PLOS ONE

RESEARCH ARTICLE

Predicting the occurrence of surgical site infections using text mining and machine learning

Daniel A. da Silva[1], Carla S. ten Caten[1], Rodrigo P. dos Santos[2], Flavio S. Fogliatto[1]*, Juliana Hsuan[3]

1 Industrial Engineering Department, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil, **2** Hospital de Clinicas de Porto Alegre, Porto Alegre, Brazil, **3** Copenhagen Business School, Copenhagen, Denmark

* ffogliatto@producao.ufrgs.br

Citation: da Silva DA, ten Caten CS, dos Santos RP, Fogliatto FS, Hsuan J (2019) Predicting the occurrence of surgical site infections using text mining and machine learning. PLoS ONE 14(12): e0226272. https://doi.org/10.1371/journal.pone.0226272

Editor: Patryk Orzechowski, University of Pennsylvania, UNITED STATES

Received: April 29, 2019

Accepted: November 22, 2019

Published: December 13, 2019

Data Availability Statement: Datasets and computacional codes are available as Supporting Information files (Supplement A). Datasets were coded to preserve patients' privacy and adhere to the hospital's ethics committee requirements. For example, an original description such as: original_text = ['the patient has a fever and also has a cough'] Was coded to the following: encoded_text = [0 1 2 3 4 5 6 2 3 7] Using the coded datasets reviewers should be able to reproduce results reported in the paper within the error margin.

Abstract

In this study we propose the use of text mining and machine learning methods to predict and detect Surgical Site Infections (SSIs) using textual descriptions of surgeries and post-operative patients' records, mined from the database of a high complexity University hospital. SSIs are among the most common adverse events experienced by hospitalized patients; preventing such events is fundamental to ensure patients' safety. Knowledge on SSI occurrence rates may also be useful in preventing future episodes. We analyzed 15,479 surgery descriptions and post-operative records testing different preprocessing strategies and the following machine learning algorithms: Linear SVC, Logistic Regression, Multinomial Naive Bayes, Nearest Centroid, Random Forest, Stochastic Gradient Descent, and Support Vector Classification (SVC). For prediction purposes, the best result was obtained using the Stochastic Gradient Descent method (79.7% ROC-AUC); for detection, Logistic Regression yielded the best performance (80.6% ROC-AUC).

1. Introduction

Surgical Site Infections (SSIs) are one of the predominant types of infection in Brazilian hospitals [1]. About one in thirty "clean" surgeries will suffer from complications due to SSIs. The rate is significantly higher if we consider "dirty" (i.e. contaminated), emergency, and prolonged surgeries, or procedures performed on patients with clinical comorbidities [2]. SSIs are also among the most frequent Adverse Events (AEs) reported on hospitalized patients, causing a substantial increase in mortality, re-hospitalization rates, and care costs [2,3].

Traditional methods for the prevention and detection of infections typically use resources (mostly human) in an intensive and time-consuming way. Computerized techniques, mainly based on Artificial Intelligence, may provide expedite and cost-efficient alternatives to the analysis of infections [4–6]. For that, it is necessary to verify the applicability of those techniques in the detection of AEs and control of hospital infections, particularly in large scale, data-rich environments such as the Brazilian healthcare system [7].

Funding: The authors received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Health surveillance has been described as an essential part of infection prevention and control programs due to its ability to promote a decrease in infection rates [8–10]. In healthcare institutions, patient information is stored mainly in the form of narrative texts and clinical reports [11]. Passive search for infections is usually carried out analyzing spontaneous reports made by health professionals (i.e. healthcare providers report clinical signs and a possible infection diagnosis in the patient's medical record, but not necessarily make the statement of a hospital infection to the surveillance entities). However, most mild and self-limiting infections are likely to remain unreported. On the other hand, surveillance and the active search for infections usually do not provide real-time information, since data collection, analysis and feedback traditionally rely on time and resource consuming methods [12]. Data Mining (DM) and Machine Learning (ML) techniques provide an alternative for that.

The use of DM to support health surveillance has been reported in the literature [13–15]. When applied to unstructured textual data, DM is referred to as Text Mining (TM); DM and TM share the same process and goal of identifying non-trivial patterns in data that are both meaningful and useful to users [16]. Both groups of techniques often use ML algorithms [17,18], enabling the prediction and classification of new records based on knowledge gathered from existing records.

There is some evidence in the literature reporting the successful use of TM and ML methods in the analysis of events that cause harm to patients; see [4–6]. Machine learning has been shown to be an effective tool for predicting infections [19,20]. In the same way, significant advances were also reported on the subject of adverse events' extraction and detection using free text to improve patients' diagnosis [21,22]. However, there is a gap in the literature regarding the joint application of TM and ML to predict SSI mining textual records of surgical descriptions, which we aim to bridge with our study.

2. Materials and methods

TM methods were used to process surgeries' and post-operative patients' records of a Brazilian hospital aiming to set the best practices for predicting and detecting SSIs using ML algorithms. An optimization of hyperparameters has also been performed for each algorithm. We analyzed a dataset comprised of textual descriptions of surgeries and post-operatives patients' records up to 30 days after the procedure.

The dataset was obtained from Hospital de Clínicas de Porto Alegre (HCPA), an 842-bed, tertiary care teaching hospital located in the city of Porto Alegre, Brazil. The hospital is deemed best in the country in its category, providing average and high complexity care through the Brazilian Unified Health System (SUS). There are 14 surgical specialties considered in the analysis.

The 30-day observation period established in this study to monitor the occurrence of post-surgical infections is grounded on empirical evidence as shown in Table 1, which was obtained mining a database of surgeries performed by each specialty in the past 5 years.

In Table 1, n_i denotes the number of records found in the 5-year period, and t_i denotes the average number of postoperative days of hospital stay demanded by specialty i. Statistic t_i is a weighted average, which considered the average number of postoperative hours of hospital stay demanded by each type of procedure within specialty i and their frequency of occurrence in the database. Most specialties require post-operative stays shorter than 30 days (the overall mean is 20.5 days), justifying the observation period established here.

Data were used to train and compare classification algorithms and text preprocessing techniques. The study was conducted in four stages (Fig 1), which were adapted from [23].

Data were managed using PostgreSQL 9.6 [24]. Text preprocessing was carried out in Python 3.5 [25], which was also used to run TM and ML methods, and evaluate their

Table 1. List of surgical specialties and associated post-operative length-of-stay.

Specialty, i	n_i	t_i
1. Pediatrics	2,963	58.5
2. Colorectal	2,516	32.8
3. Neurosurgery	2,392	27.1
4. Digestive System	9,930	26.6
5. Urology	11,136	22.0
6. Vascular	3,435	21.9
7. Plastic	1,663	21.8
8. Thoracic	3,213	20.6
9. General	9,294	15.7
10. Orthopedics and Traumatology	6,468	13.0
11. Gynecology and Obstetrics	5,398	8.8
12. Otorhino	5,652	8.5
13. Oral and Maxillofacial	278	7.5
14. Mastology	2,034	3.0

https://doi.org/10.1371/journal.pone.0226272.t001

performance. Python provides libraries to support processing of records, information retrieval, application and validation of methods; we used the NLTK [26] and scikit-learn [27] libraries.

In the *data retrieval* stage, we retrieved textual information on surgeries and patients' post-operative records from HCPA's unified database, which was then combined with inputs from the hospital's Internal Committee for Infection Control (ICIC). The committee retrospectively reviews records identifying the ones that led to infections, following an active search strategy. Records reviewed are manually selected among those more likely to display a patient infection outcome; i.e. patients to whom antibiotics were prescribed, cases in which positive wound cultures are reported or those associated to patients displaying signs and symptoms of fever, hyperemia or presence of secretion in their evolutions, older or obese patients, and those carrying chronic diseases such as diabetes. The committee reviewed the selected post-operative records and reached a conclusion, assigning one of two possible outcomes (patient infected or not infected).

There are two parts to the text records used here: (*i*) a technical description of the surgical procedure, and (*ii*) follow-ups on the evolution of patients during hospitalization and consultations, up to 30 days after surgery. Records analyzed were written in Brazilian Portuguese. The occurrence of infections is reported in both parts by those providing care to patients,

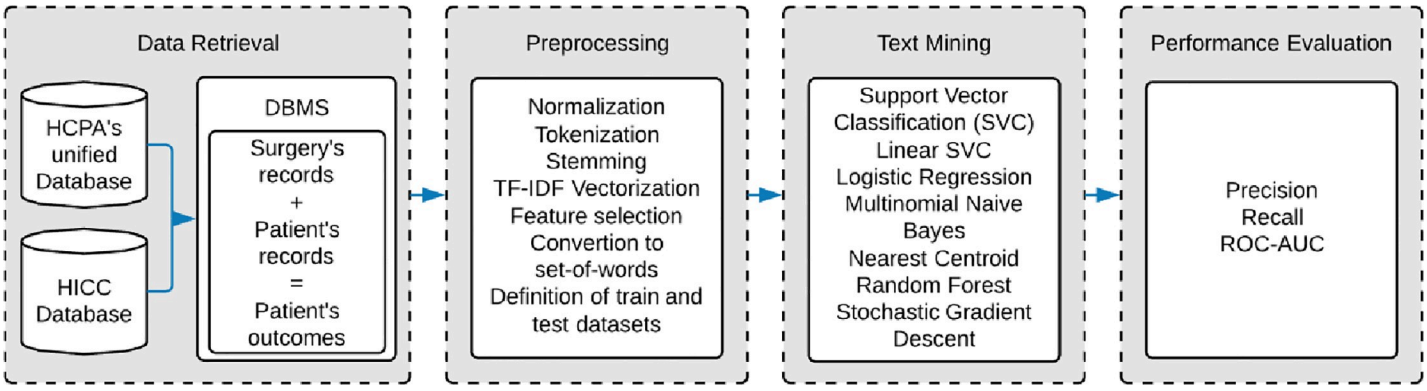

Fig 1. Overview of the proposed method.

https://doi.org/10.1371/journal.pone.0226272.g001

leading to a highly unbalanced dataset: 1.2% of the records report infections in part (*i*), while 1.6% report infections in part (*ii*). Records reporting infections on the 31st day after surgery or later were disregarded, as well as those of patients who had more than one surgery performed on the same day with different infection outcomes.

The response variable used in the text mining step to obtain predictions of outcomes was the "gold standard" established by the ICIC, and not the caretakers' reports. Thus, we may have situations in which the surgery was reported infected by caretakers, but no infection was reported by them during the post-operative period, although the case was considered infected since that was the conclusion issued by the ICIC. Note that the ICIC issued one conclusion per case; thus, whenever the ICIC concluded that a case was infected, the conclusion was valid for both surgery and post-operative descriptions.

The second stage is the *Text Mining*. There are six steps in this stage: normalization, tokenization and stemming, vectorization, feature selection, conversion to set-of-words, and definition of the train and test sets. Textual description of cases was obtained in the previous stage in *comma-separated values* (.csv) format and inserted into the PostgreSQL database. The dataset was structured with three fields of information: (*i*) the outcome of a binary variable representing the final status of the patient (1 = infected or 0 = not infected) obtained from the ICIC; (*ii*) free-form text entered directly by healthcare providers describing the surgery; and (*iii*) free-form text describing the post-operative record, also entered by caretakers. Data from the second and third fields were treated in the pre-processing module. Numerical entries were excluded from fields (*ii*) and (*iii*).

In the normalization step, stop words and punctuation were removed and the text was rewritten with no capital letters. In the tokenization step, continuous text was reduced to tokens, which are linguistic units such as words and sentences [28]. Morphological normalization was also carried out with words reduced to root form such that gender and grade information was excluded. After this step, each record was comprised of a set of tokens, delimited by blank spaces. Once tokens are identified, and prior to the feature selection step, features must be defined. In general, an *n*-gram is a sequence of *n* tokens [29]. In this work, we used unigrams, bigrams, and trigrams as features.

In the feature selection step, features were ranked according to two indices. The first is based on the χ^2 test, commonly used to verify the independence between a pair of events; in the context of feature selection, we test the occurrence of features in classes and their dependence using Eq (1).

$$\chi^2(d,t,c) = \sum_{e_t\in\{0,1\}}\sum_{e_c\in\{0,1\}} \frac{(N_{e_te_c} - E_{e_t,e_c})^2}{E_{e_t,e_c}} \quad (1)$$

where N and E are the observed and expected occurrence frequencies in document d, e_t is a binary variable indicating if feature t occurs in d, and e_c is a binary variable indicating if d is in class c. For the independence hypothesis to hold, $N_{e_te_c}$ and E_{e_t,e_c} should converge to 0.5; when that is not the case and χ^2 values are large, feature t should be selected [30].

The second feature selection index is based on F-value, which is calculated as follows [31]:

$$F\ value(t) = \frac{(X_t^{inf} - X_t')^2 + (X_t^{n\_inf} - X_t)^2}{\frac{1}{n_{inf}-1}\sum_{k=1}^{n_{inf}}(X_{k,t}^{inf} - X_t^{inf})^2 + \frac{1}{n_{n\_inf}-1}\sum_{k=1}^{n_{n\_inf}}(X_{k,t}^{n\_inf} - X_t^{n\_inf})^2} \quad (2)$$

where X_t, X_t^{inf} and $X_t^{n\_inf}$ is the average of the t-th feature in the complete, infected, and non-infected datasets, respectively; $X_{k,t}^{inf}$ is the t-th feature of the k-th infected instance, and $X_{k,t}^{n\_inf}$ is the t-th feature of k-th non-infected instance; n_{inf} is the number of infected instances, and

n_{n_inf} is the number of non-infected instances. Eq (2) gives a measure of discrimination between the two sets (infected and non-infected); whenever the F-value of a feature is greater than a threshold value the feature is inserted into the selected feature space; otherwise, it is removed from that space. High F-values indicate discriminating features [31].

Features retained were those above a given percentile of largest values; percentile and index chosen varied according to the Text Mining method tested. Indices were chosen based on their good performance in previous studies [5,23], and adequacy to the case under analysis.

In the conversion to set-of-words step, the occurrence of features (selected in the previous step) in records is evaluated regarding a given indicator, and results are organized in a matrix. Text records are listed in matrix rows and selected features in matrix columns. Two indicators were tested here: term frequency (TF), and term frequency-inverse document frequency (TF-IDF). TF gives the frequency in which a feature appears in a record. TF-IDF reflects the importance of a feature in a record from a collection of records, increasing proportionally to the feature's frequency in a record, but being compensated by the feature's frequency in the collection of records, as given next [32]:

$$TFIDF(t,d,D) = \frac{f(t,d)}{\max\{f(t,d) : t \in d\}} \times \log\frac{|D|}{|\{d \in D : t \in d\}|} \quad (3)$$

where t denotes the feature, d denotes the record, D is the total number of records considered, and $f(t,d)$ is the number of occurrences of feature t in record d.

TFIDF values for each matrix term were normalized using the unitary Euclidian norm [32], as follows:

$$\mathbf{v}_{norm} = \frac{\mathbf{v}}{||\mathbf{v}||_p} = \frac{\mathbf{v}}{\sqrt{v_1^p + v_2^p + \cdots + v_n^p}} \quad (4)$$

where $\mathbf{v}_{norm}$ is the normalized vector, $\mathbf{v}$ is the vector to be normalized and $||\mathbf{v}||^p$ is the norm used to promote normalization, such that $p = 1$ for $\ell_1 norm$ and $p = 2$ for $\ell_2 norm$ [33]. Large values of *TFIDF* are obtained whenever a term displays high frequency in a document and low frequency in the complete set of documents.

The last step of the *pre-processing* stage is the definition of training and testing sets, in which the set-of-words is divided to allow k-fold cross-validation. We divided the dataset into k mutually exclusive subsets of equal size, and used one subset for testing and $k-1$ subsets for parameter estimation. The process was carried out $k = 10$ times alternating the test subset, and performance statistics were calculated from the results [34].

In the *Machine Learning* stage, we tested different supervised classification algorithms. In supervised learning the outcome of each analyzed record in known beforehand. Records are in the format ($\mathbf{x}$,y), where $\mathbf{x}$ is the vector of features defined in the preprocessing stage and y is a binary class identifier, with outcomes 0 for not infected (or clean), and 1 for infected.

We tested the performance of ML algorithms used in similar studies; they are: Support Vector Machines (SVM) [13,14,35–45], Logistic Regression [37,42], Naive Bayes [35,38,43,44,46], Boosted Trees [38], Random Forest [38], and Nearest Neighbors [38]. The dataset was split into training and testing sets. Due to the highly unbalanced dataset, we used a stratified ($k-1$) cross-validation strategy, preserving the percentage of samples in each class in each fold, with k set to 10 [47]; that means the learning process is executed 10 times in different training sets, and the average of 10 scores is used to obtain an overall accuracy estimate. The goal was to improve the algorithms' performance on the classification of records resulting in infection. Records were randomly picked in the training set such that each class is represented in the

same proportions observed in the complete sample of records (~98.6% clean and ~1.4% infected outcomes).

Each ML algorithm has parameters that are not directly learned by the classifier; e.g., C, kernel, and γ in SVM [48]. Parameters' definition was performed through a random search in a grid of parameters [49]. The search had the objective of finding the best combination of parameters to maximize the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) [50]. The grid search was also used to select the percentile and metric (χ^2 and F-score) for attribute selection, and norm for TF-IDF ($l1$ or $l2$).

The final stage in our proposed method is *performance evaluation*. For that, we analyzed precision, sensitivity, ROC-AUC, and confusion matrices. These indicators are suitable for classification problems with unbalanced datasets, being used several similar studies reported in the literature [13,14,43–46,51,35–42].

3. Results

The original database was comprised of 27,648 surgical descriptions and 15,714 post-operative records (the use of the dataset was approved by HCPA's Ethics Committee under project number CAAE 33705014.8.0000.5327). HCPA's Ethics Committee is coordinated by Drs. Temis Maria Felix and Marcia Mocellin Raymundo; the complete list of Committee members is available at https://www.hcpa.edu.br/downloads/pesquisa/ato_n_188-2019.pdf. After excluding empty records and those that did not fit the criteria of the study, the number of records was reduced to 15,479 surgical descriptions and 12,637 post-operative records, with 98.6% of the records negative and 1.4% positive on average, according to Table 2. Table 3 provides a descriptive view of the final dataset (datasets and codes used in this analysis are given in S1 Supplement). Some remarks are noteworthy. Records in the database cover an 8-month period starting in 12/2015. During that period: (*i*) 27,648 surgical descriptions were made; of those, the ICIC audited a sample and detected 247 infections; and (*ii*) 15,714 post-operative descriptions were made; of those, the ICIC audited a sample and detected 233 infections. We excluded records of patients who had more than one surgery in the same day and only one of them was infected, since there is a single post-operative record in such situation.

In results to follow, text mining classifiers were used in the two parts of the dataset (surgical descriptions and post-operative descriptions) separately. We refer to results in the first part (surgical descriptions) as *prediction*, and to results in the second part (post-operative descriptions) as *detection*. In both cases, TM pre-processing and ML algorithms were used to classify cases as clean or infected.

Table 4 presents the performance of each ML algorithm in predicting infections and the respective TM settings to achieve the results. ROCs for the prediction algorithms are shown in

Table 2. SSI database analyzed in this study.

	Prediction			Detection		
Description	**Infected surgeries**	**Clean surgeries**	**Total**	**Infected surgeries**	**Clean surgeries**	**Total**
Initial sample	247	27,401	27,648	233	15,481	15,714
Empty records	-29	-12,103	-12,132	-2	-3,037	-3,039
Records of patients that had more than one surgery, one of which was reported clean	-4	-7	-11	-3	-9	-12
Infections reported more than 30 days after surgery	-26	0	-26	-26	0	-26
Records used in the study (final sample)	188 (1.21%)	15,291 (98.79%)	15,479 (100%)	202 (1.6%)	12,435 (98.4%)	12,637 (100%)

https://doi.org/10.1371/journal.pone.0226272.t002

Table 3. Descriptive view of the dataset.

Characteristic	Value
Number of patients	12,483
Mean (and SD) of patients' age	48.31 (22.03)
Average number of surgeries per patient	1.24
Female patients	7,107
Mean (and SD) of female patients' age	47.13 (20.44)
Male patients	5,376
Mean (and SD) of male patients' age	49.88 (23.87)
Number of surgical procedures	18,062
Elective procedures	13,027
Urgent procedures	3,239
Emergency procedures	1,796
Average size (and SD) of surgical team	5.84 (2.81)

https://doi.org/10.1371/journal.pone.0226272.t003

Fig 2, Precision-Recall boxplots in Fig 3 and Precision-Recall curves for all tested methods in Fig 4. The best performance considering the relationship between true positives and false positives represented by the area under the ROC curve was obtained by the Stochastic Gradient Descent (SGD) method. The best result was achieved using 75% of the terms selected by the χ^2 test in a set of features without normalization, assigning a weight of 0.01 to the negative (clean) class. Pre-processing strategies for each method were determined from a grid of parameter options through random search. Chosen Feature Selection method, Percentile, Transformation, Normalization, and Class_Weight options were those yielding the best ROC-AUC.

SGD classification with the pre-processing settings displayed in Table 4 reached a mean of 79.7% ROC-AUC (SD = 3.3%), mean sensitivity of 89.3% (SD = 6.8%) for positive classes, and 65% (SD = 1.4%) for negative classes. Considering the unbalance between classes, mean precision values obtained were 3.5% (SD = 0.3%) for the positive class and 99.8% (SD = 0.1%) for the negative class.

Table 5 presents the performance of each ML algorithm in detecting infections and the corresponding TM settings. ROCs for the detection algorithms are shown in Fig 5, Precision-Recall boxplots in Fig 6 and Precision-Recall curves for all tested methods in Fig 7. Logistic regression was the method yielding the best results considering the relationship between true positives and false positives. This result was achieved using 40% of the terms selected from the χ^2 test, in a set of TF-IDF terms normalized using the norm $l1$, with a weight of 0.01 for the

Table 4. Algorithms' performance in predicting SSI.

						ROC-AUC	
Method	**T**	**P**	**FS**	**N**	**CW**	**Mean**	**SD**
Random Forest (RF)	TF-IDF	85%	F	$l1$	1/0.01	76.3%	3.3%
Logistic Regression (LR)	TF-IDF	55%	F	$l1$	1/0.005	75.9%	2.5%
Linear SVC (LSVC)	TF	85%	F	$l1$	1/0.005	79.0%	4.7%
SVC	TF	10%	F	$l2$	1/0.001	75.3%	4.6%
Nearest Centroid (NC)	TF	20%	χ^2	$l1$	-	78.2%	4.3%
SGD	TF	75%	χ^2	-	1/0.01	79.7%	3.3%
M-Naive Bayes (MNB)	TF	40%	χ^2	-	20/80	75.0%	4.6%

T: Transformation; P: Percentile; FS: Feature Selection; N: Normalization; CW: Class_Weight / Prior Probability; SD: Standard Deviation

https://doi.org/10.1371/journal.pone.0226272.t004

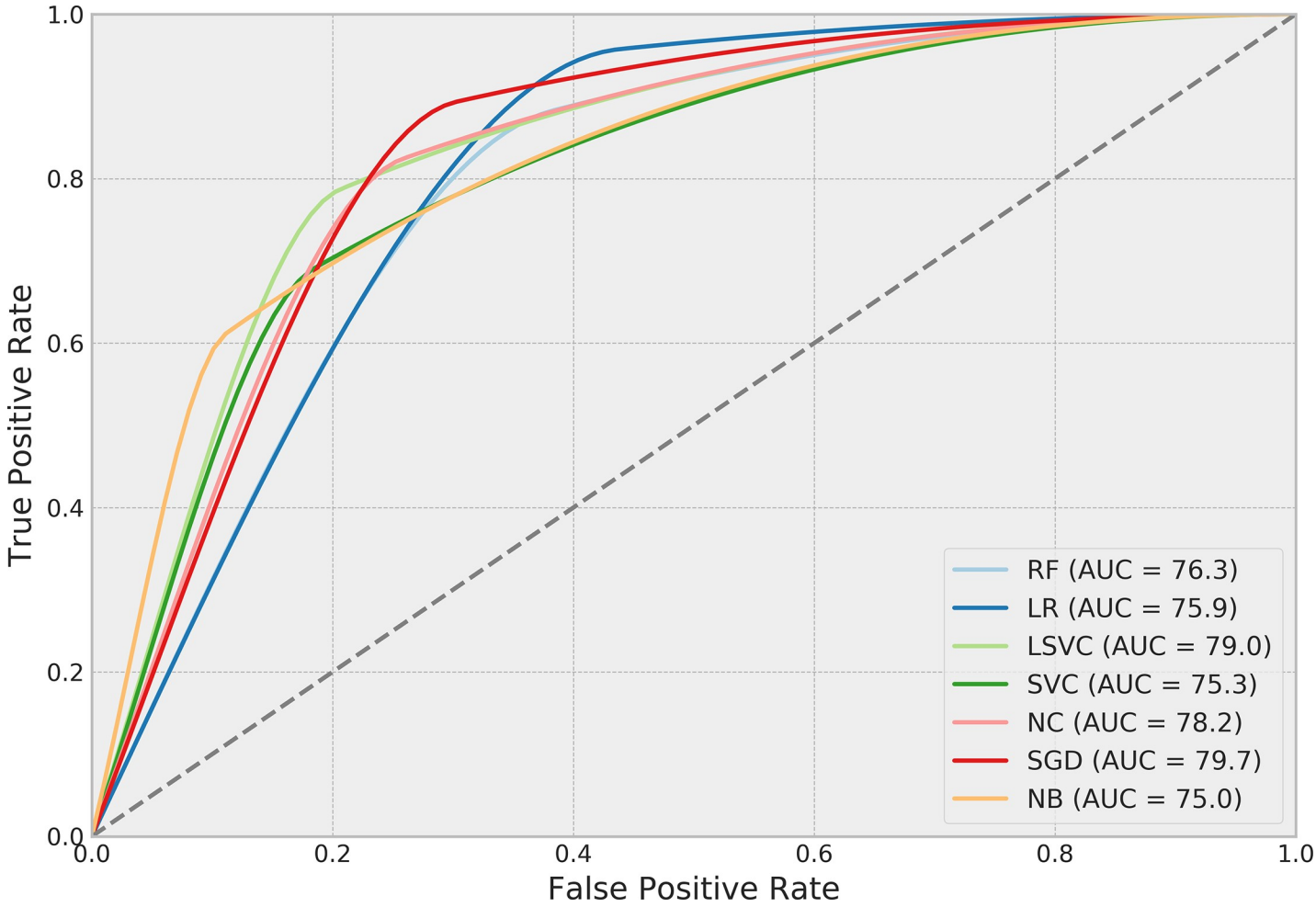

Fig 2. ROC-AUC performance of algorithms in predicting SSIs.

https://doi.org/10.1371/journal.pone.0226272.g002

negative class. The method yielded an ROC-AUC of 80.60% (SD = 2.4%), mean sensitivity of 75.7% (SD = 5.4%) for positive classes, and 85.5% (SD = 1.5%) for negative classes. Mean precision values were 7.9% (SD = 0.8%) for positive class, and 99.5% (SD = 0.1%) for negative class.

The objective of our study is to increase sensitivity in the search for infections, given that the manual procedure adopted by the surveillance team yields high precision and low sensitivity. The high sensitivity (and consequent low precision) searched in this study are depicted in the curves in Figs 4 and 7. Sensitivity was also prioritized here due to the fact that the gold-standard used in the analysis was based on sampling and selecting patients with higher potential risk; therefore, patients with lower risk that presented infection were less likely to have been investigated and accounted for in the gold-standard. Using text mining and machine learning to direct a more effective sampling by the surveillance team may lead to more infected patients being detected, yielding more reliable infection indicators and improving the gold-standard for future studies. Using the best algorithms for predicting and detecting SSIs may reduce the number of cases to be monitored in the post-operation period by more than 50% with less than 5% false negatives.

For predicting and detecting SSIs, the parameter CW (Class_Weight) was used to account for the unbalanced dataset, working similarly to a cost function with the objective of minimizing the bias between clean and infected classes. A CW = 1/0.01, for instance, implies in

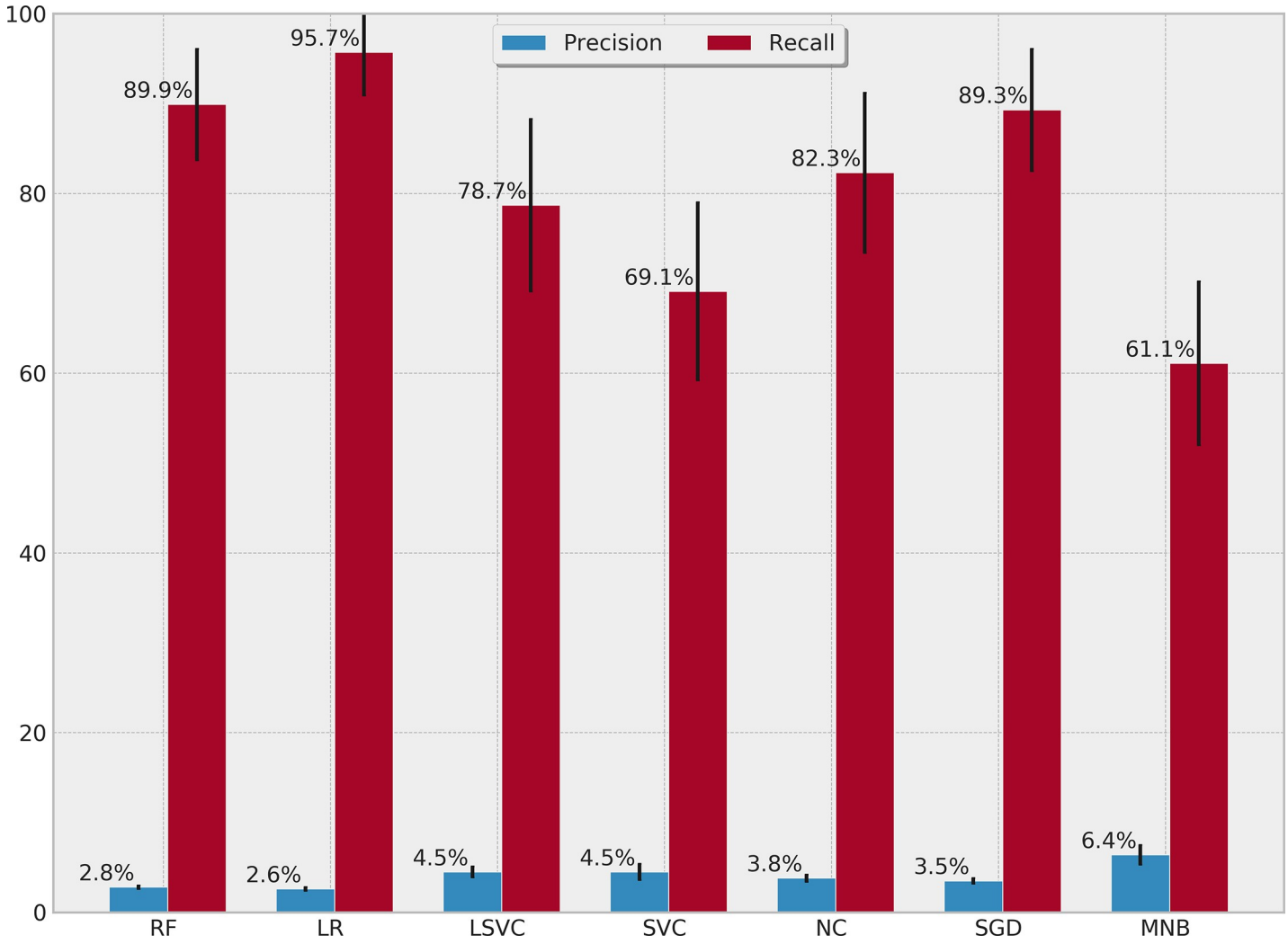

Fig 3. Precision-recall percentages and boxplots for surgical descriptions.

https://doi.org/10.1371/journal.pone.0226272.g003

assigning a weight of 1 to a positive (infected) classification and 0.01 to a negative (clean) classification.

Oversampling and undersampling approaches were also tested to account for imbalance in the dataset (results are presented as S2 Supplement). We were unable to avoid model overfitting when using oversampling, regardless of optimizing hyperparameters. That did not occur when undersampling, however none of the tested algorithms was able to outperform results obtained through the Class_Weight method.

A reduction in the number of surgeries to be monitored in the post-operative period represents a gain in terms of cost reduction and personal involvement in SSI surveillance. To attain such benefits, we should look for a compromise between the reduction in the number of events to be surveilled and the number of false negatives (infected surgeries classified as clean). That is attained analyzing the ROC-AUC mean values of each classification method in predicting infections, and the confusion matrix associated with the best method.

The best ROC-AUC in Table 4 is given by the SGD method (mean = 79.70%), with confusion matrix displayed in Table 6. Adopting the SGD for prediction would lead in a reduction

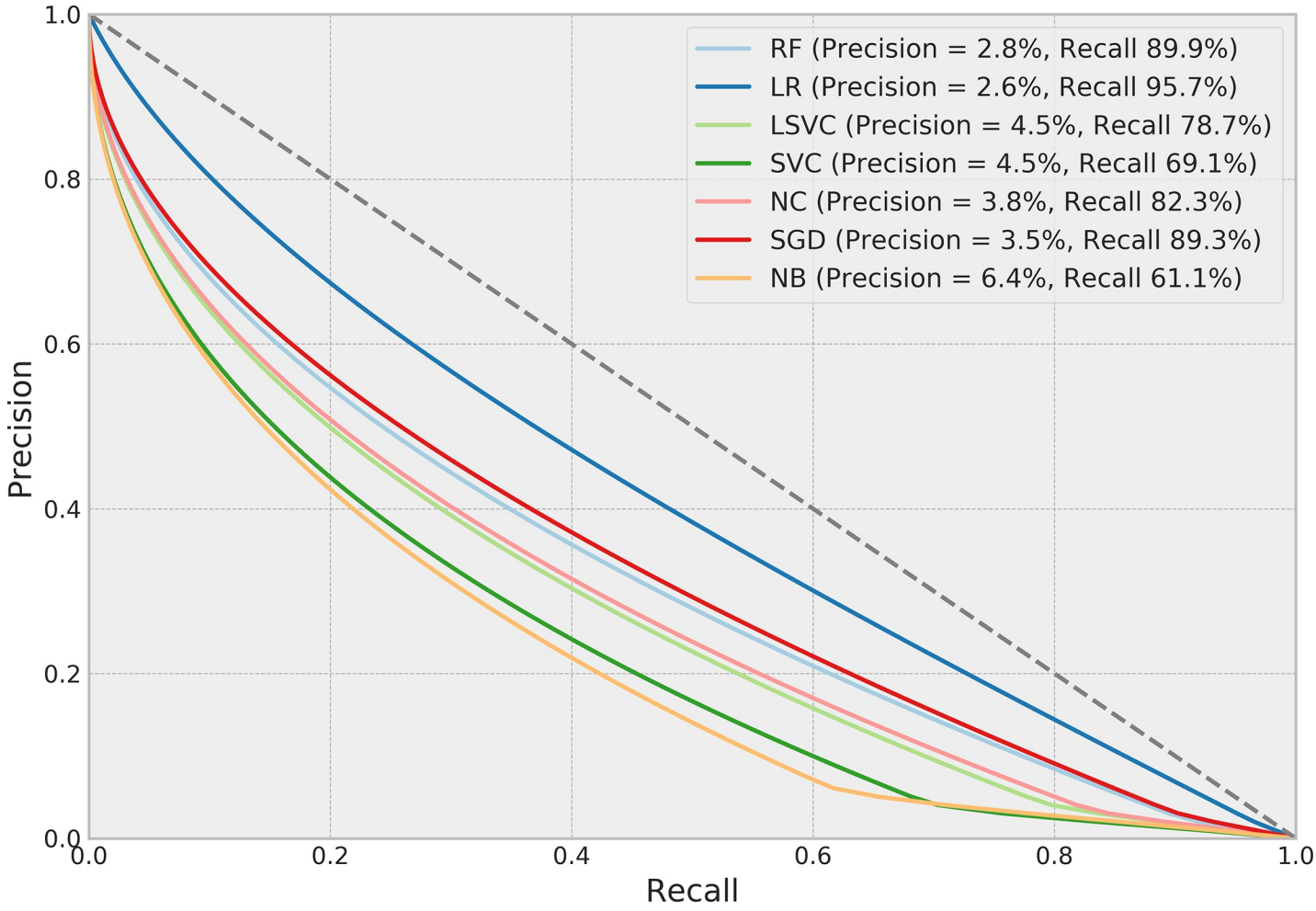

Fig 4. Precision-recall curves of methods tested for predicting SSIs.

https://doi.org/10.1371/journal.pone.0226272.g004

of 64.3% [= (n –TP–FP)/n] in the number of records to be analyzed by the ICIC, with (FP =) 20 surgeries reported as clean but actually infected not surveilled (approximately 10% of all infected surgeries). On the other hand, if the objective was to reduce the incidence of false negative classifications, the best method would be Logistic Regression: with a reduction of 55.59%

Table 5. Algorithms' performance in detecting SSI.

						ROC-AUC	
Method	**T**	**P**	**FS**	**N**	**CW**	**Mean**	**SD**
Random Forest (RF)	TF	20%	χ^2	$l2$	1/0.01	76.1%	3.4%
Logistic Regression (LR)	TF-IDF	40%	χ^2	$l1$	1/0.01	80.6%	2.4%
Linear SVC (LSVC)	TF-IDF	45%	χ^2	$l1$	1/0.01	78.1%	2.5%
SVC	TF	25%	F	$l2$	1/0.1	61.0%	6.3%
Nearest Centroid (NC)	TF-IDF	80%	χ^2	$l1$	-	76.4%	2.9%
SGD	TF	55%	F	-	1/0.05	63.6%	1.0%
M-Naive Bayes (MNB)	TF-IDF	80%	F	-	20/80	64.1%	6.5%

T: Transformation; P: Percentile; FS: Feature Selection; N: Normalization; CW: Class_Weight / Prior Probability; SD: Standard Deviation

https://doi.org/10.1371/journal.pone.0226272.t005

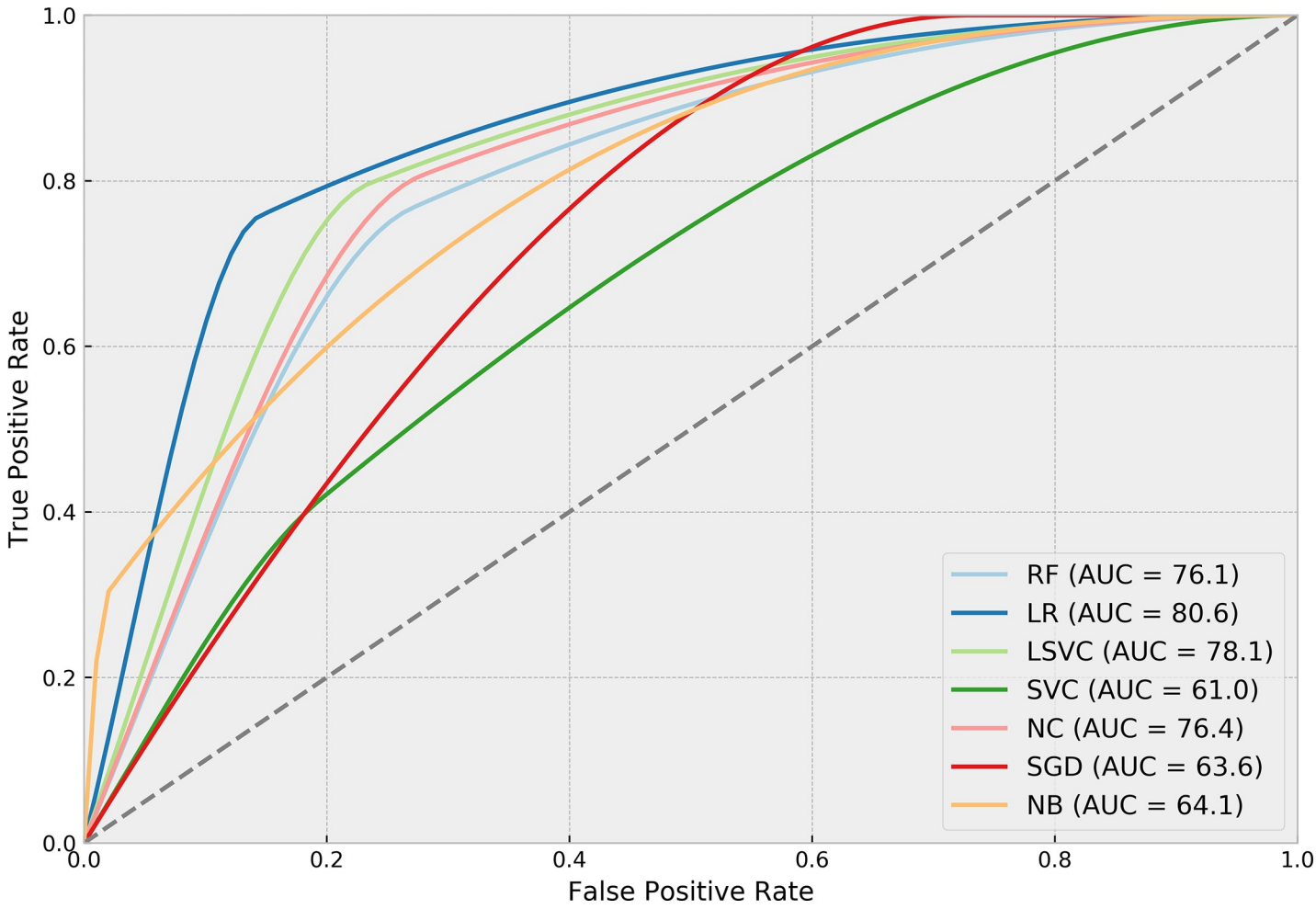

Fig 5. ROC-AUC performance of algorithms in detecting SSIs.

https://doi.org/10.1371/journal.pone.0226272.g005

in the number of records to be analyzed by the ICIC, it would imply in only 4.26% false negatives. These numbers illustrate the potential of using TM and ML methods to rationalize SSI surveillance activities.

4. Discussion

In this paper, we addressed patient safety surveillance through the use of text mining and machine learning methods using a database of surgical descriptions and post-operative follow-ups obtained from a high complexity University hospital. Our main goal was to establish the best TM and ML techniques for SSI prediction and detection using only textual data. For that, different methods of TM and ML were tested based on similar applications reported in the literature. Our results demonstrated that TM and ML are effective tools to support surveillance teams in the prediction and detection of SSIs, leading to improved patient care and safety.

Based on the TM and ML methods applied to our database of surgical descriptions it was possible to optimize surveillance efforts by reducing 55.59% of the volume of surgeries to be followed preventively, with only 4.26% of infections not detected using the Logistic Regression method. Using the SGD method, it was possible to reduce the volume of surgeries to be followed by 68.98%, although with a higher number of infections going undetected (10.64%).

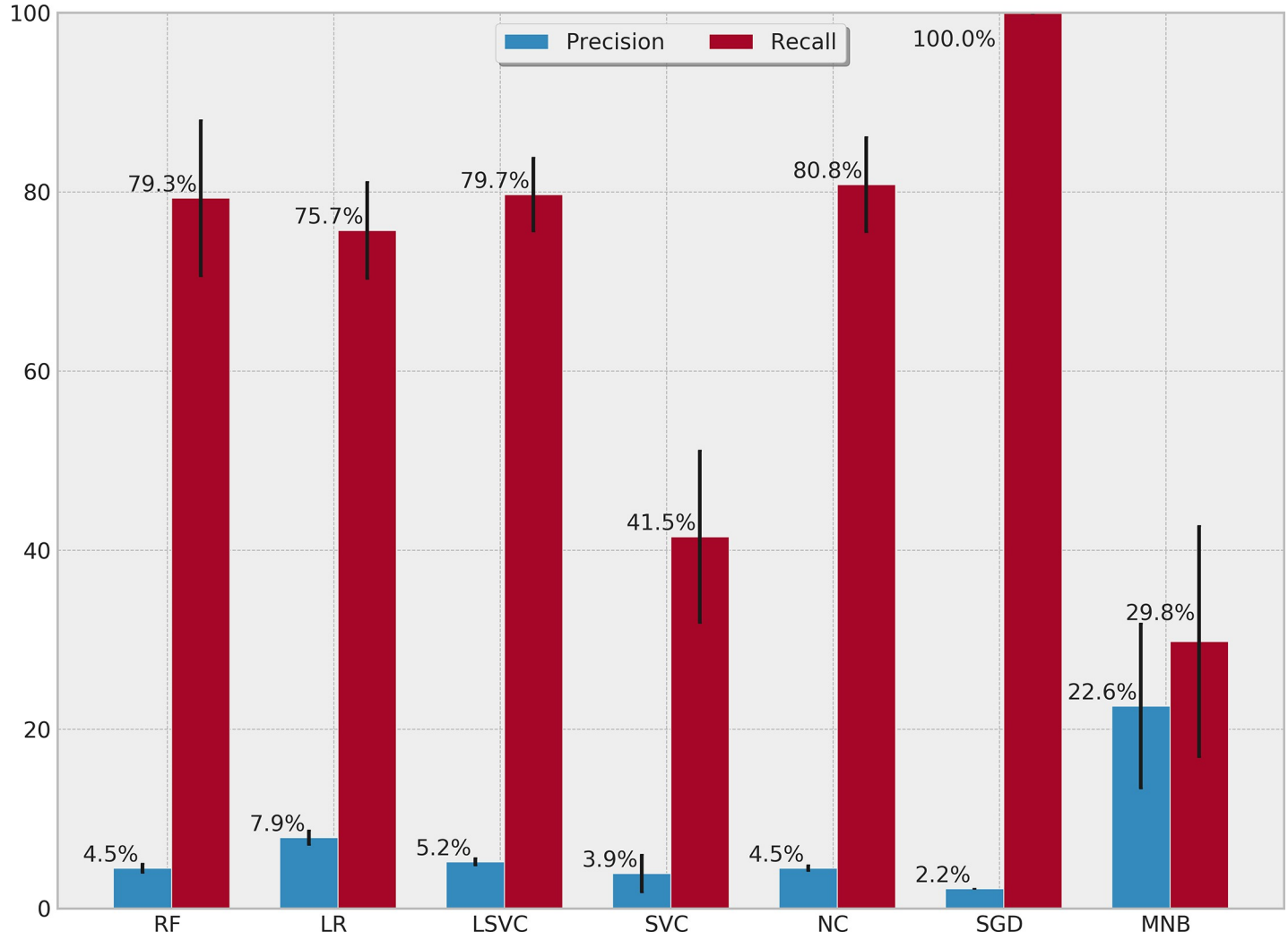

Fig 6. Precision-recall percentages and boxplots for post-operative notes.

https://doi.org/10.1371/journal.pone.0226272.g006

Independent of the ML method chosen, it is possible to optimize the time and resources invested in SSI surveillance, potentially increasing the number of SSIs that are currently undetected by the ICIC.

TM and ML methods rely on accurate classification of clean and infected surgeries by the surveillance team. Some factors may affect such classification, having a direct impact on the precision and sensitivity of TM methods. The identification of SSI requires interpretation of clinical and laboratory findings [52]. However, some surgical patients experience a short period (or no period) of hospitalization after surgery. The identification of patients developing infections after early discharge has been one of the challenges faced by infection surveillance methods [12].

The high unbalance in the dataset analyzed also imposes a challenge to the performance of TM and ML methods. The large number of negative SSI cases contributes to the increase in the false positive rate. We tried to compensate that using TM techniques combined with feature selection, TF-IDF transformation, and analysis of bigrams and trigrams. Another factor that contributes to increasing the number of false positives concerns the medical narrative described in the post-operative record, in which the patient is alerted to the risk of SSI. When

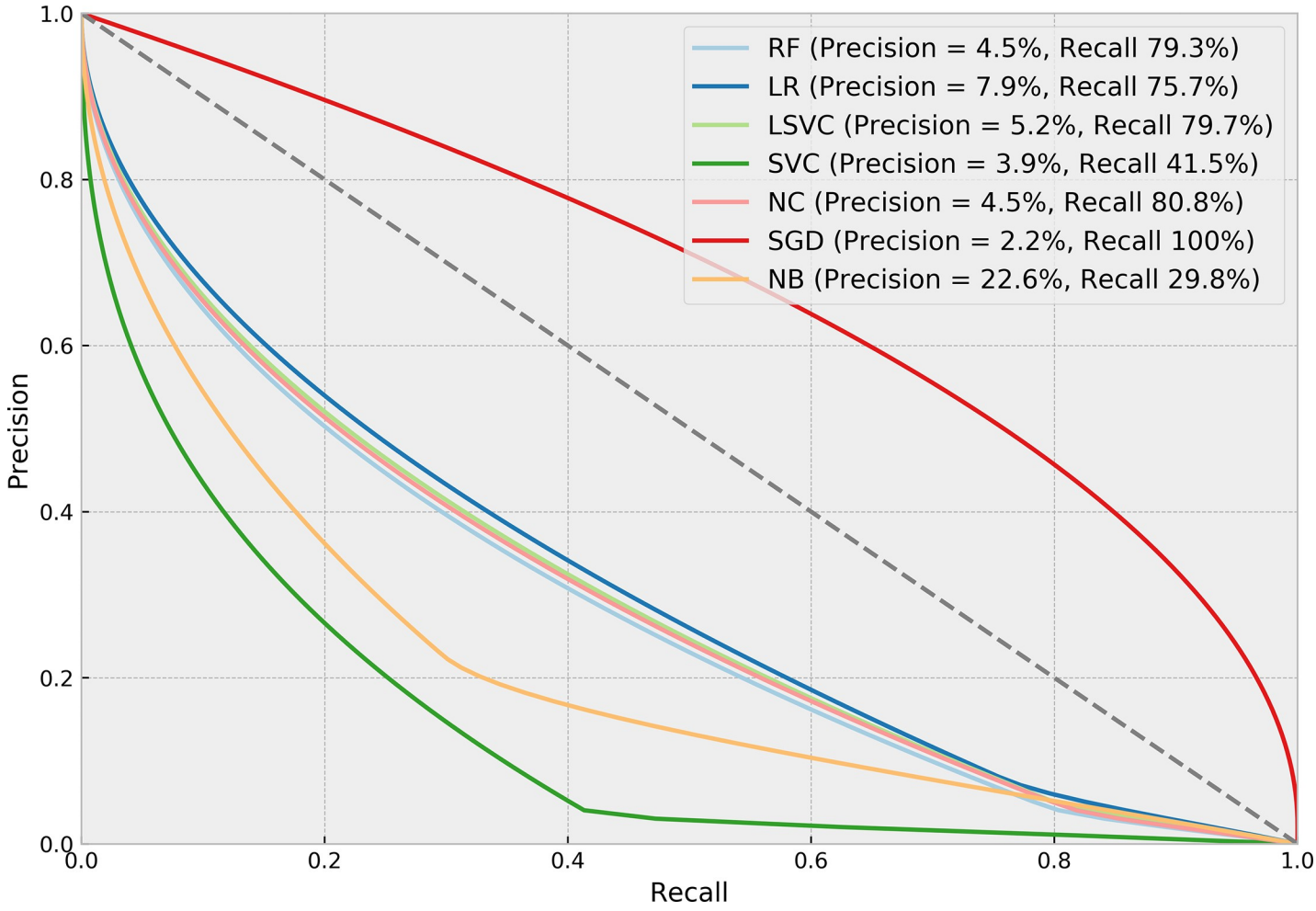

Fig 7. Precision-recall curves of methods tested for detecting SSIs.

https://doi.org/10.1371/journal.pone.0226272.g007

constructing the bag of words, such alert may confuse classifiers since records of clean surgeries have terms that are usually related to the occurrence of infection. Alerting physicians about the impacts of misleading narratives on TM performance may help overcome the problem.

TM and ML methods have the potential to play an important role in adverse events' surveillance, as pointed out in the literature [2], in the context of machine learning. Specifically regarding SSI, we demonstrated that TM and ML may be applied on reports created shortly after surgery to predict the occurrence of infections and on post-operative narrative records, to detect infections (and therefore develop preventive measures for future patients). The knowledge on SSI rates may be used as part of a feedback mechanism to decrease the future incidence of such infections [53], as well as in the training of ML algorithms in TM.

Table 6. SGD method for prediction–confusion matrix.

n = 15,479	Predicted clean	Predicted infected	
Actual clean	True negatives (TN): 9,930	False positives (FP): 5,352	15,291
Actual Infected	False negatives (FN): 20	True positives (TP): 168	188
	9,959	5,520	

https://doi.org/10.1371/journal.pone.0226272.t006

As future study we plan to explore the use of TM and ML to follow the post-operative records of specific groups of patients, selected by medical condition or age group, for example. We also view the use of additional information, such as examinations and prescriptions of medications available in the computerized system of the hospital, as potentially beneficial to improve the performance of TM and ML methods in the detection of SSIs. Finally, the literature dealing with the study of unbalanced datasets in text mining is constantly evolving. In our study, we followed the bootstrap strategy proposed by [54] to handle sample imbalance and tested binary classifiers suitable for unbalanced datasets; however, the study of alternative sample pre-treatment and classifiers is also a promising research direction.

Supporting information

S1 Supplement. Data and codes used in the analysis.
(ZIP)

S2 Supplement. Oversampling and undersampling approaches tested to account for imbalance in the dataset.
(DOCX)

Author Contributions

Conceptualization: Daniel A. da Silva.

Formal analysis: Daniel A. da Silva, Carla S. ten Caten, Rodrigo P. dos Santos, Flavio S. Fogliatto.

Methodology: Daniel A. da Silva, Carla S. ten Caten, Flavio S. Fogliatto, Juliana Hsuan.

Resources: Rodrigo P. dos Santos.

Software: Daniel A. da Silva.

Supervision: Rodrigo P. dos Santos.

Validation: Juliana Hsuan.

Writing – original draft: Daniel A. da Silva, Flavio S. Fogliatto.

Writing – review & editing: Flavio S. Fogliatto, Juliana Hsuan.

References

1. Anvisa/Brasil. Infection Diagnostic Criteria Related to Healthcare. In: Agência Nacional de Vigilância Sanitária [Internet]. 2017. p. 13–88. Available from: www.anvisa.gov.br
2. Wachter RM. Understanding Patient Safety. In: AMGH Editora. 2013. 479 p.
3. Stone PW, Kunches L, Hirschhorn L. Cost of hospital-associated infections in Massachusetts. Am J Infect Control. 2009; 37(3):210–214. https://doi.org/10.1016/j.ajic.2008.07.011 PMID: 19111366
4. Bouzbid S, Gicquel Q, Gerbier S, Chomarat M, Pradat E, Fabry J, et al. Automated detection of nosocomial infections: Evaluation of different strategies in an intensive care unit 2000–2006. J Hosp Infect [Internet]. 2011; 79(1):38–43. Available from: http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=emed10&NEWS=N&AN=2011437473 https://doi.org/10.1016/j.jhin.2011.05.006 PMID: 21742413
5. Michelson JD, Pariseau JS, Paganelli WC. Assessing surgical site infection risk factors using electronic medical records and text mining. Am J Infect Control [Internet]. 2014; 42(3):333–336. Available from: http://www.scopus.com/inward/record.url?eid=2-s2.0-84896735530&partnerID=40&md5=dad6f7d7e12724529376cb7afb4d9edc https://doi.org/10.1016/j.ajic.2013.09.007 PMID: 24406258

6. Campillo-Gimenez B, Garcelon N, Jarno P, Chapplain JM, Cuggia M. Full-text automated detection of surgical site infections secondary to neurosurgery in Rennes, France. In: Studies in Health Technology and Informatics. 2013. p. 572–575. PMID: 23920620
7. Daltoé T, Breier A, dos Santos HB, Wagner MB, Kuchenbecker R de S. Hospital Infection Control Services: Characteristics, Dimensioning and Related Activities. Rev Soc Bras Clin Med. 2014; 12(1):35–45.
8. Haley RW, Culver DH, White JW, Morgan WM, Emori TG, Munn VP, et al. The Efficacy of Infection Surveillance and Control Programs in Preventing Nosocomial Infections in Us Hospitals. Am J Epidemiol [Internet]. 1985; 121(2):182–205. Available from: http://aje.oxfordjournals.org/content/121/2/182 https://doi.org/10.1093/oxfordjournals.aje.a113990 PMID: 4014115
9. Brown KL, Ridout DA, Shaw M, Dodkins I, Smith LC, O'Callaghan MA, et al. Healthcare-associated infection in pediatric patients on extracorporeal life support: The role of multidisciplinary surveillance. Pediatr Crit Care Med [Internet]. 2006; 7(6):546–550. Available from: http://www.embase.com/search/results?subaction=viewrecord&from=export&id=L44715293%0Ahttp://dx.doi.org/10.1097/01.PCC.0000243748.74264.CE%0Ahttp://sfx.library.uu.nl/utrecht?sid=EMBASE&issn=15297535&id=doi:10.1097%2F01.PCC.0000243748.74264.CE&atitle=Health https://doi.org/10.1097/01.PCC.0000243748.74264.CE PMID: 17006389
10. Curran ET, Coia JE, Gilmour H, McNamee S, Hood J. Multi-centre research surveillance project to reduce infections/phlebitis associated with peripheral vascular catheters. J Hosp Infect. 2000; 46 (3):194–202. https://doi.org/10.1053/jhin.2000.0831 PMID: 11073728
11. Friedman C, Elhadad N. Natural Language Processing in Health Care and Biomedicine. In: Shortliffe EH, Cimino JJ, editors. Biomedical Informatics: Computer Applications in Health Care and Biomedicine [Internet]. London: Springer London; 2014. p. 255–284. Available from: https://doi.org/10.1007/978-1-4471-4474-8_8
12. Freeman R, Moore LSP, García Álvarez L, Charlett A, Holmes A. Advances in electronic surveillance for healthcare-associated infections in the 21st Century: A systematic review. J Hosp Infect. 2013; 84 (2):106–119. https://doi.org/10.1016/j.jhin.2012.11.031 PMID: 23648216
13. Aramaki E, Miura Y, Tonoike M, Ohkuma T, Masuichi H, Waki K, et al. Extraction of adverse drug effects from clinical records. Stud Health Technol Inform [Internet]. 2010; 160(Parte 1):739–743. Available from: http://ovidsp.ovid.com/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=emed9&AN=20841784%5Cnhttp://sfxhosted.exlibrisgroup.com/cmc?sid=OVID:embase&id=pmid:20841784&id=doi:10.3233/978-1-60750-588-4-739&issn=0926-9630&isbn=9781607505877&volume=160&issue=PART+
14. Bian J, Topaloglu U, Yu F. Towards large-scale twitter mining for drug-related adverse events. In: Proceedings of the 2012 international workshop on Smart health and wellbeing—SHB '12 [Internet]. 2012. p. 25. Available from: http://dl.acm.org/citation.cfm?doid=2389707.2389713
15. Silva A, Cortez P, Santos MF, Gomes L, Neves J. Rating organ failure via adverse events using data mining in the intensive care unit. Artif Intell Med. 2008; 43(0933–3657 Print):179–193.
16. Tan AH. Text Mining: The state of the art and the challenges. In: Proceedings of the PAKDD Workshop on Knowledge discovery from Advanced Databases. 1999. p. 71–76.
17. Han J, Kamber M. Data Mining: Concepts and Techniques. 2ª ed. Soft Computing. 2006. 800 p.
18. Zafarani R, Abbasi MA, Liu H. Social media mining: An introduction. Cambridge University Press; 2014. 320 p.
19. Taylor RA, Moore CL, Cheung KH, Brandt C. Predicting urinary tract infections in the emergency department with machine learning. PLoS One. 2018; 13(3).
20. Bartz-Kurycki MA, Green C, Anderson KT, Alder AC, Bucher BT, Cina RA, et al. Enhanced neonatal surgical site infection prediction model utilizing statistically and clinically significant variables in combination with a machine learning algorithm. Am J Surg. 2018; 216(4):764–777. https://doi.org/10.1016/j.amjsurg.2018.07.041 PMID: 30078669
21. Wang Z, Shah AD, Tate AR, Denaxas S, Shawe-Taylor J, Hemingway H. Extracting diagnoses and investigation results from unstructured text in electronic health records by semi-supervised machine learning. PLoS One. 2012; 7(1).
22. Zhang Y, Liu Z, Zhou W. Event recognition based on deep learning in Chinese texts. PLoS One. 2016; 11(8).
23. Lucini FR, Fogliatto FS, da Silveira GJC, Neyeloff JL, Anzanello MJ, Kuchenbecker RDS, et al. Text mining approach to predict hospital admissions using early medical records from the emergency department. Int J Med Inform [Internet]. 2017; 100:1–8. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85009230698&doi=10.1016%2Fj.ijmedinf.2017.01.001&partnerID=40&md5=19fe17844ba5835727064ff45f39f13b https://doi.org/10.1016/j.ijmedinf.2017.01.001 PMID: 28241931

24. PostgreSQL. PostgreSQL [Internet]. http://www.postgresql.org/. 2017. Available from: http://www.postgresql.org/
25. Python Software Foundation. Comparing Python to Other Languages. https://www.python.org/. 2017.
26. Bird S, Klein E, Loper E. Natural Language Processing with Python. O'Reilly Media, Inc. 2009; 43:479.
27. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res [Internet]. 2011; 12(112–113):2825–2830. Available from: http://chrhc.revues.org/2190%5Cnhttp://chrhc.revues.org/pdf/2190
28. Feldmann R, Sanger J. The text mining handbook: Advanced approaches in analyzing unstructured data. New York: Cambridge Press; 2006. 257–300 p.
29. Perkins J. Python 3 Text Processing With NLTK 3 Cookbook [Internet]. Packt Publishing. Birmingham, UK; 2014. 304 p. Available from: http://www.amazon.com/s/ref=nb_sb_noss?url=search-alias%3Daps&field-keywords=9781782167853
30. Sebastiani F. Machine learning in automated text categorization. ACM Comput Surv. 2002; 34(1):1–47.
31. Guyon I, Gunn S, Nikravesh M, Zadeh L. Feature Extraction, Foundations and Applications. Springer. New York; 2008. 778 p.
32. Sklearn. No Title. http://scikit-learn.org/stable/modules/feature_extraction.html. 2019.
33. Yi B-K, Faloutsos C. Fast time sequence indexing for arbitrary Lp norms. In: Proceedings of the 26st International Conference on VLDB. 2000. p. 385–394.
34. Witten, Frank, Hall. Data Mining: Practical Machine Learning Tools and Techniques (Google eBook). Complementary literature None. 2011. 664 p.
35. Ross M, Truong K, Lin K, Kumar A, Conway M. Text Categorization of Heart, Lung, and Blood Studies in the Database of Genotypes and Phenotypes (dbGaP) Utilizing n-grams and Metadata Features. Biomed Inform Insights [Internet]. 2013; 6:35. Available from: http://www.la-press.com/text-categorization-of-heart-lung-and-blood-studies-in-the-database-of-article-a3785 https://doi.org/10.4137/BII.S11987 PMID: 23926434
36. Yang M, Kiang M, Shang W. Filtering big data from social media—Building an early warning system for adverse drug reactions. J Biomed Inform. 2015; 54:230–240. https://doi.org/10.1016/j.jbi.2015.01.011 PMID: 25688695
37. McCart JA, Berndt DJ, Jarman J, Finch DK, Luther SL. Finding falls in ambulatory care clinical documents using statistical text mining. J Am Med Informatics Assoc. 2013; 20(5):906–914.
38. Botsis T, Nguyen MD, Woo EJ, Markatou M, Ball R. Text mining for the Vaccine Adverse Event Reporting System: medical text classification using informative feature selection. J Am Med Inform Assoc. 2011; 18(5):631–638. https://doi.org/10.1136/amiajnl-2010-000022 PMID: 21709163
39. de Bruijn B, Cherry C, Kiritchenko S, Martin J, Zhu X. Machine-learned solutions for three stages of clinical information extraction: The state of the art at i2b2 2010. J Am Med Informatics Assoc. 2011; 18(5):557–562.
40. Chee BW, Berlin R, Schatz B. Predicting adverse drug events from personal health messages. AMIA Annu Symp Proc [Internet]. 2011; 2011:217–226. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3243174&tool=pmcentrez&rendertype=abstract PMID: 22195073
41. Hur J, Özgür A, Xiang Z, He Y. Identification of fever and vaccine-associated gene interaction networks using ontology-based literature mining. J Biomed Semantics. 2012; 3(1).
42. Genkin A, Lewis DD, Madigan D. Large-scale bayesian logistic regression for text categorization. Technometrics. 2007; 49(3):291–304.
43. Ong MS, Magrabi F, Coiera E. Automated identification of extreme-risk events in clinical incident reports. J Am Med Informatics Assoc. 2012; 19(E1).
44. Ramesh BP, Belknap SM, Li Z, Frid N, West DP, Yu H. Automatically recognizing medication and adverse event information from food and drug administration's adverse event reporting system narratives. J Med Internet Res. 2014; 16(6).
45. Rochefort CM, Verma AD, Eguale T, Lee TC, Buckeridge DL. A novel method of adverse event detection can accurately identify venous thromboembolisms (VTEs) from narrative electronic health record data. J Am Med Informatics Assoc. 2015; 22(1):155–165.
46. Gurulingappa H, Rajput AM, Roberts A, Fluck J, Hofmann-Apitius M, Toldo L. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. J Biomed Inform. 2012; 45(5):885–892. https://doi.org/10.1016/j.jbi.2012.04.008 PMID: 22554702
47. Sklearn. No Title. http://scikit-learn.org/stable/modules/cross_validation.html. 2019.
48. Sklearn. No Title [Internet]. http://scikit-learn.org/stable/modules/grid_search.html. 2019 [cited 2017 Feb 4]. Available from: http://scikit-learn.org/stable/modules/grid_search.html

49. Bergstra J, Bengio Y. Random Search for Hyper-Parameter Optimization. J Mach Learn Res. 2012; 13:281–305.

50. Dong Y, Li X, Li J, Zhao H. Analysis on weighted AUC for imbalanced data learning through isometrics. J Comput Inf Syst [Internet]. 2012; 1(January):371–8. Available from: http://www.jofcis.com/publishedpapers/2012_8_1_371_378.pdf

51. Freifeld CC, Brownstein JS, Menone CM, Bao W, Filice R, Kass-Hout T, et al. Digital drug safety surveillance: Monitoring pharmaceutical products in Twitter. Drug Saf. 2014; 37(5):343–350. https://doi.org/10.1007/s40264-014-0155-x PMID: 24777653

52. Mangram AJ, Horan TC, Pearson ML, Silver LC, Jarvis WR. Guideline for Prevention of Surgical Site Infection, 1999. Hospital Infection Control Practices Advisory Committee. Infect Control Hosp Epidemiol [Internet]. 1999; 20(04):250–78; quiz 279–80. Available from: https://www.cambridge.org/core/product/identifier/S0195941700070223/type/journal_article

53. Michelson J. Improved detection of orthopaedic surgical site infections occurring in outpatients. Clin Orthop Relat Res. 2005;(433):218–224. https://doi.org/10.1097/01.blo.0000150666.06175.6b PMID: 15805961

54. Chandrasekhar CK, Srinivasan MR, Ramesh Babu B. Bootstrapping in text mining applications. Int J Science and Research 2016; 5(1): 337–344.