A Prognostic Model of Surgical Site Infection Using Daily Clinical Wound Assessment

Patrick C Sanger, PhD[1], **Gabrielle H van Ramshorst, MD, PhD**[2], **Ezgi Mercan, BSc**[3], **Shuai Huang, PhD**[4], **Andrea Hartzler, PhD**[5], **Cheryl AL Armstrong, BSN, RN, MPH**[6], **Ross J Lordon, BS**[1], **William B Lober, MD, MS**[7], and **Heather L Evans, MD, MS, FACS**[6]

[1]Department of Biomedical Informatics and Medical Education, University of Washington, Seattle, WA [2]Department of Surgery, VU University Medical Center, Amsterdam, Netherlands [3]Department of Computer Science, University of Washington, Seattle, WA [4]Department of Industrial and Systems Engineering, University of Washington, Seattle, WA [5]Group Health Research Institute, Group Health Cooperative, Seattle, WA [6]Department of Surgery, University of Washington, Seattle, WA [7]Departments of Biobehavioral Nursing and Health Systems, and Biomedical Informatics & Medical Education, University of Washington, Seattle, WA

Abstract

Background—Surgical site infection (SSI) remains a common, costly and morbid healthcare-associated infection. Early detection may improve outcomes, yet previous risk models consider only baseline risk factors ("BF"), not incorporating a proximate and timely data source: the wound itself. We hypothesize that incorporation of daily wound assessment improves the accuracy of SSI identification compared to traditional BF alone.

Methods—A prospective cohort of 1,000 post-open abdominal surgery patients at an academic teaching hospital were examined daily for serial features ("SF"), e.g. wound characteristics and vital signs, in addition to standard BF, e.g. wound class. Using supervised machine learning, we trained three Naïve Bayes classifiers (BF, SF, BF+SF) using patient data from 1-5 days before diagnosis to classify SSI on the following day. For comparison, we also created a simplified SF model that used logistic regression. Control patients without SSI were matched on 5 similar consecutive post-operative days to avoid confounding by length of stay. Accuracy, sensitivity/specificity, and AUC were calculated on a training and hold-out testing set.

Corresponding Author Address: Patrick C Sanger, PhD, University of Washington, 1959 NE Pacific Street, Seattle, WA 98195, Phone: 425-954-6102, psanger@uw.edu.

Results—Of 851 patients, 19.4% had inpatient SSI. Univariate analysis showed differences in CRP, surgery duration and contamination, but no differences in ASA scores, diabetes or emergency surgery. The BF/SF/BF+SF classifiers had AUC of 0.67/0.76/0.76. The best performing classifier (SF) had optimal sensitivity of 0.80, specificity of 0.64, PPV of 0.35, and NPV of 0.93. Features most associated with subsequent SSI diagnosis were granulation degree, exudate amount, nasogastric tube presence, and heart rate.

Conclusions—Serial features provided moderate PPV and high NPV for early identification of SSI. Addition of baseline risk factors did not improve identification. Features of evolving wound infection are discernable prior to the day of diagnosis primarily based on visual inspection.

Keywords

surgical wound infection; machine learning; early diagnosis; risk factors; statistical model

Introduction

Surgical site infections (SSI) occur in 3-5% of all surgical patients, and up to 33% of patients undergoing abdominal surgery.(1, 2) More than 500,000 SSIs are estimated to occur in the US annually, resulting in worse outcomes including length of stay, mortality, and health-related quality of life, and additional average costs of as much as $20,000 per infection.(3–7) SSI is the overall costliest healthcare-associated infection, yet many of its associated costs are non-reimbursable.(8, 9)

Many risk scores for SSI have been developed over the years, ranging from simple (e.g., National Nosocomial Infections Surveillance, which includes only 3 predictors) to complex (e.g., Surgical Site Infection Risk Score, with 12 covariates and 4 interactions).(10–15) These risk score models have three main limitations.

First, methodically, existing models frequently use univariate variable selection combined with stepwise logistic regression. These traditional methods do not ideally handle nonlinear data and may result in both selection of a suboptimal variable set and overfitting, especially when the number of potential variables is high.(16) An alternative is the application of machine learning, an approach to data analysis that has evolved from pattern recognition and computational learning theory. It involves the construction of algorithms (analogous to statistical models) that can both learn from an existing dataset and predict, or "classify" outcomes using "features" (analogous to variables) as input. Modern feature selection techniques developed to avoid overfitting in the context of high-dimensional "big data" (e.g., genomic studies) provide a more robust and reliable alternative.(17–19) Machine learning is especially well suited to biological systems because they tend to evolve large, noisy, non-linear and complex data sets. Widely used in other domains, machine learning is increasingly being explored in healthcare research and practice for pneumonia prediction, genomics, and cancer diagnosis and prediction.(20–22)

Second, these models do not provide a time-specific prediction. Predictions generally apply over a 30-day post-operative time horizon, if specified at all, leaving providers without

clinically-actionable data. These models facilitate risk-adjustment more than clinical decision support.

Third, existing models only incorporate static variables known as of the end of the operation, for example, patient characteristics, pre-operative laboratory results, comorbidities, and operative factors. These models do not incorporate a rich and continuing source of data: serial observations of the patient and their wound that may serve as markers for changing risk of SSI over time.

We hypothesize that incorporation of daily wound assessment data (i.e., SF) improves the performance of SSI classification compared to traditional baseline risk factors alone (i.e., BF). To test this hypothesis, we employed machine learning techniques to develop and test SSI classifiers for SF and BF data.

Methods

Ethics approval was obtained for the parent study,(23) briefly described below, from which the dataset was derived; the present analysis was deemed exempt from review by the University of Washington IRB due to the deidentified nature of the dataset.

Study population

Data was drawn from a prospective cohort study of 1,000 open abdominal surgery patients conducted at a 1,200-bed academic teaching hospital in the Netherlands, described previously.(23) Patients who did not ultimately undergo open surgery (n=33) or with <2 days of wound observations (n=116) were excluded from analysis, leaving 851 patients in total.

Data collection

The parent dataset had been created by a trained research team that tracked infections independently from the clinical care providers, and the observations were supervised daily and adjudicated weekly by the principle investigator (GHvR). BF collected included demographics, preoperative labs, procedure characteristics, other risk factors, and outcomes (see Results Table 1 for a full list of BF collected). SF collected daily from post-operative day 2 included abdominal wound characteristics and vital signs (see Supplemental Table 1 for a full list, including definitions of categorical wound score variables).(24) The primary outcome was SSI using the Centers for Disease Control and Prevention (CDC) criteria for superficial, deep and/or organ space infections. Follow-up was done at the outpatient clinic on postoperative day 30, or alternatively by telephone or letter. Patient charts, discharge letters, wound photographs, and culture results were reviewed by GHvR after a minimum period of three months following discharge.

For the purpose of the present analysis, we defined the SSI group as having any of the 3 types of SSI, as have prior studies.(12–15, 25, 26) Our initial analyses showed that baseline and serial features generally did not vary significantly among different classes of SSI, and we could not separately model subtypes of SSI due to the small numbers of deep and organ-space infections. In addition, though a patient may have developed multiple types of SSI

during their hospital stay, we only include their first infection in this analysis. The non-SSI group was defined as having no inpatient SSI; post-discharge SSI were not disqualifying. We grouped patients in this way because initial analyses of the data showed that patients with post-discharge SSI, while they were in the hospital, closely resembled patients who never developed SSI in terms of both BF and SF. Post-discharge SSI rates are included for descriptive purposes, but due to lack of post-discharge serial wound observations, we focus on modeling inpatient SSI.

Univariate analysis

For BF, we tested for differences between SSI and non-SSI groups using ANOVA for continuous variables and Pearson's chi-squared for binary and categorical variables. For serial features, we used reversed time analysis to examine symptom trends leading up to SSI diagnosis. Specifically, we normalized observation times to the initial day of SSI diagnosis as "Day 0", and then looked backwards in time from days -5 through -1 (i.e., 5 days before SSI diagnosis to 1 day before SSI diagnosis). Patients without SSI were matched so as to have similar post-operative days included for comparison, resulting in an equal distribution (mean, SD) of SSI and non-SSI post-operative days in the analysis to avoid confounding by length of stay (patients with SSI had longer LOS). Statistical analysis was conducted using Stata 13 (StataCorp).

Model development

Overview—First, we transformed the existing dataset through "feature generation"(27) (described below) to create potential features for inclusion in the model. Next, we used stratified randomization to divide the dataset into training (2/3) and testing (1/3) sets to ensure balanced SSI outcomes in the training and testing sets. Then, we used supervised machine learning to train and optimize classifiers using only baseline features (BF), only serial features (SF), or both baseline and serial features (BF+SF). Next, we evaluated the best performing model on the testing set. Finally, we created and evaluated a simplified model using logistic regression intended to be more clinically applicable.

Microsoft Excel 2013 was used for feature generation. WEKA 3.7.12, an open source data mining package (http://www.cs.waikato.ac.nz/ml/weka/), was used to optimize and evaluate classifiers. For model training and evaluation, missing data were imputed using means for continuous variables and modes for binary/categorical variables. For serial variables (e.g. wound characteristics, vital signs), we used the same process as just described sequentially on each day leading up to "Day 0". In general, for baseline features, we had <5% missing data except for pre-operative labs, which had anywhere between 17% (Hb) to 30-50% (Creatinine, BUN, platelets, WBC, CRP) to 50-60% (total bili, PT, PTT, albumin, total protein). For serial features, we had <5% missing data, with vital signs generally 2-3% missing and wound observations 3-5% missing. The most common reason for missing wound data was clinical need to retain an intact dressing. Less than 1% of data on outcomes (e.g. SSI, mortality) was missing.

Feature generation—Feature generation refers to the process of taking raw, unstructured data to define "features" (conceptually similar to variables) for potential use in

the analysis. For each patient, we generated both BF (Results Table 1) and SF (Supplemental Table 1). For BF, we included raw values and discretized versions using established clinical cutoffs (e.g., duration of surgery >3 hours). For SF, we generated features from raw values, differences in values from day to day, maximums/minimums/averages over time, coefficients of variation over time, rates of change over time, and deviations from linear trendlines (i.e., expected value minus observed value). For each of these features, we included varying lookback periods, from 1 to 5 days prior to diagnosis. Lookback periods are cumulative, for example, a lookback period of 5 days includes data from days -5 through -1, e.g. the maximum heart rate over the previous 5 days. Data from the day of diagnosis was not included in any model. We describe which features, over which lookback periods, were most influential in the model in the Results section.

Model training—We used a Naïve Bayes classifier, which is a conditional probability model that employs both the prior probability and the likelihood of the outcome to form a posterior probability,(28, 29) to build models with BF, SF, and BF+SF. We also explored models using support vector machines (SVM), logistic regression, and ensemble techniques (bagging, boosting, etc.). We initially chose the Naïve Bayes classifier based on its combination of simplicity and performance. Feature selection was performed by using a forward wrapper-based method(30) using an Information Gain(31) heuristic to optimize area under the ROC curve (AUC). In simpler terms, features were added to each model until they no longer resulted in improvements in AUC, a process analogous to stepwise selection employed in creation of logistic regression models. Within the training set, each classifier was trained and evaluated using 10-fold cross validation to avoid overfitting. In this procedure, the entire dataset is first randomly divided into ten equal parts, the models are trained on nine parts and tested on the tenth part, and this process is repeated ten times each time using a different part for testing (i.e., the same data was never used to both train and evaluate a classifier). The results of all these ten folds are then combined to compute the evaluation scores.

In an effort to create a simplified, more clinically-relevant model, we reduced the dataset to include only simple, easy-to-calculate serial features (minimum, average, maximum) limited to a 2 day look-back period (i.e. using only Day -1 and Day -2 data), and including only the top 5 features. We then used the same techniques described above with the exception of using a logistic regression model, which has the benefit of allowing quick clinical calculation of probabilities. We report odds ratios associated with the selected features; WEKA does not allow calculations of confidence intervals.

Performance evaluation—Model performance was evaluated based on accuracy, Kappa, and AUC. All values are based on averages over 10 cross-validation runs on the training set and 1 run on the testing set. AUCs were tested for significance based on paired t-test.

Results

Of 851 participants included in analysis, 167 (19.6%) had at least one inpatient SSI and 62 (7.3%) had at least one post-discharge SSI, for an overall SSI rate of 26.9%. Of inpatient

SSIs, the first infection was superficial for 126 (75%), deep for 22 (13%) and organ-space for 19 (11%). Figure 1 shows the overall distribution of SSIs based on post-operative day of diagnosis. Among first inpatient SSIs, the median SSI was diagnosed at day 8 (IQR 6-11), the median post-discharge SSI was diagnosed at day 18 (IQR 10-22), for overall median 9 (IQR 6-13).

Post-discharge infections are included for descriptive purposes, but because of their small numbers and our lack of post-discharge wound observations, they are not a focus of this paper. Of note, 17 of 62 of post-discharge infections occurred within 3 days of discharge and 26 of 62 occurred within 5 days of discharge. Table 1 shows differences in BF, and Table 2 shows differences in SF, among patients with and without inpatient SSI.

Baseline features (BF)

In the group that developed inpatient SSI, hemoglobin, albumin, and creatinine were significantly lower, while C-reactive protein was higher. Patients with SSI had procedures that were, on average, 59 minutes longer, had more wounds classified as "dirty", more bowel and pancreatic operations, more ostomies created, more peri-operative blood transfusions, and fewer kidney or liver transplants. Patients with SSI had lengths of stay 9.9 days longer and an in-hospital mortality 5.3% greater.

Serial features (SF) shows that the prevalence of abnormal wound symptoms (defined as a score of >0 on any of the scales depicted in Supplemental Table 1) in the 5 days and 1 day prior to SSI is higher in the SSI group, except for wound edge color which was not associated with SSI until the day of diagnosis (Day 0; p=0.048). The bottom of Table 2 also shows, among patients who subsequently developed SSI, more use of NG decompression, more wound cultures ordered, higher heart rate, larger difference in body minus wound temperature, longer wound length, higher tympanic temperature, lower diastolic BP, and similar pain scores (except on day of SSI).

Classifier performance

The following results relate to the performance of the Naïve Bayes classifiers trained on the baseline features (BF), serial features (SF) and the BF+SF datasets, and a "simplified SF" classifier using logistic regression. Table 3 demonstrates that the SF, SF+BF, and SF (simplified) classifiers perform best on both the training and testing sets. The differences between training AUC in the SF, BF+SF, and SF (simplified) classifiers were not statistically significant. Differences between training and testing AUC among all the classifiers were also not statistically significant, though this is likely due to lack of power given the small N of the holdout testing set; the simplified SF classifier shows a markedly lower testing set AUC. We chose the simper SF and SF (simplified) models for further evaluation.

Figure 2 shows the ROC curve for the SF classifiers, and Table 4 shows the resulting sensitivity/specificity combinations. The ROC curve is an average over 10 cross-validation runs on the training set. The points chosen on the curve were selected by eye for illustrative purposes.

Table 5 shows the features selected by the SF models (both the original, full SF model and a simplified SF model) using a wrapper-based feature selection method. The order of the table represents the order in which the features were added to the model, with decreasing classification discrimination towards the bottom. We indicate which original data elements the selected features were based on, as well as the lookback period (ranging from cumulative 5 days to 1 day prior to SSI), and the type of transformation used to generate the feature.

The final simplified SF model was: logit(p) = -3.087 + [NGtube.max2] * 0.645 + [degreeexudate.max2] * 0.213 + [pulse.max2] * 0.017 + [typeslough.max2] * 0.165 + [woundlength.max2] * 0.016

We did a post-hoc analysis of the performance of the full SF classifier among various subgroups to assess potential limitations in performance due to type and timing of SSI, shown in Figure 3. We demonstrate that the classifier performs equally well in identifying superficial, deep, and organ-space SSIs (sensitivity, left side of figure). We also show that performance tends to be consistent across post-operative days, with a possible trend toward higher sensitivity early in hospitalization (middle of figure). Finally, on the right side of the figure, we show that among patients who later go on to develop a post-discharge SSI, classifier specificity *prior to discharge* (i.e., detecting inpatients who do not *currently* have an SSI) is comparable, though tends to be lower than among patients who never develop SSI, likely due to some shared risk factors and/or slowly developing SSI while inpatient.

Discussion

We demonstrate with a large inpatient prospectively collected dataset that serial physical examination and vital sign data is more informative than baseline risk factor data in a prognostic model of SSI after open abdominal surgery. In addition, we showed that patients with SSI differ from patients without SSI with regard to the prevalence of abnormal wound symptoms, especially in the 1-3 days prior to diagnosis.

Baseline data from our population supports numerous other studies identifying risk factors for SSI, for example, differences in hemoglobin, CRP, surgery duration, wound class, and surgery type.(32) Yet, we demonstrate that, in our dataset, baseline features provide relatively poor classification of SSI (AUC 0.670), while serial features have significantly better performance (AUC 0.760; p<0.0001). When the SF classifier was applied to the hold-out test set, it achieved similar performance (AUC 0.741) to the training set, indicating that overfitting was not a significant concern. The addition of BF to the SF classifier did not improve performance significantly, indicating that serial features were both necessary and likely sufficient for optimal classification. We suggest that the SF classifier could be reasonably used as a screening tool, with 80% sensitivity, and 64% specificity (PPV 35%, NPV 93%; see Table 4).

The simplified SF model which used logistic regression and fewer, simpler features also performed well, with 75% sensitivity and 64% specificity (PPV 33%, NPV 92%), though showed decreased performance on the hold-out testing set. Such a model shows promise for quick bedside assessment of wounds (i.e., not the "black box" approach of many machine

learning approaches), though should be prospectively validated alongside the full, gold standard model to more confidently assess the potential performance-usability tradeoff.

In developing our models, we identified features that were most highly associated with diagnosis of subsequent SSI. Patients with SSI have an increased prevalence of abnormal wound symptoms in the 5 days prior to diagnosis. Yet, many of the most highly associated symptoms are not included in current definitions of SSI, and several of the least associated symptoms are. For example, the least associated wound-related features in Table 2 were wound edge color and amount of induration, calling into question whether these signs, currently included in the CDC definition of SSI, are ideal—or even reliable—indicators of infection. Upwards of 90% of both SSI and non-SSI patients in our dataset were deemed to have bright red skin surrounding their wound.

On the other hand, we found many good indicators of infection, including several that are not part of current definitions. For example, we found degree of granulation (i.e., a scale consisting of: closed wound, >75% filled, 50-75% filled, 25-50% filled, <25% filled) to be most associated with subsequent SSI diagnosis, and we found amount of exudate to be more associated with SSI than type of exudate. Wound edge distance was an excellent early indicator, as was heart rate, morning body temperature and, somewhat unexpectedly, nasogastric tube presence. Prolonged post-operative use of nasogastric decompression is an indicator of delayed recovery of gastrointestinal function, and has been demonstrated by others to be associated with inpatient SSI, likely as a marker for predisposition to complications rather than causative of SSI.(33) By incorporating objective elements into SSI definitions, infections may be identified earlier, reducing associated costs and morbidity and improving quality of care.

One key limitation of current risk models of SSI, including our own, is that they have been developed using data from inpatients. Yet, as economic and other legitimate concerns encourage rapid discharge of patients, 60% of SSIs now occur post-discharge, at a time when adequate follow-up is difficult.(34) Post-discharge SSIs are especially challenging because patients are ill-prepared to identify them, no standardized or reliable methods of post-discharge surveillance exist, and relatively few risk factors have been identified.(35–38) Delayed diagnosis of post-discharge SSIs has significant financial and quality costs, with more than half of patients who develop them readmitted to the hospital.(26, 37) Our dataset is unique in that, coming from a hospital in the Netherlands, length of stay is significantly longer than in the US (median 12 days in our dataset vs 5-6 days in the US), allowing us to observe events that might occur after discharge in the US.(3, 39) Yet, our dataset also has inherent limitations which likely decreased our model's ultimate performance—e.g., due to lack of data on patients after discharge, we combined patients with post-discharge SSI with patients who never developed SSI, and due to low numbers of deep and organ-space SSI, we combined all types of SSI into a single endpoint; with both of these decisions, we took a conservative, intention-to-treat approach rather than selectively excluding populations who do not fit neatly—to better show the prospective, "real-world" performance of our model. Also, our dataset represents only open or converted laparoscopic abdominal surgery patients treated in an academic setting, and includes different kinds of abdominal operations, including those among immunosuppressed patients. And finally, despite a concerted effort at

objective collection and scoring of wound data in the parent study, our secondary analysis of this data is ultimately dependent on the standard, inherently subjective CDC definition of SSI, which has been shown to be highly variable among surgeons and between surgeons and other providers.(23, 40–42) In light of our findings that wound features such as redness and induration are not reliable signs of evolving SSI, we suggest both a re-evaluation of the diagnostic process and a strengthening of the objective criteria are in order.

Future work will address many of the limitations of the current study: we aim to evaluate the full and simplified SF models' generalizability in a variety of real-world settings, prospectively testing the daily classification accuracy in a cohort of post-operative patients. In addition, we plan to initiate outpatient data collection, using our findings here to inform which wound features and symptoms we systematically track. Our work shows that looking at the wound daily has value; while this is a minor challenge for hospitalized patients, increasingly shorter hospitalizations demand new techniques to facilitate post-discharge surveillance to capture events when they happen. Our ultimate goal is to leverage mobile health technology to collect patient-generated data from multiple sources such as questionnaires, photos and other sensors, using automated image analysis to identify evolving infections in real time. Counterintuitively, we may never have a single, static model; the strength of machine learning is that it is iterative and data-driven, continuing to adapt and learn from additional populations and settings over time as the data source grows in volume and complexity.

As we continue this line of investigation to refine and apply the algorithm prospectively, this new kind of dynamic surveillance has potential to affect clinical practice in a number of ways. Used as a screening tool in unselected patients, an elevated calculated risk might increase vigilance in both nurses and doctors (e.g., in inpatients, closer lab monitoring or confirmatory imaging studies; in outpatients, shorter follow-up window), or decrease the risk that frank SSIs go unrecognized. Conversely, a low risk score could reassure patients that their post-operative wound is healing within acceptable norms, and avert unnecessary emergency evaluation. In the context of a patient with existing clinical suspicion of SSI, it may provide corroborating evidence of progression of infection, allowing earlier intervention, e.g. wound opening.

Conclusions

Using a novel machine learning-based computational method, we show that serial features (i.e., daily wound observations and vital signs) outperform traditional baseline patient/ operative risk-factor data, providing moderate PPV and high NPV for identification of SSI in advance of clinical diagnosis. We demonstrate that features of evolving wound infection are discernable prior to the day of diagnosis, primarily based on vital signs and visual inspection, proving the value of objective daily wound assessment. Existing definitions of SSI may be made more reliable and more timely by incorporation of such objective features.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

References

1. Pinkney TD, Calvert M, Bartlett DC, et al. Impact of wound edge protection devices on surgical site infection after laparotomy: multicentre randomised controlled trial (ROSSINI Trial). BMJ. 2013; 347:f4305. [PubMed: 23903454]

2. Krieger BR, Davis DM, Sanchez JE, et al. The use of silver nylon in preventing surgical site infections following colon and rectal surgery. Dis Colon Rectum. 2011; 54:1014–9. [PubMed: 21730792]

3. Mahmoud NN, Turpin RS, Yang G, Saunders WB. Impact of surgical site infections on length of stay and costs in selected colorectal procedures. Surg Infect (Larchmt). 2009; 10:539–544. [PubMed: 19708769]

4. Kirkland KB, Briggs JP, Trivette SL, et al. The impact of surgical-site infections in the 1990s: attributable mortality, excess length of hospitalization, and extra costs. Infect Control Hosp Epidemiol. 1999; 20:725–30. [PubMed: 10580621]

5. Perencevich, EN.; Sands, KE.; Cosgrove, SE., et al. Emerg Infect Dis. Vol. 9. Centers for Disease Control and Prevention; 2003. Health and economic impact of surgical site infections diagnosed after hospital discharge; p. 196-203.

6. Dipiro JT, Martindale RG, Bakst A, et al. Infection in surgical patients: Effects on mortality, hospitalization, and postdischarge care. Am J Heal Pharm. 1998; 55:777–781.

7. Anthony T, Long J, Hynan LS, et al. Surgical complications exert a lasting effect on disease-specific health-related quality of life for patients with colorectal cancer. Surgery. 2003; 134:119–125. [PubMed: 12947307]

8. Zimlichman E, Henderson D, Tamir O, et al. Health Care–Associated Infections. JAMA Intern Med. 2013; 173:2039. [PubMed: 23999949]

9. Stone PW, Braccia D, Larson E. Systematic review of economic analyses of health care-associated infections. Am J Infect Control. 2005; 33:501–509. [PubMed: 16260325]

10. Ho VP, Stein SL, Trencheva K, et al. Differing risk factors for incisional and organ/space surgical site infections following abdominal colorectal surgery. Dis Colon Rectum. 2011; 54:818–25. [PubMed: 21654248]

11. Lawson EH, Hall BL, Louie R, et al. Association between occurrence of a postoperative complication and readmission: implications for quality improvement and cost savings. Ann Surg. 2013; 258:10–18. [PubMed: 23579579]

12. Berger, RL.; Li, LT.; Hicks, SC., et al. J Am Coll Surg. Vol. 217. Elsevier Inc; 2013. Development and validation of a risk-stratification score for surgical site occurrence and surgical site infection after open ventral hernia repair; p. 974-82.

13. van Walraven C, Musselman R. The surgical site infection risk score (SSIRS): a model to predict the risk of surgical site infections. PLoS One. 2013; 8:e67167. [PubMed: 23826224]

14. Saunders L, Perennec-Olivier M, Jarno P, et al. Improving prediction of surgical site infection risk with multilevel modeling. PLoS One. 2014; 9:e95295. [PubMed: 24835189]

15. Horan TC, Gaynes RP, Martone WJ, Jarvis WR. CDC Definitions of Nosocomial Surgical Site Infections, 1992: A Modification of CDC Definitions of Surgical Wound Infections. Infect Control Hosp Epidemiol. 1992; 13:606–608. [PubMed: 1334988]

16. Mitchell, TM. Machine Learning. New York: McGraw Hill; 1997.

17. Steyerberg EW, Eijkemans MJC, Habbema JDF. Stepwise selection in small data sets: A simulation study of bias in logistic regression analysis. J Clin Epidemiol. 1999; 52:935–942. [PubMed: 10513756]

18. Hastie, T.; Tibshirani, R.; Friedman, J. The Elements of Statistical Learning. 2nd. New York, NY: Springer; 2009. p. 1-282.

19. Harrell, F. Regression modeling strategies: With application to linear models, logistic regression, and survival analysis. New York: Springer-Verlag; 2001.

20. Bejan CA, Xia F, Vanderwende L, et al. Pneumonia identification using statistical feature selection. J Am Med Informatics Assoc. 2012; 19:817–23.

21. Cruz JA, Wishart DS. Applications of machine learning in cancer prediction and prognosis. Cancer Inform. 2006; 2:59–77. [PubMed: 19458758]

22. Yip KY, Cheng C, Gerstein M. Machine learning and genome annotation: a match meant to be? Genome Biol. 2013; 14:205. [PubMed: 23731483]

23. van Ramshorst GH, Vos MC, den Hartog D, et al. A comparative assessment of surgeons' tracking methods for surgical site infections. Surg Infect (Larchmt). 2013; 14:181–7. [PubMed: 23485257]

24. van Ramshorst, GH. Dissertation: Wound Failure in Laparotomy: New Insights. 2014. Available at: http://repub.eur.nl/pub/50266/

25. Gibbons C, Bruce J, Carpenter J, et al. Identification of risk factors by systematic review and development of risk-adjusted models for surgical site infection. Health Technol Assess. 2011; 15:1–156, iii–iv. [PubMed: 21884656]

26. Limón E, Shaw E, Badia JM, et al. Post-discharge surgical site infections after uncomplicated elective colorectal surgery: impact and risk factors. The experience of the VINCat Program. J Hosp Infect. 2014; 86:127–32. [PubMed: 24393830]

27. Yu H, Lo H, Hsieh H. Feature engineering and classifier ensemble for KDD cup 2010. J Mach Learn Res. 2010:1–12.

28. John, GHG.; Langley, P. Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence. Montreal, Quebec, Canada: 1995. p. 338-345.

29. Rish I. An empirical study of the naive Bayes classifier. IJCAI 2001 Work. Empir methods Artif Intell. 2001; 22230:41–46.

30. Kohavi R, John G. Wrappers for feature subset selection. Artif Intell. 1997; 97:273–324.

31. Kullback, S. Information Theory and Statistics. New York: Wiley; 1959.

32. Mangram AJ, Horan TC, Pearson ML, et al. Guideline for prevention of surgical site infection, 1999. Hospital Infection Control Practices Advisory Committee. Infect Control Hosp Epidemiol. 1999; 20:250–78. quiz 279–80. [PubMed: 10219875]

33. Delgado-Rodríguez M, Gómez-Ortega A, Sillero-Arenas M, Llorca J. Epidemiology of surgical-site infections diagnosed after hospital discharge: a prospective cohort study. Infect Control Hosp Epidemiol. 2001; 22:24–30. [PubMed: 11198018]

34. Woelber E, Schrick E, Gessner B, Evans HL. Proportion of surgical site infection occurring after hospital discharge: A systematic review. Surg Infect (Larchmt). 2016 In press.

35. Mangram AJ, Horan TC, Pearson ML, et al. Guideline for prevention of surgical site infection, 1999. Infect Control Hosp Epidemiol. 1999; 27:97–134.

36. Daneman N, Lu H, Redelmeier DA. Discharge after discharge: predicting surgical site infections after patients leave hospital. J Hosp Infect. 2010; 75:188–194. [PubMed: 20435375]

37. Gibson, A.; Tevis, S.; Kennedy, G. Am J Surg. Vol. 207. Elsevier Inc; 2014. Readmission after delayed diagnosis of surgical site infection: a focus on prevention using the American College of Surgeons National Surgical Quality Improvement Program; p. 832-839.

38. Kazaure HS, Roman SA, Sosa JA. Association of postdischarge complications with reoperation and mortality in general surgery. Arch Surg. 2012; 147:1000–7. [PubMed: 23165614]

39. Bratzler DW, Houck PM, Richards C, et al. Use of antimicrobial prophylaxis for major surgery: baseline results from the National Surgical Infection Prevention Project. Arch Surg. 2005; 140:174–82. [PubMed: 15724000]

40. van Ramshorst G, Vrijland W. Validity of Diagnosis of Superficial Infection of Laparotomy Wounds Using Digital Photography: Inter-and Intra-observer Agreement Among Surgeons. Wounds. 2010; 22:38–43. [PubMed: 25901725]

41. Hedrick TL, Sawyer RG, Hennessy Sa, et al. Can we define surgical site infection accurately in colorectal surgery? Surg Infect (Larchmt). 2014; 15:372–6. [PubMed: 24811074]

42. Bruce J, Russell EM, Mollison J, Krukowski ZH. The quality of measurement of surgical wound infection as the basis for monitoring: a systematic review. J Hosp Infect. 2001; 49:99–108. [PubMed: 11567554]

Abbreviations

SSI	surgical site infection
BF	baseline features/baseline risk factors
SF	serial features
AUC	area under the ROC curve
CRP	C-reactive protein
ASA	American Society of Anesthesiologists
CDC	Center for Disease Control
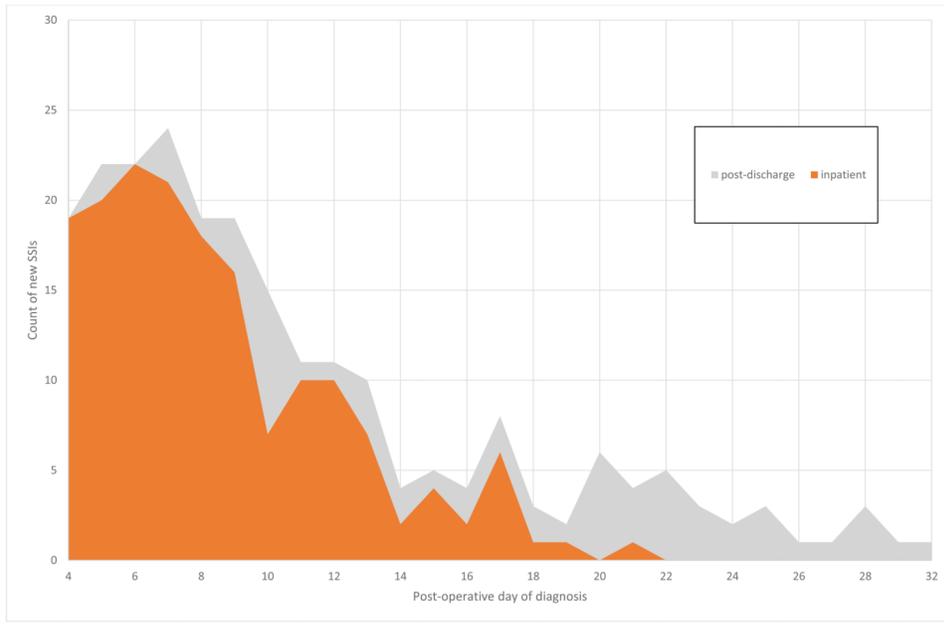
Figure 1.
Daily count of new inpatient (orange) and post-discharge (grey) SSIs.
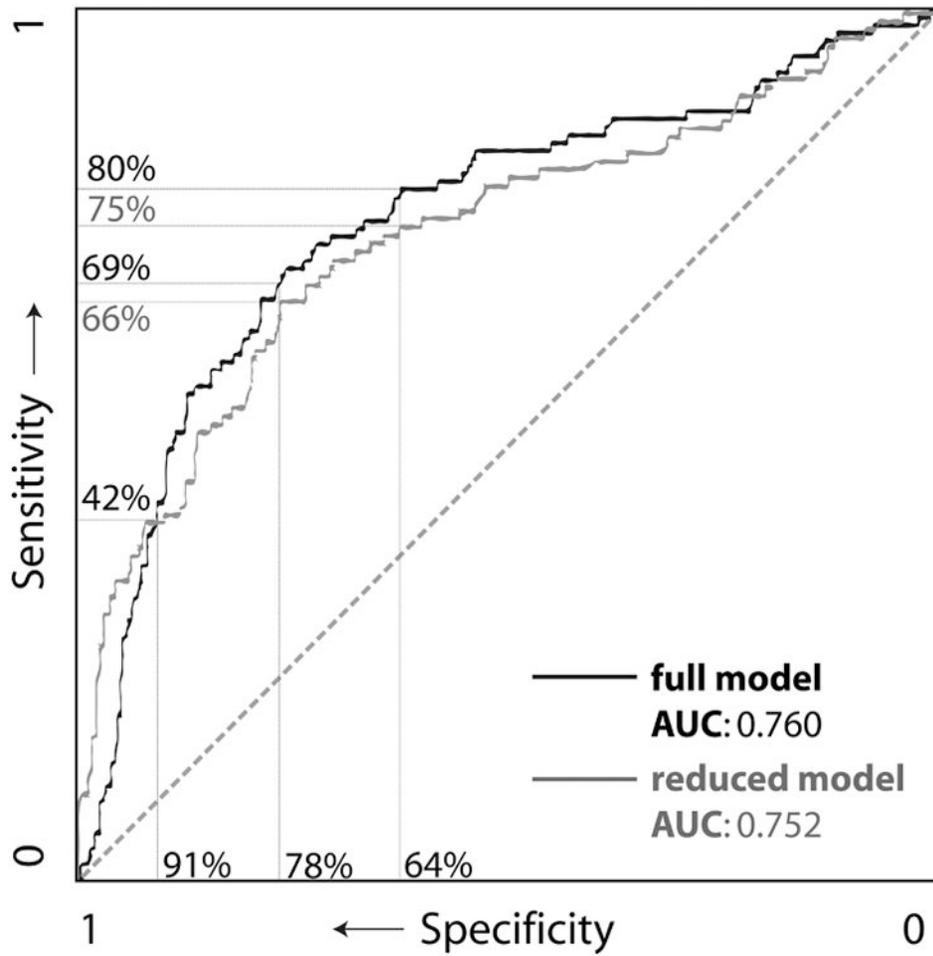
Figure 2.
Receiver operating curve (ROC) of Serial Features classifiers with example sensitivity/specificity pairs.
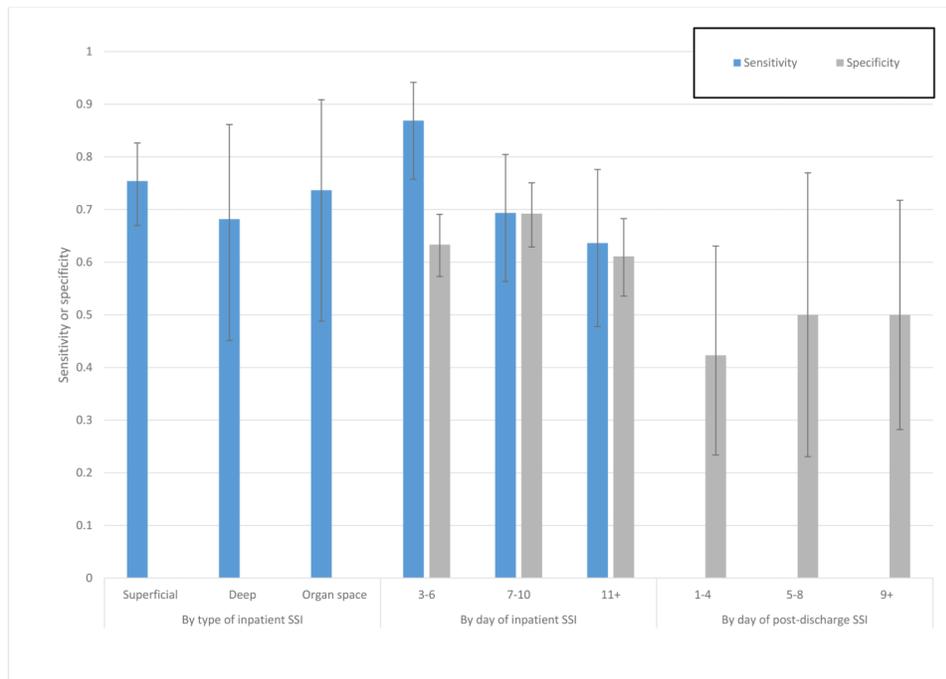
Figure 3.
Post-hoc analysis of SF classifier performance by type and day of SSI. Error bars show 95% CI.

Table 1

Baseline Features from Patient Cohorts With and Without Inpatient SSI

	Without SSI (N=684; 80.4%)	With SSI (N=167; 19.6%)	p Value
Demographics			
Age, y, mean, (95% CI)	56.2 (55.1-57.3)	57.5 (55.5-59.5)	0.29
Male sex			
n (%)	247 (36.1)	62 (37.1)	
95% CI	0.33-0.40	0.30-0.45	0.81
Pre-operative labs			
Hemoglobin, g/dL, mean, (95% CI)	12.7 (12.6-12.9)	12.1 (11.6-12.6)	<0.001
Total protein, g/dL, mean, (95% CI)	6.8 (6.7-6.9)	6.7 (6.4-7.0)	0.82
Albumin, g/dL, mean, (95% CI)	4.1 (4.0-4.2)	3.8 (3.6-4.0)	0.002
Blood urea nitrogen, mg/dL, mean, (95% CI)	12.4 (11.4-13.3)	11.2 (8.9-13.4)	0.29
Creatinine, mg/dL, mean, (95% CI)	3.3 (2.9-3.6)	2.2 (1.7-2.8)	0.006
C-reactive protein, mg/L, mean, (95% CI)	2.7 (2.0-3.4)	5.5 (3.4-7.6)	0.002
Platelet count, 103/uL, mean, (95% CI)	236 (226-246)	257 (230-283)	0.097
WBC count, 103/uL, mean, (95% CI)	8.2 (7.8-8.6)	8.9 (8.0-9.8)	0.15
PT, seconds, mean, (95% CI)	14.3 (13.5-15.1)	15.7 (13.4-18)	0.15
aPTT, seconds, mean, (95% CI)	34.5 (33.3-35.6)	37.7 (34.0-41.4)	0.036
Procedure-related			
Duration of surgery, minutes, mean, (95% CI)	253 (244-262)	312 (289-335)	<0.001
Wound class, n (%)			<0.001
Clean	135 (19.7)	12 (7.2)	
Clean-contaminated	497 (72.7)	131 (78.4)	
Contaminated	21 (3.1)	7 (4.2)	
Dirty	31 (4.5)	17 (10.2)	
Type of operation, n (%)			<0.001
Abdominal wall	44 (6.4)	3 (1.8)	
Gastroduodenum	27 (3.9)	4 (2.4)	
Gall bladder/bile duct	31 (4.5)	4 (2.4)	
Liver	101 (14.8)	19 (11.4)	
Spleen/adrenal gland & other	29 (4.2)	4 (2.4)	
Small bowel	35 (5.1)	18 (10.8)	
Kidney	179 (26.2)	23 (13.8)	
Vascular	50 (7.3)	6 (3.6)	
Esophagus	75 (11.0)	25 (15.0)	
Large bowel	69 (10.1)	35 (21.0)	
Pancreas	44 (6.4)	26 (15.6)	
Emergency surgery			
n (%)	149 (21.8)	41 (24.6)	0.45

	Without SSI (N=684; 80.4%)	With SSI (N=167; 19.6%)	p Value
95% CI	0.19-0.25	0.18-0.32	
Kidney or liver transplantation			
n (%)	187 (29.2)	29 (17.8)	
95% CI	0.26-0.33	0.12-0.25	0.003
Ostomy created			
n (%)	40 (5.8)	18 (10.8)	
95% CI	0.04-0.08	0.07-0.16	0.023
Blood transfusion peri-op			
n (%)	146 (21.4)	50 (30.1)	
95% CI	0.18-0.25	0.23-0.38	0.017
Risk factors			
Smoking			
n (%)	287 (42.0)	59 (35.3)	
95% CI	0.38-0.46	0.28-0.43	0.12
Diabetes mellitus type I or II			
n (%)	83 (12.2)	21 (12.6)	
95% CI	0.10-0.15	0.08-0.19	0.88
Chronic lung disease			
n (%)	58 (8.5)	22 (13.2)	
95% CI	0.07-0.11	0.08-0.19	0.063
Systemic corticosteroid use			
n (%)	79 (11.6)	25 (15.0)	
95% CI	0.09-0.14	0.10-0.21	0.23
Chemotherapy in 3 months pre-op			
n (%)	46 (6.7)	12 (7.2)	
95% CI	0.05-0.09	0.04-0.12	0.83
Radiotherapy in 3 months pre-op			
n (%)	12 (1.8)	3 (1.8)	
95% CI	0.01-0.03	0.00-0.05	0.97
Ascites present			
n (%)	16 (2.3)	10 (6.0)	
95% CI	0.01-0.04	0.03-0.11	0.014
Infection (non-SSI) at intake			
n (%)	75 (11.0)	14 (8.4)	
95% CI	0.09-0.14	0.05-0.14	0.33
Alcohol use			
n (%)	311 (47.1)	70 (45.5)	
95% CI	0.43-0.51	0.37-0.54	0.71
Alcohol quantity, units per week, mean, (95% CI)	4.5 (3.8-5.1)	5.2 (3.5-6.9)	0.34

	Without SSI (N=684; 80.4%)	With SSI (N=167; 19.6%)	p Value
ASA score, n (%)			0.29
ASA 1	76 (11.1)	15 (9.0)	
ASA 2	309 (45.2)	80 (47.9)	
ASA 3	280 (40.9)	67 (40.1)	
ASA 4	19 (2.8)	4 (2.4)	
ASA 5	0 (0.0)	1 (0.6)	
BMI, n (%)			0.65
Underweight	19 (2.8)	6 (3.6)	
Normal	317 (46.3)	67 (40.1)	
Overweight	220 (32.2)	59 (35.3)	
Class 1 obesity	80 (11.7)	20 (12.0)	
Class 2 obesity	20 (2.9)	8 (4.8)	
Class 3 obesity	8 (1.2)	2 (1.2)	
Outcomes			
Length of stay, d, mean, (95% CI)	15.0 (13.8-16.2)	24.9 (22.0-27.8)	<0.001
30 day mortality			
n (%)	16 (2.3)	9 (5.4)	
95% CI	0.01-0.04	0.02-0.10	0.036
In-hospital mortality			
n (%)	25 (3.7)	15 (9.0)	
95% CI	0.02-0.05	0.05-0.14	0.004

aPTT, activated partial thromboplastin time; ASA, American Society of Anesthesiologists; PT, prothrombin time; WBC, white blood cell

Table 2

Prevalence of Wound Symptoms and Other Serial Features in 5 Days Prior, 1 Day Prior, and Day of SSI

	Days -5 to -1 (cumulative/maximum)†			Day -1 (only)			Day 0 (only)		
	No SSI	SSI	p Value	No SSI	SSI	p Value	No SSI	SSI	p Value
Wound symptoms, % abnormal*									
Granulation amount	12.5	41.8	<0.001	7.1	35.0	<0.001	6.2	62.7	<0.001
Exudate amount	52.2	77.6	<0.001	19.5	56.5	<0.001	15.1	77.7	<0.001
Slough amount	9.3	30.3	<0.001	5.0	24.5	<0.001	5.2	44.3	<0.001
Edge distance	20.3	47.1	<0.001	12.2	38.7	<0.001	9.8	69.7	<0.001
Odor	1.5	4.2	0.022	0.2	2.5	<0.001	0.5	7.1	<0.001
Exudate type	38.9	66.7	<0.001	13.7	46.0	<0.001	12.2	70.8	<0.001
Slough type	9.4	30.9	<0.001	5.3	25.2	<0.001	5.2	44.3	<0.001
Wound edge color	95.4	93.9	0.410	84.6	88.9	0.170	83.2	89.5	0.048
Induration amount	94.6	93.9	0.750	86.3	79.1	0.022	86.5	73.5	<0.001
Other serial features									
NG tube, %	24.6	56.6	<0.001	9.2	26.8	<0.001	6.4	21.2	<0.001
Wound culture, %	0.60	7.8	<0.001	0.10	4.4	<0.001	0.80	23.8	<0.001
Heart rate, bpm, mean	92.6	102.5	<0.001	82.2	90.1	<0.001	81.9	89.7	<0.001
Temp. of body minus wound, °C, mean	6.8	7.3	0.014	5.8	6.3	0.004	5.9	6.4	0.005
Wound length, cm, mean	23.0	24.8	0.001	22.0	24.0	<0.001	22.2	24.1	<0.001
Tympanic temp., °C, mean	37.6	37.9	<0.001	37.2	37.3	0.003	37.1	37.4	<0.001
Diastolic BP, mmHg, mean	84.9	81.8	0.005	76.6	73.6	0.008	76	72.7	0.008
Pain rating (0-100), mean	32.6	33.8	0.590	18.5	19.4	0.640	15.3	24.7	<0.001

*
Symptoms were considered abnormal if they had a score >0 (see Supplemental Table 1)

†
For wound symptoms and other serial features with binary values, symptom was considered present if it was noted on any of the 5 days leading up to infection; for other serial features with continuous values, the maximum value over the 5 days was averaged for each group (No SSI, SSI).

BP, blood pressure; NG, nasogastric; SSI, surgical site infection

Table 3

Classifier Performance

Classifier	Correct %		Kappa		AUC	
	Training	Testing	Training	Testing	Training	Testing
BF	75.1	72.5	0.158	0.133	0.670	0.634 [†]
SF	81.7	81.0	0.397	0.340	0.760 [*]	0.741 [†]
BF+SF	79.8	81.3	0.351	0.354	0.759 [*]	0.749 [†]
SF (simplified)	82.7	81.3	0.334	0.268	0.752 [*]	0.716 [†]

[*]
BF vs SF, BF vs BF+SF, BF vs SF (simplified); SF vs BF+SF: ns; SF vs SF (simplified): p<0.0001; SF vs SF (simplified): ns

[†]
Training vs testing sets (all): ns

AUC, area under the ROC curve; BF, baseline features/baseline risk factors; SF, serial features

Table 4

Performance Characteristics of SF Classifiers with Varying Cutoff Values

Goals of use	Full SF model				Simplified SF model			
	PPV	NPV	Sens	Spec	PPV	NPV	Sens	Spec
Higher specificity	0.53	0.87	0.42	0.91	0.53	0.87	0.42	0.91
Balanced	0.43	0.91	0.69	0.78	0.42	0.91	0.66	0.78
Higher sensitivity	0.35	0.93	0.80	0.64	0.33	0.92	0.75	0.64

NPV, negative predictive value; PPV, positive predictive value; Sens, sensitivity; SF, serial features; Spec, specificity

Table 5

Features Selected for Final Complex SF Model and Simplified SF Model, In Order of Decreasing Predictive Importance

Original data element from which feature was derived	Lookback period (days)	Transformation type	Odds ratio
Complex SF model (Naïve Bayes)			
Granulation score	2	Mean value	
Exudate amount score	3	Maximum value	
Nasogastric tube presence	2	Maximum value	
Granulation score	5	Maximum value	
Nasogastric tube presence	5	Maximum value	
Heart rate	3	Maximum value	
Heart rate	4	Daily change	
Temperature of wound minus skin	5	Deviation from trend	
Wound length	2	Maximum value	
Wound culture ordered	5	Maximum value	
Body temperature	5	Maximum value	
Diastolic blood pressure	2	Raw value	
Simplified SF model (Logistic regression)			
Nasogastric tube presence	2	Maximum value	1.91 (binary)
Exudate amount score	2	Maximum value	1.24 (per score increment)
Heart rate	2	Maximum value	1.18 (per 10 bpm)
Slough type score	2	Maximum value	1.18 (per score increment)
Wound length	2	Maximum value	1.02 (per cm)

SF, serial features