boldsymbol{\beta}_r) = -\frac{1}{N}\sum_{i=1}^{N}[y_{r,i}(\boldsymbol{x}_i^T \boldsymbol{\beta}_r) - \log\,(1 + e^{\boldsymbol{x}_i^T \boldsymbol{\beta}_r})] \quad \text{(Eq. 5-13)}$$

and $r(\boldsymbol{\beta}_r)$ is the lasso regulation term,

$$r(\boldsymbol{\beta}_r) = \lambda \|\boldsymbol{\beta}_r\|_1 = \lambda \sum_{j=1}^{p} |\beta_{r,j}| \quad \text{(Eq. 5-14)}$$

where λ is a tuning parameter which is estimated by cross validation.

For the secondary level task s, the model below estimates the probability of a given subject, *i*, having the event:

$$\hat{y}_{s,i} = logit^{-1}(\boldsymbol{x}_i^T \boldsymbol{\beta}_s) \quad \text{(Eq. 5-15)}$$

To estimate $\boldsymbol{\beta}_s$, we need solve the optimization problem below:

$$\underset{\boldsymbol{\beta}_s \in \mathbb{R}^p}{\operatorname{argmin}}\, l'(\boldsymbol{\beta}_s) + r(\boldsymbol{\beta}_s) \qquad \text{(Eq. 5-16)}$$

where $l'(\boldsymbol{\beta}_s)$ is the propensity weighted negative log likelihood function defined below:

$$l'(\beta_s) = -\frac{1}{N}\sum_{i=1}^{N} \hat{y}_{r,i}[y_{s,i}(\boldsymbol{x}_i^T \boldsymbol{\beta}_s) - \log\,(1 + e^{\boldsymbol{x}_i^T \boldsymbol{\beta}_s})] \qquad \text{(Eq. 5-17)}$$

$r(\boldsymbol{\beta}_s)$ is the same lasso regulation term as the parent model.

For the tertiary level task t, the model below estimates the probability of a given subject, *i*, having the event:

$$\hat{y}_{t,i} = logit^{-1}(\boldsymbol{x}_i^T \boldsymbol{\beta}_t) \qquad \text{(Eq. 5-18)}$$

To estimate $\boldsymbol{\beta}_t$, we need solve the optimization problem below:

$$\underset{\boldsymbol{\beta}_t \in \mathbb{R}^p}{\operatorname{argmin}}\, l''(\boldsymbol{\beta}_t) + r(\boldsymbol{\beta}_t) \qquad \text{(Eq. 5-19)}$$

where $l''(\boldsymbol{\beta}_t)$ is the negative log likelihood function defined below:

$$l''(\boldsymbol{\beta}_t) = -\frac{1}{N}\sum_{i=1}^{N} \hat{y}_{s,i}[y_{t,i}(\boldsymbol{x}_i^T \boldsymbol{\beta}_t) - \log\,(1 + e^{\boldsymbol{x}_i^T \boldsymbol{\beta}_t})] \qquad \text{(Eq. 5-20)}$$

$r(\boldsymbol{\beta}_t)$ is the same lasso regulation term as the parent model.

5.3.4 Multi-task learning with penalties

Unlike the previous methods, the objective of multi-task learning with penalty (MTLP) method is to learn the regression coefficients $\boldsymbol{\beta}_t$ for all tasks simultaneously. In the MTLP method, we assume that the parent task and its child tasks share some features and the respective models should have similar coefficients for those features. Similarly, to the offset method, this similarity is enforced through penalizing the child model against the parent model. Unlike the offset method, which builds models in a top-down manner, MTLP builds all models simultaneously. Specifically, the objective function is

$$\underset{\{\beta_t \in \mathbb{R}^D\}}{argmin}\, l(\beta_t) + r_1(\beta_{level\_2}) + r_2(\beta_{level\_3}) \quad \text{(Eq. 5-21)}$$

It consists of three parts, the negative log likelihood of logistic regression, $l(\beta_t)$, and two regularization terms, $r_1(\beta_t)$ and $r_2(\beta_t)$, as shown below.

$$l(\beta_t) = -\left[\frac{1}{T \cdot N_t}\sum_{t=1}^{T}\sum_{i=1}^{N_t}\left(y_{t,i} \cdot \left(x_{t,i}^T \beta_t\right) - \log\left(1 + e^{x_{t,i}^T \beta_t}\right)\right)\right] \quad \text{(Eq. 5-22)}$$

$$r_1(\beta_t) = \lambda_1 \sum \left\| \beta_{\text{level1 parent task}} - \beta_{\text{level2 children tasks}} \right\| \quad \text{(Eq. 5-23)}$$

$$r_2(\beta_t) = \lambda_2 \sum \left\| \beta_{\text{level2 parent task}} - \beta_{\text{level3 children tasks}} \right\| \quad \text{(Eq. 5-24)}$$

where T and N_t are the number of tasks and training set for each task, respectively; $\boldsymbol{x}_{t,i}$ and $y_{t,i}$ are the feature vector and the label for the subject i in task t, respectively; $\boldsymbol{\beta}_t$ is the coefficient vector for the task t. The two regularization terms, $r_1(\boldsymbol{\beta}_t)$ and $r_2(\boldsymbol{\beta}_t)$, restrict the difference in coefficients between the level 1 parent task and its level 2 child tasks; and the difference between the level 2 parent task and its level 3 child tasks, respectively. Penalizing the difference between the parent and child models make them similar. The MTLP method addresses heterogeneity by explicitly making the parent and child models similar, thereby essentially only modeling the difference between them; and it addresses the small sample size through the use of the entire population and regularization.

5.3.5 Partial least squares regression

As with the MTLP method, partial least squares (PLS) regression models all tasks simultaneously. PLS regression is similar to principal components regression in the sense that both methods reduce the dimension of input data by projecting the outcomes and predictors into new spaces and then build regression models in those new spaces. PLS

differs from MTLP in that the task hierarchy is not explicitly given to the fitting algorithm; the algorithm has to autonomously learn the relationships among the tasks. The table below shows the algorithm of the PLS.

Table 5.2. PLS Algorithm

1. Standardize each x_j to have mean zero and variance one. Set $\widehat{\boldsymbol{y}}^{(0)} = \bar{y}\times Ones_{n\times 1}$, and $\boldsymbol{x}_j^{(0)} = \boldsymbol{x}_j, j = 1, \dots, p$. Note: p is the number of original predictors; n is the number of observations; $\boldsymbol{x}_j$ is all the n observations of the jth predictor.
2. For $m = 1,2, \dots, p$ (a) $\boldsymbol{z}_m = \sum_{j=1}^{p} \hat{\varphi}_{mj}\boldsymbol{x}_j^{(m-1)}$, where $\hat{\varphi}_{mj} =< \boldsymbol{x}_j^{(m-1)}, \boldsymbol{y} >$. (b) $\widehat{\boldsymbol{\theta}}_m = \frac{<\boldsymbol{z}_m, \boldsymbol{y}>}{<\boldsymbol{z}_m, \boldsymbol{z}_m>}$. (c) $\widehat{\boldsymbol{y}}^{(m)} = \widehat{\boldsymbol{y}}^{(m-1)} + \widehat{\boldsymbol{\theta}}_m\boldsymbol{z}_m$. (d) Orthogonalize each $\boldsymbol{x}_j^{(m)}$ with respect to $\boldsymbol{z}_m$: $\boldsymbol{x}_j^{(m)} = \boldsymbol{x}_j^{(m-1)} - \left[\frac{<\boldsymbol{z}_m, \boldsymbol{x}_j^{(m-1)}>}{<\boldsymbol{z}_m, \boldsymbol{z}_m>}\right]\boldsymbol{z}_m$.
3. Output the sequence of fitted vectors $\{\widehat{\boldsymbol{y}}^{(m)}\}_1^p$. Since the $\{\boldsymbol{z}_l\}_1^m$ are linear in the original $\boldsymbol{x}_j$, so is $\widehat{\boldsymbol{y}}^{(m)} = \widehat{\boldsymbol{X}}\widehat{\boldsymbol{\beta}}^{pls}(m)$. These linear coefficients can be recovered from the sequence of PLS transformations.

5.4 Evaluation

Outcomes based on MCR from ACS-NSQIP were used as gold standard to be

compared with the results of postoperative complication detection models. The evaluation metric is area under the curve (AUC), which is commonly used to compare detection models. The range for AUC is between .5 and 1, .5 indicating a random model and 1 indicating perfect discrimination among the outcomes. We report the cross-validated AUC on the training set. To assess the variability of the detection performances on the test, bootstrap replication was applied and the 95% (empirical) confidential interval (CI) and mean AUC scores are reported, as well. Since all methods were evaluated on the same bootstrap samples, paired t-test was used to compare each pair of methods and assess the statistical significance of the observed differences in performance.

5.5 Results

5.5.1 Evaluation results of six detection methods

Figure 5-5 depicts the performances of the six methods. Each plot in Figure 2 corresponds to a task (complication) and each column in each plot corresponds to a method. Methods are numbered in the same order as they appear in the Methods section: #1 corresponds to Single-task, #2 to Hierarchical, #3 to Offset, #4 to Propensity Weighted Observations (PWO), #5 to Multi-Task Learning with Penalty (MTLP), and #6 corresponds to Partial Least Squares (PLS). The vertical axis is AUC. For each method, the mean AUC (across the bootstrapped test samples) is represented by a disk and lines extending out of the disk correspond to the 95% CI.

To assess the statistical difference between some of the methods, in **Table 5-3**, we show the results of pairwise (paired) t-tests among the various methods. The rows of the table correspond to tasks, the columns to a comparison between two methods. Each cell

contains a number, which indicates which method has a significantly better performance and we also provide the p-value in brackets. 'NS' means 'not significant'. The methods are numbered in the same way as above.

For the detection of all complications, Single-task, Hierarchical, Offset, PWO, and MTLP have the same good performance, and are significantly better than PLS. To detect superficial SSI, Offset and PWO have virtually identical performance (difference is not significant) and they perform significantly better than the other four methods. PWO performs best for detecting deep SSI and overall SSI. Single-task, PWO, and MTLP all perform similarly (no statistically significant difference) in detecting organ space SSI but perform significantly better than the other methods. To detect pneumonia and UTI, Single-task and MTLP are not significantly different from each other but are significantly better than other four methods. Single-task, PWO, and MTLP are the top three in detecting sepsis and PWO is also the best method for detecting septic shock. In general, PWO is the best method for detecting most complications. Single-task and MTLP are close seconds (and they have virtually identical performance) and PLS is the method with the worst overall performance.

Detailed information about the performance of the methods is depicted in **Figure 5-5** and the statistical significance of the pairwise comparisons between the various methods is shown in **Table 5-3**.

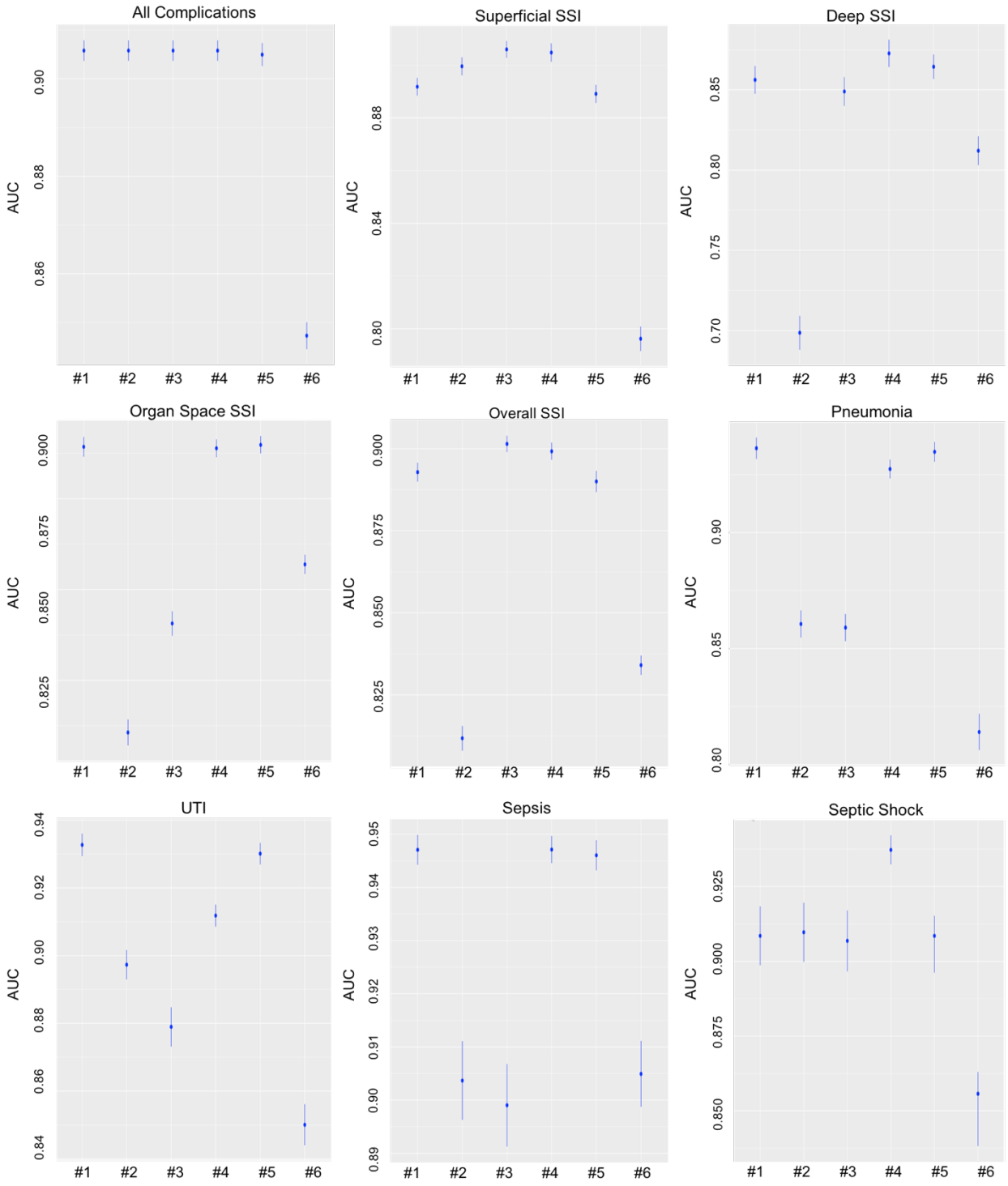

Figure 5-5. Detection performance of six models for all nine tasks, showing the mean and 95% CI

Table 5-3. Paired t-test results to compare different methods

	Method 1 vs. 2	**Method 1 vs. 3**	**Method 1 vs. 4**	**Method 1 vs. 5**	**Method 2 vs. 3**
Superficial SSI	2 (<2.2e-16)	3 (<2.2e-16)	4 (<2.2e-16)	NS	3 (=1.533e-14)
Deep SSI	1 (< 2.2e-16)	1 (= 5.073e-07)	4 (< 9.285e-10)	NS	3 (< 2.2e-16)
Organ Space SSI	1 (< 2.2e-16)	1 (< 2.2e-16)	NS	NS	3 (< 2.2e-16)
Overall SSI	1 (< 2.2e-16)	1 (< 2.011e-13)	4 (< 3.572e-15)	NS	3 (< 2.2e-16)
Pneumonia	1 (< 2.2e-16)	1 (< 2.2e-16)	1 (=2.353e-8)	NS	2 (< 2.2e-16)
UTI	1 (< 2.2e-16)	1 (< 2.2e-16)	1 (< 2.2e-16)	NS	2 (< 2.2e-16)
Sepsis	1 (< 2.2e-16)	1 (< 2.2e-16)	NS	NS	2 (< 2.2e-16)
Septic Shock	NS	1 (< 2.2e-16)	4 (<1.671e-12)	NS	2 (< 2.2e-16)
	Method 2 vs. 4	**Method 2 vs. 5**	**Method 3 vs. 4**	**Method 3 vs. 5**	**Method 4 vs. 5**
Superficial SSI	4 (< 5.879e-16)	2 (< 2.2e-16)	3 (=0.001203)	3 (< 2.2e-16)	4 (< 2.2e-16)
Deep SSI	4 (< 2.2e-16)	5 (< 2.2e-16)	4 (= 1.151e-14)	5 (= 5.073e-07)	4 (= 9.285e-10)
Organ Space SSI	4 (< 2.2e-16)	5 (< 2.2e-16)	4 (< 2.2e-16)	5 (< 2.2e-16)	NS
Overall SSI	4 (= 1.607e-14)	5 (< 2.2e-16)	3 (= 0.02337)	3 (= 2.011e-13)	4 (< 3.572e-15)
Pneumonia	4 (<2.2e-16)	5 (< 2.2e-16)	4 (< 2.2e-16)	5 (< 2.2e-16)	5 (= 2.353e-8)
UTI	4 (=1.607e-14)	5 (< 2.2e-16)	4 (< 2.2e-16)	5 (< 2.2e-16)	5 (< 2.2e-16)
Sepsis	4 (<2.2e-16)	5 (< 2.2e-16)	4 (< 2.2e-16)	5 (< 2.2e-16)	NS
Septic Shock	4 (<2.2e-16)	NS	4 (< 2.2e-16)	NS	4 (< 2.2e-16)

5.5.2 Significant variables selected

Lasso-penalized regression performs automatic feature selection. In **Table 5-4**, we provide a list of the most important features selected by the model that aims to identify whether a patient has a complication. This model is common across most methods. Due to space limitation, we cannot provide a list for all methods and all complications. Below we provide some examples of features selected by the best performing method for each of the complications.

Superficial SSI detection model based on Offset and Propensity Weighted Observations methods selected antibiotic use, gram stain ordered, and the ICD_9 code of SSI.

Besides diagnosis codes, antibiotics use, gram stain culture, **deep SSI** detection model based on PWO selected more features from laboratory results (mean value of creatinine and maximum value of WBC) and two more microbiology tests (tissue and wound culture).

Organ space SSI models based on Single-task, PWO, and MTLP methods have quite similar detection performance and important variables. The selected variables include four features of bacteria type (streptococcus, gram positive cocci, enterococcus, and escherichia. coli), three microbiology cultures (abscess and fluid culture), and the imaging orders of treatment. Interestingly, the diagnosis code of sepsis and imaging orders of sepsis treatment are selected as well. These can be explained by the fact that there are over 30 patients in our cohort have sepsis and organ space SSI together.

Table 5-4. Selected important variables for all complications and their descriptions

Category	Name	Description
	In/Out patient	Inpatient or outpatient surgery
Diagnosis code	ICD_9 code: 998.59 and 997.32	Diagnosis code of postoperative SSI and pneumonia
Microbiology test order	Abscess culture Blood culture Gram stain culture Sputum culture Urine culture	These are binary features to indicate if such microbiology test ordered or not during day 3 to day 30 after operation.
Microbiology test result	Escherichia. Coli Staphylococcus	These are binary features to indicate if the type of bacteria is positive or not no matter in which kind of microbiology test during day 3 to day 30 after operation.
Antibiotic use	Antibiotic_Superficial_SSI Antibiotic_Pneumonia Antibiotic_UTI	They are binary features to indicate if antibiotics is placed to patients during day 3 to day 30 after operation.
Laboratory results	Measurement_CR Measurement_PLT Measurement_PREALAB Measurement_WBCU	The number of measurements for creatinine, platelet count test, prealbumin, and urine white blood cells (WBCU).

Overall SSI detection models based on weighted observation and offset methods perform with no significant difference. The important variables selected are antibiotic use (UTI, superficial and deep SSI), microbiology cultures (abscess, fluid, and wound culture,

and the gram stain test), two types of bacteria (escherichia. coli and staphylococcus) and two relevant order features (imaging orders for diagnosis and procedures of treatment).

Besides the diagnosis code and antibiotic use, **pneumonia** models based on Single-task and MTLP include two features from microbiology test (bronchial and sputum culture), one binary feature of image-guided diagnosis orders, and two aggregate features from lab tests (the mean value of PCAL and PH).

UTI models based on Single-task and MTLP selected diagnosis codes (for UTI), antibiotic use, placement of urine culture, and two bacteria types (proteus and escherichia. coli).

For **sepsis** models, the top performing methods, Single-task, PWO, and MTLP, selected features including antibiotic use, microbiology cultures (abscess, blood, fluid, and urine culture, and the gram stain), two bacterial types (enterococcus and escherichia. coli), and the image guided orders of treatment.

For **septic shock** detection, most models only selected diagnosis codes. However, PWO included more variables, such as laboratory tests (maximum value of partial thromboplastic time, PH, mean value of lactate), bacteria types (stenotrophomonas and staphylococcus) and the tracheal culture.

5.6 Discussion

Manual chart review for post-operative complications is very resource intensive. In this work, we examined whether EHR-based state-of-the-art predictive modeling approaches can learn characteristics of various types of complications and subsequently detect them reliably. With detection performances (measured as AUC) exceeding 0.8 for

all complications and even 0.9 for some complications, the answer is affirmative: machine learning detection models definitely have the potential to help detect post-operative complications automatically. The question is which modeling approach is best suited for this application.

Post-operative complications are heterogeneous; they cover a wide-range of conditions, each having their own diagnostic methods, diagnoses codes, laboratory tests, and diagnostic and therapeutic procedures. They can be organized into a hierarchy and complications on the same level of the hierarchy are more similar to each other than to complications on a higher level of the hierarchy. Multi-task learning methods have the ability to exploit such similarities towards achieving better detection performance and more stable models even when the sample sizes are small.

We compared six approaches to building post-complication detection models. One of them was single-task learning, where we build independent models for each task; four methods were multi-task learning methods that can utilize the hierarchy of complications; and finally, we also utilized Partial Least Squares (PLS), which can simultaneously model multiple outcomes, but it tries to autonomously detect the relationship among the outcomes. PLS thus stands in sharp contrast with the other multi-task learning methods, as PLS automatically infers the relationships among the complications, while the other multi-task learning methods receive this information from an expert.

We found PLS to have the overall worst performance. This is not surprising, since PLS receives less information than the other multi-task learning methods. We expect PLS to bias the models based on the relationships among the outcomes, but we do provide it

with these relationships. If PLS infers the relationships among outcomes incorrectly, it will bias the models incorrectly, eroding detection performance. With some of the complications having small sample sizes, it is unsurprising the PLS failed to infer the correct relationships. If we had substantially more samples, PLS could have inferred the relationship among complications possibly better than what the expert can provide, but our sample size, albeit relatively large, was insufficient for this purpose. Single task learning managed to (significantly) outperform PLS, because we did not "force" it to bias the models beyond applying Lasso-penalty which is virtually mandatory given our sample sizes for some of the complications.

Hierarchical modeling also had disappointing performance. The essence of hierarchical modeling is to build classifiers in a subpopulation that is greatly enriched in the outcome of interest. For example, distinguishing among the three types of SSI is easier in a subpopulation of SSI patients than it is in the general population. The performance of the method did not live up to our expectation for two reasons. First, while these outcomes are rare (deep SSI occurred in 76 patients out of 5280), they still occur in sufficient numbers for a Lasso-penalized logistic regression model. The second reason concerns the way the subpopulations were constructed. If the higher-level classifier (does this patient has SSI?) predicts the patient to be free of SSI, then this patient does not enter the subpopulation and the deep SSI detector has no opportunity to learn from this sample. We could have built the deep classifier on the true SSI patients (rather than the predicted SSI patients), but then the distributions of the training SSI patients (true SSI patients) and the test SSI patients (predicted SSI patients) would be different, leading to degraded detection performance.

Our results with the Propensity Weighted Observations method tell us that the concept of enriching patients with SSI for (say) the deep SSI classifier is valid; the hierarchical method simply implemented this concept suboptimally.

The Propensity Weighted Observations (PWO) method achieved the overall highest performance with a margin that is statistically significant. PWO is closely related to the hierarchical method in that it enriches the training sample with patients who have the outcome of interest. In contrast to the hierarchical method, it achieves this enrichment through constructing a new sample, which is a propensity weighted version of the original population. For example, to identify patients with deep SSI, PWO uses the entire population, but patients with high propensity for SSI receive high weight and patients with low propensity for SSI receive low weight. The hierarchical method is a binary version of PWO, where the weights are either 0 or 1. Having the propensity weighted population removes the problem of excluding patients based on an incorrect prediction. Suppose our SSI classifier misclassifies a deep SSI patient as not having SSI. This patient will still be included in the training set for the deep SSI classifier; this observation will receive a slightly lower weight. Admittedly, using propensity score weighing for multi-task learning is rather unusual; we did not expect this method to perform so well.

The offset method, like PWO, always uses the entire population for classification. Its performance falls short of that of PWO, because its ability to remove heterogeneity is limited. It can bias a child model against the parent model, which helps with small sample sizes (it performed well on superficial SSI), but has limited effect on removing the variation

in (say) normal patients. PWO is more effective at removing heterogeneity: patients with unrelated complications receive a low weight and contribute to the model only minimally.

MLTP is essentially identical to the offset method, except MLTP optimizes all outcomes simultaneously (as opposed to sequentially in a top-down manner). In a top-down construction scheme, only the parent task can influence the child task; the model from the child task cannot influence the parent model. When all tasks are carried out simultaneously, the child models can influence the parents, as well. As a result, MLTP was either the best or second best method for almost all rare (<3% of patients) outcomes. The caveat of simultaneous optimization is the increased potential for overfitting. Indeed, comparing MTLP's cross-validated AUC scores on training set to those on the test set, reveal signs of overfitting. For example, to detect sepsis, it has a very high training AUC (>0.96), but the 95% CI of AUC on test set is only (0.8986, 0.9183).

5.7 Conclusion

Developing machine learned models to automatically detect post-operative complications definitely has the potential to accelerate the manual chart review process. We found that multi-task learning, specifically, the propensity weighted observations method, statistically significantly outperformed the single-task learning approach. While the difference in detection performance was relatively modest (albeit significant), the additional cost of implementing this method over the standard single-task learning method is minimal. Thus, we would recommend trying both single-task learning and PWO.

Our application was relatively easy: we had sufficiently many samples for Lasso-penalized logistic regression to construct a good model even for the most infrequent

outcome. In an application, where fewer samples are available or outcome distributions are more skewed, we would expect the performance gap between multi-task learning and single-task learning to open up, providing a more attractive implementation cost versus detection performance proposition for multi-task learning.

Our future work includes building postoperative complications detection models using both structured and unstructured EHR data. We hypothesize that the combination of structured and unstructured clinical data would include more significant indicators and signals of postoperative complications, and improve the performance of detection. The performance of the models with only structured data and that of the models with both structured and unstructured data will be compared and evaluated.

Chapter 6 Using EHR data and clinical notes to automatically detect surgical adverse events

6.1 Introduction

To improve the efficiency of surgical adverse events reporting, we have explored the feasibility of computerizing the process of retrospective chart review. Our previous work was conducted based on only structured EHR data. The purpose of this study was to design and evaluate an automatic HAI event detection platform based on both structured EHR data and unstructured clinical notes.

We hypothesized that the combination of structured EHR data and unstructured clinical notes with NLP approaches would improve detection performance compared to solely using either structured EHR data or unstructured clinical notes.

6.2 Materials and Methods

6.2.1 Data collection

We collected EHR data and clinical notes from the UMN CDR for the surgical patients in our cohort along with NSQIP postoperative adverse event outcomes from the registry. The patient's medical record number and date of surgery were used to link CDR data to the NSQIP registry. We used data from 2011 onward following implementation of the current Enterprise EHR system (Epic systems). Patients without matching records in the CDR (22 total, from incorrectly entered medical record numbers in NSQIP) and who

opted out of using their EHR data for research research purposes (215 total) were removed from the study cohort.

Unstructured clinical notes include discharge summaries, progress notes, and operative notes. Surgeons and surgical residents at the department of surgery helped compile potential keywords which might correspond to each type of adverse event. For example, keywords related to the diagnosis and treatment of an SSI include abscess, anastomotic leak, or wound dehiscence. A natural language processing tool, NLP-PIER (Natural Language Processing-Patient Information Extraction for Research) developed by the natural language processing/information extraction (NLP/IE) program at the University of Minnesota enables the free-text and semantic searches in the clinical notes (77). For example, searching for the keyword and UMLS concept *abscess* (i.e., CUI: C0024110) generates a list of patients who have this word in the notes with no negation, and detailed information about the mention (i.e., note date and the type of notes). Therefore, keywords and concepts were extracted from clinical notes and binary features were created to indicate the presence or absence of a positive mention of a specific keyword or concept.

Using Organ Space SSI as an example (**Table 6-1**), the standard definition used by NSQIP is provided. Relevant structured EHR data, and related keywords or concepts were searched from clinical notes.

Table 6-1. NSQIP Definition for Organ Space SSI, relevant data elements used from EHR data, and keywords or concepts from clinical notes

Organ Space SSI is an infection that occurs within 30 days after the principal operative procedure. The criteria include:	Relevant data elements from EHR database	Relevant keywords / concepts from clinical notes
AND involves any of the anatomy (e.g., organs or spaces), other than the incision, which was opened or manipulated during the operation **AND** at least ONE of the following: A. Purulent drainage from a drain that is placed through a stab wound into the organ/space. This does not apply to drains placed during the principal operative procedure, which are continually in place, with continual evidence of drainage/infection since the time of the principal operative procedure B. Organisms isolated from an aseptically obtained culture of fluid or tissue in the organ space C. An abscess or other evidence of infection involving the organ/space that is found on direct examination, during reoperation, or by histopathologic or radiologic examination D. Diagnosis of an organ/space SSI by a surgeon or attending physician	(1) Demographics (e.g., age, gender, race) (2) Histories (e.g., anemia, diabetes mellitus, BMI, etc.) (3) Lab tests (e.g., WBC, HGB, PLT, CR, etc.) (4) Microbiological results (blood culture, drainage culture, fluid culture, skin culture, tissue culture, wound culture) (5) Vitals (temperature, heart rate, blood pressure, respiratory rate, pain scale) (6) Antibiotics (e.g. Amoxicillin, Clindamycin, Levofloxacin, Vancomycin, etc.) (7) Imaging orders (e.g., CT abdomen, CT chest, X-ray colon, CT guided abscess drainage, IR abscess tube check, etc.) (8) Procedures (e.g., drain care, wound care referral, infectious disease referral, etc.) (9) Diagnosis codes (998.59 for ICD-9; K68.11 and T81.4 for ICD-10)	Abscess, Anastomotic dehiscence, Anastomotic leak, Cellulitis/cellulitic, Cloudy, Dehiscence, Demarcated/demarcation, Drain care, Drainage, Drain placement, Dressing/dressing change, Empyema, Erythema, Evisceration, Extraluminal, Extravasation, Fistula, Foul-smelling, Hartmann's blowout, Induration/Indurated, Infected/infection, Intra-abdominal abscess IV Antibiotics, Joint abscess, Leak, Interventional radiology, Malodorous, Murky, Open wound, Packing/packing change, Pelvic abscess, Pelvic collection, Pelvic sepsis, Phlegmon, Presacral abscess, Purulent, Rectal stump blowout, Presacral abscess, Rim enhancing, Wet to dry, Wound dehiscence, Wound infection, Wound packing, Vac dressing

6.2.2 Data analysis

Data for analysis from the EHR and NSQIP were integrated into our database with NSQIP outcomes used as the gold-standard. The whole dataset was divided by calendar year rather than random sampling with the most recent events closer to our target (understanding that our models would be used to detect future events). Data from 2011 to 2013 were used as training set for model development and 2014 to 2015 data were used to test detection performance. Since the input features were high-dimensional and the data matrix was sparse, LASSO regression method was adopted for modeling since this type of analysis could automatically select the most important features for an adverse event. The performance of our models is provided via a receiver operating characteristic (ROC) curve. Evaluation metrics used in this study are the AUC score, sensitivity and specificity for a particular threshold. In this case, Youden's index was used to maximum both sensitivity and specificity. To test the significance between different models, 95% CIs were calculated by performing a 1,000 replications of the Bootstrap procedure in R version 3.3.3 (2017-03-06).

6.3 Results

The mean AUC score and the 95% CI were calculated. **Figure 6-1** summarizes the mean AUC and corresponding 95% CI of each model for every HAI event. Generally speaking, AUC sores are quite promising. When 95% CIs of two models do not overlap, there will indeed be a statistically significant difference between the means (at the 0.05 level of significance). Except for superficial SSI and sepsis, the performance of models

based on both EHR data and clinical notes work significant better for other five types of HAIs (i.e., Deep SSI, Organ Space SSI, PNA, UTI, Septic Shock).

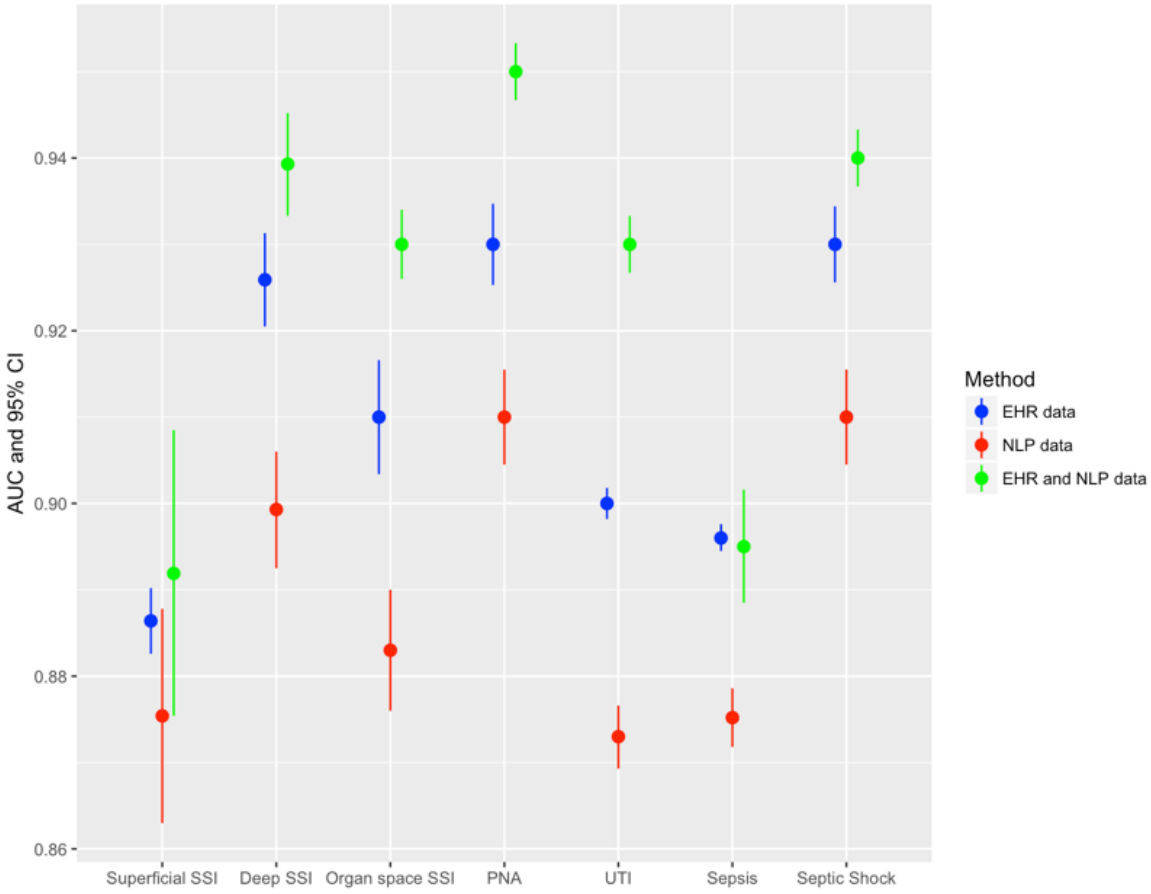

Figure 6-1. AUC score and 95% CI for HAIs based three different datasets.

For the purposes of this study, we varied the cut-off of the sensitivity and specificity to maximize both since cases with and without complications are important. Youden's index was used for finding the optimal cut-off (**Table 6-1**) (78-79). While statistically similar, the model using only EHR data has the highest sensitivity, whereas the model using both EHR data and clinical notes and model using only clinical notes was better in terms of specificity.

Table 6-2. Using Youden's index to find the cutoff maximizing both sensitivity and specificity for organ space SSI

	EHR data	Clinical notes	EHR and clinical not
Sensitivity	0.9628 (0.9486, 0.9770)	0.8432 (0. 8231, 0.8633)	0.9204 (0.9063, 0.9345)
Specificity	0.7636 (0.7513, 0.7760)	0.8505 (0.8356, 0.8654)	0.8708 (0.8618, 0.8797)

6.4 Discussion and Conclusion

Overall, this study demonstrates the feasibility of automated detection of validated surgical NSQIP adverse event occurrences for seven HAIs and five non HAIs following surgery using EHR data. This report represents one of the most robust and comprehensive analyses to date using EHR data and advanced machine-learning models. Compared to previous reports, our analyses include all NSQIP surgical patients and a hold out set of two years of test data to validate the associated models. Our results demonstrate that automated approaches are feasible for adverse event detection and may, at the very least, be an effective method for focusing abstraction on patients with a greater chance of an adverse event and excluding those with a high probability of no event. While we found that models using both EHR data and clinical notes do not always clearly perform better than models using only either EHR data or clinical notes, using this combination of data resulted in models with statistically similar performance and for several HAIs, performance was improved significantly when both data sources were included.

Like other diagnostic curves, the ROC curve cut-off can be varied resulting in different levels of sensitivity and specificity. We used Youden's index for this study to achieve a high sensitivity and specificity, but this is not the only solution for adjusting the associated cut-offs. In this study, still take Organ Space SSI as example, the highest specificity is about 0.87 and the corresponding NPV values are higher than 0.99 and very close to 1, which means 87% of all true negative cases in our study population could be identified and more than 99% of cases detected as negative are true negatives. The sensitivities are very high, but the corresponding PPV values are quite low, therefore, with true positive cases detected many false positive cases are included as well (one in 10 positive patients were positive). With over 95% of charts eliminated as negatives, the chart review process may be greatly accelerated using this approach.

In designing this study and its associated evaluation, our dataset was divided into a training set and a test set by calendar year rather than randomly sampling, since we are interested in investigating how robust our generated models are over time and if there would be any degradation of performance over time. It is inevitable that due to institutional changes and changes in practice (e.g., change of patient population, distribution of procedure, changes in diagnostic studies) the model performance could drift over time. At some point in the future, model will likely require recalibration or even outright reconstruction. It is undesirable rebuild our model every year. Our method of dividing the data set by year allows us to assess how resilient the models are to such temporal changes.

Although EHR systems were not designed for secondary purposes like research or quality improvement, the data generated and stored in the UMN CDR is a good resource

for these types of secondary use. Our use of the NLP-PIER information extraction tools was lightweight and was not restricted to the specific EHR system used at a medical institution. As such, while some minimal preprocessing was required, this work was streamlined and robust for the task of surgical adverse event detection.

The main limitation in this study is that it is conducted at a single site and the results of a single teaching hospital with a single type of EHR record may not be generalizable. For instance, patient populations, diagnostic and treatment practices, and language used to express information about adverse events may vary by site. To test the reproducibility, model validation based on other sites and EHR systems is necessary.

In conclusion, we observed that machine learning methods with EHR data and NLP applied to clinical notes for automatic adverse event detection was an effective approach. These approaches resulted in good AUC scores and demonstrate very good potential for accelerating the manual postoperative complication identification process. Ultimately, we anticipate that the maturation of this research will have a positive impact on surgical care and surgical quality improvement.

Chapter 7 Summary and Future Directions

7. 1 Summary

This dissertation explored the automated detection of surgical adverse events through machine learned models trained on structured EHR data and clinical notes at the University of Minnesota. Four studies were conducted to achieve this overall objective.

The first study aimed to test the feasibility of this idea using only structured EHR data and we limited our outcomes to the three subtypes of surgical site infection (SSI). Our SSI detection models achieved high classification accuracy proving the possibility of automated detection of surgical adverse events. In the two subsequent studies, we aimed to concentrate on the technical challenges posed by EHR data.

The second study focused on missing data, a common challenge in the secondary use of EHR data, and explored several methods for handling missing data. We compared a number of commonly-used simple imputation methods and some more advanced methods. We found that imputation improved the overall detection performance significantly and that some of the simplest imputation methods yielded excellent performance.

In the third study, we exploited the relationship between the various surgical adverse events through multi-task learning. Multi-task learning can capture the common aspects of related outcomes, potentially leading to better detection performance. We proposed five multi-task learning methods and compared them with the single-task method

(as baseline), which simply ignores the relationship among the outcomes. Surprisingly, complicated methods didn't always perform better for all HAIs.

In our last study, we investigated the use of structured EHR data, clinical notes and the combination of these data types. Models using different types of data were compared on their detection performance.

We have successfully demonstrated that with reliably labeled adverse events as a gold standard, supervised machine learning methods can be effective for automatic detection of surgical adverse events. The best model for each HAI event obtained a very good overall detection performance, with very high AUC score. We have also demonstrated that detection models can achieve very good performance either with structured EHR data alone or with features extracted from clinical notes; but the combined use of these two data types improves detection performance significantly.

7. 2 Future directions

Given the great results our detection models have achieved, a question naturally arises: how can these models help annotators? Though completely automated detection still needs more efforts, and it may not even be possible or practical, our models will be able to significantly accelerate the manual chart review process in identifying surgical adverse events. Researchers in our group have been starting to work on the application prototype. For each patient the models can compute the probability of having a particular surgical adverse event. By setting two probability cut-offs, patients could be divided into three layers: patients definitely presenting with the complication will fall into the top layer

(having high predicted probability of the adverse event in question); patients who definitely do not have the complication fall into the bottom layer (having very low predicted probability); and patients for whom the evidence in either direction was insufficient for a confident classification fall into the middle layer. Given that our models have very high accuracy in the top and bottom layers, chart reviewers don't have to review these patients and they just need to focus on screening patients in the middle layer.

In this thesis, we explored several technical challenges, missing data, including high-dimensional dataset, and imbalanced distribution. And addition direction that would be worth exploring is using more sophisticated approaches (e.g., longitudinal analysis, time series analysis) to summarize the repeated lab results and vital signs. Some specific pattern related with disease might be recognized and help improve the performance of detection model.

The work described in this thesis shows excellent potential and we see no reason why the same concepts could not be applied at other locations, but we acknowledge that this is a single-site study, which puts limitations on the generalizability of the findings. Further validation using data from other sites is necessary for wider dissertation. It should have less trouble to apply the application developed in this study to hospitals using the same EHR system (Epic locally used), however, for hospitals using different EHR, some mapping or transforming might be needed.

In conclusion, we lay the technical and conceptual foundation of a system that help detect postoperative adverse events from structured EHR data and unstructured clinical

notes, which can be used to assist surgical clinical reviewers to extract and document the surgical occurrences. Our automated system can accelerate the creation of registry date lowering the barrier of entry for providers without sacrificing the high quality. In addition, this cost-effective and efficient method can be applied to high sample sizes that enable to benefit more surgical patients in term of personalized medicine and other relevant applications.

BIBLIOGRAPHY

1. Brennan T.A., Leape L.L., Laird N.M., et al. *Incidence of adverse events and negligence in hospitalized patients.* Results of the Harvard Medical Practice Study I. N Engl J Med 1991;324:370–6.

2. Vries E N de, ramrattan M.A., Smorenburg S.M., gouma, D.J., Boermeester M.A., *The incidence and nature of in-hospital adverse events: a systematic review.* Quality and Safety in Healthcare. 2008. 17(3): 216-23.

3. Morris J.A. Jr, Carrillo Y., Jenkins J.M., Smith P.W., Bledsoe S., Pichert J., and White A., Surgical adverse events, risk management, and malpractice outcome: morbidity and mortality review is not enough. Annals of Surgery.2003. 237(6): 844-852.

4. Levinson D., Adverse event in hospitals: national incidence among Medicare beneficiaries. Washington, DC: Office of the Inspector General, Department of Health and Human Services, 2010.

5. Griffin F. A., Classen D. C., Detection of adverse events in surgical patients using the Trigger Tool approach. Quality and Safety in Health Care. 2008 Aug. 17(4):253-8.

6. Gawande A.A., Thomas E.J., Zinner M.J., and Brennan T.A., The incidence and nature of surgical adverse events in Colorado and Utah in 1992. Surgery. 1999. 126 (1):66-75.

7. Anderson O, Davis R, Hanna G.B., Vincent C.A., Surgical adverse events: a systematic review. The American Journal of Surgery. 2013. 206(2):253-62.

8. Zegers M., Bruijne M., Keizer B, Merten H., et al. The incidence, root-cause, and outcomes of adverse events in surgical units: implication for potential prevention strategies. Patient Safety in Surgery. 2011. P. 5-13.

9. Bates D.W., Evans R.S., Stetson P.D., Pizziferri L., and Hripcsak G., Detecting adverse events using information technology. Journal of American Medical Informatics Association. 2003. 10(2):115-118.

10. Murff H.J., Patel V.L., Hripcsak G., Bates D.W., Detecting adverse events for patient safety research: a review of current methodologies. Journal of Biomedical Informatics. 2003 Feb-Apr;36(1-2):131-43.

11. Tinoco A., Evans R.S., Stadverse events C.J., Lioyd J.F., Rothschild J.M., Haug P.J., Comparison of computerized survaiillance and manual chart review for adverse events. Journal of the American Medical Informatics Association. 2011. 18(4): 491-7.

12. Ana Isabel Pérez Zapata, María Gutiérrez Samaniego, Elías Rodríguez Cuéllar, Eva María Andrés Esteban, Agustín Gómez de la Cámara, Pedro Ruiz López, Detection of Adverse Events in General Surgery Using the "Trigger Tool" Methodology, Cirugía Española (English Edition), Volume 93, Issue 2, 2015, Pages 84-90.

13. Mull HJ, Borzecki AM, Loveland S, Hickson K, Chen Q, MacDonald S, Shin MH, Cevasco M, Itani KM, Rosen AK. Detecting adverse events in surgery: comparing events detected by the Veterans Health Administration Surgical Quality Improvement Program and the Patient Safety Indicators. Am J Surg. 2014.

14. Wu J, Roy J, Stewart WF. Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches. Med Care. 2010 Jun;48(6 Suppl):S106-13.

15. Melton GB, Hripcsak G. (2005). Automated detection of adverse events using natural language processing of discharge summaries. J Am Med Inform Assoc. 12(4), p.448-57.

16. Pakhomov S., Weston S., Jacobsen S., Chute C., Meverden R., Roger V. (2007). Electronic Medical Records for Clinical Research: Application to the Identification of Heart Failure. American Journal of Managed Care. 13, p.281-288.

17. Institute of Medicine. Leadership by Example: Coordinating Government Roles in Improving Health Care Quality. National Academies Press. 2002.

18. Mrdutt MM, Isbell CL, Regner JL, Hodges BR, Munoz-Maldonado Y, Thomas JS, Papaconstantinou HT. NSQIP-Based Quality Improvement Curriculum for Surgical Residents. J Am Coll Surg. 2017 May;224(5):868-874.

19. Zhang JX, Song D, Bedford J, Bucevska M, Courtemanche DJ, Arneja JS. What Is the Best Way to Measure Surgical Quality? Comparing the American College of Surgeons National Surgical Quality Improvement Program versus Traditional Morbidity and Mortality Conferences. Plast Reconstr Surg. 2016 Apr;137(4):1242-50.

20. Cohen ME, Ko CY, Bilimoria KY, et al. Optimizing ACS NSQIP modeling for evaluation of surgical quality and risk: patient risk adjustment, procedure mix adjustment, shrinkage adjustment, and surgical focus. J Am Coll Surg 2013; 217(2):336-46.

21. Bilimoria KY, Liu Y, Paruch JL, Zhou L, Kmiecik TE, Ko CY, et al. Development and evaluation of the universal ACS NSQIP surgical risk calculator: a decision aid and informed consent tool for patients and surgeons. J Am Coll Surg 2013; 217(5):833-42.

22. Kuy S, Romero RAL. Decreasing 30-day surgical mortality in a VA Medical Center utilizing the ACS NSQIP Surgical Risk Calculator. J Surg Res. 2017 Jul;215:28-33.

23. Jiang HY, Kohtakangas EL, Asai K, Shum JB. Predictive Power of the NSQIP Risk Calculator for Early Post-Operative Outcomes After Whipple: Experience from a Regional Center in Northern Ontario. J Gastrointest Cancer. 2017 May 2.

24. Khan NA, Quan H, Bugar JM, Lemaire JB, Brant R, Ghali WA. Association of Postoperative Complications with Hospital Costs and Length of Stay in a Tertiary Care Center, J Gen Intern Med. 2006 Feb; 21(2): 177–180.

25. Lawson EH1, Hall BL, Louie R, Ettner SL, Zingmond DS, Han L, Rapp M, Ko CY. Association between occurrence of a postoperative complication and readmission: implications for quality improvement and cost savings. Ann Surg. 2013 Jul;258(1):10-13.

26. Englesbe MJ, Dimick JB, Sonnenday CJ, Share DA, Campbell DA Jr. The Michigan Surgical Quality Collaborative: will a statewide quality improvement initiative pay for itself? Ann Surg 2007; 246(6):1100-3.

27. Horan TC, Andrus M, Dudeck MA. CDC/NHSN surveillance definition of health care-associated infection and criteria for specific types of infections in the acute care setting. Am J Infect Control 2008; 36(5):309-32.

28. Dimick JB, Pronovost PJ, Cowan JA Jr., Lipsett PA, Stanley JC, Upchurch JR Jr. Variation in postoperative complication rates after high-risk surgery in the United States. Surgery, Volume 134, Issue 4, October 2003.

29. Biscione FM. Rates of surgical site infection as a performance measure: Are we ready? World Journal of Gastrointestinal Surgery. 2009;1(1):11-15.

30. Tinoco A, Evans RS, Staes CJ, Lloyd JF, Rothschild JM, Haug PJ. Comparison of computerized surveillance and manual chart review for adverse events.Journal of the American Medical Informatics Association : JAMIA. 2011;18(4):491-497.

31. Wu ST, Sohn S, Ravikumar KE, et al. Automated chart review for asthma cohort identification using natural language processing: an exploratory study.Annals of allergy, asthma & immunology : official publication of the American College of Allergy, Asthma, & Immunology. 2013;111(5):10.

32. Nguyen SQ, Mwakalindile E, Booth JS, et al. Automated electronic medical record sepsis detection in the emergency department. Eisen J, ed. PeerJ. 2014;2:e343.

33. Nachimuthu SK, Haug PJ. Early Detection of Sepsis in the Emergency Department using Dynamic Bayesian Networks. AMIA Annual Symposium Proceedings. 2012;2012:653-662.

34. Hu Z, Simon GJ, Arsoniadis EG, Wang Y, Kwaan MR, Melton GB. Automated Detection of Postoperative Surgical Site Infections Using Supervised Methods with Electronic Health Record Data. Stud Health Technol Inform. 2015;216:706-10. (Medinfo 2015)

35. Nekkab N, Astagneau P, Temime L, Crépey P. Spread of hospital-acquired infections: A comparison of healthcare networks. PLoS Comput Biol. 2017 Aug 24;13(8):e1005666.

36. Sijia L, Liwei W, Ihrke D, Chaudhary V, Tao C, Weng C, Liu H. Correlating Lab Test Results in Clinical Notes with Structured Lab Data: A Case Study in HbA1c and Glucose. AMIA Jt Summits Transl Sci Proc. 2017 Jul 26;2017:221-228.

37. Esteban S, Rodríguez Tablado M, Ricci RI, Terrasa S, Kopitowski K. A rule-based electronic phenotyping algorithm for detecting clinically relevant cardiovascular disease cases. BMC Res Notes. 2017 Jul 14;10(1):281.

38. Liu Z, Tang B, Wang X, Chen Q. De-identification of clinical notes via recurrent neural network and conditional random field. J Biomed Inform. 2017 Jun 1. pii: S1532-0464(17)30122-3.

39. http://athena.ahc.umn.edu/

40. Rutala WA, Weber DJ. Cleaning, disinfection, and sterilization in healthcare facilities: What clinics need to know. Clin Infect Dis. 2004 Sep 1; 39(5):702-9.

41. http://www.cdc.gov/nhsn/acute-care-hospital/ssi/

42.http://site.acsnsqip.org/wp-content/uploads/2012/03/ACS-NSQIP-Participant-User-Data-File-UserGuide_06.pdf.

43. Mu Y, Edwards JR, Horan TC, Berrios-Torres SI, Fridkin SK. Improving risk-adjusted measures of surgical site infection for the national healthcare safety network. Infect Control Hosp Epidemiol. 2011 Oct; 32(10):970-86.

44. Levine PJ, Elman MR, Kullar R, Townes JM, Bearden DT, Vilches-Tran R, McClellan I, McGregor JC. Use of electronic health record data to identify skin and soft tissue infections in primary care settings: a validation study. BMC Infectious Disease. 2013 Apr 10; 13:171.

45. Warren DK, Nickel KB, Wallace AE, Mines D, Frasesr VJ, Olsen MA. Can additional information be obtained from claims data to support surgical site infection diagnosis codes? Infect Control Hosp Epidemiol. 2014 Oct; 35 Suppl 3: S124-32.

46. Price CS, Satitz LA. Improving the measurement of surgical site infection risk stratification/outcome detection. Final report (prepared by Denve health and its partners under contract No. 290-2006-00-20). AHRQ publication No. 12-0046-EF. Rockville, MD: Agency for Healthcare Research and Quality. March 2012.

47. Keurentjes JHM, Briët JM, de Bock GH, Mourits MJE. Surgical volume and conversion rate in laparoscopic hysterectomy: does volume matter? A multicenter retrospective cohort study. Surg Endosc. 2017 Aug 25.

48. Kaiho Y, Masuda H, Takei M, Hirayama T, Mitsui T, Yokoyama M, Kitta T, Kawamorita N, Nakagawa H, Iwamura M, Arai Y. Surgical and patient-reported outcomes of artificial urinary sphincter implantation: A multicenter, prospective, observational study. J Urol. 2017 Aug 17.

49. Borel F, Ouaissi M, Merdrignac A, Venara A, De Franco V, Sulpice L, Hamy A, Regenet N. Pancreatico-jejunostomy decreases post-operative pancreatic fistula incidence and severity after central pancreatectomy. ANZ J Surg. 2017 Aug 15.

50. Mukka S, Knutsson B, Majeed A, Sayed-Noor AS. Reduced revision rate and maintained function after hip arthroplasty for femoral neck fractures after transition from posterolateral to direct lateral approach. Acta Orthop. 2017 Aug 10:1-7.

51. Chan KS, Fowles JB, Weiner JP. Electronic health records and reliability and validity of quality measures: a review of the literature. Med Care Res Rev 2010; 67(5):503-27.

52. Kharrazi H, Wang C, Scharfstein D. Prospective EHR-Based Clinical Trials: The Challenge of Missing Data. J Gen Intern Med 2014; 29(7):976-8.

53. Little, R.J.A. & Rubin, D.B. Statistical Analysis with Missing Data, John Wiley & Sons (1987).

54. Wells BJ, Chagin KM, Nowacki AS, Kattan MW. Strategies for handling missing data in electronic health record derived data. EGEMS (Wash DC) 2013; 1(3):1035.

55. SAS/STAT software.

Available: https://www.sas.com/content/dam/SAS/en_us/doc/factsheet/stat-101372.pdf

56. SPSS missing values.

Available: http://www-03.ibm.com/software/products/en/spss-missing-values

57. Pigott TD. A review of the methods for missing data. Educ Res Eval 2001;7(4):353-383. http://dx.doi.org/10.1076/edre.7.4.353.8937

58. He Y. Missing data analysis using multiple imputation: Getting to the heart of the matter. Circ Cardiovasc Qual Outcomes 2010; 3(1):98-105.

59. Krysiak-Baltyn K, Nordahl Petersen T, Audouze K, et al. Compass: a hybrid method for clinical and biobank data mining. J Biomed Inform. 2014; 47:160-70.

60. Carpenter JR, Kenward MG. Missing Data in Randomised Controlled Trials: A Practical Guide.

Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.468.9391&rep=rep1&type=pdf

61. Kang H. The prevention and handling of the missing data. Korean J Anesthesiol 2013; 64(5): 402–406.

62. Romero V, Salmerón A. Multivariate imputation of qualitative missing data using Bayesian networks. Soft Methodology and Random Information Systems, Springer (2004), pp. 605–612.

63. Wesonga R. On multivariate imputation and forecasting of decadal wind speed missing data. Springerplus 2015; 4:12.

64. Jerez JM, Molina I, García-Laencina PJ, et al. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. Artif Intell Med 2010; 105-15.

65. Rahman M, and Davis DN. Fuzzy Unordered Rules Induction Algorithm Used as Missing Value Imputation Methods for K-Mean Clustering on Real Cardiovascular Data. Lect Notes Eng Comput Sci 2012; Vol I, London, U.K., 4-6.

Available:http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.362.9952&rep=rep1&type=pdf [Access 25 October 2016]

66. Heckman J. The common structure of statistical models of truncation, sample selection and limited dependent variables and a sample estimator for such models. Ann Econ Soc Meas 1976; 5(4): 475-492.

Available: http://econpapers.repec.org/bookchap/nbrnberch/10491.htm

67. Little RJA. Pattern-mixture models for multivariate incomplete data. J of Am Stat Assoc 1993; 88(421): 125-134.

68. Enders C. Applied Missing Data Analysis. Guilford Press, 2010.

69. Rahman SA, Huang Y, Claassen J, Heintzman N, Kleinberg S. Combining Fourier and lagged k-nearest neighbor imputation for biomedical time series data. J Biomed Inform 2015;58:198-207.

70. Surgical Site Infection (SSI) Event

Available: http://www.cdc.gov/nhsn/PDFs/pscmanual/9pscssicurrent.pdf

71. ASA PHYSICAL STATUS CLASSIFICATION SYSTEM.

https://www.asahq.org/resources/clinical-information/asa-physical-status-classification-system

72. Hu Z, Simon G, Arsoniadis E, Wang Y, Kwaan M, Melton G, "Automated Detection of Postoperative Surgical Site Infections Using Supervised Methods with Electronic Health Record Data," MedInfo 2015: 706-710.

73. Azur MJ, Stuart EA, Frangakis C, Leaf PJ (2011) Multiple imputation by chained equations: what is it and how does it work? Int J Methods Psychiatr Res 2011; 20(1): 40–49.

74. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recogn 1997; 30(7):1145–1159.

75. Kunkel D, Kaizar EE. A comparison of existing methods for multiple imputation in individual participant data meta-analysis. Stat Med. 2017 Jul 10.

76. Friedman J, Hastie T, Tibshirani R. The elements of statistical learning. New York: Springer series in statistics; 2001.

77. McEwan R, Melton GB, Knoll BC, Wang Y, Hultman G, Dale JL, Meyer T, Pakhomov SV. NLP-PIER: A Scalable Natural Language Processing, Indexing, and Searching Architecture for Clinical Notes. AMIA Jt Summits Transl Sci Proc. 2016 Jul 20;2016:150-9.