Machine Learning Models for Surgical Site Infection Prediction

Prathyusha Mandagani[1], Shaun Coleman[1], Anam Zahid[1]
Annie Pugel Ehlers[2], Senjuti Basu Roy[1], Martine De Cock[1]
[1]Institute of Technology, University of Washington, Tacoma, Washington, U.S.
[2]UW Medicine, Seattle, Washington, U.S.

Abstract

Surgical Site Infections (SSIs) are estimated to represent 17% of all hospital-acquired infections. Their impact on morbidity, mortality, and cost of care calls for techniques to risk stratify patients ahead of surgery. In this paper we train and evaluate machine learning models (logistic regression, decision trees, and boosted decision trees) that can flag – ahead of surgery – those patients who are likely to develop a SSI. In addition to the gender and the age of the patient at the time of surgery, our models rely only on available results from blood tests done prior to the surgery. We obtain the best results with boosted decision trees (AdaBoost), i.e. an AUC-score of 86% on a benchmark patient dataset of the University Hospital of North Norway. The outcomes of Glukose and CRP tests stand out in particular as indicative features for predicting a future SSI.

Introduction

Surgical Site Infections (SSIs) represent a substantial fraction of all hospital-acquired infections, resulting in prolonged hospital stays and a healthcare cost increase, up to 34,670 USD/per patient[1]. Being able to flag at-risk patients in advance can help clinicians to take appropriate interventions to reduce the number of SSIs, thereby reducing cost and improving quality of care. The existence of a significant number of blood tests and their corresponding results, which are captured in electronic medical records, enables the use of knowledge discovery techniques to train predictive models on historical patient data. In this paper we build and test machine learning (ML) models, namely logistic regression, decision trees, and boosted decision trees, to risk stratify patients ahead of surgery. We treat the problem as a binary classification task: given demographic features, i.e. gender and age at surgery, as well as clinical features, i.e. frequency and results of blood tests done prior to surgery, predict whether the patient will develop an infection (class label 1) or not (class label 0). We evaluate our approach on a benchmark patient dataset of the University Hospital of North Norway, provided in the context of the AMIA KDDM 2016 competition[2]. We obtain the best results (AUC = 86%) with a boosted decision trees model in which the values for Glukose, CRP, Albumin and Leukocytter play an important role, next to age and gender. Standalone decision trees can achieve comparable performance but are less stable, i.e. their performance depends more on the split between train and test data in the k-fold cross-validation set-up. Below we describe the dataset, our data extraction steps, as well as the feature construction and engineering details. Next we present detailed results as well as insights derived from the trained models. We conclude with interesting directions for future research.

Dataset and Features

We obtained data from the AMIA KDDM 2016 competition[2]. The dataset contains data of 909 patients who have undergone gastrointestinal surgery. There are a total of 504,675 blood test results for 811 different blood test type names. Not every patient has taken every test, and many patients have taken the same test multiple times. The data contains results from blood tests done before as well as after surgery. As explained below, we only use results from blood tests done *prior* to surgery to closely mimic a setting in which a clinician has to decide ahead of surgery whether a particular patient is at risk of developing a SSI, and if appropriate interventions are required.

183 of the 909 patients in the dataset have developed a SSI (Cases) while 726 have not (Controls). Out of the 909 patients, only 879 have at least one blood test done prior to their surgery date, 181 of which have developed a SSI, and 698 have not. In other words, only 2 out of 30 patients with no blood tests prior to surgery have developed an infection. In the remainder of this study we use the cohort of 879 patients who have at least one blood test prior to surgery. All our models are trained and tested on this cohort. Our hypothesis is that if a patient does not have any blood tests prior to surgery, then there is too little information to make a meaningful prediction, other than expecting that the patient is at no particular risk and will not develop an infection.

The frequency at which blood tests are taken varies widely with the blood test type. We narrowed down the original list of 811 different blood test types to 76 kinds of blood tests that are clinically relevant according to a domain expert *and* that are taken by at least two patients prior to surgery (see Table 1). For each patient, we counted how many times the patient took each of these 76 tests prior to surgery, resulting in 76 numerical features per patient, called the "Blood test frequencies". If a patient did not take a particular blood test before surgery, then the value for that blood test frequency indicator would be zero.

Table 1. Non-infected, infected and total number of patients who have taken the blood test type at least once before surgery, and an indication of whether the blood test type frequency is important in the AdaBoost model. Features that are important in the decision tree model are shown in Figure 1.

Blood test type	Non-infected	Infected	Total	AdaBoost
Leukocytter	657	178	835	
CRP	590	160	750	x
Albumin	529	165	694	
Glukose	410	140	550	
Neutrofile%	194	94	288	x
HbA1c	201	70	271	
Neutrofile	150	86	236	
IgA Kvantitering	147	53	200	x
Lymfocytter	118	70	188	
Monocytter%	102	69	171	x
Eosinofile%	103	68	171	
Basofile%	102	68	170	
Basofile	105	65	170	
Eosinofile	101	64	165	
IgE tota	125	39	164	
CRP (H,N)	100	54	154	x
IgG Kvantitering	106	40	146	
IgM Kvantitering	106	38	144	
Glukose stiks	92	40	132	x
Leukocytter stiks	94	37	131	
Hematokrit (EVF)	61	39	100	
Tissue transglutaminase (IgA) antistoff	64	26	90	
Kortisol	42	18	60	
Lactalbumin IgA	40	13	53	
Lactoglobulin IgA	40	13	53	
Prealbumin	34	15	49	
pH (PNA)	27	21	48	
Glukose (PNA)	25	21	46	
Eosinofile gr	33	12	45	
Umodne granulocytter	19	14	33	
C4 Kvantitering	19	7	26	
Leukocytter gl.1 spv	12	7	19	
Protein total sp.vÌ_ske	12	6	18	
Leukocytter gl.2 spv	12	6	18	
Gentamicin	4	13	17	
Eosinofile granulocytter (N)	12	4	16	x
HbA1C (PNA)	13	3	16	
Glukose diabur	11	2	13	
Glukose sp.vÌ_ske	8	5	13	
HbA1c (glykohemoglobin) (N)	11	2	13	
CRP sensitiv	3	8	11	
HbA1c (H)	6	4	10	
Protein (N)	7	2	9	
PCR-TNF alfa i biopsi	4	5	9	
Glukose (N)	7	2	9	
Laktat (N)	8	0	8	
Penicillium M1	5	2	7	
CD4 Lymfocytter	4	2	6	
*Eosinofile# (H)	4	2	6	
CD8 Lymfocytter	4	2	6	
Granulocytter (N)	5	0	5	
Eosinofile# (H)	4	1	5	
Lymfocytter (N)	5	0	5	
Bakterier (H)	3	2	5	
Monocytter1 (N)	5	0	5	
IgE total	2	2	4	
Ciklosporin A	3	1	4	
Ampicillin (H)	1	2	3	
Nitrofurantoin (H)	1	2	3	
Cephalothin (H)	1	2	3	
CD4/CD8 ratio	3	0	3	
T-celler totalt	3	0	3	
*Glukose (H)	3	0	3	
Laktat (H)	3	0	3	
Vancomycin	0	3	3	
Mecillinam (H)	1	2	3	
Blodkultur	2	1	3	
Trim. + Sulph (H)	1	2	3	
CD34 Leukaferese produkt	1	1	2	
Leukocytter (TMS)	2	0	2	
HIV antistoff (N)	2	0	2	
CRP, semikvantitativt (N)	2	0	2	
Gentamycin (H)	1	1	2	
*CRP (H)	2	0	2	
CD34 Stamceller	1	1	2	
CRP (TMS)	2	0	2	

In addition, for the 4 tests that were taken by at least 500 patients (Leukocytter, CRP, Albumin, and Glukose), for each patient, we recorded the minimum, maximum, mean and standard deviation value of the patient's results for these blood tests, resulting in 16 numerical features per patient. If a patient had taken the test only once, then we assigned 0 as the standard deviation. We replaced missing values with the average of all non-missing values of that particular test type from the patients of the same infection class, i.e. if a patient had an infection and he did not take a particular blood test (missing value) then we assigned the average of all the mean values of other infected patients who had taken that particular blood test as a value for the mean, the minimum and the maximum for the patient with the missing blood test (and standard deviation 0). Similarly, if a patient did not have an infection and did not take a particular blood test, then we used the average of all the mean values of other non-infected patients who had taken that test. Table 2 contains an overview of all features used by our ML models. Note that the 4 "Blood test values" features for a particular blood test type are presented to the ML algorithms as 4 independent variables, i.e. the ML models do not know that e.g. CRP_mean, CRP_min, CRP_max, and CRP_SD originate from the same test.

Methods

Using the features described above, we trained 3 different ML models. *Logistic regression*[3] models are discriminative classifiers that model the posterior of the class given the input features by fitting a logistic curve. As such, logistic regression model outputs can be interpreted as probabilities of the occurrence of a class. The class decision for the given probability is then made based on a threshold value, often set to 0.5. *Decision trees* are popular ML models[4] (see Fig.1). They are constructed with a recursive algorithm that grows the tree in a top-down manner, iteratively selecting the best split feature for each node. At classification time, at each internal node in the tree, a test is applied to one of the inputs, and depending on the outcome, a sub-branch of the tree is then selected. In a decision tree, when a leaf node is reached, a prediction is made. *Boosted decision trees* are collections ("ensembles") of trees. They are trained using a boosting process in which each subsequent tree is built with weighted instances which were misclassified by the previous tree[5]. When classifying new instances, the predictions made by the individual trees are combined with weighted voting. To reduce the risk for overfitting, in our experiments, we grew the standalone decision trees to a maximum depth of 5, and we limited the number of trees in the ensemble, i.e. the rounds of boosting, to 50.

Table 2. Overview of features.

Feature	**Description**
Gender	1 (male); 2 (female)
Age	Age at time of surgery Constructed from "year of birth" and "date of surgery"
76 Blood test frequencies	Number of times that the patient took the blood test prior to surgery, for 76 blood test types
Leukocytter_mean, Leukocytter_min, Leukocytter_max, Leukocytter_SD CRP_mean, CRP_min, CRP_max, CRP_SD Albumin_mean, Albumin_min, Albumin_max, Albumin_SD Glukose_mean, Glukose_min, Glukose_max, Glukose_SD	Blood test values
Infection	Response variable 0 (no infection); 1 (infection or deep infection)

Results

Table 3 contains the results of our models, evaluated with the following measures: area under the curve (AUC), sensitivity (the percentage of correctly identified infected patients among all the real infected patients), and positive predictive value (PPV: the percentage of correctly identified infected patients among all the predicted infected patients). We used a 10-fold stratified cross validation set-up in which each partition contains approximately the same percentage of infected and non-infected patients as the complete set. The decision tree and the boosted decision trees model (AdaBoost) clearly outperform logistic regression. The important features in the final decision tree, trained on all 879 patients, can be seen in Figure 1. Gender and age at surgery do not play a role in this tree. The determining features in the AdaBoost model are Glukose_min, Glukose_mean, Glukose_SD, CRP_min, CRP_mean, CRP_SD, Albumin_max, Albumin_min, Albumin_mean, Albumin_SD, Leukocytter_max, Leukocytter_min, the 7 frequency features indicated in Table 1, age at surgery, and gender.

Conclusion and Future Work

In this paper we have presented ML models for predicting ahead of surgery which patients are at risk for developing a SSI. Our decision tree and Adaboost models outperform the logistic regression model, achieving AUC values of 85% and higher. The set of features used for training the models is very modest: only gender, age, and results and frequencies of blood tests prior to surgery are taken into account. All predictions made by our models are done based on information that can be collected before surgery, and that is usually readily available in electronic medical records. This makes our models valuable tools that are easy to implement in health systems that are taking care of large populations of individuals, allowing to better identify patients who need additional interventions or care.

Table 3. Results of machine learning models (in percentage), obtained using a 10-fold cross validation set-up

Method	**PPV (Precision)**	**Sensitivity (Recall)**	**AUC**
Logistic Regression	59	29	73
Decision Tree	75	52	85
AdaBoost	73	54	86

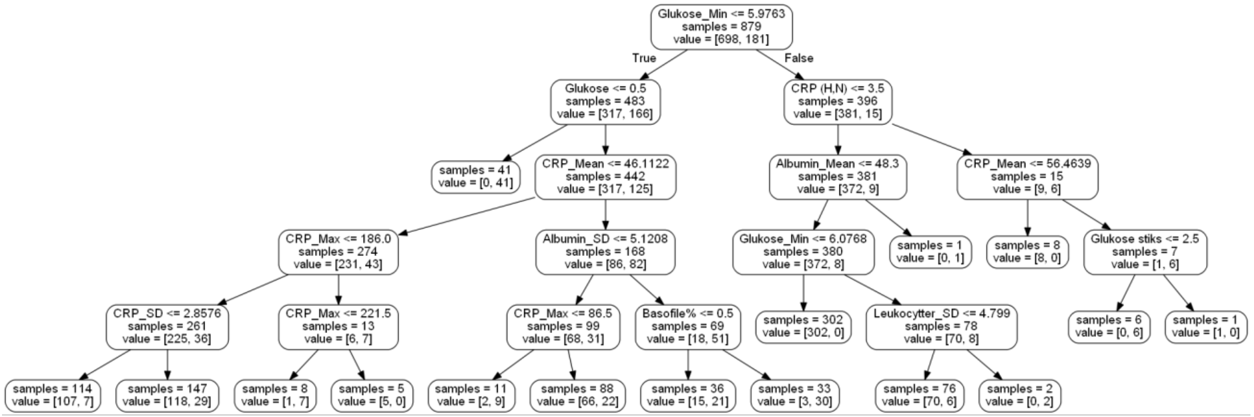

Figure 1. Decision tree for prediction of SSI, trained on all 879 patients with at least one blood test prior to surgery. The first line in each node contains the predicate evaluated at that node. For instance, “Glukose_Min <= 5.9763” verifies whether the minimum value across all Glukose tests for the patient did not exceed 5.9763, while “Glukose <= 0.5” verifies whether the number of times the patient took the Glukose test was smaller than 0.5 (i.e., the patient never took the test). The “samples=*a*” line in each node indicates the number of patients (i.e. *a*) from the training data who reached that node, while the “value=[*b*,*c*]” line indicates how many of these satisfy the predicate in the node and how many don’t, i.e. *b* patients will be processed further in the left branch, and *c* in the right branch. At the leaf level, *b* in “value=[*b*,*c*]” is the number of non-infected patients while *c* is the number of infected patients.

The data contains many missing values: for each patient usually only results for a select set of blood test types is available, which makes it more challenging to compare across patients. We circumvented this problem by only using blood test values (mean, min, max, SD) for the 4 test types taken by at least 500 patients before surgery. It is interesting to see that the decision tree and the the boosted decision trees model picked up all 4 of them (Glukose, CRP, Albumin, and Leukocytter). A valuable direction for future research would be to encorporate values from more blood test types as features, by mitigating the missing value problem through identifying duplicate blood test types. The documentation of the AMIA KDDM 2016 data mentions that there are 787 blood test types, while the data contains 811 unique strings in the test type column. The names suggest that some of these test types could be duplicates: as can be seen in Table 1, the data contains e.g. both test types “IgE total” and “IgE tota”, and both test types “Eosinofile gr” and “Eosinofile granulocytter”. If these indeed indicate the same test types, then collapsing them into one can further alleviate the missing value problem and increase accuracy of the predictive models.

References

1. Scott RD, The direct medical costs of healthcare-associated infections in U.S. hospitals and the benefits of prevention. National Center for Preparedness, Detection, and Control of Infectious Diseases, Centers for Disease Control and Prevention, Mar 2009.
2. AMIA KDDM 2016 Competition, http://kddmdata.hi.gmu.edu/, accessed on Aug 1, 2016.
3. Ng AY, Jordan MI. On discriminative vs. generative classifiers: a comparison of logistic regression and naive Bayes. Advances in Neural Information Processing Systems 2001;14:841-848.
4. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. Wadsworth Publishing Company, 1984.
5. Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences 1997;55(1):119-139.