FULL/REGULAR RESEARCH (CSCI-RTHI)

Rethinking Population Construction for Clinical AI: Patient Similarity for Surgical Site Infection Prediction

Stephen Russell[a*], Fabio Montes Suros[b], Kamal Babaei Sonbolabadi[b], Jia Zeng[b]

a Intelligent Systems and Robotics Department, University of West Florida, Pensacola, FL, USA
b Data Science Department, Jackson Health System, Miami, FL, USA
** Corresponding author.*
E-mail address: russell@uwf.edu (S. Russell).

Abstract
Surgical site infections (SSIs) are among the most costly and dangerous postoperative complications, yet early identification remains difficult due to delayed onset and complex risk factors. While machine learning has shown promise in SSI prediction, many models rely on retrospective data and non-contemporaneous features, limiting clinical impact. This study proposes a practical SSI prediction approach using only structured electronic medical record (EMR) data. We introduce a novel population sampling method based on patient similarity, employing Jaccard and cosine metrics to identify clinically comparable non-SSI cases. This enhances data balance and training relevance without relying on synthetic data or over sampling. Our model, trained on similarity-filtered data and using no lab, imaging, or free-text features, achieves performance that modestly exceeds state-of-the-art while maintaining real-time feasibility. This work shifts the focus from model complexity to cohort construction, offering a scalable, interpretable solution for infection surveillance and hospital quality improvement.

Keywords: artificial intelligence, machine learning, surgical site infections, healthcare

1. Introduction

Surgical site infections (SSIs) are a serious and common complication of surgery that poses significant challenges to patient health. In the United States alone, SSIs occur in approximately 2-10% of inpatient surgical procedures, an incidence of up to 300,000 cases per year and account for about 20% of all healthcare-associated infections (HAIs), and SSIs are the single most costly type of HAI [1]. Compared to patients who do not develop an SSI, patients who do have significantly worse outcomes and longer hospital stays; each infection prolongs hospitalization by an average of 7-10 days and is associated with a 2-11 times higher risk of post-operative death [2]. Beyond mortality, SSIs substantially impact patient recovery, often causing pain, delayed wound healing, and other complications. Further, SSIs are a leading cause of unplanned 30-day hospital readmissions following surgery [3], undermining quality of care and burdening patients with additional treatments.

SSIs also carry heavy operational and financial consequences for hospitals. Treating an SSI requires significant resources, including additional surgeries, antibiotics, and extended care, increasing the average care-cost, per infection, by tens of thousands of dollars. In aggregate, SSIs contribute an estimated $3.5 - 10 billion in excess health expenditures annually in the U.S. [1]. Hospitals must absorb many of these costs, especially when insurers refuse reimbursement for complications deemed preventable. Furthermore, SSIs adversely affect institutional performance metrics. Elevated SSI rates increase overall resource utilization and can jeopardize a hospital's quality ratings. Regulatory agencies closely monitor SSI occurrences as indicators of care quality and patient safety. For example, the Centers for Medicare & Medicaid Services (CMS) publicly reports hospital SSI rates and may impose financial penalties or reduced Medicare

reimbursements on facilities with poor infection rates. Likewise, accreditation organizations (e.g., The Joint Commission) scrutinize infection control practices. From an accreditation perspective, a pattern of frequent SSIs can put a hospital's accreditation status at risk [4]. Beyond the direct clinical harm to patients, untreated or frequent SSIs can damage a hospital's reputation, financial stability, and compliance with healthcare standards.

Early detection and prevention of SSIs are paramount to improving surgical outcomes and care provisioning overall. Timely identification of an SSI allows clinicians to intervene sooner (for instance, with targeted antibiotics or surgical debridement), limiting the infection's severity and prevent further complications. However, early detection of SSIs is challenging. Many SSIs present after the patient has been discharged, and their initial signs (e.g., mild fever or redness) may be subtle or mistaken for normal post-operative recovery [5]. Traditional surveillance methods often rely on manual chart reviews or laboratory and culture results that can be temporally-offset from the infection's onset, which means infections are frequently confirmed only days or weeks after surgery. This gap creates an urgent need for predictive models that can flag high-risk patients before an SSI fully develops. By leveraging electronic medical record (EMR) data and artificial intelligence (AI), hospitals can potentially monitor post-surgical patients continuously and receive early warnings of infection risk. Such real-time SSI detection systems could prompt proactive clinical evaluations or treatments, thereby reducing morbidity, preventing avoidable readmissions, and ultimately saving lives [5].

Developing an accurate and practical SSI prediction model, however, poses several challenges. SSIs are multifactorial events influenced by a complex interplay of patient factors (e.g., diabetes, obesity), surgical factors (procedure type, duration, technique), and perioperative care processes [6]. This complexity often necessitates large, detailed datasets, including laboratory values, microbiology results, or imaging findings, to achieve good predictive performance. Yet in many real-world settings, especially in the immediate post-operative period, such detailed data may not be readily available. These challenges are not new and there has been extensive research to develop exceptionally precise AI models that rely on detailed clinical information and often place constraints on generalized use and population variances.

The objective of this research seeks to investigate whether a simple, generalizable pre-modeling population refinement method can be used with minimal input features to create an AI model that is viable for early SSI detection, useful for SSI surveillance, and simplifies quality management and reporting. In the following sections of this paper, a background of the current literature on predicting SSIs is presented. Following this background, the development of an AI model for real-time SSI prediction using only structured EMR data and patient-similarity-driven preprocessing is presented. The results of testing the model are presented, concluding with a discussion and conclusion.

2. Related Work

Machine learning (ML) has emerged as a powerful tool for improving SSI risk prediction. A systematic review identified 24 studies describing 85 ML-based AI models for SSI detection [7]. ML algorithms commonly used in prior work include logistic regression, decision trees, random forest, support vector machines, gradient boosting methods (such as XGBoost), and artificial neural networks. Early studies applying ML to SSI prediction achieved modest performance improvements. Bartz-Kurycki et al. [8] used random forest and logistic regression models for neonatal surgery patients, reaching an area under the receiver operating characteristic (AUC) of 0.68. Similarly, Song et al. [9] achieved an AUC of 0.73 using tree-based methods for cardiac device-associated SSIs. AUC values of 0.5 for binary predictors are considered as randomly guessing; and so, the early work though pioneering and showing the feasibility of using ML to predict SSIs did not produce reliable results. With larger datasets, more advanced models showed stronger results: Chen et al. [10] employed a neural network to predict colorectal surgery SSIs, achieving 0.76 AUC. More recent efforts like Al Mamlook et al. [11] applied deep neural networks to 2.88 million surgical cases, achieving an AUC of ~0.85. Wu et al. [12] similarly developed an XGBoost-based model for arthroplasty-related SSI prediction, attaining an AUC above 0.90.

These prior efforts highlight that AI can be used to predict SSI, particularly when large-scale, diverse data is available. However, most of the prior studies included a broad array of clinical features e.g., laboratory, imaging, microbiology or bedside observation, which are not always readily accessible or lag the onset of the infection. Moreover, despite the favorable results and accuracy, there was not many studies that gave emphasis on false positives. One of the main barriers to effective SSI prediction is severe class imbalance. While the humanitarian and fiscal impact of SSIs cannot be understated, from a total hospital population perspective, SSIs typically occur in a relatively small percentage of surgical cases [11], resulting in datasets heavily skewed toward the negative class. Without some form of correction, ML models will tend to predict the majority class, yielding deceptively high accuracy but poor sensitivity and precision for true SSIs [7].

There is ample research that has shown that models trained without addressing imbalance suffer from low positive predictive value despite acceptable AUCs [11], [13], [14]. Petrosyan et al. [15] reported an AUC of 0.89 using random forest but achieved a precision of only 34%, meaning two-thirds of predicted SSI cases were false positives. Xiong et al. [14] faced similar issues with an AdaBoost model for spinal surgery SSIs, achieving only 0.23 precision and applied Synthetic Minority Over-sampling Technique (SMOTE) [16] to improve precision scores to 0.6250 with only minor reduction in recall. Oversampling methods such as SMOTE have proven effective in addressing imbalanced dataset for predicting SSI [17]. Al-Ahmari et al. [17] found that applying SMOTE improved Matthews correlation coefficient (MCC) scores and yielded better balanced recall and precision for SSI models. Other approaches, including over/under-sampling and hybrid strategies, have also been employed but with mixed success depending on dataset characteristics [18]. In addition to resampling techniques, improving the composition of the training dataset through better population selection has gained attention. Restricting modeling to specific surgery types (e.g., colorectal surgery [19], [20], orthopedic procedures [21]) can help by increasing the prevalence of SSIs and improving model relevance. These studies indicate that restrictive sampling can yield better predictive ML performance by constraining training data.

There is a considerable amount of prior work on using AI and ML models to predict SSIs with varying and promising degrees of success. The aforementioned review of the literature shows that incorporating observational (i.e., bedside visuals and clinician notes) and clinical (i.e., labs/microbiology, imaging, and pharmacy) data have a correlation to model performance. While such data is certainly an important consideration, one limitation in much of the prior work is the dependence on such observable and clinical features that, as noted previously, often lag the SSI onset or are available later in treatment timelines. It is also noteworthy that much of the prior work prioritizes the AI model in the research which, while essential, lowers the emphasis on training data selectivity and sampling. A data-first, versus model-first philosophy can elevate concerns that would only emerge *after* model training and may provide tractable guarantees on reliability. Many of these techniques use categorical segmentation (e.g. specific surgery types or demographics, such as patient age) combined with over/under-sampling or synthetic data methods to diffuse imbalanced training data effects. However, these training data manipulation methods introduce unreal (synthetic) data and/or directed emphasis on the minority (positive) class that is not representative of true operational/production distributions. This is not to disparage adaptive sampling or approaches like SMOTE; they can positively impact ML model performance. Yet, their broader efficacy is likely tied to the severity of the organic data imbalance in which the resulting model is employed.

In contrast to the rich body of prior work, the objective of the research presented in this paper is to investigate if leveraging readily available patient data, coupled with a simple, understandable, and adjustable patient-similarity approach to handle data imbalances, can yield an effective AI model for SSI surveillance. If achievable, such a model has the potential to enhance patient safety, support clinical decision-making, provide enriched SSI-related quality management, and improve surgical care outcomes on a broad scale. The following section outlines the technical approach towards this objective.

3. Methods

To address the previous mentioned objective, our approach focuses on using only structured EMR data that are routinely documented during care (excluding labs, notes, imaging, and pharmacy records). These structured data include information such as patient demographics, comorbidities, and surgical details, which are available in near real-time for every patient. Patient comorbidities are captured through the accumulation of diagnosis codes as encounters progress. This allows the early creation of derived features such as the Medicare Severity Diagnostic-Related Grouping (DRG) and maximum acuity diagnosis (maxAcuityDx). DRG is based on principal diagnosis, secondary diagnoses (including complications and comorbidities), procedures performed, and patient demographics. The maxAcuityDx feature is derived using a crosswalk (e.g., Charlson/Elixhauser) that associates diagnosis codes and their related acuity (major comorbid condition, comorbid condition, or neither) to ascertain a maximum acuity level for each encounter. While final diagnosis codes and related metrics are typically assigned retrospectively after patient discharge, through the use these derived features our model exploits these features for prospective use. In such a (close to) real-time operational window, features like DRG and maxAcuityDx would leverage their most current, provisionally available values documented up to the moment of prediction during the patient's active hospital stay. Notably, previous SSI risk models have largely focused on static preoperative risk factors and did not incorporate ongoing clinical observations [6]. In contrast, our model leverages temporal clinical data streams and simple but robust feature engineering to capture the evolving patient condition.

Another key challenge of the objective is the relatively small number of SSIs cases compared to uncomplicated surgeries, which leads to highly imbalanced data -- far fewer cases of “infection” than “no infection.” Class imbalance can impair machine learning algorithms, causing them to favor the majority class (no infection) and miss the SSI events (minority class). The technical approach seeks to mitigate this issue through a population similarity technique. In essence, we identify and include in the training dataset those past surgery cases that are most similar to the patient population of interest (for example, using distance metrics like Jaccard and cosine similarity on patient feature profiles). By selecting a cohort of clinically comparable patients in the majority class, this method enriches the training data for relevant patterns and reduces noise from dissimilar cases. It is important to note that there is prior work on patient-similarity metrics [22], [23], [24]. However, improving the nature of patient-case similarity is not the focus of this research. Rather the technical approach here selects simple, but reliable, similarity metrics to maintain a focus on model explainability and understanding.

3.1 Data Preparation and Exploratory Data Analysis

All data were collected passively, and de-identified in compliance with the Health Insurance Portability and Accountability Act (HIPAA). Studies performed on deidentified data constitute non-human subject studies. Data access was governed by an internal data use agreement and is not publicly available. The initial dataset consisted of 151,733 inpatient encounters that spanned two years (2022 & 2023) of inpatient cases (encounters) drawn from a large health care system’s EMR system. Within the dataset 20,062 (7.6%) were elective (versus emergency) surgeries; 73,300 (48%) patients were males; and there were 109,822 unique patients.

SSI positive encounters were identified by SSI-related ICD-10 diagnosis (Dx) codes (e.g. T81.4XXX) in the encounter final coding. International Classification of Diseases version 10 (ICD-10) diagnosis codes are standardized alphanumeric codes used internationally to classify and record diseases, symptoms, and other health conditions for clinical and billing purposes. Table 1 shows a list of relevant ICD-10 Dx codes. These core codes include the T81.4XXX series (e.g., T81.41XA for superficial incisional surgical site infection), K65.0 (Generalized (acute) peritonitis), K65.1 (Peritoneal abscess), K68.11 (Postprocedural infection of other intra-abdominal organ or structure), and O86.03 (Infection of obstetric surgical wound, organ and space site). While Table 1 presents a broader list of ICD-10 diagnosis codes, it is important to note that the presence of general symptoms like R10.0 (Acute unspecified abdominal pain) alone were not used to classify an encounter as SSI positive. These broader codes, such as R10.0, are included in Table 1 because they were found to be consistently present and frequently associated with the final coding of encounters that were definitively confirmed as SSI positive by the more specific diagnostic codes. This

reflects the nuanced real-world clinical documentation where general symptoms often accompany or are documented alongside specific diagnoses of surgical site infections, particularly following procedures like colon or hysterectomy. The stringent and clinically specific definition of SSI employed, as evidenced by the relatively low prevalence of SSI cases (3%) in our dataset, further underscores this approach.

A specific class of SSI, those related to colon or hysterectomy procedures, was chosen to better align with the research in the literature and SSIs related to these surgeries are of high importance to hospital quality metrics [25], [26]. One might argue that the larger population could be filtered simply on encounters who had a colon or hysterectomy related surgery and doing so would provide sufficient fidelity to restrict the majority class population. However, filtering on patients who had a colon or hysterectomy surgery assumes that all patients undergoing the same surgery are clinically comparable, but this is false. Two patients may undergo a colectomy but may differ significantly in many ways: comorbidities (e.g., diabetes, immunosuppression, malnutrition); preoperative infection status; emergency vs. elective context; and concurrent procedures or complications. These differences can affect SSI risk [27]. While potentially improving majority class imbalance issues, filtering only on surgical type can aggregates dissimilar patients and introduces heterogeneity that can confuse a machine learning model during training.

Table 1. ICD-10 Diagnosis codes frequently associated with colon or abdominal hysterectomy related SSI encounters

Code	Description
T81.41XA	Infection following a procedure, superficial incisional surgical site, initial encounter
T81.43XA	Infection following a procedure, organ and space surgical site, initial encounter
T81.44XA	Sepsis following a procedure
T81.49XA	Infection following a procedure
T81.32XA	Disruption of internal operation (surgical) wound, not elsewhere classified, initial encounter
T81.9XXA	Unspecified complication of procedure, initial encounter
K65.0	Generalized (acute) peritonitis
K65.1	Peritoneal abscess
K65.9	Peritonitis, unspecified
K68.11	Postprocedural retroperitoneal abscess
K91.89	Other postprocedural complications and disorders of digestive system
Z98.890	Other specified postprocedural states
L02.211	Cutaneous abscess of abdominal wall
R10.0	Acute unspecified abdominal pain
O86.03	Infection of obstetric surgical wound, organ and space site

There are practical considerations as well. Encounters in the minority class will include: 1) patients who have surgery at the provider's hospital and get a SSI in the same encounter, 2) patients who have a surgery at the provider's hospital and *return* with an SSI (thus, surgical data associated with a previous encounter); and 3) patients who did not have a surgery at provider hospital and came in with an SSI (thus, less available surgical data). Thus, post discharge EMR information would be available for SSI encounters that fell into category 2 and be "none" for the other two scenarios. This pragmatic problem required an additional task of locating and associating encounter records using the patient identifier (medical record number/MRN), ultimately creating a single row for each "unique" pair of encounters (if available).

Table 2 shows the list features employed to develop the AI model, in addition to the unique identifiers (MRN and account numbers). To address the aforementioned 3 scenarios of SSI encounters, the notion of "AcctNo_Best" was introduced. The "best" account number was used as the encounter identifier most likely to return values for surgical procedure information, e.g. operating room (OR) procedure quantities, anesthesia minutes, and other surgical details. Table 2 also shows the features associated with the best encounter. These features are primarily numeric and were thusly set to zero when no data was available. Textual features were one-hot encoded [28] to address missing or null values. One-hot encoding is a method of converting categorical variables into a binary vector representation, where only one element is "hot" (set to 1) and all others are set to zero, uniquely identifying each text-value category. At the end of this process, the dataset consisted of 146,451non-SSI encounters and 5,282 SSI encounters. It is fairly evident that that before doing any machine learning better population sampling is necessary to address the imbalance.

3.2 Overview of The Technical Approach

To address the challenge of constructing a clinically meaningful negative class for surgical site infection (SSI) prediction, we developed a similarity-based filtering strategy anchored to confirmed SSI cases. This approach avoids reliance on random resampling, synthetic oversampling, or procedure-based rules, each of which can distort cohort composition or reduce generalizability. To create a more balanced training dataset, first we compute diagnosis and procedure similarities. For each negative (candidate) encounter, we compute Jaccard distance diagnosis codes (JD) and procedure codes (JP) individually-pairwise with respect to each encounter in a *reference pool of clinically confirmed* SSI cases. Each candidate encounter was represented as a two-dimensional vector [JD, JP], where JD and JP quantify overlap with a specific SSI-positive encounter. We then measure the cosine similarity between each candidate vector [JD, JP] and every SSI-

Table 2. Model feature list

Feature Name	Feature Description	Data Type	Incl. In Model
MRN	Individual Medical Record Number	Numeric	No
AcctNo	Current encounter Acct number	Numeric	No
DischDate	Discharge date for current encounter	Datetime	No
LOS	Lenth of stay for current encounter	Numeric	Yes
admissionType_key	Type of admission (elective, emergency, trauma, etc.)	Text	Yes
SSI	SSI Flag associated with current encounter (0/1; 1 = Positive)	Binary	Target
encounter_id_Prev	Previous encounter Acct number	Numeric	No
DaysBtwPrevVisit	Days between last and current visit	Numeric	Yes
QtyPriorIPvisits	Number of prior inpatient visits	Numeric	Yes
VolProcs	Volume of procedures, current encounter	Numeric	Yes
VolProcs_SSI	Volume of SSI-specific procedures, current encounter	Numeric	Yes
VolProcs_Prev	Volume of SSI-specific procedures, previous encounter	Numeric	Yes
VolProcs_SSI_Prev	Volume of procedures, previous encounter	Numeric	Yes
proc1	First procedure code (prioritizing SSI-Specific PCS code) current encounter	Text	Yes
proc2	Second procedure code (prioritizing SSI-Specific PCS code) current encounter	Text	Yes
proc1Prev	First procedure code (prioritizing SSI-Specific PCS code) previous encounter	Text	Yes
proc2Prev	Second procedure code (prioritizing SSI-Specific PCS code) previous encounter	Text	Yes
AcctNo_Best	Logical "SSI-relevant" encounter - first or previous	Numeric	No
Age	Age at time of "AcctNo_Best" encounter	Numeric	Yes
Sex	Patient Gender	Text	Yes
Race	Patient Race	Text	Yes
drg	"AcctNo_Best" Diagnosis Related Grouping Code	Numeric	Yes
AcuityDRG	"AcctNo_Best" DRG acuity, e.g. comorbid condition, major comorbid condition or neither	Numeric	Yes
maxAcuityDx	"AcctNo_Best" maximum Dx Code acuity, e.g. comorbid condition, major comorbid condition or neither	Numeric	Yes
BMI	"AcctNo_Best" Body Mass Index	Float	Yes
preOrHours	"AcctNo_Best" Hours between registration and 1st OR procedure	Numeric	Yes
AnesMinsMaxSurgDur	"AcctNo_Best" Anesthesia duration (minutes) for longest surgery	Numeric	Yes
IntubateMins	"AcctNo_Best" Duration of intubation in OR	Numeric	Yes
CardioPulClampMins	"AcctNo_Best" on of cardiopulmonary clamp (minutes)	Numeric	Yes
dischargeDisposition	"AcctNo_Best" Details of discharge disposition	Text	No
SurgHospSvc	"AcctNo_Best" Hospital service	Text	Yes
SurgSpecialty	"AcctNo_Best" Surgical specialty	Text	Yes
ClinHandOff	"AcctNo_Best" Clinical handoff in/after surgery (0/1; 1 = True)	Binary	Yes
ANEPostOp	"AcctNo_Best" Anesthesia Post Op Power Plan (0/1; 1 = True)	Binary	Yes
orQty	"AcctNo_Best" Quantity of OR records (surgeries)	Numeric	Yes
orProcQty	"AcctNo_Best" Quantity of OR procedures	Numeric	Yes
orProcCancelQty	"AcctNo_Best" Quantity of canceled OR procedures	Numeric	Yes
maxOrProcMinutes	"AcctNo_Best" Longest OR procedure duration (minutes)	Numeric	Yes
maxAnesDurationMins	"AcctNo_Best" Longest OR Anesthesia duration (minutes)	Numeric	Yes

positive vector [JD_k, JP_k]. The maximum cosine similarity across all SSI-positive cases was retained. Encounters exceeding a predefined similarity threshold were included in the negative training cohort. This procedure emphasizes balanced structural similarity between diagnoses and procedures, penalizing cases with disproportionate overlap in only one domain. The composite set of negative-SSI encounters that have a cosine score above the threshold and the reference SSI-positive encounters form the ML model training set. This method does not serve as a classifier itself; rather, it is a preprocessing heuristic designed to curate a more realistic and challenging negative cohort. The downstream predictive ML model subsequently learns the decision boundary between SSI-positive and filtered negative cases.

We selected cosine similarity over alternatives (e.g., averaging JD and JP, independent thresholding) because its inductive bias emphasizes balance between diagnosis and procedure similarity. In this setting, the objective is not to predict SSI directly but to curate a negative cohort that is both clinically plausible and sufficiently challenging for the classifier. By penalizing cases with extreme asymmetry (e.g., strong comorbidity match but unrelated procedure, or vice versa), cosine discourages inclusion of encounters unlikely to provide meaningful counterexamples. Importantly, cosine similarity is not applied as a decision function for SSI risk. Its role is limited to preprocessing: constructing a negative class anchored to the structure of known SSI-positive cases. While cosine is insensitive to magnitude, this limitation is acceptable in our design because the final predictive model learns the discriminative boundary between true SSIs and curated negatives. Model performance, not the filter alone, validates the utility of the approach. In this respect, cosine similarity functions analogously to hard-negative mining strategies used in contrastive learning — an intentionally imperfect mechanism for data construction that nonetheless produces superior downstream models. The similarity portion of our method does not claim to find the "most at-risk" cases, nor to define patient similarity in the absolute sense. Rather, it creates a clinically anchored cohort of counterexamples, centered on known infections, for use in training a discriminative model.

3.3 Encounter Similarity Determination

To create the model, ideally majority class observations need to be from the same population and have a balanced number of observations compared to the minority class examples. Besides the methods discussed in the imbalance portion of this paper, it is possible to filter of patients with surgeries and then randomly select from that sub sample. However, selecting a population that is "similar" to the minority-class examples would likely be more representative of the true relevant population and thus, give better performance for AI model training. The key questions here are: similar based on what and how "similar" is similar?

There are many methods in the literature for determining the clinical similarity of encounters including techniques like clustering, e.g. KNN or K-Means [29], [30]; fuzzy approaches [31], and patient similarity networks [23]. These methods have demonstrated effective approaches for determining population similarity, but they generally required detailed clinical details and none of the prior work we found focused on patient similarity applied to SSI prediction. Given the focus on SSI prediction of our work, we adopt two straightforward and reliable similarity metrics, Jaccard and cosine distance. These similarity metrics have the added benefit of keeping the approach simple and easily explainable [32]. To determine similarity, encounters with like final ICD-10 diagnosis codes and like ICD-10 procedure (PCS) codes are compared, using the clinically validated minority class encounters as the basis set.

The construct of AcctNo_Best created in the data preparation step is essential to determining encounter similarity because similarity is defined as two encounters that share like Dx and PCS codes. To do that each record must be associated with the proper encounter, containing the SSI-relevant diagnosis and procedures. Using the known SSI encounters and excluding any SSI Dx codes, non-SSI encounters can be identified by shared diagnosis and procedures codes, given a threshold-degree of distance between the non-SSI encounter and any one of the SSI encounters. Equation 1 shows the expression of Jaccard similarity where two sets, a and b, are measured in terms of similarity; the resulting Jaccard value ranges between 0 to 1. Equation 2 expresses the cosine similarity calculation where X and Y are the components of two vectors. The cosine distance values are in the range of -1 to 1, and subtracting the absolute value from 1 gives the similarity. Each encounter-pair will generate a vector consisting of the Jaccard score for Dx codes and the Jaccard score for PCS codes. This vector is then pairwise applied with a self-similar vector to provide a single

measure of similarity. Each feature comparison (Dx and PCS codes) is best measured by Jaccard because their nature mandates a set comparison, but a single metric that summarizes the two encounters' similarity is necessary for a thresholding-filter.

$$Jaccard\_Sim(a,b) = \frac{a \cap b}{a \cup b} \quad (1) \qquad Cosine\_Sim(X,Y) = \frac{X \cdot Y}{\| X \| \| Y \|} \quad (2)$$

Figure 1 illustrates the overall framework flow. At the center is a notional example of the similarity-based selection process. In this example, encounter *a* is a known SSI encounter and encounter *b* is a known non-SSI encounter. It is important to clarify that the Jaccard index is not applied to the numeric features (e.g., Age, LOS) or one-hot encoded textual features listed in Table 2. These features serve as the direct input to the downstream machine learning prediction model. Instead, the Jaccard similarity metric is exclusively utilized as a pre-modeling population construction step for comparing set-based features: specifically, ICD-10 diagnosis (Dx) codes and ICD-10 procedure (PCS) codes between patient encounters. Jaccard similarity is calculated separately for the Dx codes and the PCS codes. Each feature comparison (Dx and PCS codes) is best measured by Jaccard because their nature mandates a set comparison. The resulting Jaccard scores for Dx and PCS codes then form a vector, which is subsequently used to calculate a cosine similarity score—the ultimate metric for filtering non-SSI cases to construct our balanced training set. For ease of understanding, in Figure 1, a text description of the respective ICD-10 codes for each encounter are shown. After Jaccard similarity has been determined, pairwise cosine similarity is calculated. In this calculation, the SSI encounter self-similarity vector is used resulting in a single score for each encounter-pair. A "fully self-similar vector" represents an encounter's perfect similarity to itself. This means that when an encounter's codes are compared to themselves, the Jaccard similarity is 1.0. In Cosine similarity, the [1.0, 1.0] vector acts as a benchmark against which the similarity of other encounters is measured, allowing a single cosine similarity score to be derived for each encounter-pair.

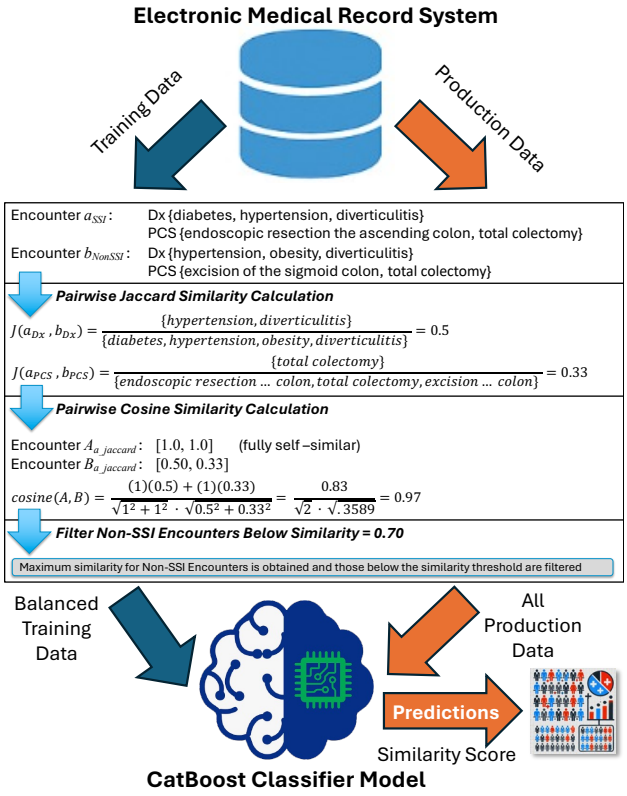

Figure 1. Notional similarity-based majority class record selection

In all pairwise similarity calculations, the non-SSI encounters are only compared to the SSI encounters allowing the maximum cosine similarity to be elicited and used for filtering. To achieve this, for each non-SSI encounter, Jaccard similarities for both diagnosis (Dx) and procedure (PCS) codes are first computed against every individual SSI encounter in the minority class. The Jaccard similarities form vectors, for each encounter, that are used to compute cosine scores against every individual SSI encounter. This process generates a set of pairwise cosine similarity scores for each non-SSI encounter. The highest score from this set, representing the maximum similarity of that non-SSI encounter to any SSI case, is then selected. This maximum cosine similarity score is subsequently evaluated against the 0.70 threshold; if it meets or exceeds this value, the non-SSI encounter is included in the "similar" majority class for the balanced training set.

A threshold value of 0.70 was employed resulting in reducing the 151,733 records to 15,339; 5,282 inpatient SSI encounters (appx 34%) and 10,057 "similar" non-SSI inpatient encounters. This 0.70

threshold was pragmatically selected as a conservative measure to achieve a substantial initial improvement in class imbalance, effectively increasing the proportion of SSI cases from approximately 3% to 34% within the filtered dataset. The 0.70 threshold, could be adjusted higher or lower, subject to a desired clinical specificity. However, setting the threshold too high may inadvertently exclude a significant number of non-SSI encounters that are indeed clinically analogous to SSI cases. Similarly, too low of a threshold would result in allowing too many clinically dissimilar cases. We recognize that the similarity threshold could be optimized and will explore tuning this parameter in future work.

3.4 ML Model Construction

The final dataset used for model training consisted of 15,339 records, comprising 5,282 SSI cases and 10,057 similar non-SSI cases. Figure 2 shows the results of the similarity and SMOTE process. With the improved training data balance, we formulate the SSI prediction problem as a classic binary classification problem using tabular data. Here the x-input is the above list of features and the y-output is a binary representation indicating SSI present (1) or not (0). Given the x-input, several AI models were developed, using no hyperparameter optimization initially, then applied a grid search optimization approach once the best model was identified. It is important to note that despite the population similarity filtering, SMOTE was employed to balance no-SSI (negative, majority class) and SSI (positive, minority class) groups. SMOTE created additional new minority data by interpolation within the available minority data via bootstrap sampling and data generation via a k-nearest neighbor algorithm, with a *k* value of 5 [16]. The 15,339-record dataset was separated into two randomly selected subsets: 1) an 80/20 train-test split training set and 2) a hold-out validation set.

4. Results

A 10-fold validation iteration was applied to assess the average validation final results. Results were measured via accuracy, precision (selectivity), recall (sensitivity), AUC, and F1 (a measure of the balance between precision and recall, based on the harmonic mean of both). Table 3 shows the results of the test data for several AI models. CatBoost [33] performed the best, providing reasonable scores across all the metrics. Though slightly lower than the best performing models in the literature, the CatBoost model's scores are consistent and competitive with most of the prior work. It is particularly noteworthy that the best performing models from the literature relied on a greater number and more clinically specific features (such as laboratory results and/or observables such as swelling, fever, or purulent discharge); many of which are difficult to obtain in real (or close to real) time from EMR systems.

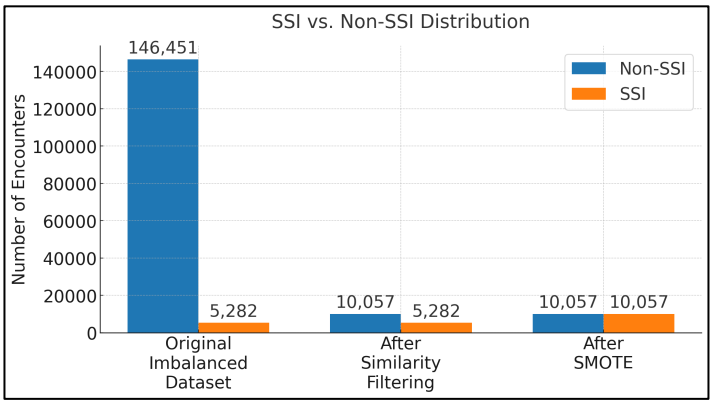

Figure 2. Distribution of data imbalance before and after.

The exclusion of such less accessible, but potentially beneficial, time lagged or clinical-detail features was deliberate. To investigate the feature selection, a SHAP (Shapley Additive exPlanations) technique [34] was conducted to identify which features contributed to the model results, and in what way. SHAP is a model-agnostic technique used to explain the output of machine learning models. It builds on cooperative game theory, particularly Shapley values, to attribute the contribution of each feature to a model's prediction in a fair and theoretically justified manner. SHAP has several desirable properties, beyond simple feature sensitivity, that help explain the predictive contribution of individual features.

Figure 3 shows the features that contribute most to the predictive power of the final CatBoost model. Interpreting the plot, each feature is scored based on its low or high (blue to red) contribution to DRG most influenced the model. This is intuitive because DRG is a system used to classify hospital inpatient stays into categories for the purpose of payment. Each DRG represents a group of clinically similar diagnoses and procedures that are expected to require similar hospital resources and costs. DRG is based on principal diagnosis, secondary diagnoses (including complications and comorbidities), procedures performed, and patient demographics (age, sex, discharge status). The inclusion of DRG might call into question the need for the similarity metric used in dealing with training data imbalance. However, DRGs have a fiscal orientation that is intended to provide financial estimates and assessment for billing purposes. Moreover, DRGs aggregate (and exclude) many encounter characteristics and thus, may have ambiguous fidelity in determining encounter similarity. Other features of note in Figure 3 include: quantity of prior visits (QtyPriorVisits); procedures (proc1 and proc2); quantity of procedures (VolProcs and VolProcs_prev); admission type (emergency and newborn); the surgeons' surgical specialty (SurgSpecialty); and anesthesia minutes occurring in the longest surgical duration (anesMinsMaxSurgDur).

Table 3. Model results for validation data

Model	Accuracy	AUC	Recall	Prec.	F1
CatBoost Classifier (best model)	.8463	.9052	.7317	.8044	.7663
Light Gradient Boosting Machine	.8416	.9029	.7349	.7906	.7616
Extreme Gradient Boosting	.8407	.9001	.7301	.7913	.7593
Random Forest Classifier	.8198	.8823	.7309	.7422	.7363
Gradient Boosting Classifier	.8179	.8819	.7357	.7358	.7356
Ada Boost Classifier	.7902	.8434	.7036	.6929	.6978
Extra Trees Classifier	.7939	.8558	.6765	.7110	.6931
Decision Tree Classifier	.7611	.7403	.6738	.6473	.6602
K Neighbors Classifier	.7254	.7788	.6911	.5863	.6342
Logistic Regression	.7287	.7814	.6660	.5951	.6283
Ridge Classifier	.7241	.7782	.6584	.5892	.6216
Linear Discriminant Analysis	.7242	.7771	.6578	.5894	.6215
SVM - Linear Kernel	.7126	.7724	.6787	.5805	.6189
Naive Bayes	.7161	.7247	.4926	.6126	.5436
Quadratic Discriminant Analysis	.3985	.5990	.9268	.3630	.5155

5. Discussion

The similarity approach was effective in improving the substantial training data imbalance. The CatBoost model also performed consistently with much of the prior work (significantly better in many cases) and utilized fewer clinical features. The prior work on SSI prediction commonly placed core focus on making an optimized and generalizable AI model, concentrating on pristine datasets with hyper-relevant features. While appropriate for advancing theoretical approaches, such as the creation of new machine learning models, this model-first focus can limit adoption and challenge real-world implementations. Unlike models reliant on retrospective or delayed data, our approach explicitly leverages contemporaneously available structured EMR data. This includes using provisional values for features such as DRG and maxAcuityDx, which would be dynamically updated throughout a patient's active hospitalization. This enables the model to provide early, pre-discharge risk assessments supporting real-time clinical decision-making and quality improvement initiatives.

The addition of an imbalance-improving similarity method for population sampling provides a tuneable means to improve an intrinsic problem with many clinical AI regimes; proportionally low volumes of the minority (positive ailment) class. The goal of the similarity method is not to perfectly distinguish positive from negative patients with cosine. It is to construct a negative training cohort that is challenging, clinically relevant, and well-aligned with the structure of known SSI cases. A 0.70 similarity threshold was used and while reasonable, the threshold did not have a clinical rationale and was not necessarily optimal. To that end, the methods in this paper could be optimized in many ways e.g., a more sophisticated similarity metric, the inclusion of additional clinical features, highly optimized ML algorithms, etc. Nonetheless, this research effort shows the utility of generally available features, perhaps creating a baseline, and the feasibility of employing even a simple similarity-based population selector can make a difference training data balance. More work is necessary in all three areas: similarity-based population selection, data imbalance correction, and SSI prediction writ large.

It is noteworthy that there may be an additional benefit at inference time. Similarity can also be computed between new encounters and the training SSI reference pool. While it not used as a hard inclusion rule for inference, this score provides an interpretable indicator of distributional alignment: high similarity values suggest structural similarity to training cases, whereas low values highlight encounters that are dissimilar and should be interpreted with caution. This dual use would enable the similarity method to function both as a training-time filter and as an inference-time signal for model explainability.

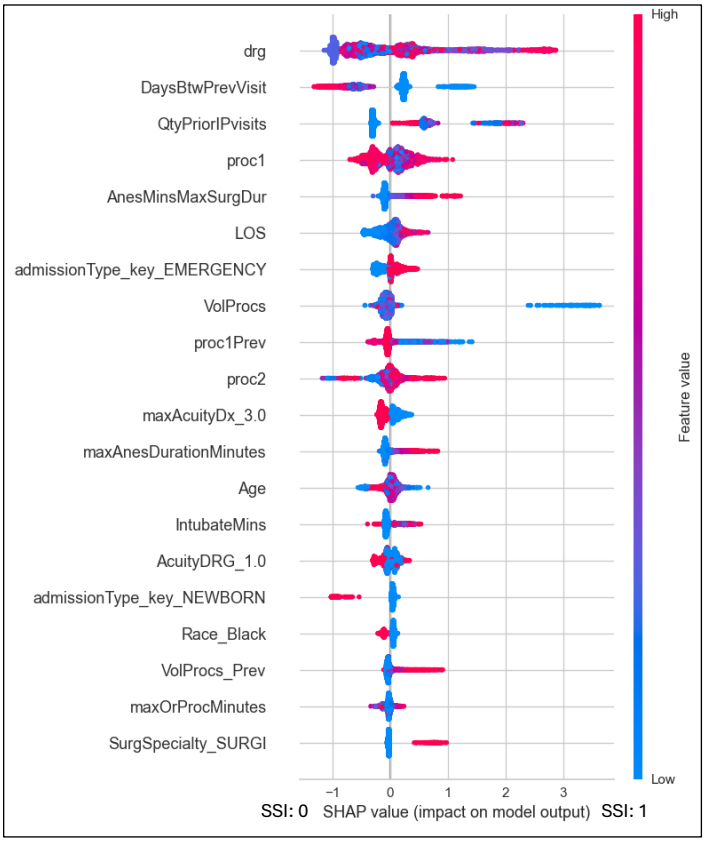

Figure 3. SHAP summary

Despite promising advances shown in this and other prior work, key gaps remain. Many published SSI prediction models lack external validation and have only been tested retrospectively. As a result, their generalizability to new settings remains an open research concern. This study also requires longitudinal study of the model's performance over time. Specifically, we acknowledge that our methodological choice to apply similarity-based population construction prior to the train-test split implies the test set is drawn from a population that has already undergone this refinement. As such, the reported performance metrics are indicative of the model's efficacy within this pre-defined, clinically analogous cohort. This approach might present an optimistic estimate if the entire framework were to be applied directly to a completely raw, un-curated, and highly imbalanced stream of EMR data without the prior application of the similarity-based selection process.

In a complete end-to-end real-world deployment scenario, any patient encounter not meeting the initial 0.7 cosine similarity threshold would effectively be considered a non-SSI case, or at least fall outside the high-risk group specifically targeted by this framework. In such instances, the prediction model itself would not process or classify these cases. This distinction clarifies the operational scope of our model, which is designed to operate on a carefully curated, clinically analogous population. While this approach enhances the model's robustness to data distribution shifts by focusing on relevant cases, it implies that the reported performance reflects the system's efficacy within this defined patient cohort. Class imbalance remains a persistent problem affecting model sensitivity and general clinical applicability. While methods like SMOTE and similarity-based sampling offer solutions, further research is needed to optimize sampling strategies and calibration approaches for highly imbalanced clinical data. Therefore, future work should involve robust external validation on independent, raw datasets, where the entire pipeline (i.e., the similarity-based population construction followed by model inference) is applied afresh. This will more comprehensively assess the framework's real-world generalizability and performance in a prospective setting, taking into account cases that are implicitly classified as non-SSI by the initial similarity filter.

6. Conclusion

The ability to predict SSIs could significantly improve clinical outcomes by enabling earlier interventions, thereby reducing postoperative morbidity, mortality, and readmissions. This work presents a pragmatic

approach to surgical site infection (SSI) prediction that departs from conventional practices in healthcare machine learning. Most prior efforts have focused on algorithmic sophistication while accepting the training data as given, often addressing class imbalance through under/oversampling or synthetic data generation. In contrast, we introduce a pre-modeling population construction strategy that intelligently selects training samples based on clinical similarity. By applying Jaccard and cosine similarity metrics to diagnosis and procedure codes, we identify non-SSI cases that are comparable to SSI cases, producing a better-balanced training set on which to build the AI model. Unlike commonly used stochastic or embedding-based sampling techniques, our method is simple, interpretable, and scalable, making it suitable for real-time hospital deployment. Importantly, our model uses only contemporaneously available structured EMR data, excluding labs, imaging, and pharmacy, making it capable of supporting early intervention and quality reporting. This design begins to enable a real-time utility in infection control workflows, a capability missing from most retrospective SSI prediction studies. Furthermore, this research shifts the focus from model selection to data engineering, emphasizing that a carefully constructed cohort can elevate the performance of even standard classifiers. This represents a subtle but significant reorientation of priorities in clinical ML research toward building better systems, not just better AI models.

The contribution in this research is not grounded in algorithmic machine learning e.g., inventing a new machine learning model. Rather, this research focuses on engineering the science of AI implementation, targeting how to better choose the right examples before training the AI and providing operational boundaries for where the AI will be most effective; a particularly critical consideration for quality management and reporting in real-world hospital settings. The novelty in this study is in a simple, but powerful rethinking of population construction for clinical ML deployment where imbalance handling aligns with the clinical intuition of comparing "like with like" and can provide a repeatable framework for other imbalanced healthcare predictive problems. Future work is planned to examine the generalizability of the approach utilizing SSI data from the National Surgical Quality Improvement Program (NSQIP) Database. This future work will also seek to expand the similarity approach and investigate more complex similarity methods and optimized thresholding.

References

[1] K. A. Ban *et al.*, "American College of Surgeons and Surgical Infection Society: surgical site infection guidelines, 2016 update," *Journal of the American College of Surgeons*, vol. 224, no. 1, pp. 59–74, 2017.

[2] K. B. Kirkland, J. P. Briggs, S. L. Trivette, W. E. Wilkinson, and D. J. Sexton, "The impact of surgical-site infections in the 1990s: attributable mortality, excess length of hospitalization, and extra costs," *Infection Control & Hospital Epidemiology*, vol. 20, no. 11, pp. 725–730, 1999.

[3] M. T. Kassin *et al.*, "Risk factors for 30-day hospital readmission among general surgery patients," *Journal of the American College of Surgeons*, vol. 215, no. 3, pp. 322–330, 2012.

[4] J. Shepard *et al.*, "Financial impact of surgical site infections on hospitals: the hospital management perspective," *JAMA surgery*, vol. 148, no. 10, 2013, Accessed: Apr. 26, 2025. [Online]. Available: https://jamanetwork.com/journals/jamasurgery/article-abstract/1730490

[5] P. C. Sanger *et al.*, "A prognostic model of surgical site infection using daily clinical wound assessment," *Journal of the American College of Surgeons*, vol. 223, no. 2, pp. 259–270, 2016.

[6] D. J. Anderson *et al.*, "Strategies to prevent surgical site infections in acute care hospitals: 2014 update," *Infection Control & Hospital Epidemiology*, vol. 35, no. S2, pp. S66–S88, 2014.

[7] A. M. van Boekel *et al.*, "Systematic evaluation of machine learning models for postoperative surgical site infection prediction," *PloS one*, vol. 19, no. 12, p. e0312968, 2024.

[8] M. A. Bartz-Kurycki *et al.*, "Enhanced neonatal surgical site infection prediction model utilizing statistically and clinically significant variables in combination with a machine learning algorithm," *The American Journal of Surgery*, vol. 216, no. 4, pp. 764–777, 2018.

[9] J. Song, E. S. Buenaventura, B. Cohen, J. Liu, D. Yao, and E. Larson, “Predictive Models for Surgical Site Infection (SSI) in Patients with a Permanent Pacemaker (PPM) Using Machine learning Methods,” Jun. 11, 2020, *Preprints*. doi: 10.22541/au.159188536.65812462.
[10] K. A. Chen, C. U. Joisa, J. M. Stem, J. G. Guillem, S. M. Gomez, and M. R. Kapadia, “Improved prediction of surgical-site infection after colorectal surgery using machine learning,” *Diseases of the Colon & Rectum*, vol. 66, no. 3, pp. 458–466, 2023.
[11] R. E. Al Mamlook, L. J. Wells, and R. Sawyer, “Machine-learning models for predicting surgical site infections using patient pre-operative risk and surgical procedure factors,” *American journal of infection control*, vol. 51, no. 5, pp. 544–550, 2023.
[12] G. Wu *et al.*, “Development of machine learning models for the detection of surgical site infections following total hip and knee arthroplasty: a multicenter cohort study,” *Antimicrob Resist Infect Control*, vol. 12, no. 1, p. 88, Sep. 2023, doi: 10.1186/s13756-023-01294-0.
[13] A. Heffernan, R. Ganguli, I. Sears, A. H. Stephen, and D. S. Heffernan, “Choice of Machine Learning Models Is Important to Predict Post-Operative Infections in Surgical Patients,” *Surgical Infections*, p. sur.2024.288, Mar. 2025, doi: 10.1089/sur.2024.288.
[14] C. Xiong *et al.*, “Construct and Validate a Predictive Model for Surgical Site Infection after Posterior Lumbar Interbody Fusion Based on Machine Learning Algorithm,” *Computational and Mathematical Methods in Medicine*, vol. 2022, pp. 1–11, Aug. 2022, doi: 10.1155/2022/2697841.
[15] Y. Petrosyan *et al.*, “Predicting postoperative surgical site infection with administrative data: a random forests algorithm,” *BMC Med Res Methodol*, vol. 21, no. 1, p. 179, Dec. 2021, doi: 10.1186/s12874-021-01369-9.
[16] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: synthetic minority over-sampling technique,” *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.
[17] S. Al-Ahmari and F. Nadeem, “Improving Surgical Site Infection Prediction Using Machine Learning: Addressing Challenges of Highly Imbalanced Data,” *Diagnostics*, vol. 15, no. 4, p. 501, 2025.
[18] K. L. Colborn, M. Bronsert, E. Amioka, K. Hammermeister, W. G. Henderson, and R. Meguid, “Identification of surgical site infections using electronic health record data,” *American Journal of Infection Control*, vol. 46, no. 11, pp. 1230–1235, 2018.
[19] S. Y. Cho *et al.*, “Development of machine learning models for the surveillance of colon surgical site infections,” *Journal of Hospital Infection*, vol. 146, pp. 224–231, 2024.
[20] Z. Shen *et al.*, “The development and validation of a novel model for predicting surgical complications in colorectal cancer of elderly patients: results from 1008 cases,” *European Journal of Surgical Oncology*, vol. 44, no. 4, pp. 490–495, 2018.
[21] A. A. A. Chowdhury, A. Sultana, A. H. Rafi, and M. Tariq, “AI-driven predictive analytics in orthopedic surgery outcomes,” *Revista Espanola de Documentacion Cientifica*, vol. 19, no. 2, pp. 104–124, 2024.
[22] A. Defilippo, P. Veltri, P. Lió, and P. H. Guzzi, “Leveraging graph neural networks for supporting automatic triage of patients,” *Scientific Reports*, vol. 14, no. 1, p. 12548, 2024.
[23] S. Pai and G. D. Bader, “Patient similarity networks for precision medicine,” *Journal of molecular biology*, vol. 430, no. 18, pp. 2924–2938, 2018.
[24] A. Sharafoddini, J. A. Dubin, and J. Lee, “Patient similarity in prediction models based on health data: a scoping review,” *JMIR medical informatics*, vol. 5, no. 1, p. e6730, 2017.
[25] D. A. Caroff *et al.*, “The limited utility of ranking hospitals based on their colon surgery infection rates,” *Clinical Infectious Diseases*, vol. 72, no. 1, pp. 90–98, 2021.
[26] D. S. Yokoe, T. R. Avery, R. Platt, K. Kleinman, and S. S. Huang, “Ranking hospitals based on colon surgery and abdominal hysterectomy surgical site infection outcomes: impact of limiting surveillance to the operative hospital,” *Clinical Infectious Diseases*, vol. 67, no. 7, pp. 1096–1102, 2018.
[27] E. Korol *et al.*, “A systematic review of risk factors associated with surgical site infections among surgical patients,” *PloS one*, vol. 8, no. 12, p. e83743, 2013.

[28] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by back-propagating errors,” *nature*, vol. 323, no. 6088, pp. 533–536, 1986.
[29] L. Dai, H. Zhu, and D. Liu, “Patient similarity: methods and applications,” Dec. 01, 2020, *arXiv*: arXiv:2012.01976. doi: 10.48550/arXiv.2012.01976.
[30] N. Wang *et al.*, “Study on the semi-supervised learning-based patient similarity from heterogeneous electronic medical records,” *BMC Med Inform Decis Mak*, vol. 21, no. S2, p. 58, Jul. 2021, doi: 10.1186/s12911-021-01432-x.
[31] P. Muthukumar and G. S. S. Krishnan, “A similarity measure of intuitionistic fuzzy soft sets and its application in medical diagnosis,” *Applied Soft Computing*, vol. 41, pp. 148–156, 2016.
[32] T. P. Rinjeni, A. Indriawan, and N. A. Rakhmawati, “Matching scientific article titles using cosine similarity and jaccard similarity algorithm,” *Procedia Computer Science*, vol. 234, pp. 553–560, 2024.
[33] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, “CatBoost: unbiased boosting with categorical features,” *Advances in neural information processing systems*, vol. 31, 2018, Accessed: May 07, 2025. [Online]. Available: https://proceedings.neurips.cc/paper/2018/hash/14491b756b3a51daac41c24863285549-Abstract.html
[34] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” *Advances in neural information processing systems*, vol. 30, 2017, Accessed: May 07, 2025. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html