Surgical Site Infection (SSI) Identification Across Multiple Facilities and Surgery Types Using Multimodal Data and Deep Learning

Arjun Chakraborty

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2024

Reading Committee:

Meliha Yetisgen, Chair

Peter Tarczy-Hornoch

Dustin Long

Program Authorized to Offer Degree:

Biomedical Informatics and Medical Education

University of Washington

Abstract

Surgical Site Infection (SSI) Identification Across Multiple Facilities and Surgery Types Using Multimodal Data and Deep Learning

Arjun Chakraborty

Chair of the Supervisory Committee:

Meliha Yetisgen

Department of Biomedical Informatics and Medical Education

Surgical site infections (SSI), infections at the surgical site that occur after surgery, impact more than a hundred thousand patients a year in the United States. They increase the risk of death after surgery, lead to complications like cellulitis and sepsis, and incur significant healthcare costs. Surveillance of SSI can guide interventions to reduce SSI rates. The current mainstay of SSI surveillance is manual chart review, which is expensive and time consuming. Automated surveillance systems addressing these drawbacks typically rely on a limited number of data modalities from the electronic health record (EHR). They predominantly use rule-based approaches or conventional machine learning algorithms to retrospectively predict whether a surgical case resulted in an SSI. This limits the performance and domain adaptation capability of published gold standard automated surveillance systems.

In contrast to previous state-of-the-art automated surveillance approaches, we employed a data-driven deep learning framework that integrated structured data, clinical text data, and temporal

information from the EHRs of surgical cases to develop an automated surveillance system. Our primary findings demonstrated several key points: a purely data-driven deep learning approach using multimodal data outperformed previously published gold standard rule-based and conventional machine learning approaches for the task of surgical site infection (SSI) prediction; the data representation and modeling strategies we utilized enabled the construction of models capable of domain adaptation across a diverse set of domains; and large language models (LLMs), specifically generalist foundation models such as Llama 3, offered previously unrealized performance gains.

Table Of Contents

1. Introduction 1

1.1 Contributions 3
1.2 Overview 4

2. Background and Significance 6

2.1 Approach of Our Research 8
2.2 Summary 9

3. Aim 1: Develop and assess an automated surgical site infection surveillance approach using deep learning and multimodal data 10

3.1 Introduction 10
3.2 Methods 10
3.3 Results 18
3.4 Discussion 28
3.5 Conclusion 30

4. Aim 2: Assess the impact of temporal data on the performance of deep learning based surgical site infection prediction 31

4.1 Introduction 31
4.2 Methods 32
4.3 Results 36
4.4 Discussion 40
4.5 Conclusion 41

5. Aim 3: Assess the domain adaptation capability of surgical site infection surveillance approach 43

5.1 Introduction 43
5.2 Methods 43
5.3 Results 51
5.4 Discussion 64
5.5 Conclusion 65

6. Aim 4: Measuring Performance of Large Language Models to Predict SSI Occurrence 67

6.1 Introduction 67
6.2 Methods 67
6.3 Results 73
6.4 Discussion 76
6.5 Conclusion 76

7. Conclusion 77

7.1 Main Findings of our Research 77
7.2 Limitations and Future Directions 79

List Of Tables

Table 3.1: Surgery counts and infection rate 11
Table 3.2: Structured data elements and their descriptions 12
Table 3.3: Statistics for top 5 most frequent note types in the dataset 13
Table 3.4: Performance conferred by different feature sets, modeling strategies, and text representations on the task of SSI prediction 20
Table 3.5: Performance conferred by various strategies to reduce numbers of false negatives compared to previous approaches presented in Table 3.4 24
Table 3.6: Notes in our dataset with the highest number in which each category of concept is mentioned at least once 24
Table 3.7: Clinical significance of the various data representation and modeling approaches we used in this chapter 27
Table 4.1: Structured data elements and their descriptions 32
Table 4.2: Performance conferred by different feature sets, modeling strategies, and text representations on the task of SSI prediction 36
Table 4.3: Clinical significance of the various data representation and modeling approaches 40
Table 5.1: Surgery counts and infection rate 43
Table 5.2: Approach to categorize surgery types into In-Domain and Out-domain 49
Table 5.3: Outline of our experimentation for Aim 3 51
Table 5.4: Performance conferred by different feature sets, modeling strategies, and text representations on the task of SSI prediction 52
Table 5.5: Performance of neural network models trained on each surgery type versus the entire training set on each surgery type 53
Table 5.6: Performance of neural network models trained on each surgery type versus the entire training set on each surgery type 55
Table 5.7: Incremental gains of adding more out-domain samples on model performance on General Surgery 57
Table 5.8: Incremental gains of adding more out-domain samples on model performance on Spine Surgery 57
Table 5.9: Performance conferred by different training strategies on different surgery types in our dataset 58
Table 5.10: Incremental gains of adding more out-domain samples on model performance on surgery types with shorter/longer timeframe of SSI occurrence 60
Table 5.11: Incremental gains of adding more out-domain samples on model performance on surgery types in which deep or organ-space SSI may represent a higher or lower percentage of total SSIs 61
Table 5.12: Summary of our findings on different groups and training strategies 63
Table 5.13: Summary of our findings on different groups and training strategies 64
Table 6.1: Surgery counts and infection rate 68
Table 6.2: Descriptions of our 4 prompt templates 71
Table 6.3: Figure depicting neural architectures used in our SSI classification approach using ClinicalBERT 72

Table 6.4: Clinical significance of the various data representation and modeling approaches 75

List Of Figures

Figure 3.1: Figure depicting our different text representation approaches 14
Figure 3.2: Figure depicting neural architectures used in our SSI classification approach 17
Figure 3.3: Examples of clinical text excerpts which drove model predictions towards a positive or negative classification 22
Figure 3.4: Shapley text plots after removal of clinical text data from the day of surgery or before 26
Figure 4.1: Method for constructing our temporal representation 33
Figure 4.2: Figure depicting neural architectures used in our SSI classification approach 35
Figure 4.3: Model explanation analyses of true and false positive and negative cases 38
Figure 5.1: Figure depicting neural architectures used in our SSI classification approach 45
Figure 5.2: Figure depicting the focus of our domain adaptation experiments 46
Figure 5.3: Figure depicting our overall experimental approach 47
Figure 6.1: Prompt we used to summarize clinical notes with Llama 3 69
Figure 6.2: Example note summary 69

List Of Abbreviations

ANC: Absolute neutrophil count

Bi-LSTM: Bidirectional long short-term memory network

CNN: Convolutional neural network

CoT: Chain-of-Thought

EHR: Electronic health record

HMC: Harborview Medical Center

LIME: Local Interpretable Model-agnostic Explanations

LLM: Large language model

LSTM: Long short-term memory network

NHSN: National Healthcare Safety Network

NLP: Natural language processing

NN: Feedforward neural network

NSQIP: National Surgical Quality Improvement Program

RF: Random forest

SOTA: State of the art

SSI: Surgical site infection

UMLS: Unified medical language system

UTI: Urinary tract infection

UWMC: University of Washington Montlake Campus

WBC count: White blood cell count

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my primary advisor, Dr. Meliha Yetisgen, for their unwavering guidance, mentorship, and encouragement throughout my PhD journey. Your expertise and insight have been instrumental in shaping both my research and my growth as a scientist. Thank you for challenging me to think critically, aim higher, and persevere in the face of challenges.

I would also like to extend my heartfelt thanks to my co-advisor, Dr. Peter Tarczy-Hornoch, whose support and leadership as the Department Chair have provided invaluable guidance. Your thoughtful advice and strategic perspective have been crucial in helping me navigate the complexities of my research and academic career. I am grateful for your belief in my potential and for your efforts to create an enriching academic environment.

A special note of appreciation goes to my third advisor, Dr. Dustin Long, whose medical expertise and innovative vision laid the foundation for this research. Thank you for conceptualizing the idea that inspired this work and for offering critical insights that bridged the gap between computational methodologies and clinical relevance. Your passion for advancing medical science has been a constant source of motivation.

I would also like to acknowledgement Dr. Gina-Anne Levow, from the Linguistics department, who served as my graduate school representative. Thank you for reviewing my dissertation and providing valuable feedback on NLP topics. Your expertise and input have significantly improved the quality of this work and ensured its rigor within the NLP domain.

I would like to thank Dr. Tony Olivas Estebane, Dr. Mahul Patel, and Dr. Judy Chen, my clinical collaborators, whose knowledge and contributions brought essential clinical context to this research. Your input has been invaluable in ensuring the practical applicability of this work.

I would like to thank Dr. Vikas O'Reilly Shah who reviewed my work and provided invaluable feedback. Your expertise and thoughtful input have been instrumental in refining this work and its broader impact.

I would also like to acknowledge the support of my collaborators, peers, and the broader academic community, whose feedback and encouragement have enriched this work. To my friends and family, thank you for your patience, understanding, and constant support throughout this challenging but rewarding journey.

Lastly, I would also like to acknowledge the generous funding provided by the National Library of Medicine, which made this research possible. Your support has been instrumental in enabling me to pursue this work at the intersection of bioinformatics, natural language processing, and deep learning.

To everyone who has contributed to my journey, thank you. This dissertation is as much a testament to your support as it is to my efforts.

Dedication

To my grandparents, Pratima Biswas, and Sisir Kumar Biswas.

1. Introduction

A surgical site infection (SSI) is defined as an infection that develops after a surgical procedure at the site of surgery[1]. SSIs impact more than a hundred thousand patients a year in the United States[2]. They increase the risk of death after surgery, lead to complications like cellulitis and sepsis, and incur significant healthcare costs. SSI increase the risk of death after surgery by 2 to 11 fold[3]. Moreover, they can cause complications like cellulitis or sepsis which can delay postsurgical recovery and even have lifelong impacts on the health of the patient. Previous studies have found that SSI lowers patients' health utility values (preference to their overall health status) by 10% to 30% over the course of their lifetime[4]. SSIs also extend postoperative length of stay by 9.7 days and incur costs of twenty thousand dollars per case on average[5]. The annual cost of managing SSI is $3.5-$10 billion[6].

Surveillance of SSI can guide interventions to reduce SSI rates, their negative health effects, and their costs. Surveillance data can be used to determine baseline infection rates and compare quality across institutions[7]. Thus, it can help healthcare institutions identify areas which need improvement. It can also identify trends over time and detect clusters of infection. In response to surveillance data indicating rising SSI rates, hospitals have taken steps which include: (1) implementing changes in the modus operandi of different departments (e.g. anesthesiology, perioperative nursing), (2) ensuring staff follow proper procedures for preoperative bathing, (3) intraoperative antisepsis of the patient's skin, and (4) monitoring compliance to antibiotic prophylaxis standards. Prior work has shown that actions taken based on surveillance data can reduce mortality associated with SSI by 27% and morbidity associated with SSI by 45%[8]. It has also shown that such actions can reduce costs associated with SSI.

Currently, most healthcare institutions conduct SSI surveillance through a set of manual methods which involve reviewing various elements of the electronic health record (EHR) to determine if a surgical case led to SSI. These methods are time consuming and costly[9]. Due to the low prevalence of SSI, infection prevention personnel have to review 33 charts on average for each SSI detected, highlighting the inefficiency of the process[10]. This means infection prevention personnel review 1,600 charts per year, each of which takes 20 minutes on average. Thus, infection prevention personnel typically spend 1 day out of every two weeks of the year performing chart reviews. Due to the high cost incurred by the process, coverage of procedures is low. Typically, only 1% of procedures are surveilled leading to doubts on the robustness of conclusions that can be drawn based on the resulting surveillance data[11]. Finally, manual surveillance suffers from a lack of objectivity, as standards for determining a case as SSI can vary at a facility or institution level[12].

Automated surveillance can address these drawbacks of manual surveillance. This form of surveillance involves using computational tools to automatically label surgical cases based on whether or not they were complicated by SSI. It is typically operationalized using a rule-based, statistical, or machine learning model that outputs a probability that a surgical case developed into SSI. Automated surveillance systems are typically used as a screening step prior to manual chart review to rule out clearly SSI negative surgical cases[10]. This can reduce the number of charts that need to be reviewed. Thus, it can reduce the time and cost associated with manual chart review. In

this way, it can enable surveillance of a higher proportion of total surgical procedures. It can also enable a consistent definition of SSI to be applied across facilities and institutions, and thus improve the objectiveness of review.

Despite the potential of automated surveillance systems to confer several benefits to the SSI surveillance process, current state of the art (SOTA) automated surveillance systems have not yet achieved the level of precision or recall required to sufficiently reduce the burden of manual chart review. Contemporary SOTA systems achieve a precision of 0.24-0.28 at a recall of 0.9[10,13]. This entails the review of 4 patient charts to detect every SSI case (as opposed to 33 required when manual chart review is done alone without automated surveillance). As the annual incidence of SSI is more than a hundred thousand, this means review of hundreds of thousands of patient charts, each taking 20 minutes[2,14]. Thus, although current automated surveillance systems allow approximately 8 (=33/4) times more cases to be detected in the current amount of time dedicated to chart review, they fail to sufficiently reduce the burden of infection prevention staff conducting manual chart review. In addition, the performance and clinical utility of these systems on data from different domains (e.g., different surgery types, different institutions) which are not well represented in the training set thus have seldom been evaluated. This is important as a goal of the work related to automated post hoc SSI prediction is the establishment of a nationwide SSI surveillance system[15]. Post hoc benefits refer to monitoring incidence of SSI after the fact as distinct from predicting prior to an SSI occurrence which patients are at increased risk for SSI and thus might warrant special attention. Data on surgical cases from around the nation will likely include domains which may have insufficient labeled data in the training set. Thus, an essential part of the development of automated SSI surveillance systems is a robust evaluation of their domain adaptation capability.

Contemporary automated SSI surveillance systems use at most 2 data modalities (e.g., structured data, clinical text data, or temporal data) the EHR and rule-based algorithms or conventional machine learning algorithms to predict each surgical case post hoc as one in which SSI did not develop or did develop[10,15–17]. However, as each data modality adds unique signals pertaining to SSI, using only 2 data modalities limits the coverage of signals pertaining to SSI that is derived from the EHR. In addition, rule-based algorithms or conventional machine learning models are not as capable as deep learning models at delineating complex relationships between features and outcome variables[18]. Nor are they as adept as deep learning models at adapting to domains not well represented in the training set. This limits the performance and domain adaptation capability of contemporary SOTA surveillance systems. Furthermore, none of the approaches, which we reviewed, utilized the recently developed foundational large language models (LLMs). These models have demonstrated near gold-standard performance without requiring costly fine-tuning strategies. Evaluating these models for post hoc SSI prediction is essential for advancing the development of highly effective and clinically valuable automated SSI surveillance methods.

In contrast to past SOTA automated surveillance approaches, we plan to use a data-driven deep learning framework using 3 modalities of data (structured data, clinical text data, and temporal data) from the EHR of surgical cases to build an automated surveillance system. We think the resulting automated surveillance system will confer better performance and have better domain adaptation capability than contemporary SOTA automated surveillance systems. In terms of performance, we aim for achieving a precision higher than that achieved by contemporary

automated surveillance systems (0.24-0.28) at a recall of 0.9. This would confer a reduction in the burden of infection preventionists conducting manual chart review greater than is possible with current automated surveillance approaches. This 11-fold decrease in number of charts to review per SSI would mean 11x more SSI detected with same effort. This would mean 11% of all surgical procedures could be surveilled, as compared to 1% with contemporary manual surveillance methods, while incurring no increase in costs and time taken for surveillance. It would enable interventions targeting SSI rates to be taken based on surveillance data on a larger, more representative sample of procedures. It would also entail a further reduction of 720,000 hours across the US in the time infection prevention staff have to spend doing manual chart review each year[18].

$$720{,}000 = 15000 * 1600 * (4 - 3)/33$$

hours of reduction = (# of infection preventionists in US) * (# of hours currently spent on doing manual chart review) * (# of charts reviewed with current automated surveillance – # of charts reviewed with our system) / (# of charts that need to be reviewed with manual only chart review)

This is on top of reductions conferred by contemporary automated surveillance methods. Thus, it would allow for faster availability of surveillance data and free up time of infection prevention staff to focus on other activities. We also think the better domain adaptation capability denoted by our deep learning approaches will allow our models to be more performant than contemporary automated surveillance approaches even on external data. Thus, they will confer better clinical utility (reduction in burden of infection prevention staff, increase in coverage of procedures surveilled) than current automated surveillance approaches even when implemented at facilities or institutions not represented in the training set. This will allow for the implementation of our system and the realization of its benefits even in facilities with data from different domains than ones in our datasets.

1.1 Contributions

We make the following contributions with our research:

- We develop and assess a framework for how state of the art (SOTA) natural language processing (NLP) approaches in combination with deep learning can be used in a data-driven way to combine structured and clinical text data to build highly performant post hoc SSI prediction models. By SOTA NLP methods, we mean methods that are on the cutting edge of research on NLP and whose ability to grant superior performance to traditional NLP methods has been established through a long body of work in the literature (Aim 1).
- We demonstrate and evaluate a proof of concept on how temporal data can be incorporated into deep learning models to improve the performance and domain adaptation capability of

such models (Aim 2).

- We perform a thorough domain adaptation capability analysis of the performance of deep learning models on surgery types on which they are not trained (Aim 3).
- We performed the first evaluation to our knowledge on the performance of several approaches using foundational LLMs (e.g., Llama 3, GPT-4) for this task (Aim 4).
- Overall, the above contributions permitted us to develop a model that is more performant than previous methods on the task of post hoc SSI prediction, allowing for a substantial reduction in the workload of infection prevention staff. This will also allow us to get closer to the goal of a nationwide SSI surveillance system. We plan to outperform SOTA performance reported on this task (0.24 - 0.28 precision at 0.9 recall) and achieve a recall of at least 0.9 and a precision of at least 0.3 using our deep learning models. By SOTA performance, we mean higher performance than has been reported in any literature on automated post hoc SSI prediction until the time of publication of the work. This implies that there will be a need of an average of only 3 chart reviews to identify an SSI case (as opposed to 4 chart reviews entailed by the current SOTA performance of automated post hoc SSI prediction algorithms) while still detecting a sufficient proportion (at least 90%) of all SSIs. Given that the annual incidence of SSI is around 110,000, and each chart review could take 20 minutes on average, this will result in a reduction of a total of thirty-six thousand hours annually in the time infection prevention staff have to spend reviewing patient charts for SSI surveillance and lead to a tenfold increase in the number of procedures that can be surveilled compared to manual surveillance methods.

1.2 Overview

The following is an overview of Chapters 2-6 of this report.

Chapter 2 Background and Significance: In Chapter 2, we introduce the problem of SSI, delineate the need for SSI surveillance, and explain the value of automated SSI surveillance. We address the current state of the art of manual and automated SSI. We introduce our novel approach to building an automated SSI surveillance system that confers higher performance and has better domain adaptation capability than those developed in the past. The number of charts needed to review per SSI detected manually is 33 and with best automated methods is 4.

Chapter 3 Aim 1: In Chapter 3, we outline our cutting-edge NLP methods to represent clinical text. We combine clinical text data with structured data into our SSI prediction deep learning models. We also share our results, and our interpretations of them.

Chapter 4 Aim 2: In Chapter 4, we outline our novel temporal data representation strategies for this task. We also combine clinical text, static structured, and temporal data for SSI prediction. As in Chapter 3, we present our results and delineate our interpretations. The best method was using a bidirectional long short-term memory network (BiLSTM) layer to process clinical text, using a long short-term memory network (LSTM) layer to process temporal data with performance being 3 charts to review per SSI detected.

Chapter 5 Aim 3: In Chapter 5, we outline experiments in which we evaluate our models in terms of their domain adaptation capability, a critical aspect of any SSI prediction model. As in Chapter 3 and 4, we present results and outline interpretations.

Chapter 6 Aim 4: In Chapter 6, we outline experiments leveraging LLMs. We try a range of approaches using LLMs for our task (post hoc SSI prediction). As in Chapter 3, 4, and 5 we present results and outline interpretations. The best method was using to generate clinical text summaries and using our best deep learning architecture from Aim 2 (Chapter 4) with performance being 2.6 charts to review per SSI detected.

Chapter 7 Conclusion: In the final Chapter, we synthesize our findings from Chapters 3, 4, and 5 and discuss their implications in terms of building useful and effective automated SSI surveillance systems. We list the limitations of our research and discuss our ideas of the most promising directions for future work.

2. Background and Significance

SSIs, infections at the surgical site that occur after surgery, impact more than a hundred thousand patients a year in the United States. They increase the risk of death after surgery, lead to complications like cellulitis and sepsis, and incur significant healthcare costs. Surveillance of SSI can guide interventions to reduce SSI rates. The current mainstay of SSI surveillance is manual chart review, which is expensive and time consuming. Automated surveillance can address drawbacks of manual surveillance. Contemporary automated SSI surveillance systems use data from the EHR and rule-based algorithms or conventional machine learning algorithms to predict each surgical case post hoc as one in which SSI did not develop or did develop. However, contemporary automated surveillance systems use a limited set of data modalities from the EHR. This limits the performance and domain adaptation capability of published gold standard surveillance systems. In contrast to past SOTA automated surveillance approaches, we used a data-driven deep learning framework using structured data, clinical text data, and temporal information from the EHR of surgical cases to build an automated surveillance system.

Current State of the Art Automated Surveillance Algorithms Do Not Achieve Sufficient Performance to Adequately Reduce the Burden of Manual Surveillance: Because of the benefits of SSI surveillance and the time and cost incurred by the manual surveillance process, many studies have attempted to develop automated surveillance systems[10,13,15]. These systems are meant to complement manual surveillance and make the surveillance process faster and less costly. However, despite these efforts, automated surveillance algorithms with sufficient recall and precision to replace or complement manual chart review are still lacking. Currently, algorithms for SSI prediction have achieved a maximum precision of 0.24-0.28 at a recall of 0.9 (A recall of 0.9 reflects an acceptable percent of cases (10%) that can be missed by automated surveillance and still lead to an adequately accurate surveillance system)[10,13] . In contrast, most studies report that the recall of manual surveillance methods is 0.60-0.74[19]. Most automated systems are meant to be a screening step, implemented prior to manual chart review to remove cases that have a low risk of representing an SSI positive case. Thus, a precision of 0.28 for such systems means 4 charts need to be reviewed to detect every SSI[10,13]. This compares to manual methods which require review of around 33 patient charts for each SSI detected[10]. Given that the annual incidence of SSIs in the United States is over a hundred thousand, this entails review of hundreds of thousands of charts, each of which could take 20 minutes on average[20,21]. Therefore, contemporary automated surveillance approaches fail to sufficiently reduce the workload of infection prevention staff conducting manual chart review. In summary, monitoring SSIs can guide intervention to reduce rates of SSI in the future, but the current mainstay of manual surveillance is time consuming, expensive, and prone to subjectivity. Automated surveillance has the potential to address these challenges, but thus far, automated surveillance systems which achieve sufficient performance to sufficiently lower the burden of manual chart review are lacking.

A Deep Learning and Multimodal Data–Based Automated Surveillance Approach Can Achieve Better Performance Than Current Automated Surveillance Systems and Further Reduce the Burden on Infection Prevention Personnel: Automated surveillance approaches to detect SSI that leverage EHR data have emerged because of the combination of widespread adoption of the EHR and advances in data science[10,16,17]. Previous efforts to develop automated SSI surveillance systems have leveraged a wide range of approaches and data types with the

objective of classifying whether a surgical case was complicated by SSI based on routinely generated EHR data. Most work to date has examined either structured data alone[14,15] or clinical notes alone[16,17]. However, both structured and clinical text data contain unique, complementary signals relevant to SSI and are both considered by human reviewers. Therefore, excluding either data type limits the performance of SSI prediction algorithms. For a more complete case representation, our approach and a smaller number of prior studies[13,18] integrated both structured data and clinical notes in model development.

Including temporality in data representations contributes critical information to SSI prediction models[14,19]. For instance, SSI is marked by an initial rise in white blood cell (WBC) count[20]. This is followed by a period of stability as the patient fights infection. Finally, there is a period of decay in WBC count as the patient recovers from infection. Previous research has represented temporal data using extensive data summarization techniques[14,19]. In contrast, we did not utilize extensive data summarization techniques. Instead, we used temporal representations with deep learning models to capture intricate patterns in the data pertinent to SSI prediction.

It has been found that a complex interplay of data modalities plays a role in determining whether an EHR record represents an SSI positive or negative case[16,22,23]. For instance, presentation of SSI symptoms after a surgical case may prompt workup for SSI. This would lead to these symptoms being recorded in clinical notes. However, temporal trends in laboratory values (e.g., WBC count) collected during this workup may be stable, indicating an absence of infection. Meanwhile, development of infections other than SSI (e.g., urinary tract infection (UTI)) in the postoperative course may lead to recording of some SSI symptoms of SSI in clinical notes (due to the similar clinical presentation of the two infections). However, as UTI typically develops later in the postoperative course, these two infections can be distinguished by temporal patterns in data like laboratory values. We realize that one limitation of our study is that the occasional patient may have an SSI diagnosis made and/or taken care of in an outside hospital and our dataset would not contain these records. Most past approaches which have used EHR data to develop automated SSI surveillance systems have used rule-based conventional machine learning approaches to predict SSI occurrence[15–17]. Deep learning has been shown to be more adept than rule-based approaches or conventional machine learning at delineating the complex relationships between modalities of data from the EHR and SSI[18]. Deep learning strategies have been shown to outperform rule-based or conventional machine learning strategies on this task when there is enough training data[23,24]. We plan to leverage our dataset of 28,864 cases from two healthcare facilities with multiple data modalities, temporal information, and deep learning methods to achieve better performance than that obtained with rule-based or conventional machine learning-based methods.

A particular type of deep learning model, LLMs have achieved close to published gold standard or surpassed published gold standard performance on a range of clinical NLP tasks, without requiring computationally costly fine-tuning approaches. This makes them an attractive choice for many clinical use cases. However, none of the studies that we reviewed leveraged such models for SSI prediction. Thus, we also perform an evaluation of a variety of approaches involving LLMs for this task to determine the relative gains that LLMs could offer for this task.

A Deep Learning and Multimodal Data–Based Automated Surveillance Approach Can Achieve Good Domain Adaptation Capability and Be Used to Conduct Surveillance In

Facilities Around the Nation: Even the most suitable data sourcing and representation and modeling strategies may yield models which are performant on resource-rich domains (domains with adequate labeled data) but fail to adapt well to resource-poor domains (domains with insufficient labeled data). This can limit the use of these models to resource-rich domains. However, a goal of the work related to automated post hoc SSI prediction is the establishment of a nationwide SSI surveillance system[15]. Data on surgical cases from around the nation will likely include domains which may have insufficient labeled data in the training set. This is why an essential part of the development of automated SSI surveillance systems is a robust evaluation of their domain adaptation capability. We hypothesized that our deep learning approaches will enable our automated surveillance systems to have better domain adaptation capability than brittle rule-based or conventional machine learning-based contemporary automated surveillance approaches. Although an assessment of the domain adaptation capability of models which predict SSI is critical, almost no past studies on post hoc SSI prediction with or without NLP have investigated the domain adaptation capability of machine learning models which predict SSI occurrence[24]. Therefore, a robust and thorough analysis of the domain adaptation capability of post hoc SSI prediction algorithms is lacking in past work. We plan to leverage our dataset, which includes two different healthcare facilities, each contributing more than 10,000 surgical cases, and encompassing 2 facilities and 9 different surgery types to conduct a thorough analysis of the domain adaptation capability of our models.

2.1 Approach of Our Research

In this project we:

- Integrated **multimodal data** including static (e.g. microbiology data) and temporal (e.g. laboratory values) structured data and unstructured clinical notes from the EHR in our patient representation to make our system more performant. Past work has shown that different modalities of data from the EHR contain unique signals pertaining to SSI which interact with each other to determine if a case is SSI positive or negative[10,16]. We leveraged our multimodal dataset to capture signals of SSI in each modality and build more performant post hoc SSI prediction models than those developed in past studies (Aim 1, Chapter 3, Aim 2, Chapter 4).
- Used **deep learning models** to capture complex relationships between different features and SSI occurrence and be more performant on post hoc SSI prediction than conventional machine learning models (Aim 1, Chapter 3, Aim 2, Chapter 4).
- Used **word embeddings in combination with deep learning models** to extract context information from text and address the issues that come with a high dimensional feature space and computational challenges associated with discrete text vectorization approaches (Aim 1, Chapter 3).
- Included **all clinical notes** within a certain time-window around surgery (7 days before to 90 days after) in our text representation to capture signals of SSI in all clinical notes and increase the domain adaptation capability and performance of our post hoc SSI prediction algorithm (Aim 1, Chapter 3).
- Input **temporal lab values and vitals in a raw format**, **without using feature engineering** methods like data summarization, into our models to improve the domain adaptation capability of our models (Aim 2, Chapter 4).

- Used a **dataset that contains 9 different surgery types, and 2 different healthcare facilities**. This allowed us to conduct a thorough evaluation of the domain adaptation capability of our models (Aim 3, Chapter 5).
- Conducted an exhaustive analysis of methods which can be used with LLMs for post hoc SSI prediction (Aim 4, Chapter 6).

2.2 Summary

This chapter outlined the value of SSI surveillance, the drawbacks of the current methods of conducting surveillance, and the potential benefits of automated surveillance. We detailed our approach to building an automated SSI surveillance system and compared it to past approaches. We outlined how it can confer better performance on the task of SSI prediction and have better domain adaptation capability than past approaches. In the next chapter, we describe our implementation of our cutting-edge NLP approaches to incorporate clinical text into our automated SSI surveillance system, share our results, and explain the implications of our findings.

The overall goal of this dissertation is to answer the question "Can an automated surveillance system based on multimodal data and deep learning outperform and have better domain adaptation capability than SOTA automated SSI surveillance approaches?" which we explore through four aims:

Aim 1: Develop and assess an automated surgical site infection surveillance approach using deep learning and multimodal data.

Aim 2: Assess the impact of temporal data on the performance of deep learning based surgical site infection prediction.

Aim 3: Assess the domain adaptation capability of surgical site infection surveillance approach.

Aim 4: Measuring the performance of LLMs to Predict SSI Occurrence.

3. Aim 1: Develop and assess an automated surgical site infection surveillance approach using deep learning and multimodal data

3.1 Introduction

My overall research question is: "*Can an automated surveillance system based on multimodal data and deep learning outperform and have better domain adaptation capability than SOTA automated SSI surveillance approaches?*" In this chapter, the focus is on Aim 1, "Develop and assess an automated surgical site infection surveillance approach using deep learning and multimodal data". We assess the effect of using structured data and clinical text from the EHR and different representations of these data modalities on SSI prediction performance of different modeling approaches. This work builds on existing work summarized in Chapter 2: Background and Significance. The goal of our work is to achieve performance comparable to or better than the reported performances of current SOTA automated SSI surveillance approaches (recall: 0.8-0.95 and precision: 0.24-0.28)[10,13]. To accomplish this, we use NLP techniques to represent clinical text and fuse clinical text and structured data from the EHR within a deep learning framework for SSI prediction.

In this chapter, we develop and assess an automated SSI surveillance approach using deep learning and multimodal data. We use structured and clinical text data and assess whether our approach outperforms published gold standard automated surveillance approaches. We first describe our methods for representing our data (structured, and clinical text) and for training and evaluating our deep learning models. Next, we assess the performance derived by using a range of data modalities, data representation methods, and machine learning models and report the performance using precision, recall, and F1-score. Specifically, we look at the benefit of multimodal data (structured data alone vs adding text), various representations of data (e.g., rule-based, bag of words (BOW), word2vec), and a range of rule-based and machine learning approaches (e.g., random forest (RF), feedforward neural network (NN), convolutional neural network (CNN), BiLSTM). We compare the performance of various combinations of feature sets, data representations, and modeling strategies against manual SSI abstraction to identify the optimal approach. Part of the assessment includes error analysis and explainability. We analyze the implication of the performance metrics in the context of our overall goal. We assess the clinical utility of our best models as compared to the manual approach and current best automated approaches.

3.2 Methods

3.2.1 Clinical Cohort and SSI Definition

SSI surveillance records from two registries were used in this study. One registry was the American College of Surgeons (ACS) National Surgical Quality Improvement Program (NSQIP)

registry at the University of Washington Montlake Campus (UWMC). The other was the Centers for Disease Control and Prevention (CDC) National Healthcare Safety Network (NHSN) registry at Harborview Medical Center (HMC). The registries were examined for the period between 2008 and 2020. Cases beyond July of 2020 were not included due to changes in surveillance and reporting programs related to COVID-19 pandemic response. Both hospitals are part of the University of Washington Medicine health system, but serve unique populations: UWMC is a tertiary care center ACS NSQIP site with a large volume of complex elective general surgery and oncology procedures, while HMC is a level-one trauma center that also maintains high volumes of elective orthopedic and neurosurgery procedures and maintains a robust NHSN surveillance program, capturing a number of procedures beyond those required by state and federal programs.

Data from UWMC NSQIP "target" procedures in general and gynecological surgery categories were included. Data from HMC NSHN records for colorectal surgery, abdominal hysterectomy, spinal fusion, craniotomy, VP shunt placement, and instrumented hip and lower-extremity procedures were included. In both data sets, cases with infection present at the time of surgery (PATOS) were excluded and all classes of SSI (superficial, deep, or organ/space) within the procedure specific 30- or 90-day surveillance period were included. Each procedure was assigned a binary value indicating whether the case was complicated by SSI according to corresponding NHSN or NQSIP criteria. The data sets consist of the subset of cases that were manually chart reviewed and annotated as to presence/absence of SSI and type of SSI from 2008-2020. The approximate annual volume at the two facilities combined is 20,000 surgical cases. Thus, from 2008-2020, approximately 240,000 surgeries were performed at both facilities combined. The registry represents about 10% of all surgical cases during this time. About 4% of the cases in the registry overall have SSI. The total number of patients in the registry at UWMC was 17,403 and at HMC was 12,577.

Counts of included procedure types and SSI event rates in our data set are presented in Table 3.1.

Procedure Class	Case Count	SSI Events (Rate)
Spine	3,724	196 (5.3%)
Orthopedic (non-spine)	2,739	105 (3.8%)
Neurosurgery (non-spine)	4,990	185 (3.7%)
General Surgery	14,200	585 (4.1%)
Gynecologic Surgery	3,211	120 (3.7%)
Cardiothoracic Surgery	328	8 (2%)
Vascular Surgery	788	7 (0.9%)
Total	29,980	1,206(4.0%)

Table 3.1: Surgery counts and infection rate

3.2.2 Structured Data Features

Clinically generated data extracted from the EHR were used to train the automated SSI identification model and included both structured and unstructured variables.

Structured data elements (Table 3.2) comprised (1) surgical procedure characteristics, (2) laboratory values including wound culture orders, (3) the presence of a recorded postoperative

fever, (4) antibiotic administration events, and (5) postoperative consultation with infectious disease teams or interventional radiology for drain placement.

Category	Features	Description
Procedural Characteristics	Surgical Procedure Class	High level classification of surgical specialty/procedure groups (ex:, spine surgery, craniotomy and other neurological procedures, general surgery, gynecologic procedures).
Procedural Characteristics	Reoperation	Binary variable indicating whether the index procedure was followed by a secondary operation within 90 days after surgery.
Laboratory Values	Culture Obtained	Binary variable indicating whether cultures (wound, tissue, or CSF culture) were collected for a patient.
Laboratory Values	White blood cell $(WBC)_{max\ postop}$	Maximum WBC count within 90 days after surgery.
Postoperative Fever	Temperature $(T)_{max\ postop}$	Maximum temperature within 90 days after surgery.
Consultations	ID Consult	Whether a postoperative infectious disease consultation was obtained (binary).
Consultations	IR Drain	Whether a procedure for postoperative drain placement by interventional radiology (IR) occurred (binary).
Antibiotics	Antibiotic Administration	Whether a clinically relevant antimicrobial agent was administered for a patient (binary).

Table 3.2: Structured data elements and their descriptions

We represented structured data as follows:

1. Procedure type was one-hot encoded ("Surgical Procedure Class").
2. Each index case was binary classified (Yes/No mapped to 1/0) regarding being followed by a repeated operation within 90 days ("Reoperation").
3. The series of WBC counts and body temperatures over time were collapsed to a single value (the maximum of the series). $WBC_{max\ postop}$ and $Temperature_{max\ postop}$ were represented as continuous variables, using the maximum value within the first 90 postoperative days. These values were processed in the following fashion: high outliers (> $150k/mm^3$ for $WBC_{max\ postop}$ count and > 42^oC for $Temperature_{max\ postop}$) and missing data points were first mean imputed by replacing missing values with the mean of that feature calculated across all surgeries in the training set. Subsequently, all values were min-max normalized, based on all values from the training set, prior to input into the model.

4. Consultation variables were represented as binary variables (Yes/No mapped to 1/0) indicating whether a postoperative infectious disease consultation ("ID Consult") was obtained or a drain was placed by interventional radiology ("IR Drain").
5. We used binary variables (Yes/No mapped to 1/0) to indicate the presence of an order for a wound, tissue, fluid, or CSF culture for a patient.
6. We used binary variables to indicate the presence of a positive culture and an order of a SSI relevant antibiotic ("Culture Obtained" and "Antibiotic Administration"). Lists of relevant cultures and antibiotics were developed in consultation with clinical content experts. Lists of relevant cultures are presented in the "Description" column of Table 3.2. Lists of relevant antibiotics are presented in Appendix Supplemental Table 1.

3.2.3 Text Data from Clinical Notes

Our clinical text corpus included all clinical text notes for a surgical case in the registry from 7 days before the surgery to 90 days after the surgery. The number of total notes in our dataset was 3,193,094. Table 3.3 lists 5 most frequent note types in our dataset. Cases with SSI had a higher mean number of notes when compared to those without SSI (213.0 versus 97.1, $p=1.5*10^{-222}$ by t-test for two independent sample means).

Note Type	# of Notes	% of total Notes
Nursing Record/Note - Inpatient	499,963	15.6%
Telephone Encounter	179,769	5.6%
Progress Note	115,118	3.6%
Operative Report	47,555	1.5%
Discharge Summary	34,465	1.1%

Table 3.3: Statistics for top 5 most frequent note types in the dataset

All clinical notes corresponding to a surgical case were concatenated. This constituted a pseudo-document for each surgical case sample in our dataset. These pseudo-documents were used in deriving our text representations. We did not have labels indicating presence/absence of SSI at the note-level, only at the surgical case-level. Thus, we needed to use this concatenation approach to attach correct labels to our clinical notes. This concatenation approach has been used successfully in several prior works exploring SSI prediction[25,26].

3.2.4 Text Data Preprocessing

We preprocessed clinical notes in the following ways:

1. Text was lowercased.
2. "Stopwords" (those with little likelihood of holding essential clinical meaning, such as "are" or "when", as defined by the nltk's stopword list[27]) were removed for discrete vectorization approaches (see Text Data Vectorization Section).
3. Numbers and punctuation were removed for discrete text vectorization approaches.

These text preprocessing approaches have been used in prior research on SSI prediction. These works have found that such approaches can reduce computational cost associated with training SSI

prediction models and improve performance of such models[17,23,25,28]. This is because these methods confer a reduction in the number of features which helps reduce noise and computational complexity of machine learning models.

3.2.5 Text Data Representation

We represented clinical text in three different ways (Figure 3.1):

1. Unigrams: We represented text using unigram features in this approach (Figure 3.1 - 1). In this representation, we used pseudo-documents as text samples for each surgical case. Some past works on SSI prediction have used unigrams to represent text[26]. These studies have pointed to the advantage of using data-driven approaches like using unigrams to represent text over using rule-based approaches. This motivated our use of unigram features.

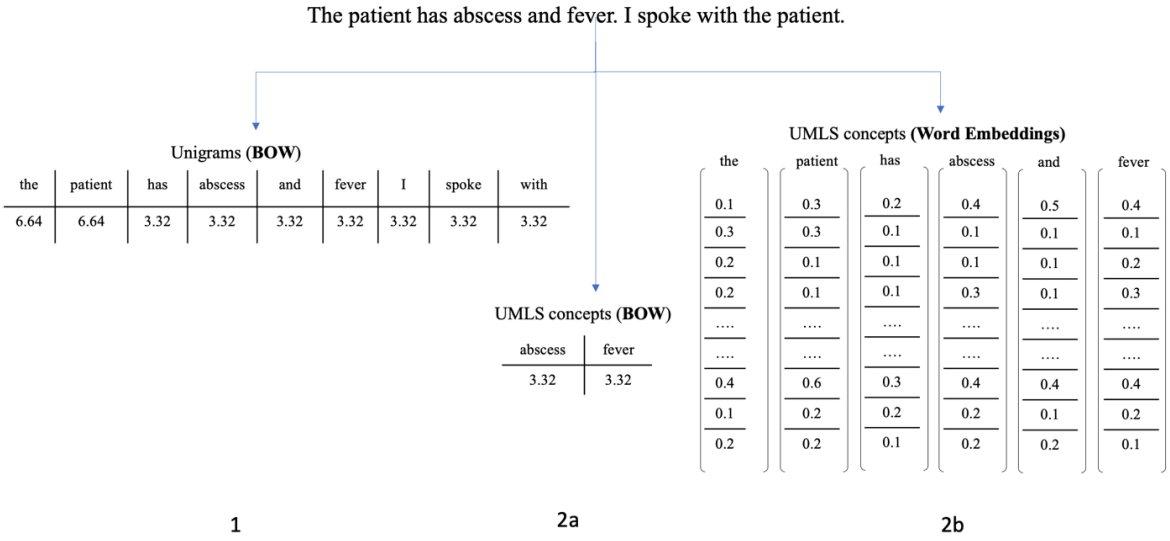


Figure 3.1: Figure depicting our different text representation approaches

2. Unified Medical Language System (UMLS) concepts: UMLS is a repository of biomedical vocabularies developed by the United States National Library of Medicine that integrates biomedical terms from over 60 different families of biomedical vocabularies (e.g., National Center for Biotechnology Information taxonomy, Gene Ontology, Diseases Database, DrugBank)[9].In UMLS, knowledge is organized by grouping synonymous terms (e.g., temperature raised, temperature elevated, body temperature above reference range, fever) into concepts (e.g., fever). All 60 vocabularies in the UMLS are mapped to these concepts. Thus, a concept signifies and equates any synonym for that concept across all the vocabularies. In this approach, we extracted disease and chemical concepts from clinical notes using scispaCy[30]. The concepts we extracted from clinical notes were those which fell within the category of disease and chemical UMLS semantic types. UMLS semantic types are expert-defined broad categorizations of UMLS concepts[31]. We represented text using these concepts in the following two ways:

    a. We considered a subset of the UMLS concepts we extracted from clinical notes as

features (Figure 3.1 - 2a). We only included the most relevant disease or chemical concepts in our feature set. We did this by limiting the set of concepts to the top 200 concepts with the highest Analysis of Variance (ANOVA) F-value between the concept and the SSI positivity label. In this representation, we used the pseudo-documents as text samples for each surgical case. We used UMLS concepts, as this would group synonymous words/phrases. There has been previous work on extracting UMLS concepts from clinical text for SSI prediction[16,32]. These studies have demonstrated the feasibility of attaining adequate performance on SSI prediction using these features as inputs to machine learning models. This motivated our use of UMLS concept features.

b. Using only UMLS concepts as features for SSI prediction may introduce noise without the surrounding context (e.g., negation). To capture context around UMLS concepts, we represented text using sentences containing UMLS concepts as inputs to our models (Figure 3.1 - 2b). We extracted sentences containing disease or chemical UMLS concepts from our pseudo-documents, ordering these sentences temporally by the date of the note in which they appear and the sequence in which they appear in the note. It was possible that many instances of sentences with certain concepts (e.g., pain) are present in almost every pseudo-document with the counts of such sentences in each pseudo-document having high variance and no correlation to whether the document represents an SSI positive case. Thus, to reduce noise, we only extracted sentences containing concepts that appeared in fewer than 15% of pseudo-documents. Moreover, we only extracted sentences with the most relevant disease or chemical concepts. We did this by limiting the set of concepts to the top 200 concepts with the highest Analysis of Variance (ANOVA) F-value between the concept and the SSI positivity label (similar to approach #2a above). We used the sequences of resulting sentences as the text representation. We chose to include information regarding the context around UMLS concepts, as many past works have shown contextual information from text to be critical to building performant post hoc SSI prediction models[16,23,33].

3.2.6 Text Data Vectorization

To vectorize our clinical text data, we used the following methods for our different text representations (Figure 3.1):

1. Unigram features: Some past works have used the BOW text vectorization technique with term frequency-inverse document frequency (TF-IDF) term weighing scheme for SSI classification and have found this approach to confer better performance than other permutations of text preprocessing and vectorization techniques[17,25]. Thus, we used the BOW text vectorization technique and the TF-IDF term weighing scheme. To derive our TF-IDF representation, smooth-inverse document frequency, sublinear term frequency, and L2 norm were used. Only words appearing in five or more documents were included as features. The highest scoring text features were selected based on an Analysis of Variance (ANOVA) F-test prior to input into models. For each model, the number of such features that were selected was a hyperparameter that was tuned (see ”Model Development” section). The final text features were concatenated with structured data

features to form the final representation of each surgical procedure as inputs for model development.

2. UMLS concepts:
    a. UMLS concept features: We used the TF-IDF values of the concepts in the clinical text samples for each surgical case as feature values (see Section 4.1.2)[25]. For the document frequency in the TF-IDF calculation, we used the number of surgical cases whose pseudo-documents contain the concept. The remaining details of our UMLS concepts representation using TF-IDF were identical to that for our unigram representation using TF-IDF.
    b. UMLS concept sentences: Some past works have used distributed representations like word embeddings for SSI prediction[23,24]. These studies have found distributed representations can outperform discrete text representations on the task of SSI prediction. Thus, we vectorized the sequence of sentences using word embeddings. We used word embeddings trained on our dataset using the word2vec method. When training our embeddings, we used the Continuous Bag of Words model architecture.

3.2.7 SSI Classification Approach

For SSI classification, we implemented two baseline approaches. One was a version of a rule-based NLP system that forms the basis of a published gold standard contemporary automated surveillance system. The other was an RF model. The rule-based system used in our study is available online at: https://github.com/jianlins/EasyCIE_GUI. The rule-based system uses an NLP pipeline with rule-based components such as section detector, named entity recognizer, and context detector to process clinical text samples. The output of this pipeline is to label each clinical text sample as indicating occurrence of SSI or no occurrence of SSI after surgery[13]. The RF baseline used structured features and unigram text features as input. RF was an apt baseline for this task as many published gold standards in the past have used this model[10,16]. We then compared the performance of these two baseline models to the following neural approaches which were applied to the text representation approaches in the previous section:

1. NN (Neural Network) – NN models can capture complex combinatorial dependencies between our many structured and text features[34]. Neural models have been shown to be more adept than conventional machine learning models at delineating complex relationships between input features and output variables[35].
2. CNN (Convolutional Neural Network) – CNN models can capture textual patterns relating to SSI that span a few words. In addition, CNN models typically take word embeddings as input, allowing them to capture semantic relationships between words. Past work has shown CNN models can outperform conventional machine learning models on the task of prospective SSI prediction[23,24].
3. BiLSTM (Bidirectional Long Short-Term Memory Network) – BiLSTM models can capture textual patterns that are longer than a few words[36]. In addition, BiLSTM models typically take word embeddings as input, allowing them to capture semantic relationships between words[34]. Past work has shown BiLSTM models can outperform conventional machine learning models and perform just as well as CNN models on the task of prospective SSI prediction using preoperative notes[23].

Figure 3.2 presents the architectures of the neural models we used for SSI classification. To assess the incremental gains of including text features and using different text representations in the task of SSI identification, we first trained each model with structured data only, then introduced various text representations derived from clinical notes.

Our different text representations included BOW with TF-IDF, and word embeddings. For our TF-IDF representation, we concatenated it directly with our structured representation. For word embeddings, we input them into different *deep* learning layers (e.g., CNN layer, LSTM layer, BiLSTM layer) prior to concatenation with structured data.

To make our predictions, our RF models took the concatenation layer as input and our neural models had a dense layer and a final output layer following the concatenation layer.

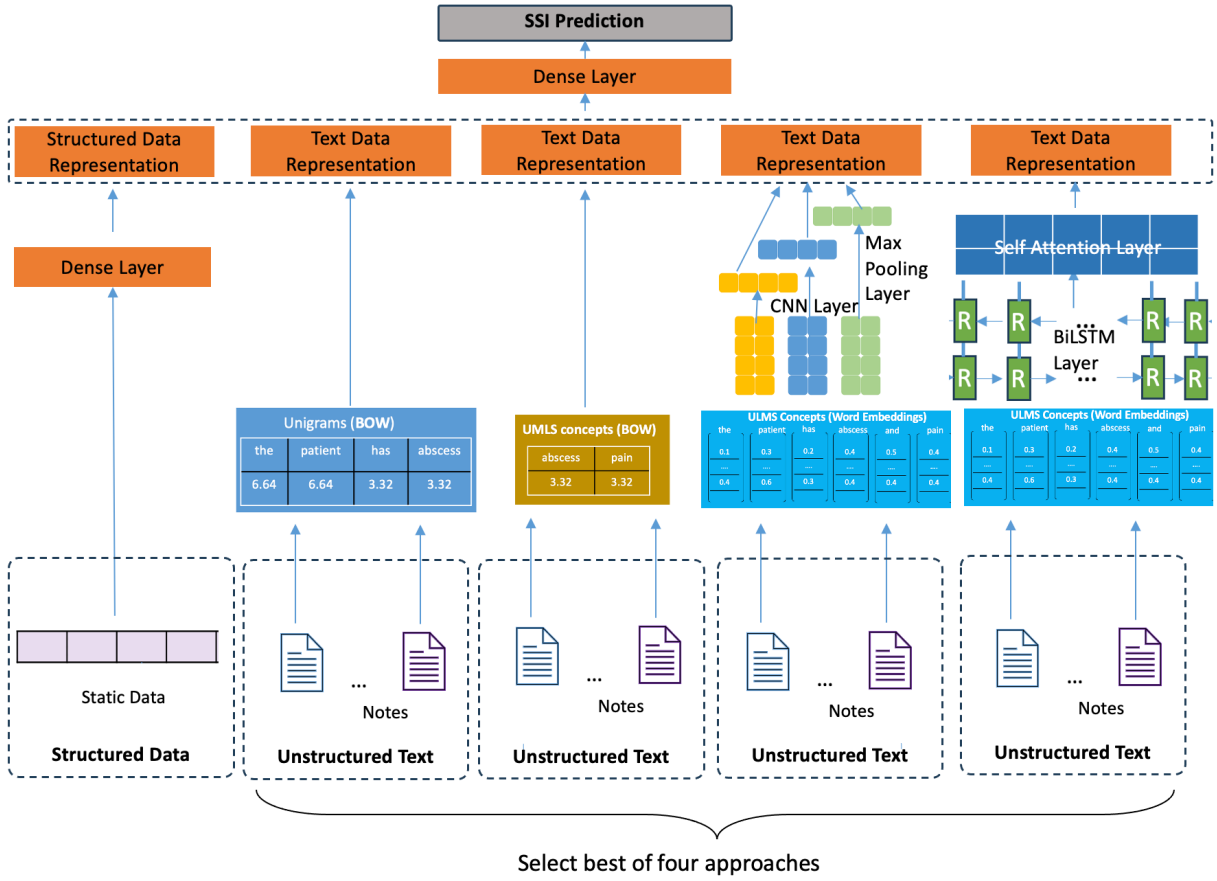


Figure 3.2: Figure depicting neural architectures used in our SSI classification approach

3.2.8 Model Development

The data set was split into training, validation, and test sets with a 7:1:2 ratio. We used the validation set to tune hyperparameters. Hyperparameter tuning was an iterative process where each step involved model training, analysis of results, and selection of hyperparameter values for the next step. The tuned hyperparameters are provided in Appendix Supplemental Table 2. Models were evaluated on the test set.

3.2.9 Model Evaluation

Model performance was evaluated using precision (positive predictive value), recall (sensitivity), and F1-score.

We tested differences between our different data representation and modeling strategies using an analysis of variance (ANOVA) test. We then performed comparisons of the performance achieved by an individual representation or strategy to another individual representation or strategy using a T-test for the means of two independent samples. We used the Bonferroni method to control the family-wise error rate.

3.2.10 Definition and calculation of clinical utility

Automated surveillance systems are typically used as a screening tool prior to manual chart review. Thus, the ultimate goal of automated surveillance systems is to reduce the burden on infection prevention staff (minimizing the number of cases that must be reviewed to detect each SSI case) without missing any SSI cases in the screening process. This can be defined as achieving as high a precision as feasible at a set high recall. Thus, to assess clinical utility, we picked 0.9 recall as the target and examined the corresponding precision values of the models. We selected a recall of 0.9 to match the target accuracy of an automated surveillance system, as suggested by our clinical collaborators[19]. The current manual SOTA system involves reviewing patient charts to determine a case as SSI positive or negative. As the incidence of SSI is 3%, this entails review of 30 charts to identify each SSI case. Additionally, the current automated SOTA system achieves 0.24 precision at 0.94 recall[13]. This entails review of 4 charts to detect every SSI case.

3.2.11 Model Explainability

To determine portions of clinical text which were relevant for SSI prediction, we generated model explanations using Shapley text plots of a random sample of true negative, true positive, false negative, and false positive cases. Shapley text plots are based on Shapley values which represent the average marginal contribution of a feature across all subsets in the total set of features[37].

3.3 Results

The results section is organized as follows. We present the performance conferred by our various feature sets, data representation methods, and modeling strategies (Section 3.3.1). We also identify the best performing model (BiLSTM - Table 3.4 exp#7). We perform model explainability and error analysis for this model (Section 3.3.2-3.3.5). We identify ways to improve performance of our best performing model based on these analyses (Section 3.3.6). We train our best performing model after implementing these strategies (Section 3.3.6). We then redo our model explainability analysis (Section 3.3.7). We then present the clinical utility of our various models (Section 3.3.8).

3.3.1 Contributions of model type, architecture and text representation on model performance

In our initial experimentation, we assessed the impact of using different feature sets, text representations, and modeling strategies on performance (F1-score) on the task of SSI prediction (Table 3.4). The column corresponding to the metric we used for statistical testing is highlighted in grey in the table. The rule-based approach was implemented by processing the pseudo-documents representing each surgical case using the rule-based NLP system to derive classifications for each case[16]. We have listed the rules that were used by the rule-based system to generate classifications in Appendix Supplemental Table 3.

Exp #	Model (Text Representation)	P	R	F1	Adjusted p-value	Comparison with
R	Rule-based NLP (Rule-based)	0.34	0.59	0.43		
1	RF (structured only)	0.64(0.63,0.65)	0.42(0.42,0.43)	0.51(0.50,0.51)		
2	NN (structured only)	0.76(0.76,0.76)	0.42(0.41,0.44)	0.54(0.52,0.56)	0.22	Exp #1
3	RF (Unigrams)	0.69(0.69,0.69)	0.47(0.47,0.48)	0.57(0.56,0.57)	$5*10^{-3}$ **	Exp #1
4	NN (Unigrams)	0.58(0.58,0.58)	**0.67(0.67,0.67)**	0.62(0.62,0.62)	$3*10^{-7}$ **	Exp #3
5	NN (UMLS Concepts)	0.60(0.59,0.60)	**0.67(0.66,0.68)**	0.63(0.63,0.63)	$6*10^{-6}$ **	Exp #4
6	CNN (UMLS Concepts)	0.62(0.60,0.64)	0.64(0.63,0.64)	0.63(0.62,0.63)	1	Exp #5
7	BiLSTM (UMLS Concept Sentences)	0.62(0.59,0.65)	**0.67(0.64,0.69)**	**0.64(0.64,0.64)**	$2*10^{-5}$ **	Exp #5

Table 3.4: Performance conferred by different feature sets, modeling strategies, and text representations on the task of SSI prediction

bold = highest value achieved on performance metrics
column with metric used for statistical testing is highlighted in grey
*** = significant p-value*

Our results showed that there were differences in the overall performance conferred by the different approaches we used (ANOVA test p-value: $1*10^{-19}$). In addition, the following were our conclusions from Table 3.4:

1. **Deep learning models vs rule-based NLP algorithms:** Our best performing deep learning models outperformed a published gold standard rule-based NLP system (Table 3.4 exp #7 vs exp #R - BiLSTM F1-score: 0.64, rule-based F1-score: 0.43). This could be due to three factors : (1) the rule-based system considered only clinical text, whereas our models considered structured data and clinical text, (2) the rule-based system was not as adept as our deep learning models at delineating complex patterns between different features and SSI, and (3) rules within the rule-based system were not as applicable to our external dataset as they were to the internal dataset based on which they were designed. Using the rule-based NLP system led to a sharp drop in precision over our best deep learning models

(Table 3.4 exp #7 vs exp #R - BiLSTM precision: 0.62 versus, rule-based precision: 0.34). This indicates that although rules in the rule-based system had high coverage of SSI positive cases, they were not specific to SSI cases in our dataset. These rules represented patterns that were also present in cases that did not develop into SSI in our dataset.

2. **Impact of adding clinical text vs using structured data alone:** Addition of clinical text to structured data as inputs to machine learning models improve performance for both RF and neural models (Table 3.4 exp #4 vs exp #2 - NN: clinical text + structured F1-score: 0.62 structured F1-score: 0.54).
3. **Impact of deep learning vs conventional machine learning (RF) models:** Deep learning models result in better overall performance than RF models if provided with a combination of structured and clinical text inputs (Table 3.4 exp #7 vs exp #3 - clinical text + structured: BiLSTM F1-score: 0.64 vs RF F1-score: 0.57).
4. **Performance of deep learning models:** A deep learning model (BiLSTM) achieves the highest F1-score on the task of post hoc SSI prediction **(**Table 3.4 exp #7 clinical text + structured: BiLSTM F1-score: 0.64**).**

3.3.2 Model explainability and error analysis

To understand the signals contributing to driving model predictions towards a positive or negative classification, we looked at Shapley text plots of our best performing model (highest F1-score – Table 3.4 exp#7) for a random sample of true negative, true positive, false negative, and false positive cases. Shapley text plots are based on Shapley values which represent the average marginal contribution of a feature across all subsets in the total set of features[37].

Figure 3.3 presents some examples of portions of our Shapley text plots. These portions were taken from our Shapley text plots for the purpose of illustrating the kinds of clinical text excerpts that drove model classifications towards a positive or negative classification. In these examples, portions of clinical text from each surgical case which drove model predictions towards a positive classification are highlighted in red. Meanwhile, portions of clinical text from each surgical case which drove model predictions towards a negative classification are highlighted in blue. Brighter red colors indicate portions of text that drove models more towards a positive classification than lighter red colors and brighter blue colors indicate portions of text that drove models more towards a negative classification than lighter blue colors.

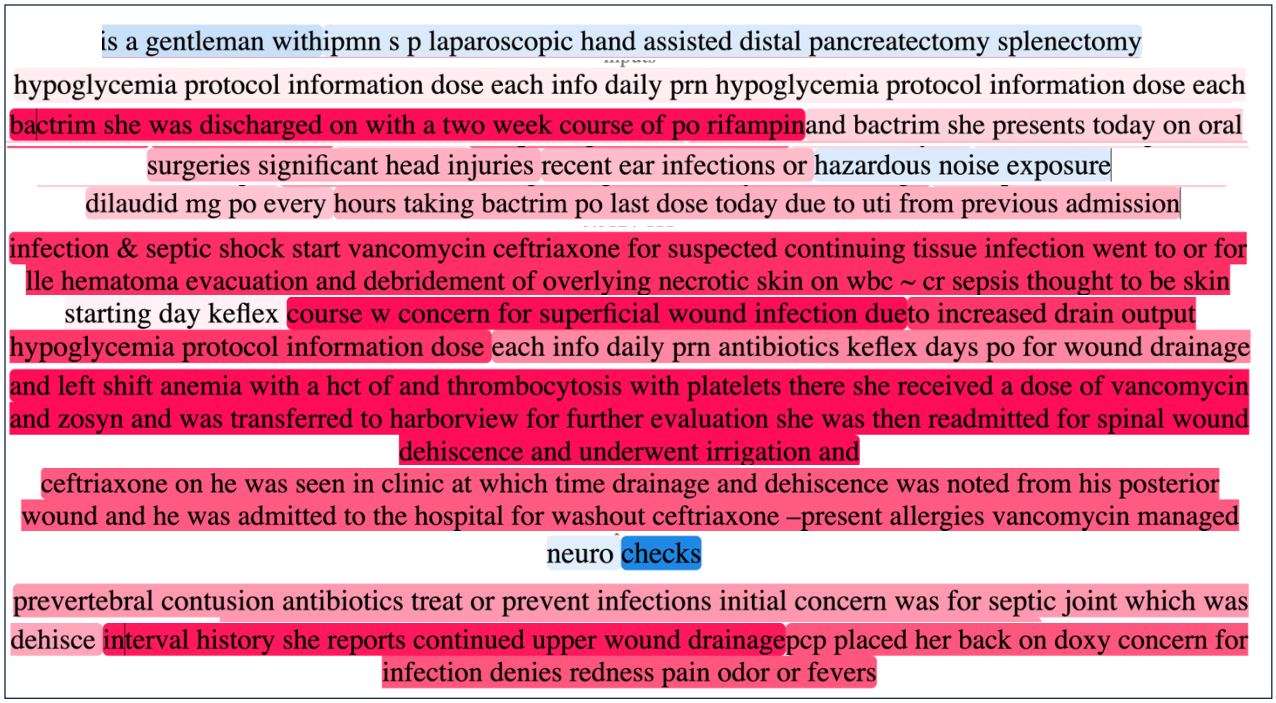


Figure 3.3: Examples of clinical text excerpts which drove model predictions towards a positive or negative classification

The following sections outline our findings from Figure 3.3. These findings include factors of our clinical text representations driving classifications for true positive, true negative, false positive, and false negative cases. Additionally, the following sections present the outline and results of experiments designed based on these findings to reduce the false negative rate. We finally present the clinical significance of some of our feature sets, data representation approaches, and modeling strategies.

3.3.3 Factors of clinical text representation driving model classifications

The clinical text excerpts in Figure 3.3 revealed that (Figure 3.3):

1. **Textual features considered important by models in making classifications:** Important factors in driving the model towards a positive classification included: signs or symptoms of SSI, treatment administered for SSI, and recordings in clinical notes that the patient had an SSI (Figure 3.3). These are also features that clinicians look at in determining whether a case led to an SSI. Thus, our model explanations make sense from a clinical perspective.
2. **Impact of contextual information from text:** Contextual information is important in making model predictions (Figure 3.3). For instance, the excerpt: "prior bowel surgery now with multiple abscesses and leukocytosis" (Figure 3.3) within clinical notes is taken as a strong indication of SSI. Meanwhile the excerpt "leukocytosis persistent but stable" (Figure 3.3) is not. The former establishes that the patient has leukocytosis and that the leukocytosis is associated with an abscess, a symptom of SSI, in the postoperative setting. This suggests the development of SSI consistent with clinical presentation described in literature[38]. Meanwhile, the latter is affirming that the patient has leukocytosis, but it is

improving and stable. This indicates that the leukocytosis is due to normal changes in laboratory values after surgery and that there is no immediate concern for infection[39].

3.3.4 Factors of clinical text representation contributing to false negatives

The clinical text excerpts in Figure 3.3 also revealed that factors leading to false negatives include (Figure 3.3):

1. **Repeated phrases:** Repeated phrases not related to SSI (e.g., repeated phrases related to collection of laboratory values) - likely due to phrases being repeated in clinical notes and the routine practice of copy-pasting when writing clinical notes.
2. **Irrelevant procedural details:** Presence of procedural details unrelated to SSI (e.g., phrases describing the surgery) - likely due to concepts corresponding to particular types of surgeries being selected as important for SSI prediction. This could be because certain surgery types lead to higher SSI rates than others.

3.3.5 Factors of clinical text representation contributing to false positives

The clinical text excerpts in Figure 3.3 revealed that factors leading to false positives include (Figure 3.3):

1. Inaccuracies in reference standard.
2. Development of infections other than SSI.
3. Clinical suspicion and workup for SSI that was later revealed to be a SSI negative case.

3.3.6 Effect on performance of using various strategies to reduce the false negative rate

Taking lessons from our error analysis, we attempted to:

1. **Remove specific note types:** Remove certain note types that may more frequently contain procedural details and less frequently contain relevant clinical signals of SSI. This was to address point #2 in Section 3.3.4.
2. **Remove repeated phrases:** Remove repeated phrases to reduce false positives to address point #1 in Section 3.3.4 (Table 3.5 exp #8).
3. **Only consider clinical notes from after surgery day:** Only consider clinical notes from after the day of surgery when constructing our text representation to address point #2 in Section 3.3.4 (Table 3.5 exp #9). Hyperparameter tuning was repeated for this approach. The tuned hyperparameters are provided in Appendix Supplemental Table 2.
4. **Use separate representations of clinical notes from before and after surgery:** Use separate BiLSTM layers to represent clinical notes from the day of or before surgery and after surgery to address point #2 in Section 3.3.4 (Table 3.5 exp #10). Hyperparameter tuning was repeated for this approach. The tuned hyperparameters are provided in Appendix Supplemental Table 2.

We implemented approaches outlined in bullet points #2-4 above and retrained and reevaluated our best performing model (highest F1-score – Table 3.4 exp#7) from Table 3.4 after using these

approaches. Table 3.5 below presents the performance of the best performing model (highest F1-score – Table 3.4 exp#7) without any changes and after implementing each strategy (bullet points #2-4) above. The column corresponding to the metric we used for statistical testing is highlighted in grey in the table.

Exp #*	Strategy #	P	R	F1	Adjusted p-value**	Comparison with
7	-	0.62(0.59,0.65)	**0.67(0.64,0.69)**	0.64(0.64,0.64)	$4*10^{-5}$ **	Exp #5
8	2	0.65(0.65,0.65)	0.65(0.64,0.65)	0.65(0.64,0.65)	1	Exp #7
9	3	0.67(0.66,0.67)	0.66(0.66,0.66)	**0.66(0.66,0.66)**	$2*10^{-5}$ **	Exp #7
10	4	0.65(0.60,0.70)	0.63(0.59,0.67)	0.64(0.64,0.65)	1	Exp #7

Table 3.5: Performance conferred by various strategies to reduce numbers of false negatives compared to previous approaches presented in Table 3.4

bold = highest value achieved on performance metric

** For all experiments presented in this table, the text representation is UMLS Concept Sentences and the model is BiLSTM*

*** = significant p-values*

Our findings on implementing each approach aimed at reducing false negatives (see bullet point #1-4 above) are listed below with its corresponding bullet point number from the list above:

1. **Remove specific note types:** Table 3.6 shows that procedural details and concepts appear in some of the same note types (e.g., Consultation – Inpt, ICU – Inpt Record). Thus, removal of note types with procedural details would also entail removing notes with potential signals of SSI.

Notes with Procedure Related Concepts	Notes with SSI Related Concepts
Nursing Record/Note – Inpt	Infectious Disease - Inpt Record
Surgery - Inpt Record	Medicine - Inpt Record
EKG/ECG Report	Nursing Record/Note - Inpt
Consultation - Inpt	ICU - Inpt Record
Progress Note	Progress Note
Nursing Record/Note	Surgery - Inpt Record
ICU - Inpt Record	Physical Therapy - Inpt Record
Physical Therapy - Inpt Record	Consultation - Inpt
Operative Report	Occupational Therapy - Inpt Record
Cardiology - Inpt Record	Neurosurgery - Inpt Record

Table 3.6: Notes in our dataset with the highest number in which each category of concept is mentioned at least once

2. **Remove repeated phrases:** Removing repetition did not help performance (Table 3.5 exp#8 vs exp#7). This indicates that phrases related to SSI were also repeated in our text samples (e.g., patient has methicillin-susceptible Staphylococcus aureus bacteremia repeated multiple times in notes each time the patient was checked up). Removal of

duplicated phrases removed these repeats and thus moved true positive cases towards a negative classification.

3. **Only consider clinical notes from after surgery day:** Only considering notes from after the day of surgery, however, led to improvements in overall performance (F1-score) (Table 3.5 exp #9 vs exp #7 - excluding notes from before surgery F1-score: 0.66, including notes from before surgery F1-score: 0.64), primarily driven by improvements in precision (Table 3.5 exp #9 vs exp #7 - excluding notes from before surgery precision: 0.67, including notes from before surgery precision: 0.62). We believe this could be because this strategy removed portions of text samples (e.g., procedural details, workup before surgery) in text samples that were irrelevant to SSI and just adding noise. The finding that only considering notes from after the surgery day improves performance has some implications. These implications concern the optimal representation and deep learning modeling strategies to use for post hoc SSI prediction. They are the following:
    a. **Impact of using discrete vs distributed representations and non-deep learning vs deep learning models:** Use of distributed representations and deep learning modeling strategies outperforms the use of discrete representations and non-deep learning modeling strategies on the task of post hoc SSI prediction (Table 3.5 exp #9 vs Table 3.4 exp #5 - word embeddings + BiLSTM F1-score: 0.66, TF-IDF + NN F1-score: 0.63).
    b. **Impact of using different deep learning layers:** For the performance gains offered by deep learning to be realized, appropriate deep learning layers should be used. A BiLSTM model outperformed a CNN model (Table 3.5 exp #9 vs Table 3.4 exp #6 - BiLSTM F1-score: 0.66, CNN F1-score: 0.63). This means that many of the textual patterns which distinguish SSI positive and SSI negative cases are long-range textual patterns.
4. **Use separate representations of clinical notes from before and after surgery:** Using separate BiLSTM layers to represent notes from the day of or before surgery and after surgery did not improve performance. Table 3.5 exp #10 vs exp #7. This indicates that information from notes from before surgery adds more noise than signal to the textual data.

3.3.7 Model explainability analyses on our models after implementing changes to reduce the false negative rate

We wished to determine the actual impact of only considering clinical notes from after the day of surgery. Thus, we looked at Shapley text plots of models trained on our new text samples (after removal of clinical text data from the day of surgery or before) (Figure 3.4). These plots show that only considering clinical notes from after surgery removed many irrelevant details from text samples.

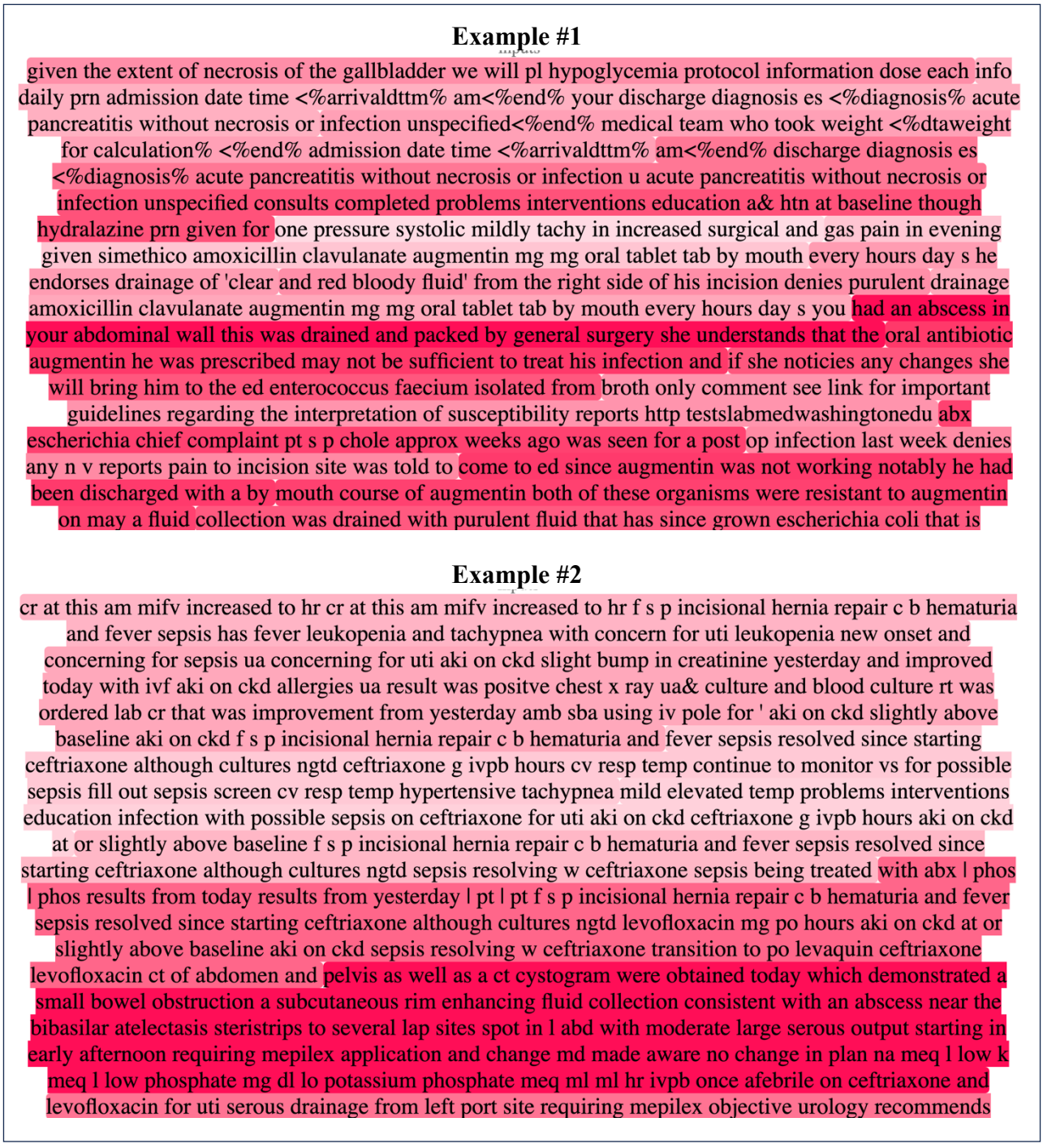


Figure 3.4: Shapley text plots after removal of clinical text data from the day of surgery or before

in bullet point #2 above (removal of clinical notes from before the day of surgery). The irrelevant procedural details have been removed from the beginning of the text samples.

3.3.8 Clinical significance

Assessment of overall performance of our various approaches is not meaningful in actual practice if it does not reflect performance in a real-world clinical setting. In such a setting, our automated

surveillance approach is likely to be used as a screening tool prior to manual chart review to rule out clearly negative cases. Thus, to assess the actual clinical importance of the various SSI prediction approaches we attempted, we assessed the precision of our models at a high recall (90%) recall (Table 3.7). We did this by tuning the models to achieve recall of 0.9 and measuring precision (Table 3.7). The column corresponding to the metric we used for statistical testing is highlighted in grey in Table 3.7. As a reminder the current literature shows best published approaches using rule-based or conventional machine learning approaches achieve precision of 0.24-0.28 with recall around 0.9.

Exp #*	Model	Text Representation	P at 0.9 R	# charts reviewed per SSI detected	Adjusted p-value	Comparison with
	Manual	Manual review	0.03	33		
R	Rule-based	Rule-based	0.34(at 0.59 R)	-***		
3	RF	Unigrams	0.17(0.15,0.19)	5.9		
5	NN	UMLS Concepts	0.26(0.25,0.27)	3.8		
9	BiLSTM	UMLS Concept Sentences – no notes before 7 days	**0.29(0.29,0.30)**	3.4	$5*10^{-6}$ **	Exp #5

Table 3.7: Clinical significance of the various data representation and modeling approaches we used in this chapter

bold = Highest value achieved on performance metric
**Exp # referenced from Table 3.4 and 3.5*
*** = significant p-value*
**** - # charts reviewed not provided for this system, as it achieved lower than 0.9 recall (0.59 recall achieved), and being a rule-based system, we could not adjust the threshold to reflect precision at 0.9 recall*

These experiments showed that modeling and text representation strategies which conferred better overall performance also showed better clinical utility (Table 3.7). Our results showed that, overall, there were differences in the clinical utility conferred by the different approaches we used (ANOVA test p-value: $2.1*10^{-22}$). In addition, the following were our conclusions from Table 3.7:

1. **Impact of rule-based or conventional machine learning-based approach vs deep learning approach:** Our best deep learning model (highest F1-score – Table 3.5 exp#9) achieved higher clinical utility than our best feedforward NN or RF models or rule-based algorithms (Table 3.7 exp #9 vs exp #R,3 – BiLSTM precision at 0.9 recall: 0.29, RF precision at 0.9 recall: 0.17, rule-based algorithm precision at 0.59 recall: 0.34).
2. **Impact of using discrete vs distributed representations and non-deep learning vs deep learning models:** Using distributed representations along with cutting edge deep learning models achieves higher clinical utility over using discrete representations with non-deep learning neural models (Table 3.7 exp #9 vs exp #5 - word embeddings + BiLSTM precision at 0.9 recall: 0.29, TF-IDF + NN (UMLS Concepts Text Representation) precision at 0.9 recall: 0.26).

3.4 Discussion

In Aim 1, we analyzed data from the EHR of almost 30,000 surgical cases from two different healthcare facilities (UWMC and HMC). We developed an algorithm based on NLP methods (e.g., methods for generating distributed text representations) and deep learning models. In contrast, prior studies used rule-based algorithms or conventional machine learning models (e.g., RF models)[10,16]. Our dataset of almost 30,000 cases made this the largest comparable study to date. It is the only one that incorporates structured and clinical text data and uses deep learning for the task of post hoc SSI prediction. As a reminder, to assess clinical utility, we picked 0.9 recall as the target and examined the corresponding precision values of the models. Combining structured data and clinical text, NLP techniques, and deep learning, we found that it was possible to achieve clinical utility (precision at 0.9 recall) higher than that achieved by contemporary published gold standard approaches reported in literature (best precision and recall of published methods being 0.24 and 0.94 vs best precision of 0.29 with recall set at 0.9 in the work presented here)[2]. This means that, whereas with contemporary approaches using conventional machine learning techniques or rule-based algorithms, 4 chart reviews are required for each SSI detected, with our approaches, only 3.5 chart reviews would be required for each SSI detected. This would result in a ninefold increase in the number of cases that can be surveilled compared with manual surveillance alone. It would also further reduce the workload of infection prevention personnel around the nation by 360,000 hours annually (for calculation details see introduction), on top of reductions conferred by contemporary automated surveillance methods.

The higher clinical utility achieved by our models over approaches representing contemporary surveillance systems is likely due to several factors such as: incorporation of clinical text data as an additional data modality on top of structured data, use of distributed representations and deep learning models, and capture of contextual information from text, some of which are supported by past work[10,23,24,33].

One such factor is the addition of clinical text data. Our findings indicated that addition of clinical text data improves performance on SSI prediction (NN structured + text data F1-score of 0.62, NN structured data without text data F1-score of 0.54). Past work has also found that addition of clinical text on top of structured data improves clinical utility of SSI prediction models[3,14]. However, we used text representations that captured important, complex, inter-modal signals relevant to SSI in our data modalities (structured data and clinical text). These representations could generalize to external datasets and had minimal noise.

To capture information relevant for SSI prediction, we used distributed text representations in combination with deep learning layers like BiLSTM. This is different from contemporary SSI prediction approaches[10,16]. These approaches use rule-based algorithms or discrete text representations in combination with conventional machine learning models. Our use of distributed text representations allowed us to capture semantic similarity between words[25]. This permitted us to achieve better test set performance with fewer training samples. Our use of deep learning models such as BiLSTM layers allowed us to capture complex relationships between different textual patterns and SSI. Some past research has shown that distributed representations in combination with deep learning models can outperform discrete text representations in combination with

conventional machine learning models[8,17]. However, these studies have found that this can only be realized if there are enough training samples[8,17]. Such work has focused on tasks similar to post hoc SSI prediction like: prospective SSI prediction and SSI prediction in French clinical text records[8,17]. However, no past work has explored distributed text representations in combination with deep learning approaches for the purpose of post hoc SSI prediction. We used the largest dataset that has been used for this task thus far. This, in combination with our use of distributed text representations and deep learning models, allowed us to achieve higher clinical utility than past approaches. Our model with the highest clinical utility achieved a precision of 0.29 precision at 0.9 recall. In contrast, the best clinical utility by past approaches to this task is 0.24 at 0.94 recall[13].

Our use of sequential BiLSTM layers also allowed us to capture contextual information from text. This type of information has been shown to be critical to the performance of SSI prediction models[6,14,16]. These studies have found that both the use of words relevant to SSI (e.g., "infection") and their context is crucial to determining whether a surgical case developed into SSI. Some examples of contextual information could be whether the clinical note is confirming infection or not, the location of the infection, the type of the infection. However, past methods of capturing contextual information from text have involved strategies such as n-grams or rule-based context detectors. These strategies suffer from the drawbacks of poor domain adaptation capability and increased computational complexity[26,27]. In contrast, our use of distributed representations in combination with sequential deep learning models do not suffer from the same drawbacks[26,27]. Together, our use of multimodal data (structured data + clinical text), distributed text representations, and deep learning layers allowed us to capture important signals of SSI in multiple data modalities. They also allowed us to capture signals which generalize to domains not well represented in the training set.

However, in order for the gains offered by distributed representations and deep learning models to be realized, we found that appropriate text representations need to be used. Our error analysis revealed that one factor contributing to false negative cases was procedural details unrelated to SSI. To address this, we demonstrated that clinical notes from or before the surgery date could be removed. This improved overall performance, primarily driven by gains in precision. We found that not considering clinical notes from the day of surgery and before removed noise related to procedural information from our text representation (see Chapter 3 Results: Effect of Using Various Strategies to Reduce False Negatives Section). Some past studies have used clinical text and/or temporal data from before surgery to capture signals relating to preexisting infections. However, our experimentation shows that doing so adds more noise than signal. -One reason for this may be that past studies which have included data from before surgery have used rule-based NLP methods or used feature selection to remove text features unrelated to SSI. This removed features in clinical notes from the day of surgery or before that were irrelevant to SSI. In contrast, we used a more data-driven approach, in which we input all sentences containing UMLS disease or chemical concepts (see Chapter 3 Methods: Text Data Representation Section). An advantage of our approach is that it should be more generalizable to external datasets that might have a different set of features relevant to SSI than in the internal dataset.

Other than conferring better clinical utility, we expected our deep learning approaches to have better domain adaptation capability than contemporary rule-based approaches (see Chapter 2 Section 7). Our findings supported this hypothesis. We found that rules in the rule-based baseline NLP approach were not able to generalize to our external dataset. The rules were not specific to SSI cases, leading to a high rate of false positives.

Our main findings in this aim were: (1) addition of clinical text data on top of structured data as inputs to machine learning models improved clinical utility on the task of SSI prediction, (2) distributed representations in combination with deep learning conferred better clinical utility than discrete representations in combination with conventional machine learning and rule-based algorithms on the task of SSI prediction, and (3) in order for the potential performance gains offered by deep learning to be realized, appropriate clinical text representations should be used.

3.5 Conclusion

My overall research question is: "*Can an automated surveillance system based on multimodal data and deep learning outperform and have better domain adaptation capability than SOTA automated SSI surveillance approaches*?" With respect to this question, the findings in this chapter suggest that an automated surveillance system based on deep learning using structured data and clinical text can achieve better clinical utility than existing automated surveillance systems. It is unknown if additional data modalities (ex. temporal data) further improve the clinical utility of deep learning models.

We did not include temporal data such as laboratory values in our list of inputs in this chapter. This meant that our models could not capture signals relevant to SSI found in changes in laboratory values such as WBC counts, absolute neutrophil counts, or temperature. SSI typically develops between the day of surgery to the 90th postsurgical day (depending on the type of surgery). During the course of SSI development, laboratory values and vitals change. A clinical suspicion of infection also prompts repeated ordering of laboratory tests and vital readings by medical practitioners. Thus, trends in temporal features like laboratory values and vitals contain important signals of infection. Moreover, our error analysis revealed factors contributing to false positives. One such factor was development of infections other than SSI. In this case, temporal trends indicative of infection in laboratory values or vitals would appear at a different time window. Another factor was clinical suspicion and workup for SSI that was later revealed to be a SSI negative case. In this case, temporal trends indicative of infection would not appear. For both factors, temporal trends would help models correctly classify the false positive cases. Therefore, in the next chapter on Aim 2, which is: Aim 2: A*ssess the impact of adding temporal data on the performance of deep learning models which predict SSI*, we will add temporal data to our list of inputs to deep learning models.

4. Aim 2: Assess the impact of temporal data on the performance of deep learning based surgical site infection prediction

4.1 Introduction

This chapter is modified and extended from a paper titled *Automated Identification of Surgical Site Infection from Electronic Medical Records* with authors Arjun Chakraborty, Peter Tarczy-Hornoch, MD, Dustin R. Long, MD, Meliha Yetisgen, PhD submitted to AMIA CRI symposium on 9/17/2024. All text in normal font is from that paper. Additions are in italics.

My overall research question is: "Can an automated surveillance system based on multimodal data and deep learning outperform and have better domain adaptation capability than state of the art automated surgical site infection (SSI) surveillance approaches?" In this chapter the focus is on Aim 2, "Assess the impact of adding temporal data on the performance of deep learning models which predict surgical site infection". We assess the effect of adding temporal data on top of static structured data and clinical text, different representations of temporal data, and different modeling approaches for temporal data on performance achieved on SSI prediction.

This work builds on Chapter 3: Develop and assess an automated surgical site infection surveillance approach using deep learning and multimodal data. In Chapter 3, we used static representations of variables (e.g., maximum values of vitals). This simple representation does not capture changes over time in laboratory values or vitals which is critical for SSI prediction. For example, SSI is marked by an initial rise in WBC count. This is followed by a period of stability as the patient fights infection. Finally, there is a period of exponential decay in WBC count as the patient recovers from infection[40]. In this chapter, we explore the impact of including temporality in our data representation on SSI prediction. We first enrich our representation for several variables (e.g., WBC count, temperature) by adding temporality. We then measure the impact of this new representation on the overall SSI prediction performance for different modeling approaches.

Our work builds on previous work which also used temporal data for SSI prediction. Such work either used extensive data summarization techniques to represent temporal data, extensive feature engineering with clinical text data, or used temporal data alone without incorporating structured data or clinical notes. In contrast to previous studies, we minimize the use of feature engineering with clinical text data, data summarization techniques with temporal data and integrate static structured data, clinical text, and temporal data within a deep learning framework to predict SSI.

This chapter is organized in the following way. We first describe our methods for representing temporal data. Next, we describe our procedures for training our deep learning models with the addition of this temporal data. Finally, we present the SSI prediction performance for a range of data modalities, data representation methods, and machine learning models. We compare the performance of various permutations of feature sets, data representations, and modeling strategies against the best automated approaches to identify the optimal approach. We define the optimal

approach as the one that maximizes precision at a recall of 0.9, using the same metric employed to evaluate clinical utility in Aim 1. Our assessment also includes error analysis and explainability.

4.2 Methods

4.2.1 Clinical Cohort and SSI Definition

Details of the dataset we used for this task are described in Section 3.2.1.

4.2.2 Structured Data Features

We enriched the structured data representation (Section 3.2.2) for laboratory values (WBC count, Absolute Neutrophil Count (ANC)) and postoperative temperature. Table 4.1 outlines the methods used to derive the structured data. The data elements we trended temporally are marked with * in the table.

Category	Features	Description
Procedural Characteristics	The surgery type of a case	E.g., Orthopedic surgery, neurosurgery
Procedural Characteristics	Reoperation	Whether the surgery was a reoperation
Laboratory Values	Culture Obtained	Binary variable indicating whether cultures (wound, tissue, or CSF culture) were collected for a patient
Laboratory Values	WBC Count Laboratory Measurements*	WBC counts from 7 days before surgery to 90 days after surgery, ordered temporally
Laboratory Values	Absolute Neutrophil Count (ANC) Laboratory Measurements*	ANC from 7 days before surgery to 90 days after surgery, ordered temporally
Postoperative Fever	Temperature Readings*	Temperature readings from 7 days before surgery to 90 days after surgery, ordered temporally
Consultations	Infectious Disease	Whether a postoperative infectious disease consultation was obtained
Consultations	Interventional Radiology Drain Placement	Whether a postoperative drain was placed by interventional radiology
Antibiotics	Antibiotic Administered	Whether a specific antibiotic was administered for a patient

Table 4.1: Structured data elements and their descriptions

*Temporal variables

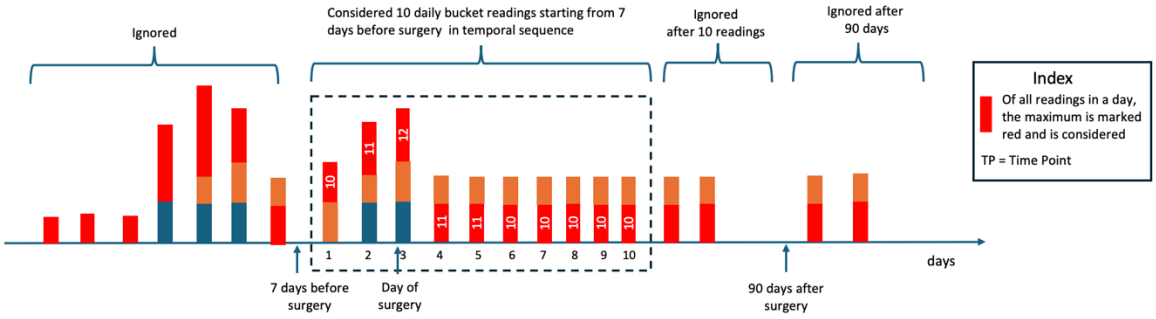
4.2.3 Text Data Representation

We used the same clinical text representation approach described in Section 3.2.4, 3.2.5, and 3.2.6. We only used postoperative notes to construct our clinical text representation. See Section 3.3.6

for a detailed explanation of why we took this approach.

4.2.4 Temporal Data Representation

To assess the effect of adding temporal data on automated SSI identification performance, we enriched our structured data representation of laboratory values (WBC count, ANC) and vitals (temperature) by using the time stamps associated with each measurement or test. *We extracted temporal data from the 7-day preoperative to 90-day postoperative period. This period was designed to capture signals related to preexisting infections.* We aggregated our data by computing the daily maximum for each variable. *To standardize the temporal representation for each patient, we fixed the sequence length by padding shorter sequences with zeros until they matched the maximum length (ten timepoints). Figure 4.1 shows an example illustrating how we constructed our temporal representation.*



Srl #	Data/Feature	TP #1	TP #2	TP #3	TP #4	TP #5	TP #6	TP #6	TP # 7	TP #8	TP # 9	TP #10
1	Collection Value	10	11	12	11	11	10	10	10	10	10	10
2	Collection Days from surgery	-6	-3	-2	0	2	3	4	29	40	71	75
3	Feature #1 (Actual value)	10	11	12	11	11	10	10	10	10	10	10
4	Feature #2 (Days since last date of collection)	1 =-6-(-7)	3 =-3-(-6)	1 =-2-(-3)	2 =0-(-2)	2 =2-0	1 =3-2	1 =4-3	25 =29-4	11 =40-29	31 =71-40	4 =75-71
5	Feature #3 (Days since date of surgery)	-6	-3	-2	0	2	3	4	29	40	71	75
6	Feature #4 (Binary value if actual value is upper limit of range=10.5)	0	1	1	1	1	0	0	0	0	0	0

Figure 4.1: Method for constructing our temporal representation

This length of ten was determined through hyperparameter tuning. We chose zero padding because SSI-negative cases in our dataset had lower laboratory values, fewer vital measurements, and higher rates of missing data compared to positive cases. Alternative methods, such as mean imputation, were unsuitable due to substantial variation in laboratory values and vital signs across patients, resulting from differences in individual physiology and the timing of data collection.

For each laboratory value or vital, we captured the following four important signals:

(1) The fold change above the upper limit of the normal range that the measured value has on a particular day: We defined fold change as the value recorded for the test or vital sign divided by the upper limit of the normal range for that test or vital (upper limit for WBC count: $10.5*10^3/mm^3$, upper limit for ANC:$7*10^3/mm^3$, upper limit for temperature: 38°C based on literature review). High values of laboratory tests or vitals may be indicative of infection.

(2) The number of days between consecutive recordings of a laboratory value or vital: A laboratory test ordered on consecutive days may indicate an ongoing clinical suspicion of infection.

(3) The number of days between the date of surgery and the measurement date: Different indications for the ordering of laboratory tests (e.g., UTI, SSI) occur at different timepoints before or after surgery.

(4) Whether the value is higher than the normal range: Our reasoning behind including this variable was similar to that for measurement representation #1.

To construct these features, we used four different transformations of the temporally ordered sequence of each laboratory value or vital. To derive temporal input for our models, we used the following approaches: (1) CNN and LSTM – We vertically stacked the sequences for all of the 4 features derived from WBC count, ANC, and temperature to constitute the temporal input to our CNN and LSTM models, (2) NN - We concatenated all of the four derived features for all of our temporal variables along the horizontal axis, averaging each feature across the time window (7 days preoperative to 90 days postoperative). This yielded a matrix of dimension: (sample size) x (four times the number of temporal variables). Each row of this matrix represented a sample. Each column represented the average of one feature over the temporal window. Each row of this matrix was used as input for linear layers.

4.2.5 SSI Classification Approach

We input our temporal data representation into different deep learning layers (NN, CNN, LSTM) to generate latent temporal representations. We concatenated these latent temporal representations with our best performing (highest F1-score) text representation and our static structured data representation for training (Figure 4.2).

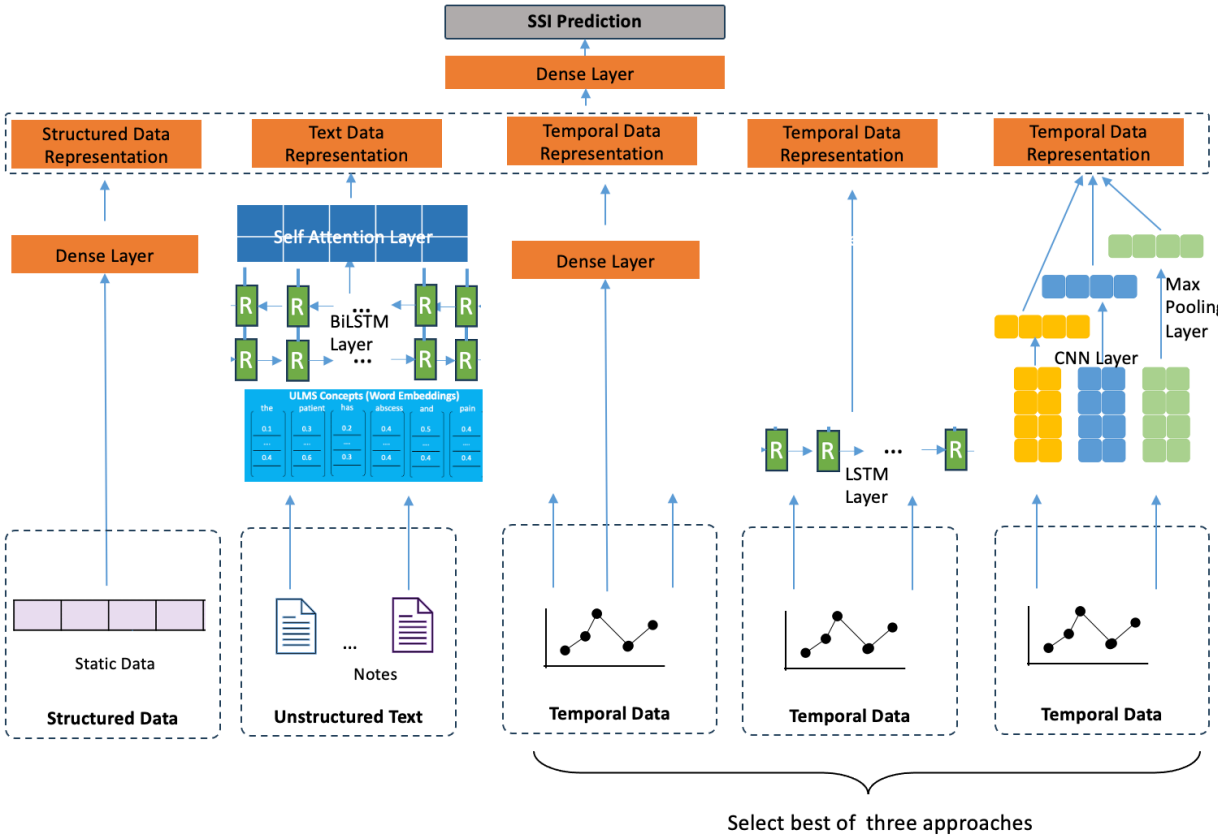


Figure 4.2: Figure depicting neural architectures used in our SSI classification approach

4.2.6 Model Development

Our hyperparameter tuning approach was previously outlined in Section 3.2.8. The hyperparameters tuned, the values considered during the tuning process, and the configurations that produced the best performance for each input set are detailed in Appendix Supplemental Table 4.

4.2.7 Model Evaluation and calculation of clinical utility

Evaluation approach and calculation of clinical utility were previously described in Chapter 3 - Section 3.2.9 and Section 3.2.10.

4.2.8 Model Explainability

Explanations for models incorporating temporal information were generated using the Local Interpretable Model-agnostic Explanations (LIME) package[41]. LIME generates explanations for a model (here considered the original model) by fitting local, interpretable models to each prediction made by the original model[6]. *First, we used the LIME package to generate values of the contribution of each temporal feature to model output for (a) a random sample of 100 true positive, and (b) a random sample of 100 true negative cases. Then, we plotted the average feature*

contribution values for each timepoint for each feature on a heatmap for each category of cases. This allowed us to explore which timepoints and which features were significant in driving predictions for each of these types of cases.

4.3 Results

4.3.1 Contributions of Model Type, Architecture and Text Representation on Model Performance

In our initial experimentation, we assessed the impact of using different feature sets, modeling strategies, and temporal data representations on performance (F1-score) on the task of SSI prediction (Table 4.2).

Exp #	Model	Temporal Representation	P	R	F1	Adjusted p-value**	Comparison with
1	NN	-	0.76(0.76,0.76)	0.42(0.41,0.44)	0.54(0.52,0.56)		
2	BiLSTM*	-	0.67(0.66,0.67)	0.66(0.66,0.66)	**0.66(0.66,0.66)**	$2*10^{-5}$ **	Exp #1
3	BiLSTM	NN	0.60(0.60,0.60)	0.69(0.69.0.70)	0.64(0.64,0.64)	$5*10^{-4}$ **	Exp #2
4	BiLSTM	CNN	0.60(0.58,0.62)	**0.70(0.67,0.73)**	0.64(0.64,0.64)	$2*10^{-6}$ **	Exp #2
5	BiLSTM	LSTM	0.65(0.65,0.65)	0.67(0.67,0.68)	**0.66(0.66,0.66)**	$2*10^{-5}$ **	Exp #2

Table 4.2: Performance conferred by different feature sets, modeling strategies, and text representations on the task of SSI prediction

bold = highest value achieved on performance metrics
column with metric used for statistical testing is highlighted in grey
** see Chapter 3 Table 3.5 exp #9 see Results: Effect of Using Various Strategies to Reduce False Negatives Section in Chapter 3 for full details on how we arrived at this model*
*** = significant p-value*

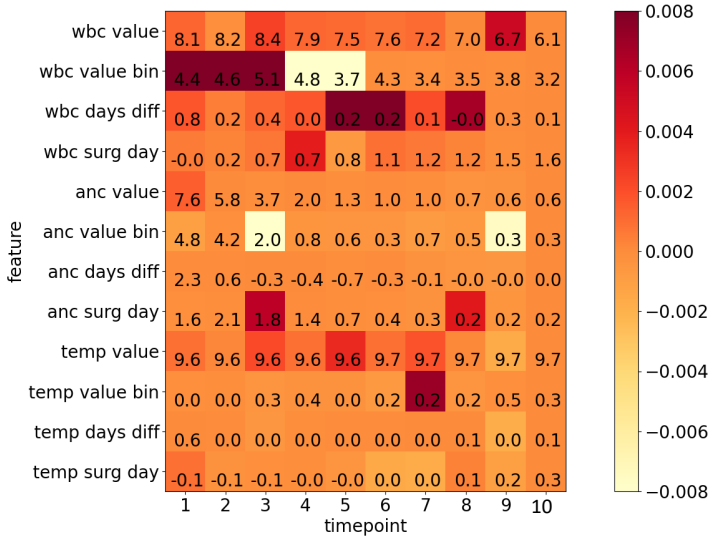
Table 4.2 shows that enriching the data representation with temporality does not affect overall performance but improves recall (Table 4.2 - Exp #5 vs. Exp #2). A plausible explanation for this is that incorporation of temporal information on laboratory values and vitals added signals pertaining to SSI (e.g., elevated laboratory values). Thus, it led to correct classification of some previously false positive cases which exhibit such signals. However, it also resulted in the misclassification of some negative cases and decreased precision. These are cases in which typical laboratory or vital signatures of infection may have existed due to development of infection other than SSI (e.g., UTI, pneumonia).

4.3.2 Model Explainability and Error Analysis

We also generated model explanations to identify temporal signals driving model predictions by employing the LIME package. This package was used to compute feature contribution values for temporal features in the model which achieved the highest overall performance (F1-score) and recall (Table 4.2 Exp #5). Specifically, LIME calculated contributions of each temporal feature to model output for (a) a random sample of 100 true positive, and (b) a random sample of 100 true negative cases. We then visualized these contributions by plotting the average feature contribution values for each timepoint and feature on a heatmap for each category (Figure 4.3). This allowed us to explore significant timepoints and features driving predictions for each category.

Heatmap of the Average Feature Contribution Across Cases for each Feature at each Time Point

Positive



Negative

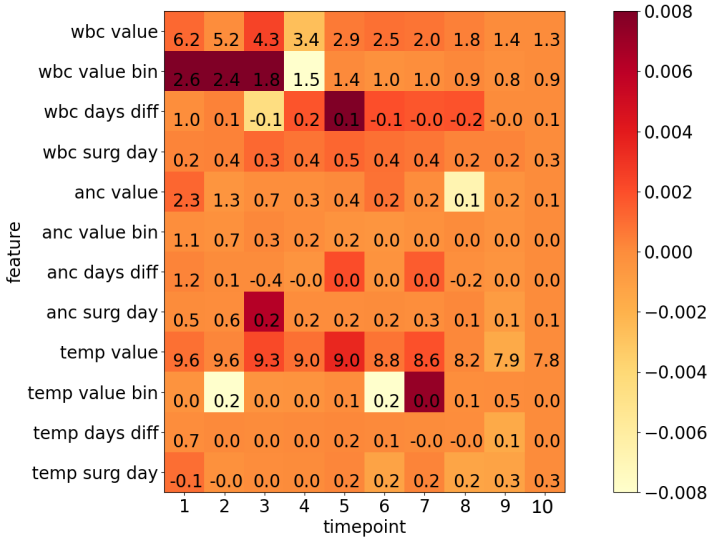


Figure 4.3: Model explanation analyses of true and false positive and negative cases
Number in cell indicates average value at that timepoint of the feature. Color indicates average feature contribution See Methods: Temporal Data Representation Section and Methods: Model Explainability and Error Analysis Section for details on how we derived our temporal data representation as well as our model explainability analyses

Figure 4.3 illustrates that the following factors in our temporal representation affected model classifications:

1. **Features of temporal changes driving negative classifications**: Sharp declines in average feature values, particularly at earlier time points, often lead to negative classifications. This may be explained by missing values, which suggest a lack of clinical suspicion for infection.

Alternatively, these drops may reflect a return of laboratory values to normal post-surgery, indicating the absence of infection.

2. **Impact of Fold Change vs. Binary Fold Change Features**: Fold change features generally lead to positive classifications and improve recall, likely because high values, such as elevated WBC counts, suggest infection. However, these features can also cause false positives, reducing precision, as high fold change values may indicate inflammatory conditions like postoperative pneumonia or community-acquired infections, which present similar laboratory profiles to true SSIs. This explains the increased recall but decreased precision when temporal data are included (see Table 4.2, Exp #5). In contrast, binary fold change features tend to result in negative classifications and higher precision, as missing or low values imply a lower likelihood of infection.
3. **Impact of Earlier vs. Later Timepoints**: Data from earlier timepoints often contribute to positive classifications, while data from later timepoints typically result in negative classifications. High values at earlier timepoints are useful for identifying true positives, whereas low or missing values at later timepoints help rule out negatives. Therefore, incorporating more time steps into our temporal representation would likely improve precision but reduce recall.

4.3.3 Clinical Significance

Assessment of overall performance of our various approaches is not meaningful in actual practice if it does not reflect performance in a real-world clinical setting. In such a setting, our automated surveillance approach is likely to be used as a screening tool prior to manual chart review to rule out clearly negative cases. Thus, to assess the actual clinical importance of our various SSI prediction approaches, we assessed the precision of our models at a high (90%) recall (Table 4.3).

Exp #	*Model* ***	*Text / Temporal Representation*	*P at 0.9 R*	*# charts reviewed per SSI detected*	*Adjusted p-value* **	*Comparison with*
	Manual	*Manual review*	*0.03*	*33*		
R	*Rule-based*	*Rule-based*	*0.34(at 0.59 R)*	*-*****		
	RF	*Unigrams*	*0.17(0.15,0.19)*	*5.9*		
1	*NN*	*UMLS Concepts*	*0.26(0.25,0.27)*	*3.8*		
2	*BiLSTM* / -*	*UMLS Concept Sentences**	*0.29(0.29,0.30)*	*3.4*	*$4*10^{-16}$ ***	*Exp #1*
5	*BiLSTM / LSTM*	*UMLS Concept Sentences / LSTM*	*0.33(0.32,0.33)*	*3*	*$5*10^{-5}$ ***	*Exp #2*

Table 4.3: Clinical significance of the various data representation and modeling approaches

bold = highest value achieved on performance metric

column with metric used for statistical testing is highlighted in grey

** Chapter 3 Table 3.5 exp #9 see Results: Effect of Using Various Strategies to Reduce False Negatives Section in Chapter 3 for full details on how we arrived at this model*

*** = significant p-value*

**** column values are formatted as text representation / temporal representation*

***** - # charts reviewed not provided for this system, as it achieved lower than 0.9 recall (0.59 recall achieved), and being a rule-based system, we could not adjust the threshold to reflect precision at 0.9 recall*

Our results showed that, overall, there were differences in the clinical utility conferred by the different approaches we used (ANOVA test p-value: $1.2*10^{-25}$) (Table 4.3). We also found that incorporation of temporal data marginally but statistically significantly improved clinical significance (Table 4.3 exp #5 vs. exp #2 - with temporal data precision at 0.90 recall: 0.33, without temporal data precision at 0.90 recall: 0.29).

4.4 Discussion

In this aim, we analyzed data from the EHR of almost 30,000 surgical cases from two different healthcare facilities (UWMC and HMC) to develop an algorithm based on static structured data, clinical text, and temporal data from the EHR. This was the only study to date which incorporated 3 different data modalities as inputs to deep learning models to build automated SSI surveillance systems. Several past studies have used structured data or clinical text data alone for SSI prediction[17,25,26]. Some others have incorporated both structured and clinical text inputs[10,16]. Other approaches have utilized temporal data for this task[42,43]. However, no past study has incorporated structured data, clinical text data, and temporal data for post-hoc SSI prediction. As a reminder, to assess clinical utility, we picked 0.9 recall as the target and examined the corresponding

precision values of the models. Our model incorporating static structured data, clinical text, and temporal data, using a BiLSTM layer to process clinical text, using a LSTM layer to process temporal data achieved 0.33 precision at 0.9 recall. This model achieved the highest clinical utility of all our data representation approaches using these 3 modalities and modeling strategies. It also achieved higher clinical utility (precision at 0.9 recall) than published methods (with temporal data precision of 0.33 at 0.9 recall and best precision and recall of published methods being 0.24 and 0.94). This means that whereas with contemporary automated surveillance systems, 4 chart reviews are required for each SSI detected, with our approaches incorporating temporal data, only 3 chart reviews would be required for each SSI detected (as discussed in Chapter 2 Background and Significance). This would result in a tenfold increase in the number of cases that can be surveilled compared with manual surveillance alone. It would also further reduce the workload of infection prevention personnel around the nation by 720,000 hours annually (for calculation details see introduction) (on top of reductions conferred by contemporary automated surveillance methods). This demonstrated that, in relation to our central question, an automated surveillance system based on deep learning and multimodal data (static structured data, clinical text, and temporal data) can achieve better clinical utility than existing automated surveillance systems. Moreover, the addition of temporal data improved clinical utility over our models using static structured data and clinical text alone. Our models incorporating temporal data achieved a precision of 0.33 at 0.9 recall, as opposed to our models which did not incorporate temporal data that achieved a precision of 0.29 at 0.9 recall. Thus, addition of temporal data entails the reduction of 0.5 charts per SSI detected (3 charts reviews per case with temporal data vs. 3.5 chart reviews per case without). This implies that addition of temporal data into SSI prediction models would further reduce the burden on infection prevention staff over exclusion of temporal data.

4.5 Conclusion

My overall research question is: "Can an automated surveillance system based on multimodal data and deep learning outperform and have better domain adaptation capability than SOTA automated SSI surveillance approaches?" Our findings in this chapter suggest that, in relation to our central question, an automated surveillance system based on deep learning and multimodal data (static structured data, clinical text, and temporal data) can achieve better clinical utility than existing automated surveillance systems. However, thus far, we have not done a detailed, definitive analysis of the domain adaptation capability of our models.

In this chapter, we explored the performance conferred by training deep learning models on static structured data, clinical text, and temporal data from the EHR on SSI prediction. Our models incorporating static structured data, clinical text, and temporal data, using a BiLSTM layer to process clinical text, and using a LSTM layer to process temporal data achieved the highest overall performance (F1-score). This approach improved clinical utility over our best approach from Aim 3 which did not incorporate temporal data. It achieved the highest clinical utility of all of our approaches. However, even SSI prediction approaches which confer high performance on internal datasets can fail to domain adapt well to external datasets. However, it is crucial that SSI prediction approaches have good domain adaptation capability. This is because SSI prediction models will be expected to conduct surveillance on facilities around the nation as part of the goal

of having a nationwide automated surveillance system. Such facilities will likely include domains which may have insufficient labeled data in the training set. This is why an essential part of the development of automated SSI surveillance systems is a robust evaluation of their domain adaptation capability. In the next chapter, we plan to explore the domain adaptation capability of our models to different domains.

5. Aim 3: Assess the domain adaptation capability of surgical site infection surveillance approach

5.1 Introduction

My overall research question is, "*Can an automated surveillance system based on multimodal data and deep learning outperform and have better domain adaptation capability than state of the art automated surgical site infection (SSI) surveillance approaches?"* In this chapter the focus is on Aim 3 "Assess the domain adaptation capability of surgical site infection surveillance approach". We evaluate the ability of our best models from Chapter 3 and 4 to adapt to different domains. In this chapter, we study domain adaptation across different surgery types. We leverage data on almost 30,000 surgical cases from two healthcare facilities within University of Washington Medical Center and seven different surgery types in order to explore the domain adaptation capability of automated SSI surveillance algorithms. We define a domain as data from any set of one or more surgery types in our dataset. We divide the surgery types in our dataset into In-domain and Out-domain subsets. We measure the performance conferred by using different training sets (in-domain, out-domain, in-domain + out-domain), data representations, modeling strategies, and training strategies on performance on different domains.

This chapter is organized in the following way. We describe our methods for using different training sets and training strategies to train our best models from Aims 1 and 2. We evaluate model performances for different domains using precision, recall, and F1-score. We analyze the implication of the performance metrics in the context of our overall goal.

5.2 Methods

5.2.1 Clinical Cohort and SSI Definition

Details of the dataset we used for this task are described in Section 3.2.1. Table 5.1 presents the case breakdown by surgery type. We did not include vascular or cardiothoracic surgeries in our experimentation in this chapter, due to the low number of samples coming from these surgery types.

Procedure Class	Case Count	SSI Events (Rate)
Spine	3,724	196 (5.3%)
Orthopedic (non-spine)	2,739	105 (3.8%)
Neurosurgery (non-spine)	4,990	185 (3.7%)
General Surgery	14,200	585 (4.1%)
Gynecologic Surgery	3,211	120 (3.7%)
Total	28,864	1,191(4.1%)

Table 5.1: Surgery counts and infection rate

5.2.2 Structured Data Features

We used the same list of structured features and structured data representation approach described in Section 4.2.2.

5.2.3 Text Data Representation

We used the same clinical text representation approach described in Section 3.2.4, 3.2.5, and 3.2.6. We only used postoperative notes to construct our clinical text representation. See Section 3.3.6 for a detailed explanation on why we used this approach.

5.2.4 Temporal Data Representation

We used the same temporal data representation approach described in Section 4.2.4. We used a time window of 7 days before surgery to 90 days after surgery to extract temporal data. Please see Section 4.2.4 for a detailed explanation on why we chose this time window.

5.2.5 SSI Classification Approach

We input the temporal structured data representation from the Methods: Temporal Data Representation Section into a LSTM layer to derive a latent representation of our temporal data (Figure 5.1). We input our clinical text representation into a BiLSTM layer to derive a latent representation of our clinical text data. We chose to use the best clinical text and temporal representation strategies from Chapters 3 and 4 in this Chapter. We concatenated the latent temporal representation and latent text representation with our static structured data representation to train machine learning models.

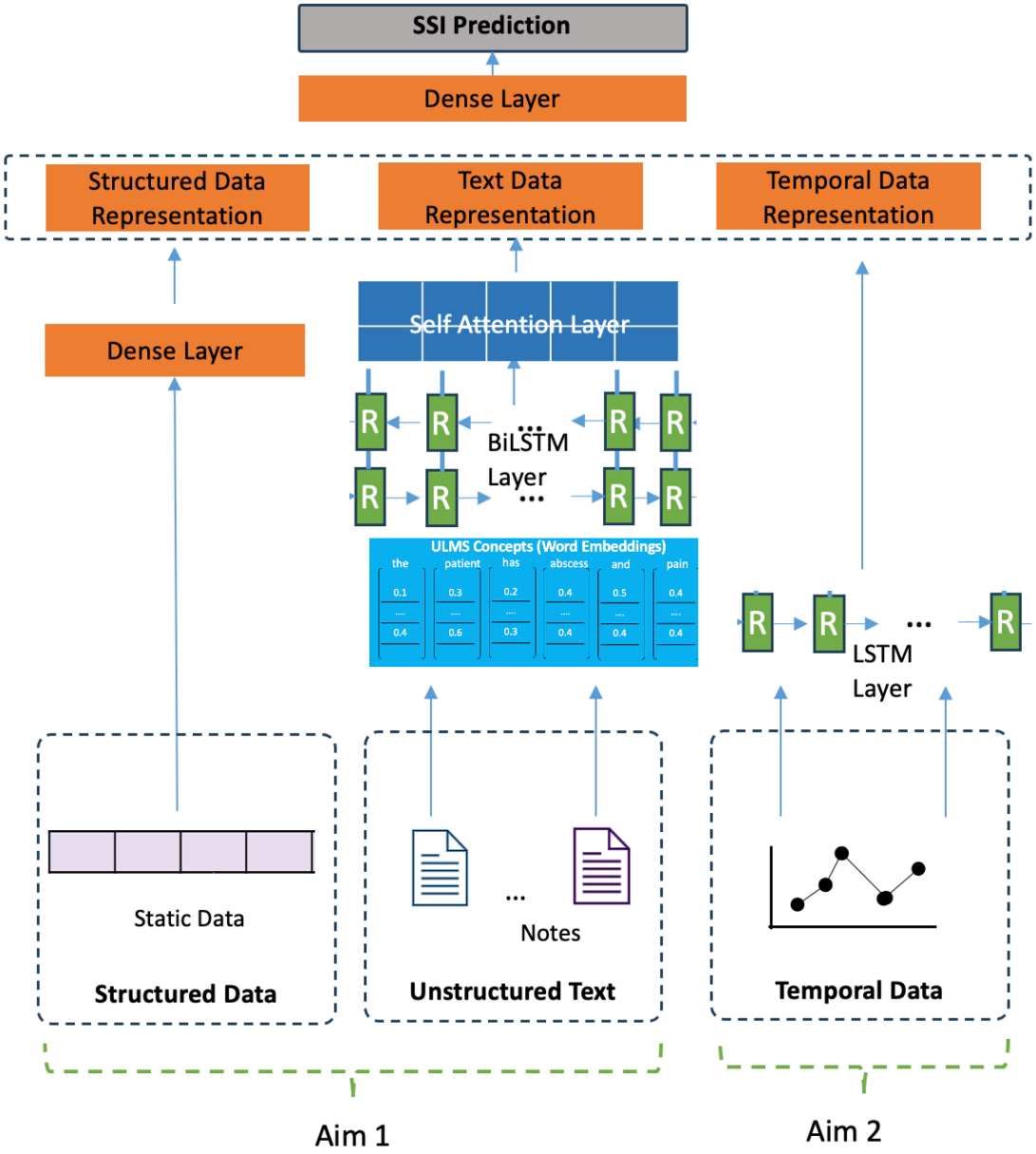


Figure 5.1: Figure depicting neural architectures used in our SSI classification approach

5.2.6 Model Development

Our hyperparameter tuning approach was similar to that outlined in Section 3.2.8.

5.2.7 Definition of Domains

The questions we wished to answer in this aim were:

1. What is the domain adaptation capability of our models on different domains of surgery types?
2. Does addition of out-domain data improve in-domain performance of our models?
3. What is the optimal training strategy to use with our models to maximize performance on each domain?

To answer these questions, the focus of our experiments in this aim was the evaluation of our post hoc SSI prediction approaches in Aims 1 and 2 on in-domain and out-domain data (we divided the surgery types in our dataset into In-domain and Out-domain subsets, explained in detail in Figure 5.3). Our approach in this aim was to (Figure 5.2):

1. Use our best data representation and modeling strategies from Chapter 3 and 4 (Chapter 3 Table 3.4 exp#5 (UMLS concept text representation + NN to construct text representation), Table 3.5 exp #9 (UMLS concept sentences text representation + BiLSTM to construct text representation), and Chapter 4 Table 4.2 exp#5 (UMLS concept sentences text representation + BiLSTM to construct text representation + LSTM to construct temporal representation).
2. Train our models on data from one set of surgery types (in-domain/out-domain/out-domain + in-domain).
3. Evaluate our models on data from another set of surgery types (in-domain).

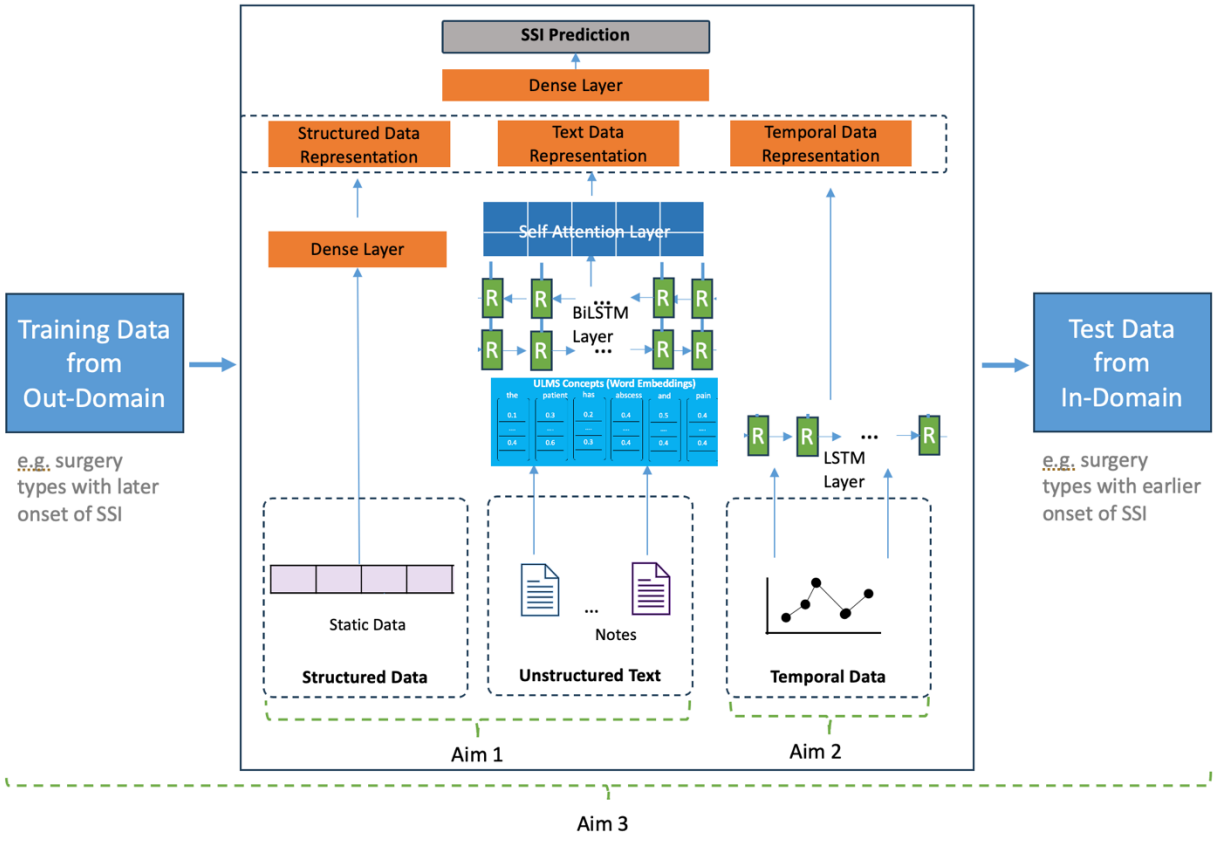


Figure 5.2: Figure depicting the focus of our domain adaptation experiments

**input in figure shows only out-domain, however in our domain adaptation experimentation, our input can be in-domain, out-domain, or in-domain + out-domain*

Our overall experimental approach is depicted in Figure 5.3. We divided the surgery types in our dataset into In-domain and Out-domain subsets, categorized data as (Figure 5.3):

- In-domain: EHR data from surgery types with certain commonalities relevant for post hoc SSI prediction.
- Out-domain: EHR data from surgery types without such commonalities.

This division was based on similarities, relevant for post hoc SSI prediction, that differentiate one subset (In-domain) from the other subset (Out-domain). These division criteria were based on literature review. Table 5.2 presents details of our approach to create the in-domain and out-domain surgery groups with clinical rationale.

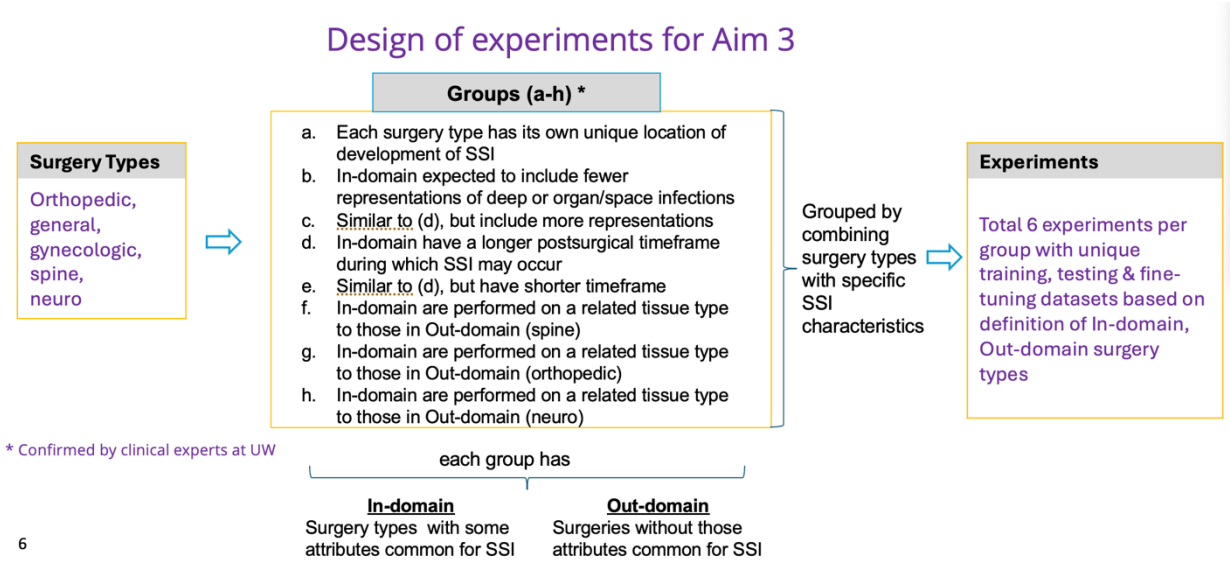


Figure 5.3: Figure depicting our overall experimental approach

Group #	Criteria for Selection*	Surgery Types in In-Domain	Surgery Types in Out-Domain	Clinical Importance
a.	Each surgery type has its own unique anatomic site(s) of SSI development, presentation of signs and symptoms of SSI, methods of diagnosis, and treatment administered for SSI.	Each individual surgery type	All other surgery types	To judge the value of training one model for all surgery types vs separate models for each surgery type.
b.	In-domain procedure groups expected to include fewer representations of deep or organ/space	Orthopedic surgery, general surgery,	Spine surgery, neurosurgery	To judge the value of training one model for all kinds of surgery types regardless of their relative

Group #	Criteria for Selection*	Surgery Types in In-Domain	Surgery Types in Out-Domain	Clinical Importance
	infections than those in Out-domain.	gynecologic surgery		representation of superficial, deep, and organ/space SSI vs training separate models represented for each of these kinds of surgery types.
c.	In-domain procedure groups expected to include more representations of deep or organ/space infections than those in Out-domain[44].	Spine surgery, neurosurgery	Orthopedic surgery, general surgery, gynecologic surgery	To judge the performance of our models on surgery types that have a relatively higher representation of deep or organ/space SSI.
d.	Procedure groups in In-domain have a longer postsurgical timeframe during which SSI may occur[45,46].	Orthopedic surgery, neurosurgery, spine surgery	General surgery, gynecologic surgery	To judge the performance of our models on surgery types in which SSI may develop later after surgery (compared to surgery types in In-domain subset).
e.	Procedure groups in In-domain have a shorter postsurgical timeframe during which SSI may occur.	General surgery, gynecologic surgery	Orthopedic surgery, neurosurgery, spine surgery	To judge the performance of our models on surgery types in which SSI may develop earlier after surgery (compared to surgery types in In-domain subset). To judge the value of training one model for all surgery types regardless of their timeframe of SSI occurrence vs training separate models, one for each kind of surgery types with a different timeframe of SSI occurrence.

Group #	Criteria for Selection*	Surgery Types in In-Domain	Surgery Types in Out-Domain	Clinical Importance
f.	Those in In-domain are performed on a related tissue type to those in Out-domain.	Neurosurgery, orthopedic surgery	Spine surgery	To judge the performance of our model trained on two procedures involving bone to a third procedure involving a related tissue type (bone)
g.	Those in In-domain are performed on a related tissue type to those in Out-domain.	Neurosurgery, spine surgery	Orthopedic surgery	To judge the performance of our model trained on two procedures involving bone to a third procedure involving a related tissue type (bone)
h.	Those in In-domain are performed on a related tissue type to those in Out-domain.	Spine surgery, orthopedic surgery	Neurosurgery	To judge the performance of our model trained on two procedures involving bone to a third procedure involving a related tissue type (bone)

Table 5.2: Approach to categorize surgery types into In-Domain and Out-domain
** Commonalities/Differences between In-domain and Out-domain data*

5.2.8 Experimentation on Domains Within In-Domain and Out-Domain

For each group in Table 5.2, we conducted a subset of 6 experiments. These experiments differed by (Figure 5.3): (1) which subset was used for training (Out-domain, In-domain or both), (2) the number of training steps, and (3) the sequence of training steps.

We performed 6 experiments, and in each, we utilized a different training strategy. For all of our experiments, we evaluated performance on all surgery types in the In-domain subset after training (Table 5.3):

1. Experiment #1 (Training Strategy: Target)[47]: In this strategy, the model is trained in a single iteration on all surgery types in the In-domain subset. The purpose of this experiment was to determine the performance of our models on surgery types with training data.
2. Experiment #2 (Training Strategy: Concurrent I)[47]: In this strategy, the model is trained in a single iteration on all surgery types in the Out-domain subset. The purpose of this experiment was to determine the performance of our models on surgery types without training data.

3. Experiment #3 (Training Strategy: Sequential)[47] : In this strategy, the model is first trained on one surgery type in the Out-domain subset and then sequentially refined on each of the other surgery types in the Out-domain subset. The purpose of this experiment was to determine the effect of sequential vs concurrent training strategies on performance of our models on surgery types without training data.
4. Experiment #4 (Training Strategy: Concurrent II)[47]: In this strategy, the model is trained in a single iteration on all surgery types in the In-domain and Out-domain subset. The purpose of this experiment was to determine the effect of adding out-domain data and using a concurrent training strategy on in-domain performance.
5. Experiment #5 (Training Strategy: Fine-tuning with concurrent)[47]: Same as Concurrent I (Experiment #2) with an additional subsequent step during which in-domain data is used to fine-tune the model. The purpose of this experiment was to determine the effect of adding out-domain data and using a Fine-tuning with concurrent training strategy on in-domain performance.
6. Experiment #6 (Training Strategy: Fine-tuning with sequential)[47]: Same as Sequential (Experiment #3) with an additional final step during which in-domain data is used to fine-tune the model. The purpose of this experiment was to determine the effect of adding out-domain data and using a Fine-tuning from sequential training strategy (Experiment #3) on in-domain performance.

For each group, we only performed a subset of these experiments. These experiments allowed us to answer the questions we asked in this aim (see Chapter 5 Methods: Definition of Domains Section). Table 5.3 delineates the takeaways from each of our experiments.

Exp #	Training Strategy	Training Set	Testing Set	Takeaway
1	Target	In-domain	In-domain	The performance of the model on surgery types with training data.
2	Concurrent I	Out-domain (all surgery types)	In-domain	The performance of the model on surgery types without training data.
3	Sequential	Out-domain (surgery type #1) -> Out-domain surgery type #2) …	In-domain	The effect of using sequential vs concurrent training strategies on in-domain performance when training on out-domain data.
4	Concurrent II	In-domain + Out-domain (all surgery types)	In-domain	The effect of adding out-domain data and using a concurrent training strategy on in-domain performance.

Exp #	Training Strategy	Training Set	Testing Set	Takeaway
5	Fine-tuning from concurrent	Out-domain (all surgery types) -> In-domain	In-domain	The effect of adding out-domain data and using a Fine-tuning from concurrent training strategy on in-domain performance.
6	Fine-tuning from sequential	Out-domain (surgery type #1) -> Out-domain surgery type #2) … -> In-domain	In-domain	The effect of adding out-domain data and using a Fine-tuning from sequential training strategy on in-domain performance.

Table 5.3: Outline of our experimentation for Aim 3

** -> Indicates the next iteration of training.*

Our training dataset and testing dataset sizes depended on the number of surgical cases in each subset. In some experiments, we varied the relative proportions of in-domain and out-domain data in the training set. To do this, we randomly sampled n # of cases from the In-domain or Out-domain subset.

5.2.9 Model Evaluation

Our model evaluation approach was similar to that outlined in Section 3.2.9. The difference between the approach described in that chapter and our evaluation approach in this chapter is, in this chapter, we did not evaluate the clinical utility of our models.

5.3 Results

5.3.1 Performance on Individual Surgery Types – Group a Exp #4 Concurrent II training Strategy

First, we assessed the overall performance (F1-score) of our models on individual surgery types. This was to determine if particular feature sets or data representations conferred better performance for particular surgery types. In these experiments, we evaluated the performance of our best models from Aims 1 and 2 (see Chapter 3 Table 3.4 exp#5, Chapter 3 Table 3.5 exp#9 and Chapter 4 Table 4.2 exp#5) at the level of individual surgery types (Table 5.4). For these experiments, we used a Concurrent II training strategy (Table 5.3). P-values calculated based on T-test for the means of two independent samples (see Chapter 5 Methods: Model Evaluation Section). They are based on comparing the performance of the highest performing feature set to the performance of the second highest performing feature set (of columns #2-4) for each surgery type.

Surgery Type	F1 for Structured + CUI	F1 for Structured + CUI Sentences	F1 for Structured + CUI Sentences + Temporal	Adjusted p-value*
General	0.53(0.53,0.54)	0.55(0.55,0.55)	**0.57(0.57,0.58)**	0.027 **
Gynecological	0.57(0.54,0.60)	0.65(0.63,0.67)	**0.69(0.67,0.71)**	$3.4*10^{-9}$ **
Neurological	0.77(0.77,0.78)	0.76(0.75,0.78)	0.76(0.73,0.79)	0.60
Spine	0.78(0.77,0.79)	**0.82(0.82,0.82)**	**0.82(0.81,0.82)**	$1.8*10^{-4}$ **
Orthopedic	0.82(0.79,0.86)	0.80(0.80,0.81)	0.80(0.80,0.81)	0.084

Table 5.4: Performance conferred by different feature sets, modeling strategies, and text representations on the task of SSI prediction

bold = highest value achieved on surgery type

** p-values calculated based on comparing the performance of the highest performing feature set to the performance of the second highest performing feature set (of columns #2-5) for each surgery type*

*** = significant p-value*

Our results showed that there were differences in the overall performance conferred by the different approaches we used (ANOVA test p-value: general surgery: $7.1*10^{-6}$, gynecological surgery: $1.5*10^{-12}$, spine surgery: $2.3*10^{-5}$). The following were our conclusions from our experiments (Table 5.4):

1. Inclusion of contextual information from text improves overall performance for surgery types with missing static structured data (e.g., general surgery, gynecological surgery) and spine surgeries (Table 5.4 row #1 col #2 vs row #1 col #1, row #2 col #2 vs row #2 col #1, row #4 col #2 vs row #4 col #1 - general surgery: with context F1-score: 0.55 without context F1-score: 0.53, gynecological surgery: with context F1-score: 0.65 without context F1-score: 0.57, spine surgery: with context F1-score: 0.82 without context F1-score: 0.78).
2. Inclusion of temporal data further improves performance for surgery types with missing static structured data (e.g., general surgery, gynecological surgery) (Table 5.4 row #1 col #3 vs row #1 col #2, row #2 col #3 vs row #2 col #2 - general surgery: with temporal data F1-score: 0.57 without temporal data F1-score: 0.55, gynecological surgery: with temporal data F1-score: 0.69 without temporal data F1-score: 0.65). This could be because, for these surgery types, information in clinical text and temporal data fill in the information gaps left by missing static structured data.
3. For neurosurgery and orthopedic surgery, neither inclusion of contextual information from text nor the incorporation of temporal laboratory data has an effect on performance. Further analysis showed that the reasons for this differ for orthopedic and neurosurgeries. In the case of orthopedic surgeries, use of structured data alone as inputs already achieves high performance. Thus, different text representations (e.g., ones which do or do not capture contextual information from text) or addition of temporal data do not have a significant effect on performance. In the case of neurosurgeries, our text representation (bullet #2b in Methods: Text Data Representation) resulted in text samples which contained many details irrelevant to SSI (e.g., details on rehabilitation after surgery, neurological health of the patient). Additionally, the presentation and management of SSI after neurosurgery is unique (e.g., cerebrospinal fluid leak, infected subdural hematoma). Thus, signals of SSI

in contextual information from clinical text for neurosurgeries were different from general textual signals of SSI. Thus, our models did not recognize the former signals.

5.3.2 Effect of Cross-Training With Data From All Surgery Types – Group a Exp#1 Target Training Strategy and Exp #4 Concurrent II Training Strategy

Next, we wished to determine the effect of jointly training on all surgery types vs training separate models for each surgery type on overall performance. Thus, in our next set of experiments, we trained our best model from Aims 1 and 2 (see Chapter 4 Table 4.2 exp#5) on portions of our training and development sets composed of one surgery type (Table 5.5). We then evaluated our models on portions of our test set composed of the same surgery type. We did this for each surgery type. This allowed us to measure the performance of surgery-type specific models. It also allowed us to compare this performance to the performance of a model trained on all surgery types. Additionally, we trained our models on the full set of surgery types (a Concurrent II training strategy) and evaluated them on testing sets composed of individual surgery types.

Exp #	Surgery Type	Train Source	P	R	F1	Adjusted p-value*	p-value compared against
1		Ortho	0.65(0.65,0.65)	0.90(0.90,0.90)	0.75(0.75,0.75)		
2	Ortho	All	0.72(0.70,0.73)	**0.95(0.91,1)**	**0.80(0.80,0.81)**	$1.3*10^{-7}$ **	Exp #1
3		Spine	0.70(0.70,0.70)	0.85(0.85,0.85)	0.76(0.76,0.76)		
4	Spine	All	0.74(0.73,0.74)	**0.90(0.90,0.90)**	**0.82(0.81,0.82)**	$1.8*10^{-6}$ **	Exp #3
5		Neurological	0.81(0.79,0.84)	0.72(0.70,0.75)	0.77(0.76,0.78)		
6	Neurological	All	0.70(0.69,0.70)	**0.84(0.78,0.89)**	0.76(0.73,0.79)	0.06	Exp #5
7		Gensurg	0.68(0.67,0.69)	0.45(0.43,0.47)	0.54(0.53,0.55)		
8	Gensurg	All	0.61(0.61,0.62)	**0.54(0.53,0.56)**	**0.57(0.57,0.58)**	$1.7*10^{-4}$ **	Exp #7
9		Gyn	0.35(0.29,0.38)	0.57(0.44,0.70)	0.43(0.41,0.44)		
10	Gyn	All	0.71(0.71,0.71)	**0.67(0.63,0.71)**	**0.69(0.67,0.71)**	$2.9*10^{-10}$ **	Exp #9

Table 5.5: Performance of neural network models trained on each surgery type versus the entire training set on each surgery type

bold = highest value achieved on performance metrics
column with metric used for statistical testing is highlighted in grey
*** = significant p-value*

Our results showed that addition of out-domain data (from other surgery types) improves overall performance (F1-score) for all surgery types except neurosurgeries. This may be due to the fact that textual signals of SSI in clinical notes of patients undergoing neurosurgeries are unique (see Chapter 5 Results: Performance on Individual Surgery Types Section). This may indicate that addition of data from other domains close to the target domain improves performance on the target domain more than addition of data from domains distant from the target domain (e.g., addition of data from other surgery types helps for orthopedic surgery and not for neurosurgery as other surgery types are closer to orthopedic surgeries than neurosurgeries).

5.3.3 Effect of Using Different Training Strategies on Performance on Surgeries in Different Anatomical Locations Than the Ones in the Training Set – Group f,g,h Exp#1 Target Training Strategy and Exp #4 Concurrent II Training Strategy

From our experiments, it appeared that neurosurgeries were a domain distant from other surgery types. Signals of SSI in clinical text of neurosurgery cases were distinct from signals of SSI in clinical text of cases from other surgery types. Furthermore, the addition of out-domain data did not improve performance for neurosurgeries. Thus, we wished to specifically look at the impact of varying the domains included in the training source on performance on neurosurgeries. We also wished to determine if training only on data from domains close to neurosurgeries conferred a benefit on performance on this surgery type.

We focused on orthopedic surgeries, spine surgeries, and neurosurgeries. This is because all of these three surgery types involve the same tissue type (bone). Therefore, looking at the domain adaptation capability of models trained on combinations of these surgery types would reveal insights on the domain adaptation capability of our models to procedures involving tissue types related to those in the training set. Additionally, orthopedic and spine surgeries are closer domains to neurosurgery than the other surgery types (as they all involve the bone tissue). Hence, focusing on orthopedic surgeries, spine surgeries, and neurosurgeries allowed us to explore whether training on these surgery types conferred a performance benefit on prediction of SSI for neurosurgery cases.

In our experimentation, we trained our best performing model from our preliminary experimentation on Aims 1 and 2 (see Chapter 4 Table 4.2 exp#5) on different combinations of neurosurgeries, orthopedic surgeries, spine surgeries, and other surgery types (Target or Concurrent II training strategies). We then evaluated our models on testing sets composed of data from each of the three surgery types (Table 5.6).

Exp #	Surgery Type	Train Source	P	R	F1	Adjusted p-value* / Compared against
1		Ortho	0.63(0.63,0.63)	0.90(0.90,0.90)	0.75(0.75,0.75)	
2	Ortho	Neurological, Orthopedic, Spine	0.77(0.77,0.77)	0.81(0.81,0.81)	0.79(0.79,0.79)	$8.5*10^{-9}$ ** / Exp #1
3		All	0.72(0.70,0.73)	**0.95(0.91,1)**	**0.80(0.80,0.81)**	$7.1*10^{-9}$ ** / Exp #2
4		Spine	0.70(0.70,0.70)	0.85(0.85,0.85)	0.77(0.77,0.77)	
5	Spine	Neurological, Orthopedic, Spine	0.72(0.71,0.72)	0.87(0.87,0.87)	0.78(0.78,0.79)	$4.3*10^{-5}$ ** / Exp #5
6		All	0.74(0.73,0.74)	**0.90(0.90,0.90)**	**0.82(0.81,0.82)**	$3.2*10^{-7}$ ** / Exp #6
7		Neurological	0.58(0.56,0.59)	0.72(0.70,0.75)	0.77(0.76,0.78)	
8	Neurological	Neurological, Orthopedic, Spine	0.74(0.74,0.74)	**0.84(0.84,0.84)**	**0.78(0.78,0.78)**	$1.4*10^{-5}$ ** / Exp #9
9		All	0.70(0.69,0.70)	**0.84(0.78,0.89)**	0.76(0.73,0.79)	0.12 / Exp #10

Table 5.6: Performance of neural network models trained on each surgery type versus the entire training set on each surgery type

bold = highest value achieved on performance metrics for each surgery type
column with metric used for statistical testing is highlighted in grey
*** = significant p-value*

Our results showed that there were differences in the overall performance conferred by the different approaches we used (ANOVA test p-value: orthopedic surgery: $3.4*10^{-14}$, spine surgery: $2.0*10^{-11}$, neurosurgery: 0.0013). From our results, we can conclude that:

1. The addition of out-domain data (a Concurrent II training strategy) improves in-domain performance for orthopedic, spine surgeries, neurosurgeries (Table 5.6 exp #2 vs exp #1, exp #5 vs exp #4, exp #8 vs exp #7 - orthopedic surgery: with out-domain data F1-score: 0.79 without out-domain data: 0.75, spine surgery: without-domain data F1-score: 0.78, without out-domain data: 0.77, neurosurgery: with out-domain data F1-score: 0.78, without out-domain data: 0.77). This is in accordance with our results in the Results: Effect of Cross-Training With Data From All Surgery Types Section.
2. Even addition of data (a Concurrent II training strategy) from domains distant from domains within in-domain improves or at least does not deteriorate in-domain performance (Table 5.6 exp #3 vs exp #2, exp #6 vs exp #5, exp #9 vs exp #8 – orthopedic surgery: with close and distant out-domain data F1-score: 0.80 with only close out-domain data: 0.79, spine surgery: with close and distant out-domain data F1-score: 0.82, with only close out-domain data: 0.78, neurosurgery: with close and distant out-domain data F1-score: 0.76, with only close out-domain data: 0.78). This strengthens our claim in the Chapter 5 Results:

Effect of Cross-Training With Data From All Surgery Types Section. The claim was: data from widely different domains can be leveraged to potentially improve performance on each domain by jointly training one model on all domains.

3. Together, these results suggest that addition of out-domain data improves, or at least does not lower, in-domain performance. Adding data from domains close to the target domain improves performance more than adding data from domains distant from the target domain.

5.3.4 Incremental gains of adding more out-of-domain samples – Group a Exp#1 Target Training Strategy, Exp #2 Concurrent I Training Strategy, Exp #4 Concurrent II Training Strategy

Our experiments showed that the addition of more out-domain samples improves performance. It also showed that addition of data from domains closer to the target domain confers a greater benefit to performance than addition of data from domains distant from the target domain (e.g., addition of data from other surgery types helps for orthopedic surgery and not for neurosurgery as other surgery types are closer to orthopedic surgeries than neurosurgeries). We wished to gain a more granular understanding of how the addition of more samples from out-domain improves performance and the magnitude of the performance gains. To explore this, we changed the relative proportions of in-domain to out-domain data and the number of total samples and determined the effect on performance (Target, Concurrent I, and Concurrent II training strategies in Table 5.3). We performed these experiments considering data from two surgery types as in-domain: spine (Table 5.8) and general surgery (Table 5.7). For each case, we considered data from all other surgery types as out-domain. We chose general surgeries as they represent a surgery type with a high percentage of missing structured data. We chose spine surgeries as they represent a surgery type with a low percentage of missing structured data. To avoid the facility from which the data came (UWMC or HMC) as a potential confounder, we only considered general surgery cases from UWMC.

Tables 5.7 and 5.8 present the results of our experiments in this section. The performance is broken down by total training set size. It is also stratified by percentage of in-domain data (data from the surgery type in the evaluation set) in the training set. P-values in the table are calculated based on T-test for the means of two independent samples (see Chapter 5 Methods: Model Evaluation Section). They are based on comparing highest to second highest performance within each total sample size category.

% Out-domain	Num Out-domain	F1	Adjusted p-value*
5000 in-domain samples			
0	0	0.53(0.53,0.53)	
50	5000	**0.57(0.57,0.57)**	$3.5*10^{-8}$ **
2500 in-domain samples			
0	0	0.50(0.50,0.50)	
50	2500	**0.55(0.55,0.56)**	$2.9*10^{-10}$ **
75	7500	**0.55(0.55,0.55)**	$4.6*10^{-8}$ **
1250 in-domain samples			
50	1250	0.38(0.38,0.38)	
75	3750	**0.54(0.54,0.54)**	$4.1*10^{-10}$ **

Table 5.7: Incremental gains of adding more out-domain samples on model performance on General Surgery

bold = best performance for each set of experiments with the same total samples size is bolded
column with metric used for statistical testing is highlighted in grey
*** = significant p-value*

% Out-domain	Num Out-domain	F1	Adjusted p-value*
3000 in-domain samples			
0	0	0.76(0.76,0.76)	
50	3000	0.80(0.80,0.80)	
75	9000	**0.81(0.81,0.81)**	$5.2*10^{-4}$ **
1500 in-domain samples			
0	0	0.76(0.76,0.76)	
50	1500	0.78(0.78,0.78)	
75	4500	**0.81(0.81,0.81)**	$9.3*10^{-7}$ **
750 in-domain samples			
0	0	0.74(0.74,0.74)	
50	750	**0.79(0.79,0.79)**	
75	2250	0.78(0.78,0.78)	$1.4*10^{-3}$ **

Table 5.8: Incremental gains of adding more out-domain samples on model performance on Spine Surgery

bold = best performance for each set of experiments with the same total in-domain sample size is bolded
column with metric used for statistical testing is highlighted in grey
*** = significant p-value*

Our results showed that there were differences in the overall performance (F1-score) conferred by the different approaches we used (ANOVA test p-value: general surgery 10K samples: $2.8*10^{-18}$, general surgery 5K samples: $3.5*10^{-8}$, general surgery 2.5K samples: $1.8*10^{-20}$, spine surgery 6K samples: $7.7*10^{-9}$, spine surgery 3K samples: $1.3*10^{-9}$, spine surgery 1.5K samples: $8.8*10^{-25}$, spine surgery 0.75K samples: $2.4*10^{-21}$). Our results also show that (Table 5.7 and 5.8):

1. Addition of out-domain data improves performance, especially for low-resource surgery types (surgery types with little data in the training set) (Table 5.7 and Table 5.8).
2. Performance incrementally improves as more out-domain data is added, provided there are a minimum number of in-domain samples in the training set (Table 5.7 and 5.8).
3. When the number of in-domain samples is low, performance improves with the addition of out-domain data up to a certain point, then stays the same (Table 5.7 and 5.8). A possible explanation for this is that at this point, the percentage of in-domain data in the training set becomes low and this detracts from performance on in-domain data, reversing the benefits of adding out-domain data.
4. Addition of out-domain data helps more for surgery types with a higher percentage of missing structured data and with a low number of training samples (Table 5.7 vs Table 5.8).
5. For a fixed set of training samples, training with a mix of in-domain and out-domain data confers better performance than training with in-domain data alone (Table 5.7 and 5.8).

5.3.5 Effect of Using Different Training Strategies on Performance on Individual Surgery Types – Group a Exp #4 Concurrent II Training Strategy, Exp #5 Fine-tuning from concurrent Training Strategy, Exp #6 Fine-tuning from sequential Training Strategy

We next explored whether training the model on one set of domains (surgery types) and then fine tuning on another set of domains (Fine-tuning from concurrent or Fine-tuning from sequential training strategies in Table 5.3) improves performance on individual domains (surgery types). This would allow us to explore the optimal sequence of training and/or fine-tuning steps to use for each surgery type in our dataset (Table 5.9). P-values in the table are calculated based on T-test for the means of two independent samples (see Chapter 5 Methods: Model Evaluation Section). They are based on comparison between best strategy involving fine-tuning with Concurrent II strategy.

Surgery Type	Concurrent II	Fine-tuning from concurrent	Fine-tuning from sequential	Adjusted p-value*
General	**0.57(0.57,0.58)**	0.55(0.53,0.57)	0.54(0.53,0.56)	$4*10^{-6}$ **
Gynecological	**0.69(0.67,0.71)**	0.54(0.51,0.56)	0.54(0.53,0.54)	$7.1*10^{-8}$ **
Neurological	0.76(0.73,0.79)	0.75(0.74,0.76)	0.74(0.74,0.74)	0.75
Spine	**0.82(0.81,0.82)**	0.79(0.78,0.79)	0.76(0.76,0.76)	$4.5*10^{-7}$ **
Orthopedic	0.80(0.80,0.81)	0.84(0.83,0.85)	**0.85(0.84,0.85)**	$1.7*10^{-6}$ **

Table 5.9: Performance conferred by different training strategies on different surgery types in our dataset

best performance for surgery type is bolded
*** = significant p-value*

Our results showed that there were differences in the overall performance (F1-score) conferred by the different approaches we used (ANOVA test p-value: general surgery: $8.3*10^{-8}$, gynecological surgery: $2*10^{-11}$, spine surgery: $8.4*10^{-12}$, orthopedic surgery: $2.9*10^{-8}$). From our results, we can also draw the following conclusions (Table 5.9):

1. Fine-tuning models on a target domain (a surgery type) deteriorates performance (Fine-tuning from concurrent or Fine-tuning from sequential training strategies in Table 5.3) for most surgery types (general surgery: concurrent best F1-score: 0.57 fine-tuning best F1-score: 0.55, gynecological surgery: concurrent best F1-score: 0.69 fine-tuning best F1-score: 0.54, spine surgery: concurrent best F1-score: 0.82 fine-tuning best F1-score: 0.79). In fact, fine-tuning reverses much of the gains derived from cross-training with data from all surgery types. Fine-tuning may cause our model to focus on more procedure-specific signals of SSI and abandon using general signals of SSI applicable to all surgery types. This may be limiting the ability of model to detect all SSI cases in the test set.
2. Fine-tuning (Fine-tuning from concurrent or Fine-tuning from sequential training strategies in Table 5.3) helps overall performance (F1-score) for surgeries which have a low percentage of missing structured data (e.g., orthopedic) (orthopedic surgery: concurrent best F1-score: 0.80 fine-tuning best F1-score: 0.85) (Table 5.9).
3. There is no definitive pattern as to whether Fine-tuning from concurrent or Fine-tuning from sequential training strategies are better.

5.3.6 Domain adaptation capability of our models to surgery types with a longer or shorter postsurgical timeframe during which SSI may occur – Group d,e Exp#1 Target Training Strategy, Exp #2 Concurrent I Training Strategy, Exp #4 Concurrent II Training Strategy

We studied the domain adaptation capability of our models to other domains such as: sets of surgery types for which SSI may occur during a longer or shorter postsurgical timeframe. The latter set more commonly includes procedures involving implanted prosthetic materials. Domains with a shorter timeframe of SSI development will have different signs and symptoms (as the SSI which occur are different kinds of infections), different medications prescribed (as SSI occur in different anatomical regions), and different periods of laboratory value changes. Thus, the patterns relevant to SSI in these two domains are different. There may be scenarios in which one of these domains does not have much data in the training set. Hence, we wished to explore whether our models can adapt to either of these domains.

Thus, we trained our models on either the domain encompassing surgery types for which SSI occur during a longer postsurgical timeframe, domains for which SSI occur during a shorter postsurgical timeframe, or a combination of both (Target, Concurrent I, and Concurrent II training strategies in Table 5.3). We then evaluated our models on both domains (Table 5.10).

Exp #	Surgery Type	Train Source	P	R	F1	Adjusted p-value* / Compared against
1		Long	0.55(0.55,0.55)	0.44(0.44,0.44)	0.49(0.49,0.49)	
2	Short	Short	0.63(0.63,0.63)	0.51(0.51,0.51)	0.56(0.56,0.56)	
3		Short + Long	0.64(0.64,0.64)	**0.51(0.51,0.51)**	**0.57(0.57,0.57)**	$1.0*10^{-3}$ ** / Exp #4
4		Short	0.67(0.67,0.67)	0.88(0.88,0.88)	0.77(0.77,0.77)	
5	Long	Long	0.68(0.67,0.69)	0.87(0.87,0.87)	0.77(0.76,0.77)	
6		Short + Long	0.74(0.74,0.74)	**0.87(0.87,0.87)**	**0.79(0.79,0.79)**	$1.3*10^{-5}$ ** / Exp #9

Table 5.10: Incremental gains of adding more out-domain samples on model performance on surgery types with shorter/longer timeframe of SSI occurrence

best performance for each category of evaluation set composition and % of in-domain data in training set is bolded
column with metric used for statistical testing is highlighted in grey
*** = significant p-value*

See the next section for our conclusions from our results in Table 5.10.

5.3.7 Domain adaptation capability of our models to surgery types expected to include fewer or more representations of deep or organ/space infections than is typical – Group b,c Exp#1 Target Training Strategy, Exp #2 Concurrent I Training Strategy, Exp #4 Concurrent II Training Strategy

We were also interested in the domain adaptation capabilities of our models to surgery types which were expected to include fewer or more representations of deep or organ/space infections than is typical. Surgeries from surgery types with higher relative proportions of deep and organ/space SSI have a different set of signs and symptoms, diagnostic tests (to diagnose SSI), and treatments. Thus, as in the previous section, we wished to explore whether our models can domain adapt to either of these domains (Table 5.11).

Exp #	Surgery Type	Train Source	P	R	F1	Adjusted p-value* /Compared against
1		Deep/OS	0.54(0.54,0.54)*	0.52(0.52,0.52)	0.53(0.53,0.53)	
2	Superficial	Superficial	0.69(0.69,0.69)	0.50(0.50,0.50)	0.58(0.58,0.58)	
3		Superficial Deep/OS	0.65(0.65,0.65)	**0.54(0.54,0.54)**	**0.59(0.59,0.59)**	$7.8*10^{-5}$ ** / Exp #4
4		Superficial	0.71(0.71,0.71)	0.81(0.81,0.81)	0.76(0.76,0.76)	
5	Deep/OS	Deep/OS	0.74(0.74,0.74)	0.81(0.81,0.81)	0.77(0.77,0.77)	
6		Superficial Deep/OS	0.75(0.75,0.75)	**0.83(0.83,0.83)**	**0.79(0.79,0.79)**	$4*10^{-5}$ ** / Exp #9

Table 5.11: Incremental gains of adding more out-domain samples on model performance on surgery types in which deep or organ-space SSI may represent a higher or lower percentage of total SSIs

best performance for each category of evaluation set composition and % of in-domain data in training set is bolded
column with metric used for statistical testing is highlighted in grey
*** = significant p-value*

Our results showed that, overall, there were differences in the overall performance (F1-score) conferred by the different approaches we used (ANOVA test p-value: long postsurgical timeframe of SSI development: $9.3*10^{-12}$, short postsurgical timeframe of SSI development: $4.5*10^{-13}$, relatively higher percentage of deep/organ space SSI: $1.1*10^{-21}$, relatively lower percentage of deep/organ space SSI: $5.1*10^{-16}$). Our conclusions from our results in Table 5.10 and Table 5.11 are:

1. Our models can domain adapt to domains which have a shorter or longer timeframe of SSI development (Table 5.10 exp#1 and exp#4).
2. Our models can domain adapt to domains which may have lower or higher proportions of deep or organ space SSI (Table 5.11 exp#1 and exp#4).
3. Addition of out-domain data contributes to in-domain performance for these four domains (Table 5.10 exp #3 vs exp#2, exp #6 vs exp #5, Table 5.11 exp #3 vs exp #2, exp #6 vs exp #5).

5.3.8 Summary of findings on domain adaptation experiments

Our experiments (Table 5.3) with different groups in Table 5.2 allowed us to determine the optimal training strategy to use for obtaining best performance on each domain (for further details please see Chapter 5 Methods: Experimentation on Domains Within In-Domain and Out-Domain Section). Table 5.12 summarizes our findings.

Finding #	Training Strategy	Group	Finding	Supporting Section/bullet
1	Concurrent II	All domains	↑ Addition of contextual information from clinical text and addition of temporal data helps	Section 5.3.1 bullet # 1,2,3

Finding #	Training Strategy	Group	Finding	Supporting Section/bullet
			performance on all domains except those with a low percentage of missing structured data and those distant from domains within the training set.	
2	Concurrent II	All domains	↑ Addition of data from other domains helps overall performance for most domains, especially low-resource ones, and does not lower performance on others. Performance is benefitted more the closer domains in the training set are to the target domain and the greater the size of the out-domain data.	Section 5.3.2, Section 5.3.3 bullet #1,2,3 Section 5.3.7 bullet #1,2,3
3	Concurrent II	All domains	↑ Training one model with data from all domains allows one to leverage data from multiple domains to improve performance on most domains.	Section 5.3.2, Section 5.3.4 bullet #1,2 Section 5.3.7 bullet #1,2,3
7	Target	Low-resource domains	↓ As the number of in-domain samples falls, performance drops sharply.	Section 5.3.4 bullet #4
8	Concurrent II	Low-resource domains	↑ Addition of out-domain data confers a greater benefit to performance on low-resource domains.	Section 5.3.4 bullet #4
9	Concurrent II	Domains with missing structured data	↑ Addition of out-domain data confers a greater benefit to performance on domains with missing structured data.	Section 5.3.2, Section 5.3.4 bullet #4
10	Concurrent II	All domains	↑ Training with a mix of in-domain and out-domain data improves performance over training with just in-domain data.	Section 5.3.4 bullet #5
11	Concurrent I	All domains	↑ Provided there are enough training samples, our models can domain adapt to spine surgeries,	Section 5.3.3 bullet #1,2,3, Section 5.3.7 bullet #1,2,3

Finding #	Training Strategy	Group	Finding	Supporting Section/bullet
			neurosurgeries, orthopedic surgeries, surgeries with a relatively higher proportion of deep/organ space SSI, surgeries without a relatively higher proportion of deep/organ space SSI, surgery types for which SSI occurs during a shorter postsurgical timeframe, and surgery types for which SSI occurs during a longer postsurgical timeframe.	
13	Fine-tuning from concurrent, Fine-tuning from sequential	Domains with missing structured data	↓Domains with missing structured data do not benefit from fine-tuning. In fact, fine-tuning reverses much of the benefit of cross-training with data from other domains for these domains.	Section 5.3.5 bullet #1
14	Fine-tuning from concurrent, Fine-tuning from sequential	Domains with a low percentage of missing structured data	↑Domains with a low percentage of missing structured data may benefit from fine-tuning. It is not clear which fine-tuning strategy (fine-tuning from concurrent, fine-tuning from sequential) works better.	Section 5.3.5 bullet #2,3

Table 5.12: Summary of our findings on different groups and training strategies
refer to Table 2 for definition of different training strategies
refer to Table 3.1 for definition of groups
↑ *=findings on factors which increase performance*
↓*=findings on factors which decrease performance*

Based on our findings, we make the following recommendations on the best training strategy to use in different scenarios to optimize performance on different domains (Table 5.13).

Scena-rio #	In-domain dataset size	Out-domain dataset size	In-domain description	Recommended feature set	Recommended training strategy
1	large	large	Domains with missing structured data	Structured, text (with context*), temporal	Concurrent II
2	large	large	Domains without missing structured data	Structured, text (with context), temporal	Fine-tuning from concurrent, Fine-tuning from sequential

3	large	small	All domains	Not enough information	Concurrent II
4	small	large	All domains	Structured, text (with context), temporal	Concurrent II
5	small	small	All domains	Not enough information	If out-domain dataset size >> in-domain dataset size, Concurrent II Else try to get more out-domain data; if this is possible, Concurrent II; if this is not possible Target
6	none	large	Domains with missing structured data	Structured, text (with context)	Concurrent I
7	none	large	Domains without missing structured data	Structured, text (with context), temporal	Concurrent I

Table 5.13: Summary of our findings on different groups and training strategies
see Table 5.3 for descriptions of various training strategies
**with context – with incorporation of contextual information from text*

5.4 Discussion

In this aim, we analyzed data from the EHR of almost 30,000 surgical cases from two different healthcare facilities (UWMC and HMC) to determine the domain adaptation capability of the deep learning models we developed in Aims 1 and 2. We also explored the impact of adding out-domain data on in-domain performance. Additionally, we determined the optimal training strategy to use with our models to maximize performance on diverse domains. This was the largest study to date which has conducted a domain adaptation analysis of SSI prediction models. We found that (1) addition of new data modalities adds unique signals pertaining to SSI prediction and improves performance of machine learning models on SSI prediction for a wide variety of domains (surgery types), (2) our models can domain adapt to a diverse set of domains provided there are enough training samples, (3) addition of more out-domain data incrementally improves overall performance for a wide variety of domains (4) addition of data from domains close to the target domain helps performance on the target domain more than addition of data from domains distant from the target domain, (5) addition of out-domain data confers greater performance gains on low-resource domains and domains with a high percentage of missing structured data, and (6) fine-tuning can help performance on surgery types with a low percentage of missing structured data.

In accordance with our findings from previous chapters, we found, in this chapter, that addition of data from additional data modalities contributes unique signals relevant for SSI prediction. In this aim, we found that incorporation of contextual information from clinical text into our SSI prediction models on top of structured data improved performance for most surgery types.

Inclusion of temporal data conferred greater benefits to performance on surgery types with missing static structured data. This indicates that adding contextual information from clinical text and temporal data fills in missing information gaps in static structured data.

Our findings showed our models can domain adapt to diverse surgery types provided there is enough training data. We also found that addition of more out-domain data incrementally improves performance on most domains without lowering performance on others. Our findings showed that addition of out-domain data from domains close to the target domain confers bigger gains in performance on the target domain than addition of out-domain data from domains distant from the target domain. This suggests the need to have a representative training set which includes sufficient data from a diverse set of domains. In fact, we found that there is a benefit to including a diverse set of domains in the training set, even if the amount of data from each domain is lower. These findings indicate the optimal training strategy to confer the highest performance on all domains is to include as much data from as many domains and as many modalities as possible (while constructing representations of these modalities to avoid noise as elaborated in Chapters 3 and 4) in the training set and train one model on data from all domains.

In a clinical setting, there will be surgery types which have less training data. Our findings show that for target domains which are low-resource or have missing structured data, the addition of out-domain data confers a greater benefit to performance than for other domains. Thus, for these kinds of domains, cross-training with data from other domains might be particularly helpful.

We found that fine-tuning on specific domains representing surgery types with a low percentage of missing structured data can help performance on these surgery types. Thus, it may be helpful to fine-tune models on these domains before making predictions for cases from these domains.

5.5 Conclusion

Our finding that our models can domain adapt to a diverse set of surgery types provided there is enough training data is relevant to our overall question, "*Can an automated surveillance system based on multimodal data and deep learning outperform and have better domain adaptation capability than published gold standard automated SSI surveillance approaches?"* Our findings support the statement that an automated surveillance system based on deep learning using static structured, clinical text, and temporal data can achieve better domain adaptation capability than existing automated surveillance systems based on rule-based or conventional machine learning-based approaches. Our findings from this Chapter and Chapters 3 and 4 show that such a system can also achieve better clinical utility and domain adaptation capability than existing automated surveillance systems. However, in order for this goal to be realized, one must gather enough training data to enable deep learning approaches. This is encouraging as the past decade has seen the availability of increasing amounts of biomedical and clinical data[48,49]. Thus far, the availability of task-specific data on SSI prediction has not followed this trend, due to concerns regarding the violation of patient privacy and other factors[50]. However, the general impetus for data sharing for healthcare knowledge discovery and the realization of the importance of better automated SSI surveillance tools is likely to increase the availability of task-specific and non-task-specific datasets for this task.

In this chapter, we explored the domain adaptation capability and optimal training strategy to use with our best models from Aims 1 and 2 to confer the best performance on the task of SSI prediction for different domains. Thus far, in predicting SSI, we have not used the most recent approaches in natural language processing: approaches using LLMs. Smaller LLMs (e.g., BERT, BioGPT, BioMedLM) have been fine-tuned to adapt them to specific tasks. This approach has achieved SOTA performance on several NLP tasks. Another category of models, known as generalist foundation models (e.g., Llama 3, GPT-4), has achieved gold-standard performance on a range of NLP tasks without the need for expensive fine-tuning approaches. We wish to determine whether using such approaches can result in performance gains on the task of SSI prediction. Thus, in the next chapter, we explore several different approaches using LLMs and assess their performance on the task of post hoc SSI prediction.

6. Aim 4: Measuring Performance of Large Language Models to Predict SSI Occurrence

6.1 Introduction

My overall research question is, "*Can an automated surveillance system based on multimodal data and deep learning outperform and have better domain adaptation capability than state of the art automated surgical site infection (SSI) surveillance approaches?"* In this Chapter, the focus is on Aim 4: Measuring Performance of Large Language Models to Predict SSI Occurrence. This work builds on Chapters 3, 4, and 5. In Chapters 3, 4, and 5, we assessed the performance and domain adaptation capability of various rule-based, conventional machine learning, and deep learning approaches on the task of post hoc SSI prediction (namely predicting presence of SSI after the fact). In this Chapter, the focus is on using LLMs, the most recent advancements in deep learning within the field of NLP, for SSI prediction. Recently, LLMs have achieved higher performance on a variety of NLP tasks (including clinical NLP tasks) than previously published gold standards. Most of these models (e.g., Llama 3, Med-PaLM 2, BERT) leverage the transformer architecture, which offers advantages over prior text representation techniques as they can be trained on larger corpora and can take a longer context window into consideration when constructing the representation of a word[51,52]. In the last few years, generalist foundation models (e.g., Llama 3, GPT-4), a special type of LLMs trained using the method of reinforcement learning from human feedback, has achieved close to gold standard performance on a wide variety of NLP tasks without incurring the computational cost of fine-tuning[53]. We wished to evaluate whether the gains in performance conferred by LLMs also applied to the task of post hoc SSI prediction. In this chapter, we tried different approaches to detection of SSI occurrence using two LLMs (BERT, Llama 3) and evaluated their performance on post hoc SSI prediction[52,54–56]. We aimed to conduct an exhaustive analysis of different approaches that can be used with LLMs for this task and their performance. We tried 3 separate approaches to SSI prediction using LLMs: (1) in-context learning approaches with a generalist foundation model (Llama 3[57]), (2) fine-tuning a smaller model, ClinicalBERT[55], and (3) using a generalist foundation model (Llama 3) to summarize clinical notes and using the resulting clinical text representation, combined with our best non-LLM deep learning model for prediction.

This chapter is organized as follows: first, we outline our fine-tuning and in-context learning approaches, detailing our methods for optimizing the hyperparameters of each of these three approaches. Next, we present our results and provide an interpretation of our findings. Finally, we discuss the implications of our findings and propose directions for future work.

6.2 Methods

We used three approaches and one baseline approach:

1. Baseline #1 (covered in section 6.2.2): Our best performing deep learning model from Chapter 4 (Chapter 4 Table 4.2 exp #5) which used BiLSTM to process clinical text data, LSTM to process temporal information, and integrated clinical text data, temporal information, and structured data using multimodal data fusion.

2. Approach #1 (section 6.2.3): In-context learning with Llama 3, a generalist foundation model.
3. Approach #2 (section 6.2.4): Fine-tuning with BERT, a smaller LLM.
4. Approach #3 (section 6.2.5): We utilized a generalist foundation model (Llama 3) to summarize clinical notes and employed these summaries to construct clinical text representations. The remainder of the approach followed the same methodology as our best-performing model from Chapters 4 (Chapter 4 Table 4.2 exp #5 – BiLSTM-LSTM) (same model as Baseline #1).

6.2.1 Clinical Cohort and SSI Definition

We selected a random sample of 7,211 cases from the dataset we used for this task, described in Chapter 3, Section 3.2.1. This was due to the high computational cost of summarizing clinical notes for each case with Llama 3. Counts of included procedure types and SSI event rates in this random sample are presented in Table 6.1.

Procedure Class	Case Count	SSI Events (Rate)
Spine	825	50 (6.1%)
Orthopedic (non-spine)	633	35 (5.5%)
Neurosurgery (non-spine)	1,147	65 (5.7%)
General Surgery	3,571	227 (6.4%)
Gynecologic Surgery	771	31 (4.0%)
Cardiothoracic Surgery	81	1 (1%)
Vascular Surgery	183	2 (1%)
Total	7,211	411(5.7%)

Table 6.1: Surgery counts and infection rate

6.2.2 Baseline: Our Best Previous Deep Learning Approach

Our baseline approach was our best performing deep learning model from Chapter 4 (Chapter 4 Table 4.2 exp#5). As a reminder, this model used BiLSTM to process clinical text data, LSTM to process temporal information, and integrated clinical text data, temporal information, and structured data using multimodal data fusion. Since we used a different dataset and test-train split than in our experimentation in Chapter 4, we tuned some of the hyperparameters (Appendix Supplemental Table 5).

6.2.3 Approach #1: Using In-Context Learning with Llama 3 to predict SSI occurrence

Our experimentation on using in-context learning approaches with Llama 3 for SSI prediction occurred in two stages: (a) clinical text summarization, (b) data preparation, and (c) in-context learning. We explored which in-context learning and fine-tuning approaches yield the best performance on SSI prediction[58]. The following subsections cover each of these stages in more detail.

Clinical Text Summarization: In the first stage of our approach, we used Llama 3 to summarize all clinical notes from the day of operation to the 90th postoperative day for each surgical case. We summarized clinical notes to forgo the challenges LLMs face on long-context tasks and the deterioration of performance beyond a certain context length[59] . The average length of each clinical note was 1772 words and each note was summarized into 100 words. We generated these summaries to present them in clinical vignettes within our prompts. Figure 6.1 presents our summarization prompt, and Figure 6.2 presents an example summary. Our choice of Llama 3 rather than other LLMs was driven by the model's adeptness at summarizing text[52].

Summarization Prompt
Summarize the following clinical note emphasizing development of infection and administration of antibiotics for surgical site infection. Limit the summary to less than 50 words. Clinical note: [note text]

Figure 6.1: Prompt we used to summarize clinical notes with Llama 3

*Legend: Task Instruction, Task Guidance

Example Summary
The patient, a 60-year-old man, developed a superficial wound infection after low anterior resection for rectal adenocarcinoma. He was administered Keflex, which showed good effect. The wound is now healing well, with only a 1-cm opening remaining.

Figure 6.2: Example note summary

Data Preparation: The second stage of our approach using Llama 3 involved preparing our multimodal data for presentation within clinical vignettes to the Llama 3 model. We presented the multimodal clinical data for each of our surgical cases within the prompt as follows:

1. Static structured data: We presented static structured data as tuples within our prompt (e.g., The surgical case was a reoperation.). This notation was chosen based on findings in recent research works that it is an effective way to present structured data within prompts to generalist foundation models in in-context learning strategies[60].
2. Temporal structured data: We represented temporal data as a list of temporally ordered values within our prompt (e.g., The series of WBC counts for the surgical case within the 7th preoperative to 90th postoperative period, ordered temporally is: 11.6 k/mm^3, 10.7 k/mm^3, 10.5 k/mm^3). This notation was chosen based on findings in recent research works that it is an effective way to present temporal information within prompts to generalist foundation models in in-context learning strategies[61].
3. Clinical text data: We presented summaries of clinical notes for each case (see Clinical Text Summarization within this section above), ordered temporally by note date, within our prompt.

In-context Learning with Llama 3: The third stage of our approach involved in-context learning. This involved engineering prompts that best suit our purpose: to accurately predict SSI occurrence

with EHR data from surgical cases. This is crucial, as the format and content of the prompt has a profound effect on the output of Llama 3[58].

We explored four prompt templates for prompt engineering, as below, to determine which one yields best performance. We used Chain-of-Thought (CoT) prompting in our one-shot and few-shot prompting approaches (bullet point #3,4), as it has been shown to be superior to regular prompting, especially for complex tasks[62]. Table 6.2 further illustrates these templates with examples.

1. QA (question answer): This prompt was a presentation of the test case and the task: to determine whether the case is SSI positive or negative.
2. QA + guide: This prompt added basic instructions (presented in Appendix Table 6) relevant to the task in addition to the content of the QA prompt.
3. QA + guide CoT (one-shot): This prompt added a one-shot example from our training set in addition to the content of the QA + guide prompt. One-shot examples were selected from a training set, that was separate from the testing set of test cases. The examples demonstrated correct completions of the task (answer whether a case is a SSI and explain your reasoning). Samples of such examples are presented in Appendix Supplemental Table 9.
4. QA + guide CoT (few-shot): This prompt added few-shot examples from our training set in addition to the content of the QA + guide prompt. This set of examples were selected from the same pool of examples as in our one-shot prompts.

Prompt	Example
QA - Present the task (Answer whether the following case is a surgical site infection)	Clinical case: [test case] The following are the summaries of the clinical notes of the case, ordered temporally: [Clinical text summaries] [Structured data presented as tuples] Answer whether the surgical case is a surgical site infection and explain your reasoning.
QA + guide (0-shot) - Present the task. Then provide basic instructions on how to perform the task.	Guide (see Appendix Supplemental Table 6) Clinical case: [test case (same format as QA)] Using this text, answer the question of whether or not the patient experienced a surgical site infection at any point during the summarized postoperative course and explain your reasoning.
QA + guide CoT (one-shot) - Present the task. Then provide basic instructions on how to perform the task. Then present a random clinical case with its gold standard training label.	Guide (see Appendix Supplemental Table 6) Clinical case: [clinical vignette of random case from training set (same format as QA)] Using this text, answer the question of whether or not the patient experienced a surgical site infection at any point during the summarized postoperative course and explain your reasoning.

	[gold standard training labels] Clinical case: [test case (same format as QA)] Using this text, answer the question of whether or not the patient experienced a surgical site infection at any point during the summarized postoperative course and explain your reasoning.
QA + guide CoT (few-shot) - Present the task. Then provide basic instructions on how to perform the task. Then present a clinical case that is similar to the test case with its gold standard training label.	Guide (see Appendix Supplemental Table 6) Clinical case: [clinical vignette of random case from training set (same format as QA)] Using this text, answer the question of whether or not the patient experienced a surgical site infection at any point during the summarized postoperative course and explain your reasoning. [gold standard training labels] Clinical case: [clinical vignette of random case from training set (same format as QA)] Using this text, answer the question of whether or not the patient experienced a surgical site infection at any point during the summarized postoperative course and explain your reasoning. [gold standard training labels] Clinical case: [test case (same format as QA)] Using this text, answer the question of whether or not the patient experienced a surgical site infection at any point during the summarized postoperative course and explain your reasoning.

Table 6.2: Descriptions of our 4 prompt templates

*Legend: Task Instruction, Task Guidance, Specific Example

6.2.4 Approach#2: Fine-Tuning with Clinical BERT

In our fine-tuning experiments, we used the same model as in our Baseline approach (our best performing deep learning model from Chapter 4 - Chapter 4 Table 4.2 exp #5 – BiLSTM-LSTM) with one modification. The only modification we made was our clinical text representation changed to using ClinicalBERT to generate word embeddings for our clinical text samples (in Chapter 4 we used static embeddings locally trained on our corpus using the word2vec method) (Chapter 3 Methods: Text Data Representation Section Point #2b). We chose ClinicalBERT, as it has embedded clinical knowledge due to its clinically focused fine-tuning corpora. In addition, it has achieved published gold standard performance on a range of clinical NLP tasks in previous works[54]. We input these representations into our best performing deep learning architecture (Chapter 4 Table 4.2 exp #5). Figure 6.3 depicts the architecture we used for these fine-tuning experiments. To handle the typical length of our sequences (~1800 tokens), we segmented our clinical text samples into segments of 512 tokens, generated ClinicalBERT embeddings for each

segment, and concatenated these embeddings. We trained this model for a fixed number of epochs, which was a hyperparameter which was tuned, to fine-tune the ClinicalBERT embeddings. Many of the hyperparameters used for this model were identical to our best performing deep learning model (Chapter 4 Table 4.2 exp #5) - as the temporal and structured representations were identical for the two models. Hyperparameters which were new or differed from our Chapter 4 approach are presented in Appendix Supplemental Table 7.

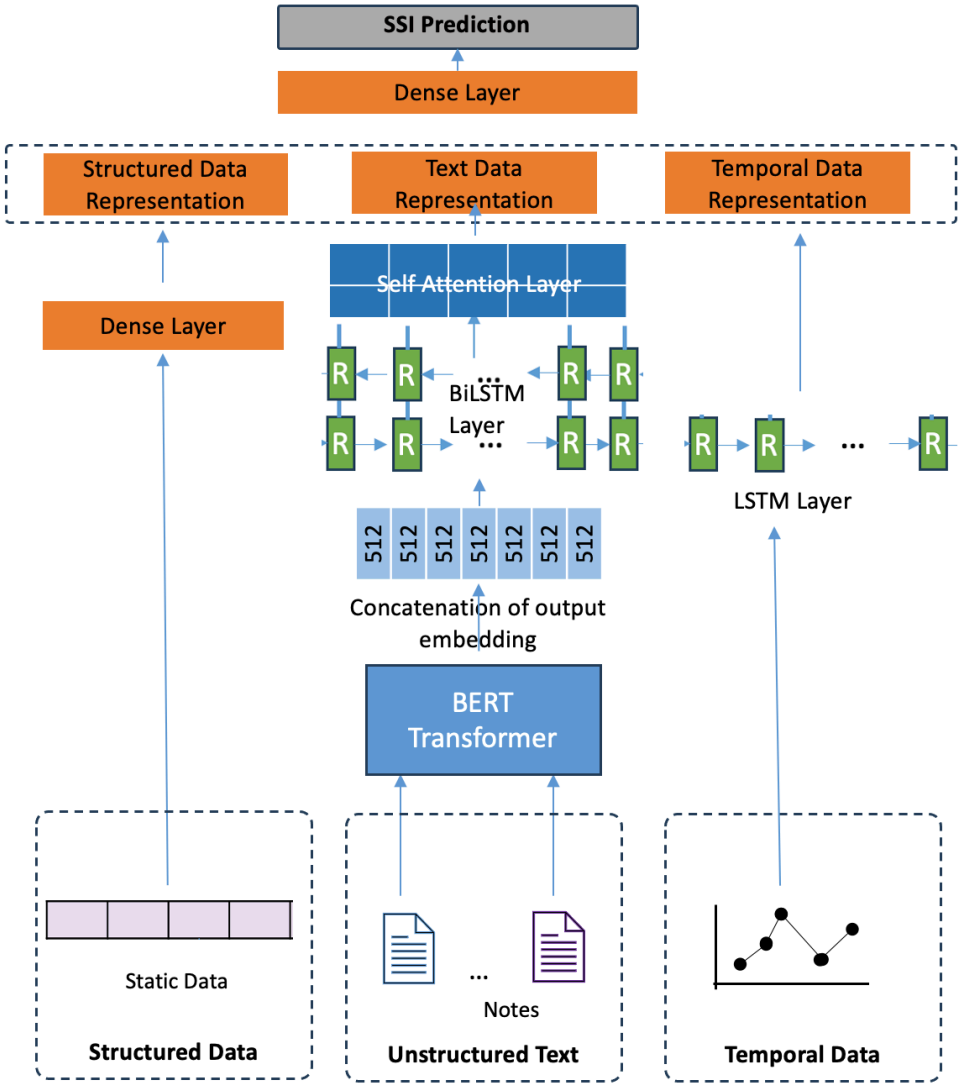


Table 6.3: Figure depicting neural architectures used in our SSI classification approach using ClinicalBERT

6.2.5 Approach #3: Summarizing clinical notes with Llama 3

We wished to conduct an exhaustive analysis of methods which can be used with LLMs for post hoc SSI prediction. In addition, it is possible that certain LLMs (e.g., Llama 3, GPT-4) are more adept at specific parts of the task of post hoc SSI prediction (summarization of clinical notes). Therefore, in a separate approach, we concatenated our clinical note summaries generated using Llama 3 (see Section 6.2.3 Clinical Text Summarization), temporally by note date, to create a

pseudo-document for each surgical case and used these pseudo-documents to construct our text representations. All other aspects of this approach were identical to those for Chapter 4 Table 4.2 exp#5. Hyperparameters of this approach which were new or differed from our Chapter 4 approach are presented in Appendix Supplemental Table 8.

6.2.6 Model Evaluation

Our model evaluation approach was similar to that outlined in Section 3.2.9. To evaluate performance of our in-context learning approaches, we considered a "yes" answer by the model to be a positive prediction and a "no" answer to be a negative prediction. Answers which did not contain a "yes" or "no" were considered to be a negative prediction.

6.2.7 Definition and calculation of clinical utility

Our definition of clinical utility was similar to that outlined in Section 3.2.10.

6.3 Results

6.3.1 Performance of Various Approaches using LLMs for SSI Prediction

In our initial experimentation, we assessed the impact of using different approaches with LLMs on performance (F1-score) on the task of SSI prediction (Table 6.3). The column corresponding to the metric we used for statistical testing is highlighted in grey in the table.

Exp #	Approach	Model (Text Representation)	P	R	F1	Adjusted p-value	Comparison with
1	Baseline	Baseline	0.75(0.73,0.77)	0.61(0.60,0.62)	0.67(0.67,0.68)		
2	#1	Llama 3 zero-shot	0.26(0.26,0.26)	**0.88(0.86,0.90)**	0.40(0.40,0.40)	$5.7*10^{-15}$**	Exp #1
3	#1	Llama 3 zero-shot + guide	0.17(0.16,0.18)	0.96(0.96,0.96)	0.29(0.27,0.30)	$8.6*10^{-11}$**	Exp #1
4	#1	Llama 3 one-shot + guide	0.19(0.18,0.19)	**0.96(0.96,0.96)**	0.31(0.29,0.32)	$5.8*10^{-17}$**	Exp #1
5	#1	Llama 3 few-shot + guide	0.22(0.21,0.22)	**0.96(0.96,0.96)**	0.36(0.35,0.36)	$1.5*10^{-16}$**	Exp #1
6	#2	Clinical BERT fine-tuning	0.72(0.70,0.74)	0.62(0.62,0.62)	0.67(0.66,0.68)	0.32	Exp #1
7	#3	Llama 3 summaries + BiLSTM	0.69(0.69,0.69)	0.71(0.69,0.73)	**0.70(0.69,0.71)**	$2.6*10^{-4}$**	Exp #1

Table 6.3: Performance conferred by different feature sets, modeling strategies, and text representations on the task of SSI prediction

bold=highest value achieved on performance metrics
column with metric used for statistical testing is highlighted in grey
***=significant p-value*

Our results showed that there were differences in the overall performance conferred by the different approaches we used (ANOVA test p-value: $6.7*10^{-39}$). In addition, the following were our conclusions from Table 6.3:

1. **ClinicalBERT vs locally trained embeddings:** Embeddings generated using the ClinicalBERT transformer architecture yield higher recall, but lower precision than static embeddings trained on our local dataset (Table 6.3 exp #6 vs exp #1 - ClinicalBERT recall: 0.62, static embeddings recall: 0.61). A plausible explanation for this phenomenon is the subword-based tokenization and the broader vocabulary coverage of ClinicalBERT, facilitated by its extensive pretraining corpus. For instance, these features of ClinicalBERT could have led to misspellings of the word infection in the test set being recognized as a sign of SSI with ClinicalBERT and not with our locally trained embeddings.
2. **In-context learning vs fine-tuning approaches:** In-context learning approaches achieve a higher recall but lower precision than fine-tuning (Table 6.3 exp #2-5 vs exp #1 & 6 - in-context learning results 0.88-0.96 recall and 0.17-0.26 precision, fine-tuning results 0.61-0.62 recall and 0.72-0.75 precision). This indicates that the knowledge embedded in LLMs (LLMs) such as Llama 3 results in a high detection rate of true SSIs, while also incorrectly classifying many non-SSI cases as SSI positive. Our analysis suggests that this may be due to the models' difficulty in distinguishing between SSIs and other types of infections that may arise in the postoperative context. Conversely, fine-tuning models for the specific task of post hoc SSI prediction enhances precision but reduces recall. This implies that while fine-tuning improves the models' capacity to differentiate SSIs from other infections, it simultaneously impairs their ability to accurately detect actual SSI cases.
3. **Zero-shot + guide vs zero-shot:** Addition of our guide to the prompt improved recall but led to a decline in precision (Table 6.3 exp #3 vs exp #2 - with guide recall: 0.96, without guide recall: 0.88, with guide precision: 0.17, without guide precision: 0.26). This implies that although the information provided in the guide facilitated Llama 3's detection of a greater percentage of SSI cases, it also resulted in the model misclassifying certain non-SSI cases as SSI. For instance, infections other than SSIs that included terminology specified in rule #1 of the guide were erroneously categorized as SSIs.
4. **Few-shot vs One-shot vs zero-shot:** Adding 1-shot and few-shot examples improved precision while keeping recall the same compared to using zero-shot prompts (Table 6.3 exp #5 & 4 vs exp #3 – few-shot precision: 0.22, one-shot precision: 0.19, zero-shot precision: 0.17). A possible explanation for this is our one-shot or few-shot examples may illustrate the use of certain broad terms such as “surgical site infection” in contexts which are indicative of SSI negativity rather than positivity (ex: administration of antibiotics for preventing SSI). Thus, they may help the model distinguish uses of such terms indicating SSI whereas indicating absence of SSI.
5. **Llama 3 summaries vs sentences containing UMLS concepts:** Using summaries of clinical notes generated using Llama 3 rather than sentences containing UMLS concepts extracted using scispaCy yielded a 3-point improvement in F1-score and a 10-point improvement in recall (Table 6.3 exp #7 vs exp #1 – Llama 3 summaries F1-score: 0.70, UMLS concept sentences: 0.67). A plausible explanation for this phenomenon is the targeted instruction provided in our summarization prompt, which emphasized the

development of infection and the administration of antibiotics pertinent to SSIs. This approach resulted in summaries generated by Llama 3 that prioritized information more relevant to SSI prediction, in contrast to sentences that contained UMLS terminology. Furthermore, it facilitated the exclusion of less relevant details related to procedural aspects and repetitive phrases concerning laboratory value collection. Consequently, this focus contributed to a reduction in false negative classifications.

6.3.2 Clinical Significance

Assessment of overall performance of our various approaches is not meaningful in actual practice if it does not reflect performance in a real-world clinical setting. In such a setting, our automated surveillance approach is likely to be used as a screening tool prior to manual chart review to rule out clearly negative cases. Thus, to assess the actual clinical importance of our various SSI prediction approaches, we assessed the precision of our models at a high (90%) recall (Table 6.4). We evaluated our baseline rule-based and conventional machine learning-based approaches, our best deep learning approach from Chapters 3 and 4, our best approach using ClinicalBERT, our best zero-shot approach with the generalist foundation model Llama 3, and our approach using Llama 3 to summarize clinical notes and a BiLSTM for prediction.

Exp #	Model ***	Text Representation	P at 0.9 R	# charts reviewed per SSI detected	Adjusted p-value	Comparison with
	Manual	Manual review	0.03	33		
R	Rule-based	Rule-based3	0.34(at 0.59 R)	-****		
RF	RF	Unigrams	0.17(0.15,0.19)	5.9		
NN	NN	UMLS Concepts	0.26(0.25,0.27)	3.8	0.18	Exp #RF
1	BiLSTM / LSTM*	UMLS Concept Sentences	0.33(0.32,0.33)	3	$3*10^{-7}$ **	Exp #NN
5	Llama 3		0.22(at 0.96 R)	4.5		
6	ClinicalBERT / LSTM	UMLS Concept Sentences	0.32(0.31,0.33)	3.1	0.11	Exp #1
7	BiLSTM / LSTM	LLM Summaries	**0.38(0.37,0.38)**	2.6	$1*10^{-8}$ **	Exp #1

Table 6.4: Clinical significance of the various data representation and modeling approaches

bold=highest value achieved on performance metric

column with metric used for statistical testing is highlighted in grey

** Chapter 4 Table 4.2 exp #5 see Results: Effect of Using Various Strategies to Reduce False Negatives Section in Chapter 3 for full details on how we arrived at this model*

*** = significant p-value*

**** column values are formatted as text representation / temporal representation*

***** - # charts reviewed not provided for this system, as it achieved lower than 0.9 recall (0.59 recall achieved), and being a rule-based system, we could not adjust the threshold to reflect precision at 0.9 recall*

Our results showed that, overall, there were differences in the clinical utility conferred by the different approaches we used (ANOVA test p-value: $2*10^{-10}$) (Table 6.4). Using ClinicalBERT embeddings did not improve clinical utility over using static, locally trained embeddings (Table 6.4 exp#6). Few-shot approaches with Llama 3 yielded high recall (0.96) but low precision (0.22) (Table 6.4 exp#5). Although the higher recall of 0.96 may lead to marginally higher percentage of SSI detected, the low precision would entail 3 more chart reviews for each SSI detected compared with our best deep learning models from Chapters 3 and 4 (Table 6.4 exp#1). Using Llama 3 to summarize clinical notes resulted in the highest clinical utility (Table 6.4 exp #7).

6.4 Discussion

In this aim, we used a subset of 7211 cases from our entire dataset of 30,000 surgical cases to evaluate the performance of a range of approaches using LLM for SSI prediction. We found that our approach using LLM's to generate clinical text summaries (approach #3) achieved higher clinical utility than any of our previous approaches (in Chapters 3 and 4). As a reminder, to assess clinical utility, we picked 0.9 recall as the target and examined the corresponding precision values of the models. Our best LLM approach (approach #3) achieved a precision of 0.38 at 0.9 recall while our previous best approach (from Chapter 4) achieved a precision of 0.33 at 0.9 recall. This would entail review of 2.6 charts for each SSI detected with the LLM approach and would yield a reduction of 0.5 charts per SSI detected over the previous best approach.

Our zero-shot experimentation led to high recall but low precision. Although the high recall is beneficial in terms of % of SSI detected, the lower precision would greatly increase the burden of manual chart review in comparison with our best performing approach (5.9 charts reviewed for each SSI case detected with zero-shot approaches vs 2.6 charts reviewed for each SSI case detected with best approach). Analysis revealed that the low precision was due to Llama 3 not having the knowledge to distinguish SSI from other types of infection.

6.5 Conclusion

My central research question is, "Can an automated surveillance system based on multimodal data and deep learning outperform and have better domain adaptation capability than published gold standard automated SSI surveillance approaches?" In regards to this question, we explored various approaches with LLMs, the most recent deep learning models used to confer gains in utility on a variety of NLP applications, for SSI prediction. We found that using LLMs to summarize clinical text enables each model to perform the task it is specialized at (e.g., text summarization for LLM (Llama 3), clinical text data processing for BiLSTM, temporal data processing for LSTM). Therefore, it improves overall performance, as well as clinical utility on the task of post hoc SSI prediction. Future work will look at whether using a wider range of prompting strategies or fine-tuning generalist foundation models (e.g. Llama 3, GPT-4) might improve the ability of these models to distinguish these other types of infection from SSI.

7. Conclusion

In this dissertation, we explored the overall question, "*Can an automated surveillance system based on multimodal data and deep learning outperform and have better domain adaptation capability than SOTA automated SSI surveillance approaches?"* We detail our main findings in regards to this question in this Chapter. We also present limitations of our research and future directions.

7.1 Main Findings of our Research

Our main findings were:

1. Finding #1: We found that a purely data-driven deep learning approach using multimodal data can outperform previous published gold standard rule-based or conventional machine learning-based approaches on the task of SSI prediction. This is the first study to date that has used deep learning with such a large dataset and so many data modalities for this task. In order to make our data representation amenable to our deep learning approaches, we constructed a data representation with limited feature engineering. In order to mitigate the risk of overfitting with our deep learning models, we had to minimize noise in our data representation (Aim 1 – Chapter 3, Aim 2 – Chapter 4).

   Implications: The implications of our finding that a purely data-driven deep learning approach using multimodal data and a large dataset can confer performance gains over the SOTA rule-based or conventional machine learning-based automated surveillance systems has the following implications:

   1. Large task-specific and non task-specific datasets can be used with data-driven approaches to confer SOTA performance on this task which scales with training dataset size. This can enable ever increasing reductions in the burden of clinical staff performing manual chart review. This is because the performance of deep learning approaches can leverage large datasets and scale with dataset size. The last two decades have seen the widespread adoption of the EHR, the emergence of machine learning in healthcare applications, and an impetus to share healthcare data for the discovery of new knowledge. This had led to the release of a number of public medical datasets and the increasing availability of biomedical and healthcare data. Some examples of such datasets are the MIMIC-III dataset, the 2010 and 2012 i2b2 datasets, and the 2015 SemEval dataset[48,49,63]. One possible use case for these datasets for the task of SSI prediction is: using a weak labeler such as generalist foundation models (e.g., GPT-3.5, GPT-4) to automatically label clinical notes in the MIMIC-III datasets for training supervised models for this task[64]. Others are using the i2b2b 2010 dataset to build named entity recognition tools or using the i2b2 2012 dataset or the 2015 SemEval dataset to build relation extraction tools specialized for this task. Thus far, there are not many publicly available SSI prediction-specific datasets. However, the realization that accurate and timely SSI surveillance can lead to reductions in SSI rates, better health

outcomes, and lower healthcare costs is likely to increase the availability of task-specific datasets. Thus, there are large non-task-specific datasets which can be leveraged to train SSI prediction models, and the size of these datasets and availability of task-specific datasets is likely to increase in the future.

2. Labor-intensive rule-design or feature engineering can be forgone for this task. This is because deep learning approaches, which do not involve such steps, can outperform conventional machine learning-based or rule-based-based approaches[65]. This implies greater efficiency and lower cost in developing and updating these systems.
3. Deep learning approaches forgo the need for rule design or feature engineering. Thus, unlike SOTA rule-based or conventional machine learning-based systems, they can be expected to confer good domain adaptation capability[65]. This implies that deep learning-based automated surveillance systems can potentially confer similar gains in performance over conventional-machine learning-based or rule-based systems on different domains (e.g., surgery types, facilities, institutions) than ones in the training set (please see following sections on Finding #2).

The findings above mean that data-driven deep learning-based SSI prediction that can reduce the burden on infection prevention personnel even more than current SOTA conventional machine learning-based or rule-based systems. Furthermore, these reductions in burden on infection prevention personnel will scale with increasing availability of training data. This is particularly advantageous as over the past two decades, there has been a substantial increase in the amount of medical and biomedical data and this trend is likely to continue in the future (please see bullet point #1 in this section)[66]. We hypothesized that another advantage of using a data-driven deep learning approach is it would have better domain adaptation capability than conventional learning-based or rule-based systems (please see bullet point #3 in this section). This pertains to our second finding (please see section below).

2. Finding #2: Using the data representation and modeling strategies above, we found it was possible to build models that could domain adapt to a diverse set of domains (Aim 3 - Chapter 5).

   Implications: One limitation of rule-based or conventional machine learning-based systems is that they are based on rules or features that may not apply to domains not well represented in the internal dataset. We hypothesized that our purely data-driven deep learning approach could overcome this limitation (please see end of Implications of Finding 1 Section above). Our findings from our domain adaptation experiments supported this hypothesis. We found that our models can domain adapt to a diverse set of domains. Thus, our research shows that the performance gains conferred by deep learning-based SSI surveillance systems apply even to domains not in the training set. Many facilities around the nation will likely have data that includes domains that are not present in the training set. Thus, deep learning-based surveillance systems can confer better clinical utility on SSI surveillance in facilities around the nation, even ones with domains not well represented in the training set, than contemporary automated surveillance approaches. Therefore, data-driven deep learning approaches can be used to build surveillance systems that bring us

closer to the realization of one automated surveillance system that conducts surveillance for facilities around the nation.

In addition, we found that performance incrementally increases for most domains with addition of more training data and additional data modalities. This means that the increasing availability of biomedical and healthcare data can be leveraged to both enable our deep learning approaches and improve performance on a diverse set of domains, even low-resource domains.

3. Finding #3: LLMs, specifically generalist foundation models (e.g., Llama 3) can be used to offer previously unrealized gains on performance (Aim 4 – Chapter 6).

   Implications: Over the last few years, a kind of LLM called generalist foundation models have achieved higher than published gold standard performance or close to published gold standard performance on a variety of NLP tasks without the need for expensive fine-tuning approaches[53]. Generalist foundation models have knowledge embedded in them that enables them to achieve high performance with zero-shot or few-shot approaches[52]. In addition, many of them are trained using reinforcement learning from human feedback, which allows them to align with the specific use cases for which they will be used. Recent initiatives have also seen generalist foundation models being integrated with the systems of EHR vendors like Epic, facilitating their use in a clinical setting[67].

   Our work shows that generalist foundation models can be used to confer performance gains on the task of post hoc SSI prediction. However, our research also showed that using generalist foundation models for the entire SSI prediction pipeline does not yield the highest performing approach. We found that using generalist foundation models specifically for tasks in which they have demonstrated superior capabilities (e.g., text summarization) and using other deep learning methods for the rest of our SSI classification approach led to the best performing approach.

7.2 Limitations and Future Directions

In this section, we state the limitations of our work and propose some potential future directions for our research. They include:

1. **Limitation 1 We did not explore the impact of different imputation strategies (Aim 1):** Some variables have variable levels of missingness based on whether the case is SSI or not. This is missingness not at random (the amount of missingness depends on the value of other variables) and was discussed above in this Chapter and in Chapter 4. Other laboratory values and vitals exhibit missingness at random (e.g., ANC). In this case, SSI positive and negative cases have equal proportions of missingness. For variables which exhibit missingness at random, imputation approaches may be appropriate as there is no meaning to the missingness itself. Past approaches have used several different methods to impute data for SSI prediction[15,43,68]. These studies have found that often, the best imputation method depends on the specific variable being imputed.
   Future Work To Address Limitation: Future work could look at which imputation strategy is best to use with specific temporal variables.
2. **Limitation 2 We did not include imaging data (Aim 1 + 2):** One data modality which we did not use that could be incorporated into SSI prediction models is imaging data. Imaging data might help in cases where there was a workup for SSI with the revelation

that, later, the case turned out to be SSI negative. In such cases, temporal data like laboratory values and vitals might be elevated, as there was a clinical suspicion of infection[69]. In such cases, incorporation of imaging data could help performance. For instance, an imaging study that is part of the workup could confirm the patient has no infection[70]. Imaging studies could also be indicative of infection other than SSI. We did not incorporate imaging data into our models. This is because imaging data was stored in a separate database in our case, making data extraction complicated. We also felt out of the 3 data modalities we used and imaging data, imaging data was least relevant for SSI prediction.

Future Work To Address Limitation: Future work could explore whether the incorporation of imaging data into SSI prediction models helps performance. It could also explore the optimal way to incorporate imaging data into SSI prediction models.

3. **Limitation 3 We did not conduct an exhaustive analysis of optimal model explainability methods (Aim 1+2):** We did not explore using different methods to generate model explanations and evaluating which method presents model explanations to healthcare workers who are likely to use the automated surveillance systems (Aim 1+2): We performed some basic model explainability analyses of our models.

   Future Work To Address Limitation: Future research could further explore the optimal way to generate model explanations to make them understandable by the clinical users of automated surveillance too.

4. **Limitation 4 We did not evaluate performance on data from other facilities/institutions (Aim 3):** Our dataset included data from 2 healthcare facilities that were part of the same institution. Thus, the domain adaptation capability of our models to data from other institutions or facilities needs to be evaluated.

   Future Work To Address Limitation: Future work could procure data from external facilities/institutions and evaluate the domain adaptation capability of our models on such data.

5. **Limitation 5 We did not explore all prompt engineering strategies (Aim 4):** We did not try all prompt engineering techniques that could be used for this task (e.g., presentation of similar examples to this task).

   Future Work To Address Limitation: Future work could explore a wider range of prompt engineering strategies on the task of post hoc SSI prediction.

6. **Limitation 6 We did not explore fine-tuning larger LLMs (e.g., Llama 3, GPT-4) (Aim 4):** We did not fine-tune the larger generalist foundation LLMs in Aim 4.

   Future Work To Address Limitation: Future work could fine-tune generalist foundation models for the task of post hoc SSI prediction.

References

1. Owens, C. D. & Stoessel, K. Surgical site infections: epidemiology, microbiology and prevention. *J Hosp Infect* 70 Suppl 2, 3–10 (2008).
2. Magill, S. S. *et al.* Changes in Prevalence of Health Care–Associated Infections in U.S. Hospitals. *New England Journal of Medicine* 379, 1732–1744 (2018).
3. Anderson, D. J. & Perl, T. M. Basics of Surgical Site Infection: Surveillance and Prevention. in *Practical Healthcare Epidemiology* 147–161 (Cambridge University Press). doi:10.1017/9781107153165.015.
4. Gheorghe, A. *et al.* Health Utility Values Associated with Surgical Site Infection: A Systematic Review. *Value Health* 18, 1126–37 (2015).
5. Pinchera, B. *et al.* Update on the Management of Surgical Site Infections. *Antibiotics* 11, 1608 (2022).
6. Seidelman, J. L., Mantyh, C. R. & Anderson, D. J. Surgical Site Infection Prevention. *JAMA* 329, 244 (2023).
7. Petherick, E. S., Dalton, J. E., Moore, P. J. & Cullum, N. Methods for identifying surgical wound infection after discharge from hospital: a systematic review. *BMC Infect Dis* 6, 170 (2006).
8. Khuri, S. F. The Comparative Assessment and Improvement of Quality of Surgical Care in the Department of Veterans Affairs. *Archives of Surgery* 137, 20 (2002).
9. Selby, L. V. *et al.* Comparing surgical infections in National Surgical Quality Improvement Project and an Institutional Database. *Journal of Surgical Research* 196, 416–420 (2015).
10. Grundmeier, R. W. *et al.* Identifying surgical site infections in electronic health data using predictive models. *Journal of the American Medical Informatics Association* 25, 1160–1166 (2018).
11. Bonde, A. *et al.* Assessing the utility of deep neural networks in predicting postoperative surgical complications: a retrospective study. *Lancet Digit Health* 3, e471–e485 (2021).
12. Rosenthal, R. *et al.* Surveillance of surgical site infections by surgeons: biased underreporting or useful epidemiological data? *Journal of Hospital Infection* 75, 178–182 (2010).
13. Bucher, B. T. *et al.* Portable Automated Surveillance of Surgical Site Infections Using Natural Language Processing. *Ann Surg* 272, 629–636 (2020).
14. Brossette, S. E. *et al.* A Laboratory-Based, Hospital-Wide, Electronic Marker for Nosocomial Infection. *Am J Clin Pathol* 125, 34–39 (2006).
15. Zhu, Y. *et al.* Applying Machine Learning Across Sites: External Validation of a Surgical Site Infection Detection Algorithm. *J Am Coll Surg* 232, 963-971e1 (2021).
16. Shi, J. *et al.* Using Natural Language Processing to improve EHR Structured Data-based Surgical Site Infection Surveillance. *AMIA Annu Symp Proc* 2019, 794–803 (2019).
17. Karhade, A. V. *et al.* Natural language processing for automated detection of incidental durotomy. *The Spine Journal* 20, 695–700 (2020).
18. Alzubaidi, L. *et al.* Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *J Big Data* 8, 53 (2021).

19. Belío-Blasco, C., Torres-Fernández-Gil, M. A., Echeverría-Echarri, J. L. & Gómez-López, L. I. Evaluation of Two Retrospective Active Surveillance Methods for the Detection of Nosocomial Infection in Surgical Patients. *Infect Control Hosp Epidemiol* 21, 24–27 (2000).
20. Magill, S. S. *et al.* Changes in Prevalence of Health Care–Associated Infections in U.S. Hospitals. *New England Journal of Medicine* 379, 1732–1744 (2018).
21. Brossette, S. E. *et al.* A Laboratory-Based, Hospital-Wide, Electronic Marker for Nosocomial Infection. *Am J Clin Pathol* 125, 34–39 (2006).
22. Verberk, J. D. M. *et al.* Automated surveillance systems for healthcare-associated infections: results from a European survey and experiences from real-life utilization. *Journal of Hospital Infection* 122, 35–43 (2022).
23. Chen, W. *et al.* Artificial Intelligence–Based Multimodal Risk Assessment Model for Surgical Site Infection (AMRAMS): Development and Validation Study. *JMIR Med Inform* 8, e18186 (2020).
24. Rabhi, S., Jakubowicz, J. & Metzger, M.-H. Deep Learning versus Conventional Machine Learning for Detection of Healthcare-Associated Infections in French Clinical Narratives. *Methods Inf Med* 58, 031–041 (2019).
25. Ehrentraut, C., Ekholm, M., Tanushi, H., Tiedemann, J. & Dalianis, H. Detecting hospital-acquired infections: A document classification approach using support vector machines and gradient tree boosting. *Health Informatics J* 24, 24–42 (2018).
26. da Silva, D. A., ten Caten, C. S., dos Santos, R. P., Fogliatto, F. S. & Hsuan, J. Predicting the occurrence of surgical site infections using text mining and machine learning. *PLoS One* 14, e0226272 (2019).
27. Loper, E. & Bird, S. NLTK. in *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics* - 63–70 (Association for Computational Linguistics, Morristown, NJ, USA, 2002). doi:10.3115/1118108.1118117.
28. Jacobson, O. & Dalianis, H. Applying deep learning on electronic health records in Swedish to predict healthcare-associated infections. in *Proceedings of the 15th Workshop on Biomedical Natural Language Processing* 191–195 (Association for Computational Linguistics, Stroudsburg, PA, USA, 2016). doi:10.18653/v1/W16-2926.
29. Bodenreider, O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 32, 267D – 270 (2004).
30. Neumann, M., King, D., Beltagy, I. & Ammar, W. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. in *Proceedings of the 18th BioNLP Workshop and Shared Task* 319–327 (Association for Computational Linguistics, Stroudsburg, PA, USA, 2019). doi:10.18653/v1/W19-5034.
31. Bodenreider, O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 32, 267D – 270 (2004).
32. Shen, F. *et al.* Detection of Surgical Site Infection Utilizing Automated Feature Generation in Clinical Notes. *J Healthc Inform Res* 3, 267–282 (2019).
33. Chapman, A. B., Mowery, D. L., Swords, D. S., Chapman, W. W. & Bucher, B. T. Detecting Evidence of Intra-abdominal Surgical Site Infections from Radiology Reports Using Natural Language Processing. *AMIA Annu Symp Proc* 2017, 515–524 (2017).

34. Asudani, D. S., Nagwani, N. K. & Singh, P. Impact of word embedding models on text analytics in deep learning environment: a review. *Artif Intell Rev* 56, 10345–10425 (2023).
35. D. K., S., N. Joshi, S., Kumar, V. H., Venkataramanan, V. & C. S., K. A Review on Neural Networks and its Applications. *Journal of Computer Technology & Applications* (2023) doi:10.37591/jocta.v14i2.1062.
36. Zhao, C., Huang, X., Li, Y. & Yousaf Iqbal, M. A Double-Channel Hybrid Deep Neural Network Based on CNN and BiLSTM for Remaining Useful Life Prediction. *Sensors* 20, 7109 (2020).
37. Lundberg, S. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. *arXiv preprint arXiv* 07874, (2017).
38. Zabaglo, M. & Sharman, T. *Postoperative Wound Infection*. (StatPearls Publishing, 2024).
39. Jung, G. H., Hwang, H. K., Lee, W. J. & Kang, C. M. Extremely high white blood cell counts on postoperative day 1 do not predict severe complications following distal pancreatectomy. *Ann Hepatobiliary Pancreat Surg* 23, 377 (2019).
40. Foy, B. H., Sundt, T. M., Carlson, J. C. T., Aguirre, A. D. & Higgins, J. M. Human acute inflammatory recovery is defined by co-regulatory dynamics of white blood cell and platelet populations. *Nat Commun* 13, 4705 (2022).
41. Ribeiro, M. T., Singh, S. & Guestrin, C. 'Why Should I Trust You?' in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1135–1144 (ACM, New York, NY, USA, 2016). doi:10.1145/2939672.2939778.
42. Soguero-Ruiz, C. *et al.* Data-driven Temporal Prediction of Surgical Site Infection. *AMIA Annu Symp Proc* 2015, 1164–73 (2015).
43. Strauman, A. S. *et al.* Classification of postoperative surgical site infections from blood measurements with missing data using recurrent neural networks. *IEEE EMBS International Conference on Biomedical & Health Informatics* (2017).
44. Hou, Y., Collinsworth, A., Hasa, F. & Griffin, L. Incidence and impact of surgical site infections on length of stay and cost of care for patients undergoing open procedures. *Surg Open Sci* 11, 1–18 (2023).
45. Prada, C. *et al.* Timing and Management of Surgical Site Infections in Patients With Open Fracture Wounds: A Fluid Lavage of Open Wounds Cohort Secondary Analysis. *J Orthop Trauma* 35, 128–135 (2021).
46. Martin, D. *et al.* Timing, diagnosis, and treatment of surgical site infections after colonic surgery: prospective surveillance of 1263 patients. *J Hosp Infect* 100, 393–399 (2018).
47. Lee, K., Dobbins, N. J., McInnes, B., Yetisgen, M. & Uzuner, Ö. Transferability of neural network clinical deidentification systems. *Journal of the American Medical Informatics Association* 28, 2661–2669 (2021).
48. Johnson, A. E. W. *et al.* MIMIC-III, a freely accessible critical care database. *Sci Data* 3, 160035 (2016).
49. Uzuner, Ö., South, B. R., Shen, S. & DuVall, S. L. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association* 18, 552–556 (2011).

50. Jensen, P. B., Jensen, L. J. & Brunak, S. Mining electronic health records: towards better research applications and clinical care. *Nat Rev Genet* 13, 395–405 (2012).
51. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ACL Anthology* (2018).
52. Dubey, A. *et al.* The Llama 3 Herd of Models. *arXiv preprint arXiv:* 21783, (2024).
53. Nori, H. *et al.* Can Generalist Foundation Models Outcompete Special-Purpose Tuning? Case Study in Medicine. *arXiv preprint arXiv* 16452, (2023).
54. Alsentzer, E. *et al.* Publicly Available Clinical BERT Embeddings. *arXiv preprint arXiv* 03323, (2019).
55. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. (2018).
56. Ouyang, L. *et al.* Training language models to follow instructions with human feedback. *Adv Neural Inf Process Syst* 27730–27744 (2022).
57. Brown, T. B. *et al.* Language Models are Few-Shot Learners. *arXiv preprint arXiv* 14165, (2020).
58. Agrawal, M., Hegselmann, S., Lang, H., Kim, Y. & Sontag, D. Large Language Models are Few-Shot Clinical Information Extractors. (2022).
59. Jiang, H. *et al.* LongLLMLingua: Accelerating and Enhancing LLMs in Long Context Scenarios via Prompt Compression. in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* 1658–1677 (Association for Computational Linguistics, Stroudsburg, PA, USA, 2024). doi:10.18653/v1/2024.acl-long.91.
60. Lee, Y., Kim, S., Yu, T., Rossi, R. A. & Chen, X. Learning to Reduce: Optimal Representations of Structured Data in Prompting Large Language Models. *arXiv preprint arXiv* 14195, (2024).
61. Jiang, Y. *et al.* Empowering Time Series Analysis with Large Language Models: A Survey. *arXiv preprint arXiv* 03182, (2024).
62. Wei, J. *et al.* Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Adv Neural Inf Process Syst* 24824–24837 (2022).
63. Elhadad, N. *et al.* SemEval-2015 Task 14: Analysis of Clinical Text. in *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)* 303–310 (Association for Computational Linguistics, Stroudsburg, PA, USA, 2015). doi:10.18653/v1/S15-2051.
64. Gallifant, J. *et al.* Peer review of GPT-4 technical report and systems card. *PLOS Digital Health* 3, e0000417 (2024).
65. Alzubaidi, L. *et al.* Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *J Big Data* 8, 53 (2021).
66. Acosta, J. N., Falcone, G. J., Rajpurkar, P. & Topol, E. J. Multimodal biomedical AI. *Nat Med* 28, 1773–1784 (2022).
67. Nashwan, A. J. & AbuJaber, A. A. Harnessing the Power of Large Language Models (LLMs) for Electronic Health Records (EHRs) Optimization. *Cureus* 15, e42634 (2023).
68. Hu, Z. *et al.* Strategies for handling missing clinical data for automated surgical site infection detection from the electronic health record. *J Biomed Inform* 68, 112–120 (2017).

69. Choi, M. K., Kim, S. B., Kim, K. D. & Ament, J. D. Sequential Changes of Plasma C-Reactive Protein, Erythrocyte Sedimentation Rate and White Blood Cell Count in Spine Surgery : Comparison between Lumbar Open Discectomy and Posterior Lumbar Interbody Fusion. *J Korean Neurosurg Soc* 56, 218 (2014).
70. Sawyer, R. G., Evans, H. L. & Hedrick, T. L. Technological Advances in Clinical Definition and Surveillance Methodology for Surgical Site Infection Incorporating Surgical Site Imaging and Patient-Generated Health Data. *Surg Infect (Larchmt)* 20, 541–545 (2019).

Appendix

List of Antibiotics Derived from Consultations with Clinicians

Generic Name	Antimicrobial Class	SSI Relevance
amoxicillin%		medium
ampicillin%		medium
augmentin		medium
azithromycin		medium
aztreonam		medium
benzathine%		medium
%penicillin%		medium
cefaclor		medium
cefadroxil		medium
cefalexin		high
cefazolin		high
cefdinir		medium
cefditoren		medium
cefepime		high
cefixime		low
cefoperazone		low
cefotaxime		medium
cefotetan		medium
cefoxitin		medium
cefpodoxime		medium
cefprozil		low
ceftaroline		medium
ceftazidime%		low
ceftibuten		low
ceftizoxime		low
ceftolozane%		low
ceftriaxone		high
cefuroxime		medium
cephalexin		high
cephalothin		low
chloramphenicol		low
ciprofloxacin		medium
clarithromycin		low
clindamycin		medium
cloxacillin		low
colistin		medium
daptomycin		high
dicloxacillin		low

Generic Name	Antimicrobial Class	SSI Relevance
dirithromycin		low
doripenem		low
doxycycline		high
enoxacin		low
ertapenem		medium
erythromycin%		low
flucloxacillin		low
fluconazole	antifungal	medium
gentamicin		high
imipenem%		medium
levofloxacin		high
linezolid		high
meropenem		medium
metronidazole		high
minocycline		low
moxifloxacin		medium
nafcillin		high
nitrofurantoin		low
piperacillin%		high
quinupristin%		low
rifampin		high
streotomycin		low
sulfamethoxazole%		high
sulfisoxazole		medium
teicoplanin		low
tetracycline		medium
ticarcillin%		low
tigecycline		medium
tobramycin		medium
trimethoprim%		high
vancomycin		high
micafungin	antifungal	medium
amikacin		medium
amphotericin%	antifungal	medium
caspofungin	antifungal	medium

Appendix Supplemental Table 1: List of Antibiotics derived from consulations with Clinicians, their antimicrobial class, and relevance to SSI

Model/Encoding	Feature	Value
BOW	Term Weighing Scheme	TF-IDF
	Min DF	**5**,20
	Max DF	23147 (corpus-size)
Word Embedding	Dimensions	**24**,32,40
	Min Count	**1**,5
	Window	5
	Method	**Continuous Bag of Words**, Skip-Gram
RF (Structured)	Tree Depth	20,**25**,30
	Maximum # Leaves	4,**6**,8
	# Trees	**25**,50
	Class Weight	**1:1,**2:1
RF (BOW unigrams)	Tree Depth	20,**25**,30
	Maximum # Leaves	4**,**6,**8**
	# Trees	**25**,50
	# Selected Text Features	200,**500,**1000
	Class Weight	**1:1,**2:1
RF (UMLS unigrams)	Tree Depth	20,**25**,30
	Maximum # Leaves	4,6,**8**
	# Trees	25,**50**
	# Selected Text Features	10,20,**50**,100
	Class Weight	**1:1**, 2:1
ANN (Structured)	# Epochs	20,25,30,**35**
	# Units in Hidden Layer	15,20,**25**
	Dropout	0.1,**0.2,**0.3
	Class Weight	**1:1,**2:1
ANN (BOW unigrams)	# Epochs	**10**,15
	# Units in Hidden Layer	200,**400**,500
	Dropout	**0.1**,0.2
	Class Weight	1:1,**2:1**
	# Selected Text Features	50,**500,**7500
ANN (UMLS unigrams)	# Epochs	5
	# Units in Hidden Layer	100**,**200,**300**
	Dropout	**0.1**,0.2,0.3
	Class Weight	2:1
	# Selected Text Features	**100**,200,300

Model/Encoding	Feature	Value
CNN (UMLS concepts (sentences))	# Epochs	**5,**10,15
	Window Size	**2**,3,4
	# CNN Units	**24**,32,48
	Max Length	200,**500**,1000
	# Units in Hidden Layer 1	**15**,20,25
	# Units in Hidden Layer 2	**20**,25,30
	Dropout Text	0.1**,0.2**
	Dropout Structured	0.1**,0.2**
	Class Weight	2:1
BiLSTM (UMLS concepts sentences)	# Epochs	2,**3,**4
	# LSTM Units	24
	Max Length	2000,**2250**,2500
	# Units in Hidden Layer 1	100
	# Units in Hidden Layer 2	20
	Dropout Text	0.1
	Dropout Structured	0.2
	Class Weight	2:1
BiLSTM (UMLS concepts sentences (after removal of presurgical and surgical notes))	# Epochs	**2**,3,4
	# LSTM Units	24
	Max Length	1200,1500,**1800,**2000
	# Units in Hidden Layer 1	100
	# Units in Hidden Layer 2	20
	Dropout Text	0.1
	Dropout Structured	0.2
	Class Weight	2:1
BiLSTM (UMLS concepts sentences (using separate representations of presurgical and postsurgical notes))	# Epochs	2,**3,**4
	# LSTM Units Preop	24
	# LSTM Units Postop	24
	Max Length Preop	1200,1500,**1800,**2000
	Max Length Postop	200,**500**,1000
	# Units in Hidden Layer 1	100
	# Units in Hidden Layer 2	**20**,25
	Dropout Text	0.1
	Dropout Structured	0.2
	Class Weight	2:1

Appendix Supplemental Table 2 : Configuration of RF, ANN, CNN, and BiLSTM models in our approaches using structured and clinical text data

*Hyperparameter values tried during hyperparameter tuning are listed, selected hyperparameter values are bolded

Rule Conditions	Rule Output
perirectal abscess, phlegmon, empyema, abscesses, abscess, abscesses, infected fluid, infected hematoma, purulence, purulent, purulent discharge, pus, puss, purulent drainage, abscesses planned procedure, abscess planned procedure, abscess drain placement plan, abscesses drain placement plan, postoperative stitch abscess, stitch abscess	Concept = SITE INFECTION
skin, erythema, wound, wounds	Concept = SKIN
incisional, incisions, surgical incision, surgical incisions, #wound, #wounds, incision, post surgical scar, surgical incision, surgical incisions, surgical site, surgical sites, surgical wound, phlebectomy site, fistulotomy site, jp drain	Concept = INCISION
discharge, drain,	Concept = DRAINAGE

Rule Conditions	Rule Output
drainage, draining, drains, ooze, oozing, serous, soupy, turbid, drainage, purulent	
infections, skin infection, infectious, infectious diarrhea, Infectious surveillance, Infectious disease, infection, infected, infectious disease management, infection control, wound infection, wound infected, Feculent drainage, Urinary tract infection, Urinary tract infections, infected wound, phlegmon, infection, cultures \w+ +, cultures \w+ \w+ +, cultures \w+ \w+ \w+ +, culture \w+ +, culture \w+ \w+ +, culture \w+ \w+ \w+ +, cultures \w+ positive, cultures \w+ \w+ positive, cultures \w+ \w+ \w+ positive, culture \w+ positive, culture \w+ \w+ positive, culture \w+ \w+ \w+ positive, panniculitis, viral infection, #dehiscence of delorme, #dehiscence of \w+ delorme,	Concept = INFECTION

Rule Conditions	Rule Output
#dehiscence of \w+ \w+ delorme, Fournier gangrene, Fournier ' s gangrene, Fournier's gangrene, Fourniers gangrene, infection prevention, wound infection prevention, infectious \w+ clinic, infectious clinic, acute infection, infectious colitis, respiratory infection, respiratory \w+ infection, Kidney infection, Kidney infections, yeast infection, yeast infections, bladder infections, bladder infection, infections :, peritoneal infection, infected \w+ removal, catheter infection	
if text = "admitted for" within 30 words of concept if text = "admitted with" within 30 words of concept if text = "admitted to \w+ with" within 30 words of concept if text = "admitted to \w+ \w+ with" within 30 words of concept if text = "admitted to \w+ \w+ \w+ with" within 30 words of concept if text = "admission for" within 30 words of concept if text = "during \w+ operation" within 30 words of concept if text = "indication for procedure" within 30 words of concept if text = "#preoperative" within 30 words of concept if text = "#postoperative diagnosis" within 30 words of concept if text = "primary encounter diagnosis" within 30 words of concept if text = "chief complaint" within 30 words of concept if text = "Operative findings" within 30 words of concept if text = "ed" within 30 words of concept	Temporality associated with Concept = “dayone”

Rule Conditions	Rule Output
if text = "to ed" within 30 words of concept if text = "to \w+ ed" within 30 words of concept	
if text="fecal" within 30 words of concept if text="purulent" within 30 words of concept if text="pulrlant" within 30 words of concept if text="no purulent" within 30 words of concept if text="no pulrlant" within 30 words of concept if text="not purulent" within 30 words of concept if text="not pulrlant" within 30 words of concept if text="non-purulent" within 30 words of concept if text="non - purulent" within 30 words of concept if text="foul" within 30 words of concept if text="stool" within 30 words of concept if text="Stool softener" within 30 words of concept if text="infected" within 30 words of concept if text="ileostomy bag" within 30 words of concept if text="without" within 30 words of concept if text="no \w+ erythema" within 30 words of concept if text="no \w+ purulent" within 30 words of concept if text="no \w+ \w+ purulent" within 30 words of concept if text="no \w+ \w+ \w+ purulent" within 30 words of concept if text="no \w+ \w+ \w+ \w+ purulent" within 30 words of concept if text="clear" within 30 words of concept	Context associated with Concept = “infected”
if text="open up" within 10 words of concept if text="opened" within 10 words of concept if text="opened by" within 10 words of concept if text="opened using" within 10 words of concept if text="opening" within 10 words of concept if text="performed" within 10 words of concept if text="dehiscence" within 10 words of concept if text="dehicence" within 10 words of concept if text="#reopened" within 10 words of concept if text="split" within 10 words of concept if text="rupture" within 10 words of concept if text="ruptured" within 10 words of concept if text="leak" within 10 words of concept if text="leaked" within 10 words of concept if text="leaking" within 10 words of concept if text="leakage" within 10 words of concept if text="dehisced" within 10 words of concept	Context associated with Concept = “open”

Rule Conditions	Rule Output
if text="dehisce" within 10 words of concept if text="dehiscing" within 10 words of concept if text="dehisces" within 10 words of concept if text="is open" within 10 words of concept if text="without" within 30 words of concept if text="disruption" within 50 words of concept if text="NG tube" within 8 words of concept if text="discharge" within 30 words of concept if text="erythema" within 30 words of concept if text="erythematous" within 30 words of concept if text="is red" within 30 words of concept if text="#purulence" within 30 words of concept if text="#pus" within 30 words of concept if text="redness" within 30 words of concept if text="soupy" within 30 words of concept if text="#drainage" within 30 words of concept if text="fascial" within 30 words of concept if text="deep incision" within 30 words of concept if text="fascia" within 30 words of concept if text="supraumbilical" within 30 words of concept if text="umbilical" within 30 words of concept if text="mastectomy incisions" within 30 words of concept if text="mastectomy incision" within 30 words of concept if text="groin" within 30 words of concept if text="groins" within 30 words of concept if text="wound" within 30 words of concept if text="wounds" within 30 words of concept if text="incision" within 30 words of concept if text="wound" within 30 words of concept if text="skin" within 30 words of concept if text="incision" within 15 words of concept if text="incisions" within 15 words of concept if text="port site" within 30 words of concept if text="jp drain" within 30 words of concept	Context associated with Concept = “superficial”
if text = "Nissen" within 30 words of concept if text = "omentum" within 30 words of concept if text = "mesenteric" within 30 words of concept if text = "Genitourinary" within 30 words of concept if text = "ileostomy" within 30 words of concept if text = "colostomy" within 30 words of concept if text = "ostomy" within 30 words of concept if text = "anast" within 30 words of concept if text = "gastrectomy" within 30 words of concept	Context associated with Concept = “deep”

Rule Conditions	Rule Output
if text = "laparoscopic" within 30 words of concept	
if text = "coloanal" within 30 words of concept	
if text = "liver" within 30 words of concept	
if text = "pneumoperitoneum" within 30 words of concept	
if text = "peritoneum" within 30 words of concept	
if text = "jejunal" within 30 words of concept	
if text = "ileum" within 30 words of concept	
if text = "colon" within 30 words of concept	
if text = "anastamosis" within 30 words of concept	
if text = "anastomosis" within 30 words of concept	
if text = "anastomotic" within 30 words of concept	
if text = "anastamotic" within 30 words of concept	
if text = "mediastinum" within 30 words of concept	
if text = "mediastinal" within 30 words of concept	
if text = "mastectomy" within 30 words of concept	
if text = "mastectomy incisions" within 30 words of concept	
if text = "mastectomy incision" within 30 words of concept	
if text = "subcostal" within 30 words of concept	
if text = "sternotomy" within 30 words of concept	
if text = "sternum" within 30 words of concept	
if text = "colon" within 30 words of concept	
if text = "perirectal" within 30 words of concept	
if text = "rectum" within 30 words of concept	
if text = "hepatic" within 30 words of concept	
if text = "perihepatic" within 30 words of concept	
if text = "bladder" within 30 words of concept	
if text = "biliary" within 30 words of concept	
if text = "gallbladder" within 30 words of concept	
if text = "pancreas" within 30 words of concept	
if text = "pancreatic" within 30 words of concept	
if text = "spleen" within 30 words of concept	
if text = "renal" within 30 words of concept	
if text = "kidney" within 30 words of concept	
if text = "kidneys" within 30 words of concept	
if text = "stomach" within 30 words of concept	
if text = "gastric" within 30 words of concept	
if text = "duodenum" within 30 words of concept	
if text = "duodenal" within 30 words of concept	
if text = "laparoscopic" within 30 words of concept	
if text = "cholecystectomy" within 30 words of concept	
if text = "cholecystitis" within 30 words of concept	

Rule Conditions	Rule Output
if text = "pancreatitis" within 30 words of concept	
if text = "bowel" within 30 words of concept	
if text = "coloatmospheric" within 30 words of concept	
if text = "soft tissue" within 30 words of concept	
if text = "discharge" within 30 words of concept	
if text = "erythema" within 30 words of concept	
if text = "erythematous" within 30 words of concept	
if text = "is red" within 30 words of concept	
if text = "redness" within 30 words of concept	
if text = "soupy" within 30 words of concept	
if text = "fascial" within 30 words of concept	
if text = "deep incision" within 30 words of concept	
if text = "fascia" within 30 words of concept	
if text = "supraumbilical" within 30 words of concept	
if text = "umbilical" within 30 words of concept	
if text = "mastectomy incisions" within 30 words of concept	
if text = "mastectomy incision" within 30 words of concept	
if text = "groin" within 30 words of concept	
if text = "groins" within 30 words of concept	
if text = "wound" within 30 words of concept	
if text = "wounds" within 30 words of concept	
if text = "incision" within 30 words of concept	
if text = "skin" within 30 words of concept	
if text = "incision" within 30 words of concept	
if text = "incisions" within 30 words of concept	
if text = "port site" within 30 words of concept	
if text = "jp drain" within 30 words of concept	
if text = "phlebectomy site" within 30 words of concept	
if text = "surface" within 30 words of concept	
if text = "superficial" within 30 words of concept	
if text = "skin" within 30 words of concept	
if text = "aortic" within 30 words of concept	
if text = "chest" within 30 words of concept if text = "pleural" within 30 words of concept if text = "Respiratory" within 30 words of concept if text = "lung" within 30 words of concept if text = "nipple" within 30 words of concept if text = "thoracic" within 30 words of concept if text = "esophagus" within 30 words of concept if text = "esophageal" within 30 words of concept if text = "thoracoscopic" within 30 words of concept	Context associated with Concept = “chest”

Rule Conditions	Rule Output
if text = "sp debridement" within 30 words of concept if text = "s / p debridement" within 30 words of concept if text = "supraumbilical" within 30 words of concept if text = "umbilical" within 30 words of concept if text = "omentum" within 30 words of concept if text = "mesenteric" within 30 words of concept if text = "Genitourinary" within 30 words of concept if text = "ileostomy" within 30 words of concept if text = "colostomy" within 30 words of concept if text = "#ostomy" within 30 words of concept	Context associated with Concept = “abdominal”
if text = "groin" within 30 words of concept if text = "groins" within 30 words of concept if text = "extremities" within 30 words of concept if text = "#axillary" within 30 words of concept if text = "ischial" within 30 words of concept if text = "limb" within 30 words of concept if text = "limbs" within 30 words of concept if text = "wrist" within 30 words of concept if text = "hand" within 30 words of concept if text = "foot" within 30 words of concept if text = "feet" within 30 words of concept if text = "toe" within 30 words of concept if text = "arm" within 30 words of concept if text = "leg" within 30 words of concept if text = "forearm" within 30 words of concept if text = "RLE" within 30 words of concept if text = "LLE" within 30 words of concept if text = "RUE" within 30 words of concept if text = "LUE" within 30 words of concept if text = "PICC line" within 30 words of concept if text = "intramuscular" within 30 words of concept if text = "thigh" within 30 words of concept if text = "limb" within 30 words of concept if text = "ble" within 30 words of concept if text = "bue" within 30 words of concept if text = "and" within 30 words of concept if text = "v" within 30 words of concept if text = "limbs of aortic" within 30 words of concept if text = "limb of aortic" within 30 words of concept if text = "\w+ limb of aortic" within 30 words of concept	Context associated with Concept = “extremities”

Rule Conditions	Rule Output
if text = "anast" within 30 words of concept if text = "gastrectomy" within 30 words of concept if text = "laparoscopic" within 30 words of concept if text = "coloanal" within 30 words of concept if text = "liver" within 30 words of concept if text = "subxiphoid" within 30 words of concept if text = "left upper quadrant" within 30 words of concept if text = "left lower quadrant" within 30 words of concept if text = "right upper quadrant" within 30 words of concept if text = "right lower quadrant" within 30 words of concept if text = "pneumoperitoneum" within 30 words of concept if text = "peritoneum" within 30 words of concept if text = "jejunal" within 30 words of concept if text = "ileum" within 30 words of concept if text = "colon" within 30 words of concept if text = "#anastamosis" within 30 words of concept if text = "#anastomosis" within 30 words of concept if text = "#anastomotic" within 30 words of concept if text = "#anastamotic" within 30 words of concept if text = "umbilical" within 30 words of concept if text = "colon" within 30 words of concept if text = "perirectal" within 30 words of concept if text = "rectum" within 30 words of concept if text = "hepatic" within 30 words of concept if text = "perihepatic" within 30 words of concept if text = "bladder" within 30 words of concept if text = "biliary" within 30 words of concept if text = "gallbladder" within 30 words of concept if text = "pancreas" within 30 words of concept if text = "peripancreatic" within 30 words of concept if text = "pancreatic" within 30 words of concept if text = "spleen" within 30 words of concept if text = "renal" within 30 words of concept if text = "kidney" within 30 words of concept if text = "kidneys" within 30 words of concept if text = "stomach" within 30 words of concept if text = "gastric" within 30 words of concept if text = "duodenum" within 30 words of concept if text = "duodenal" within 30 words of concept if text = "laparoscopic" within 30 words of concept	Context associated with Concept = “abdominal”

Rule Conditions	Rule Output
if text = "cholecystectomy" within 30 words of concept if text = "cholecystitis" within 30 words of concept if text = "pancreatitis" within 30 words of concept if text = "\w+ diet" within 30 words of concept if text = "bowel" within 30 words of concept if text = "coloatmospheric" within 30 words of concept if text = "gastroesophageal junction" within 30 words of concept if text = "ge junction" within 30 words of concept if text = "midline" within 30 words of concept if text = "ventral" within 30 words of concept if text = "iliac" within 30 words of concept if text = "after" within 30 words of concept if text = "appears to" within 30 words of concept if text = "#as" within 30 words of concept if text = "at that time" within 30 words of concept if text = "at this time" within 30 words of concept if text = "be ruled out for" within 30 words of concept if text = "be ruled out" within 30 words of concept if text = "being ruled out" within 30 words of concept if text = "borderline" within 30 words of concept if text = "c w" within 30 words of concept if text = "can be ruled out for" within 30 words of concept if text = "can be ruled out" within 30 words of concept if text = "cannot be completely excluded" within 30 words of concept if text = "cannot be excluded" within 30 words of concept if text = "cannot be fully excluded" within 30 words of concept if text = "concern for" within 30 words of concept if text = "concerns of" within 30 words of concept if text = "concerned of" within 30 words of concept if text = "concerned about" within 30 words of concept if text = "concerned for" within 30 words of concept if text = "concerning for" within 30 words of concept if text = "consistent with" within 30 words of concept if text = "could be either" within 30 words of concept if text = "could be ruled out for" within 30 words of concept if text = "could be due to" within 30 words of concept	Context associated with Concept = “uncertain”

Rule Conditions	Rule Output
if text = "could be" within 30 words of concept	
if text = "could be ruled out" within 30 words of concept	
if text = "did not rule out" within 30 words of concept	
if text = "doubt" within 30 words of concept	
if text = "exam to assess" within 30 words of concept	
if text = "examination assess" within 30 words of concept	
if text = "examination to assess" within 30 words of concept	
if text = "examinations to assess" within 30 words of concept	
if text = "exams to assess" within 30 words of concept	
if text = "final report indication" within 30 words of concept	
if text = "for presumed" within 30 words of concept	
if text = "#from" within 30 words of concept	
if text = "given \w+ history" within 30 words of concept	
if text = "given patient s history" within 30 words of concept	
if text = "in her" within 30 words of concept	
if text = "in his" within 30 words of concept	
if text = "indication" within 30 words of concept	
if text = "is to be ruled out for" within 30 words of concept	
if text = "is to be ruled out" within 30 words of concept	
if text = "likely reflecting" within 30 words of concept	
if text = "likely" within 30 words of concept	
if text = "markedly" within 30 words of concept	
if text = "may be \w+ underestimated" within 30 words of concept	
if text = "may be contributing" within 30 words of concept	
if text = "may be due to" within 30 words of concept	
if text = "may be related to" within 30 words of concept	
if text = "may be ruled out for" within 30 words of concept	
if text = "may be ruled out" within 30 words of concept	
if text = "may be underestimated" within 30 words of concept	

Rule Conditions	Rule Output
if text = "may be unmasking" within 30 words of concept	
if text = "may be" within 30 words of concept	
if text = "may be consistent with" within 30 words of concept	
if text = "may have been preceded by" within 30 words of concept	
if text = "may have" within 30 words of concept	
if text = "may represent" within 30 words of concept	
if text = "might be ruled out for" within 30 words of concept	
if text = "might be ruled out" within 30 words of concept	
if text = "might be" within 30 words of concept	
if text = "must be ruled out for" within 30 words of concept	
if text = "must be ruled out" within 30 words of concept	
if text = "not been ruled out" within 30 words of concept	
if text = "not ruled out" within 30 words of concept	
if text = "ought to be ruled out for" within 30 words of concept	
if text = "ought to be ruled out" within 30 words of concept	
if text = "possibility of" within 30 words of concept	
if text = "possible" within 30 words of concept	
if text = "possibly" within 30 words of concept	
if text = "possible" within 30 words of concept	
if text = "poss" within 30 words of concept	
if text = "presumably" within 30 words of concept	
if text = "presumed to" within 30 words of concept	
if text = "probable" within 30 words of concept	
if text = "probably" within 30 words of concept	
if text = "question was" within 30 words of concept	
if text = "questioned" within 30 words of concept	
if text = "r o" within 30 words of concept	
if text = "ro" within 30 words of concept	
if text = "rule her out for" within 30 words of concept	
if text = "rule her out" within 30 words of concept	
if text = "rule him out for" within 30 words of concept	
if text = "rule him out" within 30 words of concept	
if text = "rule out for" within 30 words of concept	
if text = "rule out" within 30 words of concept	

Rule Conditions	Rule Output
if text = "rule the patient out" within 30 words of concept	
if text = "rule the patient out for" within 30 words of concept	
if text = "s p" within 30 words of concept	
if text = "should be ruled out for" within 30 words of concept	
if text = "should be ruled out" within 30 words of concept	
if text = "show a question of" within 30 words of concept	
if text = "shown a question of" within 30 words of concept	
if text = "status post" within 30 words of concept	
if text = "studies to assess" within 30 words of concept	
if text = "study to assess" within 30 words of concept	
if text = "suggestive of" within 30 words of concept	
if text = "suggest" within 30 words of concept	
if text = "suspected" within 30 words of concept	
if text = "suspicion for" within 30 words of concept	
if text = "suspicious for" within 30 words of concept	
if text = "the presumed" within 30 words of concept	
if text = "to evaluate for any" within 30 words of concept	
if text = "evaluate for" within 30 words of concept	
if text = "treatment of" within 30 words of concept	
if text = "unclear dose" within 30 words of concept	
if text = "unclear" within 30 words of concept	
if text = "was initially suspected" within 30 words of concept	
if text = "was suspected" within 30 words of concept	
if text = "what must be ruled out is" within 30 words of concept	
if text = "whether or not" within 30 words of concept	
if text = "will be ruled out for" within 30 words of concept	
if text = "will be ruled out" within 30 words of concept	
if text = "with a question of" within 30 words of concept	
if text = "without difficulties" within 30 words of concept	
if text = "without difficulty" within 30 words of concept	

Rule Conditions	Rule Output
if text = "without diff" within 30 words of concept	
if text = "without not" within 30 words of concept	
if text = "without no" within 30 words of concept	
if text = "given" within 30 words of concept	
if text = "r / o" within 30 words of concept	
if text = "risk of" within 30 words of concept	
if text = "risks of" within 30 words of concept	
if text = "vs" within 30 words of concept	
if text = "versus" within 30 words of concept	
if text = "rule out \w+ showed" within 30 words of concept	
if text = "rule out \w+ \w+ showed" within 30 words of concept	
if text = "indication for procedure" within 30 words of concept	
if text = "unsure" within 30 words of concept	
if text = "uncertain" within 30 words of concept	
if text = "potential" within 30 words of concept	
if text = "for" within 30 words of concept	
if text = "due to concern" within 30 words of concept	
if text = "concern that" within 30 words of concept	
if text = "could be" within 30 words of concept	
if text = "most recent" within 30 words of concept	
if text = "or inflammatory" within 30 words of concept	
if text = "differential considerations" within 30 words of concept	
if text = "v" within 30 words of concept	
if text = "concerned that" within 30 words of concept	
if text = "concerns for" within 30 words of concept	
if text = "studies for" within 30 words of concept	
if text = "question of" within 30 words of concept	
if text = "reviewed signs and symptoms of" within 30 words of concept	
if text = "concern for" within 30 words of concept	
if text = "related to" within 30 words of concept	
if text = ",Ä¢" within 30 words of concept	
if text = "limited study" within 30 words of concept	
if text = "assess for" within 30 words of concept	
if text = "differential includes" within 30 words of concept	
if text = "with concern of" within 30 words of concept	
if text = "for concern of" within 30 words of concept	
if text = "not \w+ excluded" within 30 words of concept	

Rule Conditions	Rule Output
if text = "concerned about" within 30 words of concept	
if text = "he thinks" within 30 words of concept	
if text = "she thinks" within 30 words of concept	
if text = "he thought" within 30 words of concept	
if text = "she thought" within 30 words of concept	
if text = "suspicion of" within 30 words of concept	
if text = "may indicate" within 30 words of concept	
if text = "might indicate" within 30 words of concept	
if text = "will evaluate" within 30 words of concept	
if text = "will evaluate" within 30 words of concept	
if text = "evaluate \w+ for" within 30 words of concept	
if text = "evaluate \w+ \w+ for" within 30 words of concept	
if text = "evaluate \w+ \w+ \w+ for" within 30 words of concept	
if text = "evaluate \w+ \w+ \w+ \w+ for" within 30 words of concept	
if text = "evaluate \w+ \w+ \w+ \w+ \w+ for" within 30 words of concept	
if text = "evaluate \w+ \w+ \w+ \w+ \w+ \w+ for" within 30 words of concept	
if text = "evaluate \w+ \w+ \w+ \w+ \w+ \w+ \w+ for" within 30 words of concept	
if text = "evaluate \w+ \w+ \w+ \w+ \w+ \w+ \w+ \w+ for" within 30 words of concept	
if text = "evaluate \w+ \w+ \w+ \w+ \w+ \w+ \w+ \w+ \w+ for" within 30 words of concept	
if text = "evaluate \w+ \w+ \w+ \w+ \w+ \w+ \w+ \w+ \w+ \w+ for" within 30 words of concept	
if text = "evaluate \w+ \w+ \w+ \w+ \w+ \w+ \w+ \w+ \w+ \w+ \w+ for" within 30 words of concept	
if text = "evaluate \w+ \w+ \w+ \w+ \w+ \w+ \w+ \w+ \w+ \w+ \w+ \w+ for" within 30 words of concept	
if text = "evaluate \w+ \w+ \w+ \w+ \w+ \w+ \w+ \w+ \w+ \w+ \w+ \w+ \w+ for" within 30 words of concept	
if text = "evaluate \w+ \w+ \w+ \w+ \w+ \w+ \w+ \w+ \w+ \w+ \w+ \w+ \w+ \w+ for" within 30 words of concept	

Rule Conditions	Rule Output
if text = "evaluate \w+ \w+ \w+ \w+ \w+ \w+ \w+ \w+ \w+ \w+ \w+ \w+ \w+ \w+ \w+ \w+ \w+ for" within 30 words of concept if text = "evaluate \w+ \w+ \w+ \w+ \w+ \w+ \w+ \w+ \w+ \w+ \w+ \w+ \w+ \w+ \w+ \w+ \w+ \w+ for" within 30 words of concept if text = "evaluate \w+ \w+ \w+ \w+ \w+ \w+ \w+ \w+ \w+ \w+ \w+ \w+ \w+ \w+ \w+ \w+ \w+ \w+ for" within 30 words of concept if text = "may be" within 30 words of concept if text = "may \w+ be" within 30 words of concept If text = “possibility of” within 30 words of concept	
If concept = “SITE INFECTION” and concept = “SKIN” If concept = “SITE INFECTION” and concept = “INCISION”	Document label = “superficial SSI”
If concept = “SITE INFECTION” and context = “abdominal” If concept = “DRAINAGE” and context = “abdominal” and context = ‘infected” and context = “superficial” If mention = “superficial SSI” and context = “abdominal” If mention = “superficial SSI” and context = “abdominal” If concept = “INCISION” and context = “abdominal” and context = “infected” and context = “open” If concept = “INCISION” and context = “abdominal” and context = “infected” and context = “open”	mention = “abdominal site infection”
If concept = “SITE INFECTION” and context = “abdominal” If concept = “DRAINAGE” and context = “abdominal” and context = ‘infected” and context = “superficial”	Document label = “abdominal site infection”
If concept = “INFECTION” and context = “uncertain”	mention = “possible infection”
If concept = “SITE INFECTION” and concept = “SKIN” and context = “uncertain”	mention = “possible superficial surgical site infection”
If concept = “SITE INFECTION” and concept = “INCISION” and context = “uncertain”	Document label = “possible superficial surgical site infection”

Rule Conditions	Rule Output
If concept = “STITCHING” and context = “open” and context = “infected” and context = “uncertain”	
If concept = “SITE INFECTION” and context = “abdominal” and context = “uncertain” and context = “deep”	mention = “possible abdominal organ space infection”
If concept = “INFECTION” and context = “chest” and context = “uncertain” and context = “deep”	mention = “possible chest organ space infection”
If concept = “INFECTION” and context = “extremities” and context = “uncertain” and context = “deep”	mention = “possible extremity organ space infection”
If concept = “SITE INFECTION” and concept = “INCISION” and context = “abdominal” and context = “uncertain” and context = “superficial”	mention = “possible abdominal superficial infection”
If concept = “SITE INFECTION” and concept = “SKIN” and context = “chest” and context = “uncertain” and context = “superficial”	mention = “possible chest superficial infection”
If concept = “SITE INFECTION” and concept = “SKIN” and context = “extremities” and context = “uncertain” and context = “superficial”	mention = “possible extremity superficial infection”
If mention= “possible infection” and temp = “dayone” If mention= “possible superficial infection” and temp = “dayone” If mention label = “possible extremity organ space infection” and temp = “dayone” If mention= “possible abdominal organ space infection” and temp= “dayone” If mention= “possible chest organ space infection” and temp= “dayone” If mention= “possible extremity superficial infection” and temp= “dayone” If mention= “possible abdominal superficial infection” and temp= “dayone” If mention= “possible chest superficial infection” and temp= “dayone”	Document label = “POSS_INFECTION_IN_24H”
If document label = “POSS_INFECTION_IN_24H”	If Case Label = “SSI_Doc” document label = “SSI”

Rule Conditions	Rule Output
If document label = “superficial surgical site infection” If document label = “possible superficial surgical site infection” If document label = “abdominal site infection”	
If NOT (document label = “POSS_INFECTION_IN_24H”) and NOT(document label = “SSI”) and (NOT document label = “superficial SSI”) and (NOT document label = “possible superficial infection”) and (NOT document label = “abdominal surgical infection”)	Case Label = “not SSI_Doc”

Appendix Supplemental Table 3: Rules that were used by the rule-based system to generate classifications

Model/Encoding	Feature	Value
Word Embedding	Dimensions	24
	Min Count	**1**,5
	Window	5
	Method	Continuous Bag of Words
BiLSTM (UMLS concepts sentences (after removal of presurgical and surgical notes))	# Epochs	2
	# LSTM Units	24
	Max Length	1800
	# Units in Hidden Layer 1	100
	# Units in Hidden Layer 2	20
	Dropout Text	0.1
	Dropout Structured	0.2
	Class Weight	2:1
LSTM (temporal data)	# Epochs	3,**4**
	# LSTM units	**2**,3,4
	Max Length	7,**10**,13
	Dropout	**0.1**,0.2

Appendix Supplemental Table 4: Configuration of LSTM and BiLSTM models in our deep learning approaches incorporating temporal information

*Hyperparameters tried during hyperparameter tuning are listed, selected hyperparameters are bolded

Model/Encoding	Feature	Value
BiLSTM	# Epochs	**2**,3,4,5,6
	Max Length	1200,1500,**1800,**2000

Appendix Supplemental Table 5: Configuration of models using UMLS concept sentences to construct the text representation

*Hyperparameter values tried during hyperparameter tuning are listed, selected hyperparameter values are bolded.

Guide Used in Prompt Engineering Experiments

A series of text statements will be provided, each representing a summary of clinical notes written after surgery and beginning with the prefix "Summary:". Each of these text statements represents a summary of clinical notes written by clinicians to document the course on subsequent days after surgery. For example, text statement following the first instance of "Summary:" describes observations and events happening right after surgery, the text following the second instance of "Summary:" describes observations and events happening 1 day after a surgery, the text following the second instance of "Summary:" describes observations and events happening 2 days after a surgery, etc.

Rules to apply in making this determination are:

Rule 1: In order to classify a case as having surgical site infection, "specific terms" indicating a surgical site infection must be used. Specific terms used to document or describe treatment of a surgical site infection may differ according to surgical procedure type. For example after an abdominal surgery, the following terms are commonly used to document a surgical site infection: "fluid collection", "phlegmon", and "abscess". After surgery performed on the spine, pelvis, arms, legs, this is often described as "osteomyelitis", "spine infection", "infected hardware", "hardware infection", or "hardware-associated osteomyelitis". In gynecological surgery, they may be described similar to abdominal surgery, but also as "cuff infection". After neurosurgery such as a craniotomy, they may be described as “meningitis” or osteomyelitis of the skull. Across all procedure types, these may be described more generally as "surgical site infection" or "wound infection". Other slight variations, synonyms, or differences in spelling for these specific terms may be considered; however, different forms of infection or postoperative complications such as pneumonia, urinary tract infection, bloodstream infection, stroke, heart attack, or respiratory failure should not be considered relevant to the condition of surgical site infection. Non-specific descriptions of infection such as sepsis, "possible infection", or documentation of treatment of an infection with antibiotics without an specific term clearly indicating infection of the surgical site (the area anatomically near to where the surgery was performed) should not be considered as evidence supporting a diagnosis of surgical site infection. The abbreviation "SSI" often stands for "surgical site infection", but may also be used to describe "sliding scale insulin" so when SSI is used in the context of diabetes treatment or blood sugar management, it is not relevant as a term describing surgical site infection.

Rule 2: If a "specific term" describing a potential surgical site infection as outlined in Rule 1 is present in the documentation, the infection must be present near the anatomic site of the surgery that was performed. For example, an intraabdominal abscess, fluid collection, or phegmon after an abdominal surgery would be consistent with the diagnosis of a surgical site infection. However, an intraabdominal abscess, fluid collection, or phegmon after a surgery on an unrelated body region such as the head, a leg, or arm would not be consistent with the diagnosis of a surgical site infection because the infection identified by the specific term is not anatomically close to where the surgery occurred.

Rule 3: Surgical site infection might not be present until several days after a surgery and might only be mentioned in a small proportion of the text statements. So even if a specific term is used documenting surgical site infection on one day and not another, the case should still be classified as positive for surgical site infection.

Rule 4: To qualify, a surgical site infection must occur only after surgery, in the absence of pre-existing infection at the surgical site. For example, a patient may have a history of an intraabdominal abscess for which they undergo surgery. In such cases, this should not be counted as a surgical site infection because the infection (identified by the specific term "intraabdominal abscess") was present prior to the surgery and did not arise as a consequence of the surgery. Examples of terms that may be used to describe pre-existing infections include "chronic non-healing wound", "diverticulitis", "enterocutaneous fistula", "chronic infection", etc.
- In order to classify a case as positive for surgical site infection, you should be relatively confident in your determination. If a patient has cellulitis, receives antibiotic treatment, has a fever, leukocytosis, etc but there is not definitive language specifically describing a surgical site infection, it should be classified as negative.
Limit your response to less than 200 words. In explaining your reasoning, please specifically comment on the following points:
- Which "specific terms" documenting surgical site infection were present (Rule 1)
- If the infection involved the surgical site (Rule 2)
- Whether or not infection was present prior to surgery (Rule 4)

Appendix Supplemental Table 6: Guide used in prompt engineering experiments
*Legend: Task Guidance

Model/Encoding	Feature	Value
ClinicalBERT	# Epochs (for fine-tuning)	**3**,4
	# Segments	3,**4**,5
	Max Length	512

Appendix Supplemental Table 7: Configuration of models using ClinicalBERT to construct the text representation
*Hyperparameters tried during hyperparameter tuning are listed, selected hyperparameters are bolded

Model/Encoding	Feature	Value
BiLSTM	# Epochs	2,3,4,5,**6**
	Max Length	1200,**1500**,1800

Appendix Supplemental Table 8: Configuration of models using Llama 3 generated summaries to construct the text representation
*Hyperparameter values tried during hyperparameter tuning are listed, selected hyperparameter values are bolded

Guide Used in Prompt Engineering Experiments

Example #1:

The clinical text for the surgical case contains the excerpt: Summary: The patient, a [censored]-year-old male, underwent L3-5 laminectomy and L1-2 hemilaminotomy with central decompression on [censored]. He is recovering well and denies pain. The patient has a history of spinal stenosis, diabetes, hypertension, reflux, falls, and obstructive sleep apnea. He is taking various medications, including antibiotics (Ceftriaxone 1Gr 0800), pain medications (Fentanyl 100 mcg, Methadone 20 mg, Acetaminophen 1000 mg), and anticoagulants (Heparin 5,000 units Subcutaneous Q8 Hours). The patient's postoperative management includes monitoring for infection, pain management, and glycemic control. The patient is expected to be discharged home when the drain is discontinued and he is medically stable...

The structured EHR data for the surgical case is: The maximum white blood cell count for the patient involved in the surgical case from the day of surgery to 90 days after surgery is 19.040001. The maximum temperature for the patient involved in the surgical case from the day of surgery to 90 days after surgery is 37.599998. The surgical case is not a reoperation. The surgical case did not result in the order of a wound, tissue, or cerebrospinal fluid culture. The patient was not given antibiotics at any time between 7 days after surgery and 90 days after surgery. A procedure for postoperative drain placement by interventional radiology did not occur for the surgical case. A postoperative infectious disease consultation was not obtained

Answer: This case is not a surgical site infection. Here is the reasoning.
   The clinical note summary contains the term "infection" but this term is used in the context of prophylaxis or prevention and for monitoring for infection, not in the context of diagnosis of infection.
   In order for general terms ike "infection" or "surgical site infection" to be indicative of SSI, they must be used in the context of confirming a diagnosis of an infection at the surgical site after surgery.
   Additionally, negation of specific terms indicating "surgical site infection" does not indicate a surgical site infection.
   Therefore, this case is not a surgical site infection.

Example #2:

the clinical text for the surgical case contains the excerpt: Summary: The patient underwent laparoscopically-assisted excision of a retrocaval mass, which was diagnosed as recurrent pheochromocytoma. IV antibiotics were administered preoperatively. The procedure was converted to open due to friable tissue surrounding the mass. The patient was sent to the ICU postoperatively due to a prolonged operation, obesity, and obstructive sleep apnea. Summary: The patient developed no infection, received no antibiotics, and the reason for no antibiotic administration was not specified. Summary: The patient is a [censored]-year-old woman who underwent resection of a recurrent 11cm pheochromocytoma and was admitted to the SICU for postoperative hypoxia and hemodynamic monitoring. She was intubated and sedated due to low oxygen saturation. No antibiotics were administered, and there is no indication of infection. Summary: The

patient developed hypotension overnight, requiring 2.5L of IV fluids and an increased MIVF rate. She received antibiotics, likely for surgical site infection prevention, following laparoscopic converted to open excision of retrocaval pheochromocytoma. Summary: [censored]-year-old woman post-laparoscopic surgery for pheochromocytoma, experiencing minimal incision site pain, managed with oxycodone and tylenol. No signs of infection, and antibiotics not administered. Summary: The patient is a [censored]-year-old woman who underwent laparoscopic surgery for a 11cm pheochromocytoma. She developed hypoxia and was intubated, but her oxygen saturation improved after correction of right mainstem intubation and extubation. She has been experiencing relative hypotension, which is being managed with hypertonic IV fluids. The patient is also being treated for post-operative hyperglycemia and acute blood loss anemia. She is not receiving antibiotics due to a history of allergy to Augmentin. The plan is to continue monitoring her vital signs, electrolyte levels, and blood glucose, and to provide pain management and DVT prophylaxis. Summary: The [censored]-year-old woman, post-laparoscopic excision of recurrent pheochromocytoma, is experiencing minimal pain at the incision site, controlled with oxycodone and tylenol. No signs of infection, and antibiotics are not mentioned. The focus is on pain management and monitoring vital signs. Summary: The patient is a [censored]-year-old woman who underwent laparoscopic surgery converted to open excision of recurrent retrocaval pheochromocytoma. She developed an infection, for which antibiotics were administered. The reason for antibiotic administration was to treat the infection. Summary: The patient is recovering from laparoscopic converted to open resection of retrocaval pheochromocytoma. The patient has developed bilateral atelectasis, and is being treated with supplemental oxygen. The patient also has hypervolemia and is being treated with IV lasix and PO KCl. Antibiotics are not mentioned in the clinical note. Summary: The patient is recovering from laparoscopic surgery converted to open excision of a recurrent retrocaval pheochromocytoma. She has no signs of infection, and antibiotics are not mentioned in the note. The patient's pain is well-controlled, and she is showing interval improvement in her abdomen. Summary: The patient, a [censored]-year-old woman, underwent open excision of a 2.5 cm retrocaval pheochromocytoma on [censored] Post-operatively, she developed no signs of infection, and antibiotics were not administered. Summary: The patient developed acute pulmonary insufficiency following surgery, hypoxemia, and atelectasis due to right main stem intubation, requiring intubation and ICU admission; antibiotics were not administered. Summary: The patient, [censored], [censored]-year-old female, was discharged from the University of Washington Medical Center with a diagnosis of acute pulmonary insufficiency following surgery, hypoxemia, and atelectasis due to right main stem intubation, and recurrent pheochromocytoma. The patient developed an infection after surgery, which was treated with antibiotics. The patient was instructed to keep the wound clean and dry, change the dressing as directed, and take antibiotics as prescribed. Follow-up appointments were scheduled, and the patient was advised to return promptly if experiencing increasing pain, fever, or other concerning symptoms. Summary: The patient, [censored], was hospitalized with acute pulmonary insufficiency following surgery, hypoxemia, and atelectasis due to right main stem intubation, as well as a recurrent pheochromocytoma. The patient was administered antibiotics and was instructed to take them exactly as directed until they were all gone. The patient was also

advised to keep the wound clean and dry, change the dressing as directed, and monitor for signs of infection. Follow-up appointments were scheduled, and the patient was instructed to return promptly if any concerning symptoms occurred. Summary: The patient has an allergy to Augmentin, an antibiotic, which causes swelling. No antibiotics were administered, and no infection was reported...

The structured EHR data for the surgical case is: The maximum white blood cell count for the patient involved in the surgical case from the day of surgery to 90 days after surgery is 11.9. The maximum temperature for the patient involved in the surgical case from the day of surgery to 90 days after surgery is 37.200001. The surgical case is not a reoperation. The surgical case did not result in the order of a wound, tissue, or cerebrospinal fluid culture. The patient was not given antibiotics at any time between 7 days after surgery and 90 days after surgery. A procedure for postoperative drain placement by interventional radiology did not occur for the surgical case. A postoperative infectious disease consultation was not obtained

Answer: This case is not a surgical site infection. Here is the reasoning.
The clinical note summary contains the term "infection" but it does not contain the term "surgical site infection". The term "infection" is not modified by "surgical site". Thus, it cannot be determined for certain that the infection occurred at the surgical site.
Even if the term "surgical site infection" is mentioned in clinical notes, it must be used in the context of diagnosing a surgical site infection or a surgical site infection for which antibiotics are prescribed.
It is not enough for an surgical site infection specific term to only be mentioned in clinical text summaries for a case to be a surgical site infection. For a case to be a surgical site infeciton, the specific term must be used in the context of confirming the diagnosis of the surgical site infection or describing the administration of antibiotics for the infection.
Therefore, this case is not a surgical site infection.

Appendix Supplemental Table 9: Examples used for one-shot and few-shot prompting