Predictive Models for Surgical Site Infection (SSI) in Patients with a Permanent Pacemaker (PPM) Using Machine learning Methods

Jiyoun Song[1], Elioth Sanabria-Buenaventura[2], Bevin Cohen[1], Jianfang Liu[1], David Yao[2], and Elaine Larson[1]

[1]Columbia University
[2]Columbia University Fu Foundation School of Engineering and Applied Science

June 11, 2020

Abstract

Introduction Given infections in patients with PPM are responsible for adverse outcomes such as an increased rate of mortality, one important reduction strategy of the incidence of SSIs is to identify and predict patients at high risk. Methods A retrospective cohort study was conducted in patients with PPM discharged from a large academic health center in New York City from 2006 through 2016. Risk factors identified through bivariate analysis were used to build predictive models. Five-fold cross-validation was applied to build models. The performance of the three machine learning models–logistic regression, decision tree (DT), and support vector machine (SVM)– for predicting surgical site infection (SSI) in patients with a permanent pacemaker (PPM) was compared. Results A total 205/9,274 (2.16%) patients with PPMs were diagnosed with a hospital-acquired SSI. Overall, the logistic regression algorithm had the highest prediction ability with the largest AUC at 72.9%. But the SVM model showed the highest sensitivity at 43.8% and positive predictive value at 32.5%. All three models showed excellent specificity and accuracy (over 98% and 96%, respectively). Conclusion Despite that this study showed the comparison of three predictive models, it has very limited clinical implications because of the low predictability of models (i.e., low PPV). Therefore, future researchers may improve the model by incorporating text data from clinical notes through natural language processing. Each algorithm had strengths and weaknesses in terms of accurate prediction, and interpretable clinical decision support. However, logistic regression was more accurate for predicting low-prevalence diseases such as SSI.

Introduction

In the United States approximately 14.4 million patients have cardiac arrhythmias, which are responsible for about 40,700 deaths annually.[1] Permanent pacemakers (PPMs) are increasingly common as the indications for device placement expand.[2] Each year, about one million patients globally receive cardiovascular implantable electronic devices including PPM[3] which, like any foreign body, increase the risk of infection. The frequency of cardiovascular implantable electronic device-related infections has increased dramatically due to the increasing number of cardiovascular devices being implanted in the last five decades.[4] Infections in patients with PPM are responsible for prolonged lengths of hospital stay and increased rate of readmissions, re-operation, and/or mortality.[5,6]

Surgical site infections (SSIs) are one of the most common hospital-acquired infections, occurring in approximately 2% to 5% of patients who undergo surgery, resulting in 157,000 to 300,000 cases in the United States annually.[7,8] They are associated with increased pain and discomfort for patients, longer lengths of stay and risk for hospital readmissions, increased mortality, and the potential of a negative psychological impact on the subjects.[9] In addition, the cost of treatment for these infections is approximately $10 billion per year.[10]

Because of their high cost and associated adverse outcomes, extensive efforts to reduce the incidence of SSIs and other types of infections are in place. One important reduction strategy is to identify patients at high risk so that enhanced prevention and control measures can be implemented early. Machine learning methods are used in healthcare to efficiently manage datasets that would otherwise be too large to handle with a traditional analytic method.[11,12] Thus, the aim of this study was to develop and compare the ability of the three machine learning predictive models–logistic regression, decision tree (DT), and support vector machine (SVM)–to identify risk factors for SSIs in patients with PPM.

Method

Sampling and setting

The sample for this study included patient admissions in which a PPM was implanted between 01/01/2007 and 12/31/2016 to one of three hospitals in metropolitan New York City– a 196-bed community hospital, a 738-bed adult tertiary/quaternary care hospital and an 862-bed adult and pediatric tertiary/quaternary care hospital– to which more than 100,000 patients are admitted annually. The implantation of a PPM procedure was identified using procedure date and International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM), Principal Procedure Code (**Appendix 1**).[13]

Description of dataset

The dataset was derived from a federally funded grant (Nursing Intensity of Patient Care Needs and Rates of Healthcare-Associated Infections [NIC-HAI], Agency for Healthcare Research and Quality, R01 HS024915) and extracted from various electronic databases (e.g., admission-discharge-transfer system, electronic health record, a clinical data warehouse, and departmental records). This study was approved by the institution's Institutional Review Board.

Potential risk factors for SSIs

The variables included in the data analyses and modeling were selected based on the published literature regarding known or predicted risk factors associated with SSIs.[14-18] Individual-level*host* factors included: (1) age and gender; (2) comorbidities – diabetes mellitus, obesity, hypertension, cancer, renal failure, chronic pulmonary disease, transplant, and postoperative hematoma; and (3) socioeconomic status as reflected by type of health insurance (i.e., Medicare, Medicaid, or commercial insurance). *Environmental*factors were: (1) invasive procedures such as central venous catheters; (2) admission-source from healthcare facility or non-healthcare facility/home; (3) hospital-related factors such as prior hospitalization within six months, length of stay (calculated from admission date to the onset SSIs for patients with SSIs, or from admission date to discharge date for the patients without SSIs), and intensive care unit (ICU) stay; (4) nurse staffing.

Nurse staffing was measured for 2 weeks prior to the onset of SSIs for patients with SSIs, or for 2 weeks after the surgical procedure for patients without SSIs. During that time frame, we used overall median nursing hours per patient day for each unit as the standard nurse staffing.[19] If the staffing hours were below 80% of the median during the time frame examined, it was regarded as understaffing. The total hours/patient day for registered nurses (RN) and total hours for nursing support staff (i.e., licensed practical nurse [LPN] and nursing assistant [NA]) were examined to determine whether the patient experienced understaffing (yes/no) and, if so, for how many days within the 2 week time frames. In case the patients moved around multiple units in the same day, it was regarded that they experienced understaffing if understaffing was existing at least once.

Initial statistical analysis

Descriptive statistics included means with standard deviations (SD) or medians with interquartile ranges (IQR) for categorical variables. All statistical analyses were performed using R Statistical Software (Foundation for Statistical Computing, Vienna, Austria). The relationship between potential predictor variables

and SSIs was initially tested using chi-square or student t-test. Then, the variables with p-values < 0.10 were included to build the predictive models. The workflow of machine learning algorithms in this study are shown in **Figure 1** .

Dataset preparation

The full dataset was randomly divided into two groups—80% for training and 20% for testing. To minimize bias and variance in the model-building process and to avoid overfitting, a 5-fold-cross-validation was performed. That is, the total training dataset was resampled into five folds of equal size. It was then repeatedly tested by rotating five times, with four training-folds and one validation-fold. The average of the model against each of the folds was obtained. Following this, the model was evaluated against the testing dataset (See 'Model evaluation' section below).

Applying machine learning

Method 1: logistic regression

The binary logistic regression for classification was used to predict the odds of having SSIs (i.e., the probability of having an SSI divided by the probability of not having an SSI). A two-tailed $p < 0.05$ indicates statistical significance, and the point estimate (i.e., odds ratio [OR]) was used to estimate the direction and effect size in the logistic regression analysis.

Method 2: decision tree

The purpose of DT is to classify the diverse characteristics of the existing data into groups that have similar characteristics. The appropriate split rule for classification should be selected to build optimal DTs, and to classify the data into sub-nodes with similar characteristics.[20,21] The classification and regression trees (CART) algorithm was used in this study. The splitting process was continued to create the next branch of a DT until a node had 5% of the total training set. To avoid overfitting, *pruning,*that is, the removal of nodes that do not provide additional information, was done through five-fold-cross-validation.[22] The DT was pruned back to the point at which the cross-validated error was at a minimum.

Method 3: support vector machine

SVM recognizes a pattern and finds the optimal hyperplane, or decision boundaries, to classify the data into two categories and minimize misclassification or error.[23] Each datum in the dataset is considered as a point in n-dimensional shape, and the SVM classifies the data into two different categories by plot or graph (the hyperplane) at an n-1-dimensional space. Simply, the SVM starts to find the hyperplane to classify each data point into one of either side of the hyperplane. For this study, all categorical data were converted into numeric attributes with a normalized scale because of the nature of a SVM. Cost (C) controls the number of misclassified examples in the training set to balance between allowing slack variables and obtaining a large margin, and gamma (γ) value controls the number of support vectors by defining the radius of the samples selected by the model.[24] To prevent overfitting, C (cost) and γ (gamma), were adjusted several times to identify the best model through the five-fold-cross-validation. Because SVM is a black-box model, the actual structure of the model cannot be described.

Model evaluation

To compare each model's performance, a test dataset was used to calculate the following estimations: accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and the area under curve (AUC). These parameters were calculated by true positive (TP), true negative (TN), false positive (FP), and false negative (FN). In this study, TP is the number of patients with SSIs in the training set who are also correctly classified as having SSIs in the test set; TN is the number of patients without SSIs in the training set who are correctly classified as not having SSI in the test set; FP refers to the number of

patients without SSIs in the training set who are incorrectly classified as having SSIs in the test set; FN is the number of patients with SSIs who are incorrectly classified as not having SSIs in the test set.

Results

Cohort demographics

A total of 9,274 patients had a PPM implanted during the study period, 205/9,274 (2.2%) of whom were diagnosed with a hospital-acquired SSI.**Table 1** summarizes the patient characteristics. Over half of the patients had a history of hypertension (65.8%), more than a quarter had renal failure or chronic pulmonary disease. The median age for patients with SSIs was eight years younger than for patients without SSIs (68 years and 76 years, respectively, $p < 0.05$). Males were significantly more likely than females to develop SSIs (63.9% and 56.8%, respectively, $p < 0.05$), and renal failure was almost twice as frequent in the patients with SSIs (46.34% and 27.28% respectively, $p < 0.0001$). In addition, while 3.7% of the patients overall developed a postoperative hematoma, the occurrence was three times more common in patients with SSIs ($p < 0.0001$). Over 45% of patients experienced a period of under-staffing. Although fewer patients with SSIs experienced understaffing, they experienced a longer duration of understaffing (difference not significant).

Building predictive models using machine learning

Nine factors (i.e., age, gender, hypertension, renal failure, postoperative hematoma, central venous catheterization, type of health insurance, length of hospitalization, and ICU stay) that indicated differences between the patients with SSIs and those without SSIs using chi-square or t-test were included in the predictive model.

Logistic regression

Table 2 summarizes the results of the associations between the risk factors and SSIs that were identified via bivariable/multivariable logistic regressions. Younger patients were more likely to have SSIs (OR, 0.99 [95% CI, 0.98-0.99]), and males were more likely to have SSIs than females (OR, 1.35 [95% CI, 1.01 – 1.8]). Renal failure and postoperative hematoma were associated with increased risk of SSIs (OR, 2.3 [95% CI, 1.74 – 30.4], and OR, 3.97 [95% CI, 2.59 – 6.08], respectively). However, patients with hypertension were less likely to have SSIs (OR, 0.62 [95% CI, 0.47 – 0.82]). In addition, having central venous catheterization was associated with increased risk of SSIs (OR, 4.21 [95% CI, 3.1 – 5.72]). With regard to the types of health insurance, only Medicaid status was associated with an increased risk of SSIs (OR 1.96 [95% CI, 1.19 – 3.24]). An extra day of hospitalization increased the risk of SSIs by 0.8% (OR 1.008 [95% CI, 1.001 – 1.015]), and patients with an ICU stay were more likely to have SSIs than were those without SSIs (OR 1.76 [95% CI, 1.33 – 2.34]) (all p-values < 0.05). However, in the multivariable logistic regression, gender, type of health insurance and the length of stay were no longer significant (all p-values > 0.05).

Decision tree

Figure 2 presents the DT model with the highest predictive ability among the five-fold-cross validation. The optimal DT was created via the *'pruning'* process with the following parameter adjustments: The degree of complexity (i.e., size of the DT) was set at between 0.001 and 0.005, a minimum of 20 observations must exist in a node in order for a split to be attempted, and there must be seven split nodes.

In this model, the presence of central venous catheterization was the first splitting parameter, which means that this characteristic was the strongest discriminating factor. This was followed by a length of stay of seven days or more, renal failure, an age of 78 years or more, postoperative hematoma, hypertension and the type of health insurance (Medicare). As shown, 62% of patients were predicted to be at risk of SSIs solely because of the presence of a central venous catheter. For patients who had had PPM implanted and had not received central venous catheterization, the likelihood of having SSIs was 1% (see**Figure 2** bottom left box). On the other hand, the likelihood of having SSIs was increased by up to 75% when the other identified risk factors were added (see **Figure 2** bottom right box). In the model, the presence of hypertension and

type of health insurance (Medicare) did not change the likelihood of having SSIs (see**Figure 2** rightmost two boxes).

Support vector machine

In this study, the radial kernel function, which generates non-linear hyperplanes, was used to determine the presence or absence of SSIs. The parameters were tuned several times to obtain the optimal SVM model. Overall among the models, the highest prediction ability was obtained when the cost (C) was 10 and gamma (γ) was in the range of 0.5 to 2.

Evaluation of the prediction ability

Table 3 provides a comparison of the prediction ability among the three models. Overall, the logistic regression algorithm had the highest prediction ability with the largest AUC at 72.9%, which suggests acceptable discrimination, and the decision tree and the support vector machine had the least ability to discriminate based on the AUC score. The support vector machine had the highest sensitivity at 43.8%, but specificity, NPV and accuracy were similar in the three model (over 98%, over 98% and over 96%, respectively). In addition, the support vector machine had the highest PPV at 32.5%.

Discussion

Recent developments in technology have led to improvements in medical diagnosis, computer-assisted decision support, and ability to make health-related decisions. In this study, machine learning algorithms (logistic regression, DT, and SVM) were used to predict SSIs in patients with implanted PPM and the predictive ability of each algorithm was compared. Research that uses a machine learning approach to analyze large datasets can provide reliable clinical insights, with the ultimate goal of decreasing health care costs, increasing efficiency of service delivery, reducing operational time and improving patient satisfaction and clinical outcomes.[25,26]

While most of the risk factors identified in this large dataset have been previously identified (e.g., renal failure, postoperative hematoma, ICU stay),[14-18] others such as obesity and temporary pacing wires and device replacement/revision were not identified, probably because they may have been under-reported by ICD-9-CM codes (**Appendix 2**). In addition, hypertension, which is directly associated with the cardiovascular conditions leading to the need for PPM, was associated with a lower risk of SSIs in this study, perhaps because it was correlated with other measured or unmeasured factors. Nurse staffing was not associated with SSIs in this study, potentially because other factors such as surgical technique or post-operative wound care were more important or because staffing was inadequately measured. Although the authors defined understaffing as below 80% of the median of nursing hours per unit following the method in a previous study,[19] there is no standardized measure of appropriate nursing hours. Because using the metric of nursing hours/patient-day does not necessarily measure the intensity of nursing care-needs, measures of staffing are needed to account for variations in the intensity of patient care requirements.

The purpose of machine learning approaches is to construct generalizable computational models.[27] Many previous studies of machine learning to identify risk factors have used a case-control design (ratio 1:1 to 1:4),[28-30] but in this study we attempted to find appropriate methods for real-time applications of machine learning in low-prevalence conditions such as SSIs. Thus, the stratified random splitting by the number of cases of SSI, cross-validations, and sophisticated parameter adjustments were used to improve the predictive models. However, researchers should explore further strategies to improve predictive ability when the data has a large difference in proportion between case and non-case.

In this study, two machine learning algorithms in addition to the more traditional logistic regression modeling were tested, and logistic regression resulted in the best predictive ability with highest AUC. High accuracy, however, is not the best parameter to use for evaluating these models because it is useful primarily when applied to symmetrical datasets in which the false positive and false negative rates are almost the same, such

as case-control study designs.[31,32]Moreover, although both the DT and SVM models had low AUC, they had high specificity and were therefore more effective for ruling-out negative patients in low-prevalence diseases or conditions such as SSIs. Despite that this study showed the comparison of three predictive models, it has very limited clinical implications because of the low predictability of models (i.e., low PPV). This might have been related to the lack of available information within the dataset. Therefore, future researchers may improve the model by incorporating text data from clinical notes through natural language processing.

Machine learning algorithms, including DTs and SVM, also had distinctive strengths. A DT is visually intuitive, allowing comprehensible classification. In addition, as seen in the DT developed in this study, the process by which the cumulative risk factors increased the risk of SSIs was also clearly shown. Thus, in terms of usability, a DT is useful for clinical decision-making because healthcare providers are able to follow the decision pathway.[33] On the other hand, a SVM is preferable to DT in datasets which have more potential risk factors with a small sample size because it utilizes the multidimensional data space for classification.[34]Thus, a further algorithm based on the DT or SVM, or in combination with other algorithms, is warranted to improve predictive ability and to take advantage of the strengths of each model.

Limitation

As with any study using a retrospective design, associations can be identified but causality cannot be inferred. Furthermore, unidentified factors not included in the dataset might have confounded some of the associations identified. Because ICD-9-CM codes were used to identify comorbidities, it is likely that some factors (e.g., obesity) were under-reported and therefore not included in the analysis. Lastly, external validity is uncertain because these algorithms were developed and tested on data from three hospitals from the same geographic region.

Conclusion

In this study, advanced machine learning algorithms were used to build prediction models by analyzing the risk factors for SSIs. Each algorithm had its strengths and weaknesses in terms of accurate prediction, and interpretable clinical decision support. However, logistic regression was more accurate for predicting low-prevalence conditions such as healthcare-associated infections.

References

1. Benjamin EJ, Muntner P, Alonso A, et al. Heart Disease and Stroke Statistics-2019 Update: A Report From the American Heart Association. *Circulation.* 2019;139(10):e56-e528. doi: https://doi.org/10.1161/cir.0000000000000659.

2. Epstein AE, DiMarco JP, Ellenbogen KA, et al. 2012 ACCF/AHA/HRS focused update incorporated into the ACCF/AHA/HRS 2008 guidelines for device-based therapy of cardiac rhythm abnormalities: a report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines and the Heart Rhythm Society. *Journal of the American College of Cardiology.* 2013;61(3):e6-75. doi: https://doi.org/10.1016/j.jacc.2012.11.007.

3. Thompson A, Neelankavil JP, Mahajan A. Perioperative Management of Cardiovascular Implantable Electronic Devices (CIEDs). *Current Anesthesiology Reports.* 2013;3(3):139-143. doi: https://doi.org/10.1007/s40140-013-0026-5.

4. Greenspon AJ, Patel JD, Lau E, et al. 16-year trends in the infection burden for pacemakers and implantable cardioverter-defibrillators in the United States 1993 to 2008. *Journal of the American College of Cardiology.* 2011;58(10):1001-1006. doi: https://doi.org/10.1016/j.jacc.2011.04.033.

5. Ihlemann N, Moller-Hansen M, Salado-Rasmussen K, et al. CIED infection with either pocket or systemic infection presentation–complete device removal and long-term antibiotic treat-

ment; long-term outcome. *Scandinavian cardiovascular journal : SCJ.* 2016;50(1):52-57. doi: https://doi.org/10.3109/14017431.2015.1091089.

6. Deharo JC, Quatre A, Mancini J, et al. Long-term outcomes following infection of cardiac implantable electronic devices: a prospective matched cohort study. *Heart (British Cardiac Society).*2012;98(9):724-731. doi: https://doi.org/10.1136/heartjnl-2012-301627.

7. Magill SS, Edwards JR, Bamberg W, et al. Multistate Point-Prevalence Survey of Health Care–Associated Infections. *New England Journal of Medicine.* 2014;370(13):1198-1208. doi: https://doi.org/10.1056/NEJMoa1306801.

8. Anderson DJ, Podgorny K, Berríos-Torres SI, et al. Strategies to prevent surgical site infections in acute Care Hospitals: 2014 update.*Infection control and hospital epidemiology.* 2014;35(6):605-627. doi: https://doi.org/10.1086/676022.

9. Weigelt JA, Lipsky BA, Tabak YP, Derby KG, Kim M, Gupta V. Surgical site infections: causative pathogens and associated outcomes.*American Journal of Infection Control.* 2010;38(2):112-120. doi: https://doi.org/10.1016/j.ajic.2009.06.010.

10. Zimlichman E, Henderson D, Tamir O, et al. Health care-associated infections: a meta-analysis of costs and financial impact on the US health care system. *JAMA internal medicine.*2013;173(22):2039-2046. doi: https://doi.org/10.1001/jamainternmed.2013.9763.

11. Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. *Health information science and systems.*2014;2:3-3. doi: https://doi.org/10.1186/2047-2501-2-3.

12. Murdoch TB, Detsky AS. The Inevitable Application of Big Data to Health CareThe Inevitable Application of Big Data to Health Care.*JAMA.* 2013;309(13):1351-1352. doi: http://doi.org/10.1001/jama.2013.393.

13. Centers for Disease Control and Prevention. International classification of diseases, ninth revision, clinical modification (ICD-9-CM). 2013; http://www.cdc.gov/nchs/icd/icd9cm.htm. Accessed March 10, 2019.

14. Klug D, Balde M, Pavin D, et al. Risk Factors Related to Infections of Implanted Pacemakers and Cardioverter-Defibrillators.*Circulation.* 2007;116(12):1349-1355. doi: http://doi.org/10.1161/CIRCULATIONAHA.106.678664.

15. Polyzos KA, Konstantelias AA, Falagas ME. Risk factors for cardiac implantable electronic device infection: a systematic review and meta-analysis. *Europace : European pacing, arrhythmias, and cardiac electrophysiology : journal of the working groups on cardiac pacing, arrhythmias, and cardiac cellular electrophysiology of the European Society of Cardiology.* 2015;17(5):767-777. doi: https://doi.org/10.1093/europace/euv053.

16. Alfonso-Sanchez JL, Martinez IM, Martín-Moreno JM, González RS, Botía F. Analyzing the risk factors influencing surgical site infections: the site of environmental factors. *Canadian journal of surgery Journal canadien de chirurgie.* 2017;60(3):155-161. doi: http://doi.org/10.1503/cjs.017916.

17. Clarke SP, Donaldson NE. Nurse staffing and patient care quality and safety. In: *Patient safety and quality: An evidence-based handbook for nurses.* Agency for Healthcare Research and Quality (US); 2008.

18. Song J, Tark A, Larson EL. The relationship between pocket hematoma and risk of wound infection among patients with a cardiovascular implantable electronic device: An integrative review. *Heart & lung : the journal of critical care.* 2020;49(1):92-98. doi: https://doi.org/10.1016/j.hrtlng.2019.09.009.

19. Shang J, Needleman J, Liu J, Larson E, Stone PW. Nurse Staffing and Healthcare-Associated Infection, Unit-Level Analysis. *The Journal of nursing administration.* 2019;49(5):260-265. doi: https://doi.org/10.1097/nna.0000000000000748.

20. Kotsiantis SB, Zaharakis I, Pintelas P. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering.* 2007;160:3-24. doi:

21. Breiman L, Friedman J, Stone CJ, Olshen RA. *Classification and Regression Trees.* Taylor & Francis; 1984.

22. Song Y-Y, Lu Y. Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry.*2015;27(2):130-135. doi: http://doi.org/10.11919/j.issn.1002-0829.215044.

23. Burges CJC. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery.*1998;2(2):121-167. doi: http://doi.org/10.1023/A:1009715923555.

24. Ben-Hur A, Weston J. A user's guide to support vector machines.*Methods in molecular biology (Clifton, NJ).* 2010;609:223-239. doi: http://doi.org/10.1007/978-1-60327-241-4_13.

25. Wiens J, Shenoy ES. Machine Learning for Healthcare: On the Verge of a Major Shift in Healthcare Epidemiology. *Clinical Infectious Diseases.* 2017;66(1):149-153. doi: https://doi.org/10.1093/cid/cix731.

26. Corbett E. The real-world benefits of machine learning in healthcare. 2017; https://www.healthcatalyst.com/clinical-applications-of-machine-learning-in-healthcare, Feb 15th, 200.

27. Reitermanová Z. Data Splitting. 2010; https://www.mff.cuni.cz/veda/konference/wds/proc/pdf10/WDS10_-105_i1_Reitermanova.pdf. Accessed 02/15/2020.

28. Meyer A, Zverinski D, Pfahringer B, et al. Machine learning for real-time prediction of complications in critical care: a retrospective study. *The Lancet Respiratory medicine.* 2018;6(12):905-914. doi: https://doi.org/10.1016/s2213-2600(18)30300-x.

29. Chen CY, Lin WC, Yang HY. Diagnosis of ventilator-associated pneumonia using electronic nose sensor array signals: solutions to improve the application of machine learning in respiratory research.*Respiratory research.* 2020;21(1):45. doi: https://doi.org/10.1186/s12931-020-1285-6.

30. Taninaga J, Nishiyama Y, Fujibayashi K, et al. Prediction of future gastric cancer risk using a machine learning algorithm and comprehensive medical check-up data: A case-control study. *Scientific reports.*2019;9(1):12384. doi: https://doi.org/10.1038/s41598-019-48769-y.

31. Li DC, Hu SC, Lin LS, Yeh CW. Detecting representative data and generating synthetic samples to improve learning accuracy with imbalanced data sets. *PloS one.* 2017;12(8):e0181853. doi: https://doi.org/10.1371/journal.pone.0181853.

32. Sun Y, Wong AKC, Kamel MS. CLASSIFICATION OF IMBALANCED DATA: A REVIEW. *International Journal of Pattern Recognition and Artificial Intelligence.* 2009;23(04):687-719. doi: https://doi.org/10.1142/S0218001409007326.

33. de Laat PB. Algorithmic Decision-Making Based on Machine Learning from Big Data: Can Transparency Restore Accountability? *Philosophy & Technology.* 2018;31(4):525-541. doi: https://doi.org/10.1007/s13347-017-0293-z.

34. Joachims T. *Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms.* Kluwer Academic Publishers; 2002.

35. Thara T, Sakchai S-h, Thakul O, Ittichai S, Anukoon K, Chin T. Machine learning applications for the prediction of surgical site infection in neurological operations. *Neurosurgical Focus FOC.*2019;47(2):E7. doi: https://doi.org/10.3171/2019.5.FOCUS19241.

Hosted file

`SSI_ML_tables.docx` available at https://authorea.com/users/332293/articles/458733-predictive-models-for-surgical-site-infection-ssi-in-patients-with-a-permanent-pacemaker-ppm-using-machine-learning-methods

Hosted file

SSI_ML_figure.docx available at https://authorea.com/users/332293/articles/458733-predictive-models-for-surgical-site-infection-ssi-in-patients-with-a-permanent-pacemaker-ppm-using-machine-learning-methods