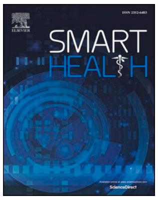
Smart Health 37 (2025) 100596

Contents lists available at ScienceDirect

Smart Health

journal homepage: www.elsevier.com/locate/smhl



ML-predicted surgical site infections: An epidemiological study utilizing machine learning on routinely collected healthcare data to predict infection risk

Davide Golinelli [a,b], Simona Rosa [c], Paola Rucci [c], Francesco Sanmarchi [c,*], Dario Tedesco [d], Carlo Biagetti [e], Alessio Gili [b], Andrea Bucci [f], Luca Romeo [f], Roberto Grilli [a]

[a] *Division of Health Services and Policy Research, AUSL Romagna, Ravenna, Italy*
[b] *Department of Life Science, Health, and Health Professions, Link University, Rome, Italy*
[c] *Department of Biomedical and Neuromotor Sciences, Alma Mater Studiorum - University of Bologna, Italy*
[d] *Department of Medicine and Surgery, University of Parma, Parma, Italy*
[e] *Division of Infectious Diseases, Infermi Hospital, AUSL Romagna, Rimini, Italy*
[f] *Department of Economics and Law, University of Macerata, Italy*

ARTICLE INFO

Keywords:
Surgical site infections
Machine learning
Orthopaedic surgery
Risk stratification
Healthcare data
XGBoost
Personalized medicine

ABSTRACT

Background: Surgical site infections (SSIs) are a major public health issue, causing increased morbidity, longer hospital stays, and higher healthcare costs. Despite progress in infection control, predicting and preventing SSIs remain crucial for improving patient outcomes. This study examines the use of machine learning (ML) on routinely collected healthcare data (RCD) to predict SSIs in orthopaedic surgery, aiming to improve risk stratification and guide interventions.

Objectives: To develop, test, and validate an ML predictive model using RCD to assess SSI risk in orthopaedic surgery patients.

Methods: A retrospective study was carried out using RCD from a 1.2 million population in an Italian Local Health Authority, covering surgeries from 2017 to 2021. The population included patients undergoing hip or knee arthroplasty and open reduction of fractures. Several ML algorithms, including eXtreme Gradient Boosting (XGBoost), were used for model development. The models' performance was assessed by recall, accuracy, and area under the receiver operating characteristic curve (AUC). A feature importance analysis identified key SSI risk predictors.

Results: The XGBoost model demonstrated superior performance, with a recall exceeding 70% and an AUC>0.70, overcoming other methods. Significant predictors included the ASA classification, opioid use, priority class of the surgery operation, and length of hospital stay.

Conclusions: ML models, particularly XGBoost, effectively predicted SSI risk in orthopaedic patients, offering a new approach to infection control and prevention. Incorporating ML and RCD highlights the potential for scalable, data-driven personalized medicine interventions. Future research will focus on model validation and integration of these tools into healthcare systems for enhanced patient management.

---

* Corresponding author. Via San Giacomo 12, 40126, Bologna, Italy.
E-mail address: francesco.sanmarchi@studio.unibo.it (F. Sanmarchi).

https://doi.org/10.1016/j.smhl.2025.100596
Received 15 March 2025; Received in revised form 6 June 2025; Accepted 25 June 2025

1. Background

Surgical site infections (SSIs) are among the most common healthcare-associated infections (Bhangu et al., 2018; Scardoni et al., 2020; Seidelman et al., 2023), with approximately 0.5%–3% of patients undergoing surgery experiencing infection at or adjacent to the surgical incision site (Bhangu et al., 2018). The assessment of risk and the surveillance of SSIs is a critical activity to quantify and minimize the clinical risk associated with surgery. The surveillance of SSIs is particularly challenging and costly, as it involves a prolonged postoperative follow-up. With an estimated cost of €10,000–20,000 per infection, SSIs nonetheless represent the most significant healthcare-associated infection in terms of economic impact (Buttazzi et al.; Owens & Stoessel, 2008). They also result in worse health outcomes for patients, leading to prolonged hospital stays, increased mortality, and compromised quality of life (Bhangu et al., 2018; Buttazzi et al.; Owens & Stoessel, 2008; Seidelman et al., 2023; Surgical Site Infection, 2008).

According to the most recent national reports, the incidence of SSIs in Italy ranges between 2% and 9% of all surgical procedures (De Oliveira and Sarmento Gama, 2017; Protocollo Sistema et al., 2011; Seidelman et al., 2023). In the Emilia-Romagna region (Northern Italy), over 145,000 surgical procedures are performed annually, of which approximately 36,000 are orthopaedic surgeries. SSIs are observed in 1–2% of all regional interventions, affecting between 1500 and 3000 people yearly (De Oliveira and Sarmento Gama, 2017; Protocollo Sistema et al., 2011; Seidelman et al., 2023). SSIs occur during postoperative hospitalization or after hospital discharge, in a critical post-discharge phase where the patient is at risk of being lost to follow-up. This crucial issue could be alleviated by a system capable on the one hand of defining and stratifying the risk of individual patients (risk stratification profiling) to develop SSIs, and on the other hand of monitoring and effectively surveilling patients in the post-discharge phase.

Advances in information technology and the progressive availability of digitalized health information offer new tools for the sector, including the automation of SSI prediction and surveillance. Although traditional automated and semi-automated SSI prediction and surveillance systems are based on fixed, a priori-defined classification algorithms built on well-defined parameters, various studies suggest that artificial intelligence (AI), and, specifically, machine learning (ML), can support the development of more precise and effective SSI prediction and surveillance tools (Bhangu et al., 2018; Buttazzi et al.; De Angelis et al., 2023; Grundmeier et al., 2018; Luz et al., 2020; Mamlook et al., 2023; Peiffer-Smadja et al., 2020; Scardoni et al., 2020; Wu et al., 2022). Indeed, the landscape of SSI prediction has been significantly enriched by the advent of ML models, which have showcased considerable success in identifying patients at substantial risk for SSIs. For instance, a study by Mamlook and colleagues (Mamlook et al., 2023) investigated a multitude of functioning models, including sophisticated deep neural network (DNN) approaches that boast high accuracy, with some achieving good prediction capabilities.

By using innovative ML techniques to calculate and assign a specific risk to each patient, it is possible to create a patient risk stratification system that allows classifying patients with different profiles, to implement more appropriate preventive interventions, prophylaxis, and treatment on one hand, and more effective and efficient surveillance on the other.

1.1. Related work

While traditional SSI surveillance systems rely on fixed-rule algorithms and clinician-driven reporting, advancements in AI and ML offer the potential for dynamic, data-driven approaches. ML has demonstrated substantial promise in predicting SSIs by identifying non-linear patterns and complex interactions in large clinical datasets (Bhangu et al., 2018; Buttazzi et al.; De Angelis et al., 2023; Grundmeier et al., 2018; Luz et al., 2020; Mamlook et al., 2023; Peiffer-Smadja et al., 2020; Scardoni et al., 2020; Wu et al., 2022). Recent studies have explored ML models for SSI prediction, particularly in specialties such as neurosurgery and head and neck surgery. A systematic literature review by Scardoni et al. (Scardoni et al., 2020) retrieved eight studies reporting on the application of ML-based models to predict, control, and assess SSI and their risk factors pre- and post-surgery in European and US surgery departments (Haas et al., 2013; Hu et al., 2017; Kuo et al., 2021; Weller et al., 2018; Soguero et al.). They found promising results of applying ML-based models to SSI prediction coming from neurosurgery, and head and neck surgery settings. However, many of the studies analyzed by Scardoni et al. (Scardoni et al., 2020) primarily utilized ad hoc collected data, emphasizing the preliminary success of machine learning models in specific surgical settings. The scalability and broader application of such predictive algorithms necessitate integration with routinely collected data (RCD), which are more standardized and prevalent across diverse healthcare systems. Indeed, the use of risk prediction and stratification tools should be based on available and reliable data, ideally validated and certified. For the development of these tools to scale from experimental models to routine use, it is essential to utilize standardized data sources present across multiple healthcare systems. In the Italian healthcare system, as well as in most Western countries' healthcare systems, RCDs feed databases used for evaluation, research, planning, and economic-financial reporting. These RCDs include discharge records, pharmaceutical usage, and inpatient and outpatient visit data, among others.

1.2. Aim of the study

This study aims to develop a predictive model based on ML to predict the risk of developing SSI and to classify patients undergoing orthopaedic surgery based on their preoperative risk, using data from the RCD. The focus will be on three orthopaedic surgical procedures: hip arthroplasty, knee arthroplasty, and open reduction of fracture. The ML predictive model will be trained, and its efficacy validated using retrospective data from databases of Azienda Unità Sanitaria Locale della Romagna (serving 1.2 million citizens), through a supervised (human-in-the-loop) process for algorithm refinement on the stages of choice, validation, and interpretation of predictors.

2. Methods

2.1. Study design, population and data

We conducted a retrospective cohort study using data derived from RCD. The study population comprises adult patients who underwent surgical procedures for hip arthroplasty, knee arthroplasty, and open fracture reduction, and were monitored for SSIs in the SiCHER data flow for the years 2017–2021.

For this study, various routinely collected healthcare databases were utilized. The SIChER database served as the core data source for constructing the ML algorithm. The regional surveillance system for surgical site infections - SIChER -, developed by the Regional Health and Social Agency of the Emilia-Romagna region (northern Italy) and based on the technical specifications provided by the European Centre for Disease Prevention and Control (Buttazzi et al.). SIChER data (2017–2021) were linked with the databases (for the same years and retrospectively until 2014): Hospital Discharge Records, Pharmaceutical prescriptions, Outpatient Specialist Visits.

Three surgical procedures were considered in this study: hip arthroplasty, knee arthroplasty, and fracture reduction. The RCD database used in this work consisted of almost 850 variables (listed in Appendix), including preoperative risk factors, intra-operative variables, and morbidity outcomes. Considering only preoperative risk factors, type of surgery, and intraoperative risk factors, a thorough literature review and expert clinical opinions were used to refine the list of potential features to include. Some of the variables included in our study were binary variables related to ICD-9 or ATC codes. For instance, the variable “ATC” considered 364 4-digit ATC codes. For each of these codes, we created a binary (dummy) variable that took value 1 if the patient has been prescribed that particular drug, and 0 otherwise. The same approach was used for the disease codes in the hospitalization data and the health services code for the ASA database. In the former case, we created 105 dummy variables for the same number of 2-digit ICD-9 primary diagnosis codes, while in the latter we had 353 binary variables for 4-digits ASA codes.

2.2. Data analysis and model development

For the classification of surgical site infections, data from all surgical interventions of patients who underwent orthopaedic surgery from 2017 to 2021 at AUSL of Romagna were used. Given that SSI is a relatively rare event, it was decided to classify the variable related to the presence or absence of infection using a machine learning model suitable for an imbalanced setting. To mitigate this problem, we introduced a hyperparameter to adjust the weighting of positive and negative samples in the loss function. The hyperparameter value was set by calculating the ratio of negative to positive instances in our training data. This procedure effectively increases the penalty for misclassifying positive instances, guiding the model to pay greater attention to these rare but clinically important cases during training (Wade and Glynn, 2020). This strategy mitigates class imbalance without altering the original data distribution by helping the model to better capture the minority (positive) class. To optimize the generalization performance while handling high-dimensional data and imbalanced classes, the eXtreme Gradient Boosting (XGBoost) algorithm was selected as a strategic compromise. This algorithm classifies the response variable by aggregating outputs from multiple weak decision models (*i.e.*, decision trees), effectively addressing the challenges posed by the dataset’s composition. The algorithm’s output for each patient is the "probability" of infection. An observation is classified as an infection when the probability is greater than or equal to 50%. This means that the algorithm can also be used to classify a patient into different infection risk bands depending on the probability of infection.

XGBoost (Chen et al., 2015) was selected as the core predictive model because of its ensemble architecture, which combines multiple decision trees and has demonstrated stronger generalization capacity, and superior robustness to overfitting when compared to other ML methods widely used for solving predictive medicine tasks (Bernardini et al., 2019). These peculiarities make XGBoost particularly suitable for real-world healthcare datasets characterized by high dimensionality, moderate noise, and class imbalance. Its ability to capture complex, nonlinear interactions enhances predictive reliability, even in the presence of inter-patient variability inherent to routinely collected healthcare data. Prior studies have demonstrated the effectiveness of XGBoost in tackling similar challenges across diverse clinical prediction tasks, including the estimation of diabetes-related complications (Nicolucci et al., 2022) and the prediction of organ dysfunction in critically ill ICU patients with COVID-19 (Montomoli et al., 2021).

The algorithm was tested using a stratified 10-fold cross-validation, which partitioned the entire dataset into ten non-overlapping windows for each cycle of cross-validation. Within each window, the a priori probability between classes was maintained. In each iteration, the algorithm was trained on nine windows of a sample of 25,715 patients and tested on the remaining 2857 patients. Consequently, ten classifications were obtained on an equal number of sub-samples. A cross-validation strategy (5-fold) and grid search focusing on maximizing recall (the proportion of infections correctly classified) were used to optimize the hyperparameters within the training dataset, aiming to regulate model complexity and avoid overfitting. As a result, each division in the external loop was trained using the ideally tuned hyperparameters established in the internal loop. Key hyperparameters for XGBoost included the number of boosting iterations, learning rate, maximum tree depth, and the ratio of subsampling training features.

Feature importance was computed using XGBoost’s built-in mechanism based on the frequency criterion, i.e. the number of times a given feature was used to split the data across all trees in the ensemble. This metric provides a global view of the relevance of each feature in the model’s prediction. The higher the frequency, the more the model relied on that feature when constructing the boosted decision trees. Specifically, we extracted the normalized importance scores (summing to 1 across all features) from each best-performing XGBoost model across the cross-validation folds. These scores reflect the relative contribution of each feature to the model’s decisions. To enhance robustness, we averaged the importance values across folds and selected the top 10 features with the highest mean importance.

Lastly, SIChER also includes a composite indicator, the Infection Risk Index (IRI), which is assigned to each monitored surgical procedure and calculated on: duration of the procedure, contamination class, ASA score, and surgical technique. The IRI was proposed by the NHSN and is based on studies conducted between 1981 and 1991 (Luz et al., 2020; Peiffer-Smadja et al., 2020). For this study, the potential correlation of the IRI variable with the outputs of the ML algorithms we implement was evaluated.

Ethics approval

This study has been approved by the Romagna CEROM research ethics committee board with identifier code: Prot. 8485/2022, I.5/271, approved on 11-11-2022.

3. Results

The study population included 28,752 surgical interventions of patients who underwent orthopaedic surgery (Table 1). The prevalence of SSI was 1.8% (518 out of 28,752). 36.2% of the sample were males, with a mean age of 69.4 (SD ± 16.5). 10,970 (38.4 %) hip replacement procedures were performed, followed by ORIF and knee replacements. The mean duration of surgery was 72.4 min (38.7), and the mean length of hospital stay was 9.2 days (SD = 6.2).

The classification results for the 10-fold cross-validation procedure are displayed in Table 2 and Fig. 1 (where we reported average values across the 10 folds): the results of the XGBoost algorithm are compared with those from a random forest, a decision tree and a logistic regression in terms of Recall (proportion of correctly identified infections among all infections), Accuracy (ACC), that represents the overall correctness of the classification, and the Area under Receiving Operative Characteristic Curve (AUC). All metrics range between 0 and 1, with higher values indicating better performance.

The XGBoost algorithm presents the highest Recall and AUC in each of the samples, with values of Recall higher than 60% and AUC above 0.60. The best performance in terms of accuracy is either provided by the random forest algorithm or the logistic regression model with values around 98%. Nevertheless, it should be recalled that in a very unbalanced sample where the outcome of interest is extremely rare, the accuracy itself could be not a valid measure since it does not account for the number of times the outcome is badly predicted.

Table 1
Descriptive statistics of the population.

Variable	
Infection – N (%)	518 (1.8 %)
Sex – Male – N (%)	10330 (36.2 %)
Anesthesiological Risk Assessment (ASA):	N (%)
1. Healthy patient	3334 (11.7 %)
2. Patient with mild systemic disease and no functional limitation	14572 (51.0 %)
3. Patient with severe systemic disease and moderate functional limitation	10103 (35.4 %)
4. Patient with a severe systemic disease	559 (2.0 %)
5. Dying patient, whose survival is not guaranteed for 24 h	4 (0.03 %)
Surgical session type:	N (%)
1. Elective (session planned at least 24 h before)	17485 (61.2 %)
2. Urgent (session planned in the last 24 h)	10194 (35.7 %)
3. Unknown	893 (3.1 %)
Antibiotic prophylaxis:	N (%)
1. Yes	16910 (59.2 %)
2. No	9660 (33.8 %)
3. Unknown	2002 (7.0 %)
Admission type:	N (%)
1. Scheduled ordinary	9619 (33.7 %)
2. Emergency admission	12178 (42.6 %)
3. Scheduled admission with pre-hospitalization	6635 (23.2 %)
4. Emergency admission from short-stay observation unit	140 (0.5 %)
Facility type:	N (%)
1. Public	9580 (33.5 %)
2. Private nursing home	18992 (66.5 %)
Surgical procedure site:	N (%)
1. Hip arthroplasty	10970 (38.4 %)
2. Knee arthroplasty	7058 (24.7 %)
3. Fracture reduction	10544 (36.9 %)
Priority class for hospitalization:	N (%)
1. Class A to be operated within 30 days	4463 (15.6 %)
2. Class B to be operated within 60 days	6728 (23.6 %)
3. Class C to be operated within 180 days	4303 (15.1 %)
4. Class D to be operated within a year	760 (2.7 %)
5. Not present	12318 (43.1 %)
Age in years – mean (standard deviation)	69.41 (16.5)
Duration of surgery in minutes – mean (standard deviation)	72.42 (38.7)
Length of stay in days – mean (standard deviation)	9.22 (6.2)

Table 2
ML’s prediction models results and comparison.

Fold	Xgboost			Random forest			Decision Tree			Logistic		
	Recall (%)	Acc (%)	AUC	Recall (%)	Acc (%)	AUC	Recall (%)	Acc (%)	AUC	Recall (%)	Acc (%)	AUC
1	74.51	70.33	0.768	50.00	98.18	0.595	50.82	97.94	0.700	50.00	98.18	0.500
2	67.64	69.80	0.675	50.00	98.18	0.700	51.78	97.97	0.642	50.00	98.18	0.501
3	63.25	67.52	0.671	50.00	98.21	0.607	49.82	97.86	0.582	49.99	97.97	0.497
4	65.51	68.18	0.687	50.00	98.21	0.658	51.80	97.97	0.645	50.09	98.20	0.500
5	70.73	74.03	0.719	50.00	98.17	0.698	51.69	97.79	0.667	50.09	98.18	0.500
6	62.63	69.23	0.679	50.00	98.17	0.593	49.89	97.97	0.630	50.09	98.18	0.500
7	66.28	70.84	0.666	49.98	98.14	0.671	49.93	98.04	0.657	50.09	98.18	0.500
8	67.29	74.69	0.691	50.00	98.18	0.688	49.84	97.86	0.631	50.09	98.18	0.500
9	63.96	68.15	0.657	50.00	98.18	0.622	51.87	98.14	0.588	50.09	98.18	0.500
10	69.80	70.35	0.721	50.00	98.18	0.629	51.71	97.83	0.663	50.09	98.14	0.499

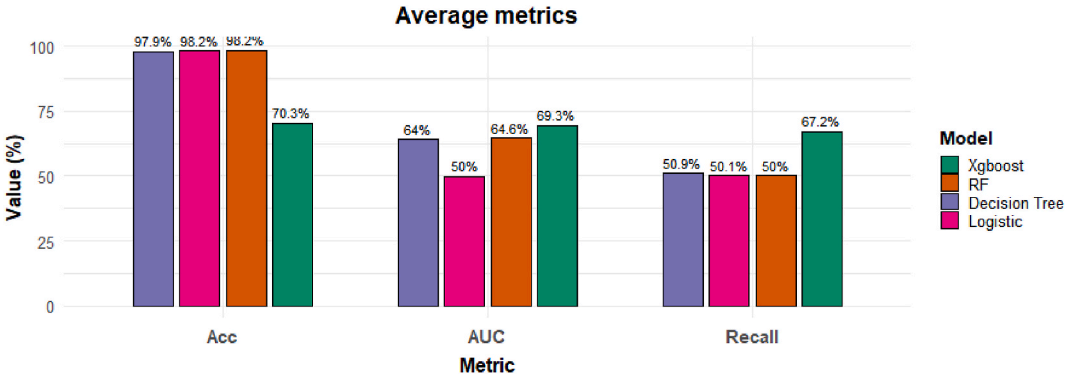


Fig. 1. Average prediction metrics among the 10 folds.

To add robustness to our results, we used statistical tests to compare the performances (Recall, ACC, and AUC) of the XGBoost with the random forest, the decision tree, and the logistic regression model. Since all metrics had a distribution departing from normality, we used non-parametric, Wilcoxon signed-rank tests ($\alpha = 0.05$).

The performance in terms of Recall for the XGBoost model was significantly better compared to both Random Forest ($W = 55$; $Z = 2.752$; $p < 0.05$) and Decision Tree models ($W = 55$; $Z = 2.752$; $p < 0.05$). Similarly, the AUC of XGBoost was significantly higher compared with Random Forest ($W = 49$; $Z = 2.141$; $p < 0.05$) and Decision Tree models ($W = 55$; $Z = 2.752$; $p < 0.05$).

We also computed the linear and nonlinear correlation of the probability obtained through XGBoost with the IRI score, finding an almost null correlation between them. This may mean that the information extracted from the RCD is different from that used in the definition of the IRI score.

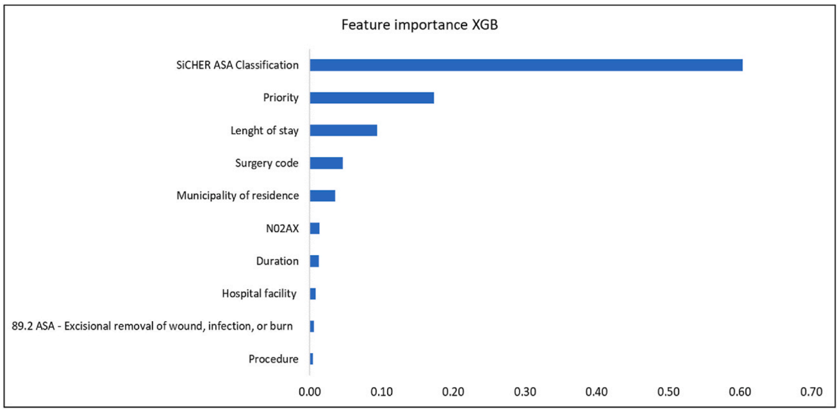


Fig. 2. Features with relative importance in the construction of the boosted decision trees.

3.1. Features relevance

To add some explainability to our approach, we have analyzed the importance of the features used as inputs in our algorithm. The features' importance was calculated by measuring the number of times a feature was used to split the data across all trees. This approach reflects the relative importance of each feature in the construction of the boosted decision trees (XGBoost model). The results are reported in Fig. 2. It can be noticed that the most important input feature was the SICHER ASA Classification, the priority class of the surgery operation was the second feature in terms of relevance while being prescribed and using ATC N02AX drugs (e.g., tramadol, dezocine, oliceridine, tilidine and naloxone) also appeared as an input feature and predictor of SSI. Among the most relevant features, length of stay has a twofold interpretation: on the one hand, an infection would inevitably lead to a longer stay and this could be somehow an endogenous effect; on the other hand, patients with longer hospital stays are more prone to developing infections.

4. Discussion

This study leverages ML techniques on routinely collected healthcare data to predict the risk of SSIs in patients undergoing orthopaedic surgery. Our findings show the potential of ML in enhancing the precision of SSI risk prediction, which is crucial for implementing targeted preventive measures, thereby improving patient outcomes, and optimizing resource allocation, thus reducing the burden on patients and healthcare systems.

The application of the eXtreme Gradient Boosting (XGBoost) algorithm demonstrated a superior ability to classify patients into their infection risk, with a recall higher than 62 % and an area under the receiver operating characteristic curve (AUC) above 0.65. This performance indicates its robustness in identifying patients at higher risk of SSIs, a critical step towards effective preventive measures. The success of XGBoost in our study aligns with emerging literature suggesting that ML can offer nuanced insights into healthcare-associated infection risks, supporting conventional statistical models in handling complex, non-linear relationships and high-dimensional data (Abbas et al., 2019; Culver et al., 1991; De Angelis et al., 2023; Haley et al., 1981; Kuczewski et al., 2020; Kwong et al., 2024; Luz et al., 2020; Ma et al., 2020; Mamlook et al., 2023; Peiffer-Smadja et al., 2020; Petrosyan et al., 2021; Scardoni et al., 2020; Thirukumaran et al., 2019; Wu et al., 2022).

For instance, Soguero-Ruiz et al. (Soguero et al.) explored different ML-based models for SSI temporal prediction and reported preoperative and postoperative accuracy to range from, respectively, 0.69 and 0.67 to 0.91 and 0.90. Also, Luz et al. (Luz et al., 2020) reviewed 11 studies on ML approaches in the field of SSI. They found that the median features/variables used for developing ML algorithms for SSI were 50, the median study size was 5214, the best model was the L1-regularized logistic regression, and the max AUROC was 0.96. Compared to their results, our model showed worse performance in terms of max AUROC, with XGBoost outperforming other algorithms, but higher feature selection and study size. Therefore, the performance of the XGBoost model was lower than the highest AUROC values reported in some of the literature, where models such as L1-regularized logistic regression or deep neural networks have achieved higher values (Luz et al., 2020; Haas et al., 2013; Hu et al., 2017; Weller et al., 2018; Soguero et al.). Several factors may account for this discrepancy. First, many of these studies were conducted in controlled research settings using prospectively collected or ad hoc datasets, with richer clinical granularity and a higher prevalence of SSIs. By contrast, our model was developed using RCD with inherent limitations in coding precision, missingness, and variability, particularly in post-discharge outcomes, i.e. where many SSIs emerge but are harder to trace without clinical annotations. Moreover, our dataset reflects a real-world population from a regional health system, which includes a wider patient mix and diverse clinical pathways compared to more homogeneous cohorts in academic centers. This broader generalizability may come at the expense of maximal predictive performance, but it enhances the model's potential for future integration into operational surveillance systems. As a pragmatic benchmark, we compared the ML-based approach based on XGBoost to other validated tools (i.e. logistic regression, random forest, decision tree; Table 2). These tools showed lower predictive accuracy when applied to the same data. Finally, it is worth emphasizing that our study does not aim to propose a single superior method, but rather to explore how multiple techniques perform on real-world data. The inclusion of standard models (such as logistic regression) provides a useful baseline to contextualize XGBoost's performance, highlighting its relative strengths in managing non-linear and high-dimensional clinical features.

The results of our study highlight the opportunity of ML tools to integrate and enhance traditional SSI surveillance methods, which rely on fixed, a-priori-defined algorithms. In fact, the model presented can represent a real-world example of how these algorithms can be built relying on data linkage of multiple sources of RCD, integrating into the regional health system of Emilia-Romagna, and potentially into the national and supra-national one. Using RCD can guarantee greater control over input data, greater homogeneity and standardization of information, and potentially greater confidence in the model's outputs.

Comparatively, our ML model's higher recall and AUC metrics indicate its efficacy in correctly identifying true positive cases of SSI, a critical advantage in a field where early detection can dramatically influence patient prognosis. However, it is important to note the context of these achievements within an unbalanced dataset, where SSIs are a relatively rare outcome. Traditional metrics such as accuracy, while high for the random forest model, may not fully capture the predictive model's performance in such scenarios. This underlines the importance of selecting appropriate performance metrics that reflect the model's operational utility in clinical settings.

Our approach underscores the importance of testing various ML techniques to identify the most effective models for specific healthcare applications. The choice of an explainable tool like XGBoost enables stakeholders to understand the model's decision-making process, fostering trust and facilitating clinical and organizational decisions. Explainability in ML models is paramount, especially in healthcare, where decisions directly impact patient outcomes and resource allocation. In our analysis, indeed, the most significant predictors identified (ASA score, the priority class of the surgery operation, and length of stay) carry some clinical and organizational implications. The ASA classification and surgery priority reflect the patient's preoperative health status and the inherent risk of the procedure, highlighting areas for focused preoperative assessment and postoperative care. Meanwhile, the length of stay serves as a complex indicator of both infection risk and its consequence, suggesting a target for interventions aimed at reducing hospitalization times without compromising care quality. Interestingly, LOS emerged as a double-edged indicator, highlighting the bidirectional relationship between hospitalization duration and infection risk. This insight aligns with existing research indicating prolonged hospital stays as both a consequence and a risk factor for SSIs, suggesting a need for strategies that minimize hospitalization time without compromising patient care (Bhangu et al., 2018; De Oliveira and Sarmento Gama, 2017; Luz et al., 2020; Owens & Stoessel, 2008; Seidelman et al., 2023). In comparison with Mamlook's findings (Mamlook et al., 2023), both our study and the referenced research highlight the ASA score as a crucial predictor for SSIs. However, our results differ in the emphasis on the surgery type (emergency vs. elective) as a predictor. In our model, the surgery type played a minor role in predicting SSIs for hip arthroplasty, knee arthroplasty, and fracture reduction procedures. This contrast might be attributable to our focused investigation on orthopaedic surgeries, which inherently carry different risk profiles compared to the broader range of surgeries examined in the other article, including cardiac, gynecology, general, interventional, otolaryngology, orthopaedic, plastic, and vascular surgeries. The diverse nature of these procedures could explain why the surgery type emerged as a more significant factor in the broader study, highlighting the nuanced impacts of surgical context on infection risks.

We also found that patients undergoing pharmacological therapy with opioids (ATCN02AX, e.g., tramadol, dezocine, naloxone) were at higher risk of SSI. This underscores the intricate relationship between preoperative medication regimes and postoperative outcomes. Patients prescribed these medications often present with more complex health profiles, including chronic pain issues and potentially compromised immune systems, which could inherently increase their vulnerability to infections, including SSIs. The relationship between opioid use and increased preoperative risk of SSIs suggests a multifaceted interaction between pain management strategies and infection risk. Therefore, opioids may be considered as a proxy for patients' overall frailty and clinical condition, and a possible predictor of SSI.

In line with Thirukumaran et al. (Thirukumaran et al., 2019) and Petrosyan et al. (Petrosyan et al., 2021), the performance of our model, which is primarily based on administrative data, could be improved by integrating clinical data (e.g., laboratory data) found in electronic health records that are not yet routinely available in Italy. Despite this, our sample, and training and testing database were significantly larger than that used in (Luz et al., 2020; Petrosyan et al., 2021; Scardoni et al., 2020; Thirukumaran et al., 2019), which reinforces our findings.

Furthermore, an effective SSI control strategy cannot be done without two well-defined integrated phases: prediction and surveillance. On one hand, our model, trained on RCD and with SSIs diagnosed and coded by a doctor based on information present during that hospital stay (in addition to a post-discharge surveillance phase), can represent a useful tool for the phase of prediction and stratification of individual patient risk (or population risk). On the other hand, it is necessary to develop surveillance models that are also able to identify the onset of SSIs automatically with AI/ML. An example of this, in a similar setting in Italy, is described by De Angelis et al. (De Angelis et al., 2023). The authors propose an automated surveillance system for SSI identification from hospital discharge letters based on natural language processing (NLP). Ideally, these two systems could be integrated: on one side, the classification algorithm based on NLP could be implemented to perform automatic surveillance on hospital discharge letters. On the other hand, our model could be refined and fed with this data to improve its predictive performance, and therefore to stratify patients before surgical intervention based on their risk, to mitigate the risk of developing SSIs.

Our study has several limitations. The retrospective design and reliance on routinely collected data may introduce biases related to data quality and completeness. Furthermore, the generalizability of our results to other regions or healthcare systems may be limited by differences in data collection practices and healthcare delivery models (Haley et al., 1981). Future research should focus on prospective studies to validate our predictive model's effectiveness in real-time SSI risk assessment and explore its integration with existing healthcare information systems for dynamic risk stratification.

5. Conclusions

In conclusion, our study illustrates the potential of ML to improve SSI prevention and surveillance, providing a dynamic approach to infection control. By leveraging the predictive capabilities of ML on RCD, we attempted to pave the way for more personalized, efficient, and effective healthcare interventions. As we move forward, the integration of AI, and specifically ML, into public health strategies will be crucial in addressing the complex challenges of infection prevention and control, ultimately improving patient outcomes and healthcare system resilience. The results of this retrospective study entail the necessity to conduct prospective studies to evaluate the real-world effectiveness of our predictive tool. Such studies will not only validate the model's predictive accuracy but also its utility in clinical practice, including its integration into existing healthcare information systems for dynamic risk stratification and intervention planning.

CRediT authorship contribution statement

Davide Golinelli: Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Methodology, Investigation, Data curation, Conceptualization. **Simona Rosa:** Methodology, Formal analysis, Data curation. **Paola Rucci:** Writing – review & editing, Validation, Supervision. **Francesco Sanmarchi:** Writing – review & editing, Writing – original draft, Investigation, Conceptualization. **Dario Tedesco:** Writing – original draft, Supervision, Project administration, Investigation, Conceptualization. **Carlo Biagetti:** Writing – original draft, Validation, Supervision, Methodology, Investigation, Data curation, Conceptualization. **Alessio Gili:** Writing – review & editing, Writing – original draft, Validation, Supervision, Resources. **Andrea**

Bucci: Writing – review & editing, Writing – original draft, Validation, Supervision, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Luca Romeo:** Writing – review & editing, Writing – original draft, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Roberto Grilli:** Writing – review & editing, Validation, Supervision, Project administration, Investigation.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used OpenAI's ChatGPT4 in order to improve readability and language of the manuscript, but not in the creation of images, graphics, tables, or for other tasks. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Funding statement

This research received no specific funding.

Declaration of competing interest

The authors declare no conflict of interest.

Acknowledgements

None.

Appendix

Variables list

Sicher.

Variable name	Legend
hospital_code	Hospital code number
session_date	Date of the surgical session
session_time	Time of the surgical session
hospitalization_year	Year of hospitalization
lha_code	Local Health Authority code number
asa_score	ASA score
session_type	Type of surgical session
session_duration	Duration of surgical session (minutes)
surgery_type	Type of surgery
infection_type	Type of infection
infection_site	Site of infection
surveillance_id_n …	Unique identifier of the surgical procedure being monitored
surgical_procedure_code_n …	Surgical procedure code
flag_infection_1	Presence of infection 1
flag_infection_2	Presence of infection 2
flag_infection_3	Presence of infection 3
antibiotic_prophylaxis_Y/N	Antibiotic prophylaxis (YES/NO)
antibiotic_prophylaxis_molecule_1	Antibiotic prophylaxis molecule administered 1.1
antibiotic_prophylaxis_molecule_n	Antibiotic prophylaxis molecule administered 2
prophylaxis_adm_start_date_1	Start date of antibiotic prophylaxis 1
prophylaxis_adm_start_date_n	Start date of antibiotic prophylaxis 2
infection_onset_date_1	Date of the infection onset 1
infection_onset_date_2	Date of the infection onset 2
infection_onset_date_3	Date of the infection onset 3
NHSN category 1	CDC/National Healthcare Safety Network (NHSN) Surveillance category
NHSN category n	CDC/National Healthcare Safety Network (NHSN) Surveillance category
Surgery type	Type of surgical intervention
hospitalization_region_code	Italian region where the patient was hospitalized
mdc	Major diagnostic category
citizenship	Patient's citizenship
priority_class	Hospitalization priority class
municipality	City of residence

(*continued on next page*)

(*continued*)

Variable name	Legend
serum_creatinine	Serum creatinine (mg/dL)
admission_discipline	Admission discipline
discharge_discipline	Discharge discipline
age_years	Patient’s age (years)
age_days	Patient’s age (days)
euroscore	Euroscore
ejection_fraction_preoperative	Preoperative ejection fraction
hospital_length_stay	Hospital length of stay (days).
preoperative_hospital_length_stay	Preoperative hospital length of stay
educational_level	Patient’s educational level
hospitalization_reason	Reason of hospitalization (day hospital)
MPR	Major Procedure Related category
AHRQ_MPR	AHRQ_MPR category
GG_DH	Number of day hospitl access
main_disease	Main disease category (ICD 9)
FLAG_PAT	Disease (YES/NO)
birth_weigth	Weight at birth (grams)
drg_weigth	DRG weigth
blood_pressure_systolic	Systolic blood pressure
pain	Pain (yes/no)
sdo	Hospital discharge form
sex	Patient’s sex
marital_status	Patient’s marital status
lha_type	Type of local health authority
hospital_type_1	Type of hospital 1
hospitalization_type	Type of hospitalization
sdo_type	Type of hospital discharge form
hospital_type_2	Type of hospital 2

ASA.

lha_code	Local Health Authority code number
hospital_code	Code of the hospital
citizenship	Patient’s cityzenship
hospitalization_type	Type of hospitalization
priority_class	Hospitalization priority class
helth_service_code	Code of the provided health service
helth_service_agg_1	Provided health service: first level aggregation
helth_service_agg_2	Provided health service: second level aggregation
helth_service_agg_3	Provided health service: third level aggregation
emergency_dep_code	Emergency department priority level
working_diagnosis	Working diagnosis

PHARMA.

atc	Anatomical Therapeutic Chemical classification system code
antibiotic_days	Length of antibiotic treatment (days)
antibiotic_amount	Amount of administered antibiotic (packages)

SDO (Hospitalization in the previous years)

mdc	Major diagnostic category
priority_class	Hospitalization priority class
birthplace	Birthplace (municipality)
municipality	City of residence
drg_rg	DRG as computed by the regional healthcare administration
drg_type	Type of DRG
age	Patient’s age
educational_level	Patient’s educational level
hospital_length_stay	Hospital length of stay (days)
preoperative_hospital_length_stay	Preoperative hospital length of stay
leave_days	Days of leave during hospitalization
mpr	Grouper: MPR
cod_pat1	Main disease
cod_pat.1	Disease code

Data availability

Data will be made available on request.

References

Abbas, M., De Kraker, M. E. A., Aghayev, E., et al. (2019). Impact of participation in a surgical site infection surveillance network: Results from a large international cohort study. *Journal of Hospital Infection, 102*(3), 267–276. https://doi.org/10.1016/j.jhin.2018.12.003

Bernardini, M., Romeo, L., Misericordia, P., & Frontoni, E. (2019). Discovering the type 2 diabetes in electronic health records using the sparse balanced support vector machine. *IEEE J Biomed Health Inform, 24*, 235–246. https://doi.org/10.1109/JBHI.2019.2899218

Bhangu, A., Ademuyiwa, A. O., Aguilera, M. L., et al. (2018). Surgical site infection after gastrointestinal surgery in high-income, middle-income, and low-income countries: A prospective, international, multicentre cohort study. *The Lancet Infectious Diseases, 18*(5), 516–525. https://doi.org/10.1016/S1473-3099(18)30101-4

Buttazzi R, Ricchizzi E, Gagliotti C, Moro ML. Sorveglianza delle infezioni del sito chirurgico in Emilia-Romagna.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., et al. (2015). Xgboost: Extreme gradient boosting. *R package version 04-2, 1*(4), 1–4.

Culver, D. H., Horan, T. C., Gaynes, R. P., et al. (1991). Surgical wound infection rates by wound class, operative procedure, and patient risk index. *The American Journal of Medicine, 91*(3), S152–S157. https://doi.org/10.1016/0002-9343(91)90361-Z

De Angelis, L., Baglivo, F., Arzilli, G., et al. (2023). Hospital-acquired infections surveillance and prevention: Using natural language processing to analyze unstructured text of hospital discharge letters for surgical site infections identification and risk-stratification. *Intelligence-Based Medicine, 8*, Article 100120. https://doi.org/10.1016/j.ibmed.2023.100120

De Oliveira, A. C., & Sarmento Gama, C. (2017). Surgical site infection prevention: An analysis of compliance with good practice in a teaching hospital. *Journal of Infection Prevention, 18*(6), 301–306. https://doi.org/10.1177/1757177417703190

Grundmeier, R. W., Xiao, R., Ross, R. K., et al. (2018). Identifying surgical site infections in electronic health data using predictive models. *Journal of the American Medical Informatics Association, 25*(9), 1160–1166. https://doi.org/10.1093/jamia/ocy075

Haas, L. R., Takahashi, P. Y., Shah, N. D., et al. (2013). Risk-stratification methods for identifying patients for care coordination. *American Journal of Managed Care, 19*(9), 725–732.

Haley, R. W., Culver, D. H., Emori, T. G., Hooton, T. M., & White, J. W. (1981). Progress report on the evaluation of the efficacy of infection surveillance and control programs. *The American Journal of Medicine, 70*(4), 971–975. https://doi.org/10.1016/0002-9343(81)90563-5

Hu, Z., Simon, G. J., Arsoniadis, E. G., Wang, Y., Kwaan, M. R., & Melton, G. B. (2017). *Automated detection of postoperative surgical site infections using supervised methods with electronic health record data.* Published online.

Kuczewski, E., Munier-Marion, E., Amour, S., et al. (2020). How to save time in surveillance without losing quality: Comparison of two methods of surgical site infection surveillance in orthopaedic prosthetics in a French university hospital. *Journal of Infection Prevention, 21*(5), 202–205. https://doi.org/10.1177/1757177420939240

Kuo, M., Schroeder, R., Barbeito, A., et al. (2021). Preoperative care assessment of need scores are associated with postoperative mortality and length of stay in veterans undergoing knee replacement. *Federal Practitioner, 38*(7), 316–324. https://doi.org/10.12788/fp.0148

Kwong, J. C. C., Nickel, G. C., Wang, S. C. Y., & Kvedar, J. C. (2024). Integrating artificial intelligence into healthcare systems: More than just the algorithm. *Npj Digit Med, 7*(1), 52. https://doi.org/10.1038/s41746-024-01066-z. s41746-024-01066-z.

Luz, C. F., Vollmer, M., Decruyenaere, J., Nijsten, M. W., Glasner, C., & Sinha, B. (2020). Machine learning in infection management using routine electronic health records: Tools, techniques, and reporting of future technologies. *Clinical Microbiology and Infection, 26*(10), 1291–1299. https://doi.org/10.1016/j.cmi.2020.02.003

Ma, R., He, J., Xu, B., et al. (2020). Nomogram prediction of surgical site infection of HIV-infected patients following orthopedic surgery: A retrospective study. *BMC Infectious Diseases, 20*(1), 896. https://doi.org/10.1186/s12879-020-05613-3

Mamlook, R. E. A., Wells, L. J., & Sawyer, R. (2023). Machine-learning models for predicting surgical site infections using patient pre-operative risk and surgical procedure factors. *American Journal of Infection Control, 51*(5), 544–550. https://doi.org/10.1016/j.ajic.2022.08.013

Montomoli, J., et al. (2021). Machine learning using the extreme gradient boosting (XGBoost) algorithm predicts 5-day delta of SOFA score at ICU admission in COVID-19 patients. *Journal of Intensive Medicine, 1*(2), 110–116.

Nicolucci, A., et al. (2022). Prediction of complications of type 2 diabetes: A machine learning approach. *Diabetes Research and Clinical Practice, 190*, Article 110013.

Owens, C. D., & Stoessel, K. (2008). Surgical site infections: Epidemiology, microbiology and prevention. *Journal of Hospital Infection, 70*, 3–10. https://doi.org/10.1016/S0195-6701(08)60017-1

Peiffer-Smadja, N., Rawson, T. M., Ahmad, R., et al. (2020). Machine learning for clinical decision support in infectious diseases: A narrative review of current applications. *Clinical Microbiology and Infection, 26*(5), 584–595. https://doi.org/10.1016/j.cmi.2019.09.009

Petrosyan, Y., Thavorn, K., Smith, G., et al. (2021). Predicting postoperative surgical site infection with administrative data: A random forests algorithm. *BMC Medical Research Methodology, 21*(1), 179. https://doi.org/10.1186/s12874-021-01369-9

Protocollo Sistema nazionale di Sorveglianza delle infezioni del sito chirurgico (SNICh). (2011). Aggiornamento dicembre. https://assr.regione.emilia-romagna.it/pubblicazioni/rapporti-documenti/protocollo-snich-aggiornamento-dic2011.

Scardoni, A., Balzarini, F., Signorelli, C., Cabitza, F., & Odone, A. (2020). Artificial intelligence-based tools to control healthcare associated infections: A systematic review of the literature. *Journal of Infection and Public Health, 13*(8), 1061–1077. https://doi.org/10.1016/j.jiph.2020.06.006

Seidelman, J. L., Mantyh, C. R., & Anderson, D. J. (2023). Surgical site infection prevention: A review. *JAMA, 329*(3), 244. https://doi.org/10.1001/jama.2022.24075

Soguero-Ruiz C, Augestad M, Lindsetmo RO, Skrøvseth SO. Data-driven temporal prediction of surgical site infection.

Surgical site infection: Prevention and treatment of surgical site infection. (2008). RCOG Press.

Thirukumaran, C. P., Zaman, A., Rubery, P. T., et al. (2019). Natural language processing for the identification of surgical site infections in orthopaedics. *The Journal of Bone and Joint Surgery, 101*(24), 2167–2174. https://doi.org/10.2106/JBJS.19.00661

Wade, C., & Glynn, K. (2020). *Hands-On Gradient Boosting with XGBoost and scikit-learn: Perform accessible machine learning and extreme gradient boosting with Python.* Packt Publishing Ltd.

Weller, G. B., Lovely, J., Larson, D. W., Earnshaw, B. A., & Huebner, M. (2018). Leveraging electronic health records for predictive modeling of post-surgical complications. *Statistical Methods in Medical Research, 27*(11), 3271–3285. https://doi.org/10.1177/0962280217696115

Wu, G., Khair, S., Yang, F., et al. (2022). Performance of machine learning algorithms for surgical site infection case detection and prediction: A systematic review and meta-analysis. *Annals of Medicine & Surgery, 84*. https://doi.org/10.1016/j.amsu.2022.104956