International Journal of Medical Informatics 129 (2019) 234–241

Contents lists available at ScienceDirect

International Journal of Medical Informatics

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

Mining patient-specific and contextual data with machine learning technologies to predict cancellation of children's surgery

Lei Liu[a,b], Yizhao Ni[a,b], Nanhua Zhang[b,c], J. "Nick" Pratap[b,d,*]

[a] Department of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA
[b] College of Medicine, University of Cincinnati, Cincinnati, OH, USA
[c] Division of Biostatistics and Epidemiology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA
[d] Department of Anesthesia, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA

ARTICLE INFO

Keywords:
Pediatric surgery cancellation
Quality improvement
Predictive modeling
Machine learning

ABSTRACT

Background: Last-minute surgery cancellation represents a major wastage of resources and can cause significant inconvenience to patients. Our objectives in this study were: 1) To develop predictive models of last-minute surgery cancellation, utilizing machine learning technologies, from patient-specific and contextual data from two distinct pediatric surgical sites of a single institution; and 2) to identify specific key predictors that impact children's risk of day-of-surgery cancellation.

Methods and findings: We extracted five-year datasets (2012–2017) from the Electronic Health Record at Cincinnati Children's Hospital Medical Center. By leveraging patient-specific information and contextual data, machine learning classifiers were developed to predict all patient-related cancellations and the most frequent four cancellation causes individually (patient illness, "no show," NPO violation and refusal to undergo surgery by either patient or family). Model performance was evaluated by the area under the receiver operating characteristic curve (AUC) using ten-fold cross-validation. The best performance for predicting all-cause surgery cancellation was generated by gradient-boosted logistic regression models, with AUC 0.781 (95% CI: [0.764,0.797]) and 0.740 (95% CI: [0.726,0.771]) for the two campuses. Of the four most frequent individual causes of cancellation, "no show" and NPO violation were predicted better than patient illness or patient/family refusal. Models showed good cross-campus generalizability (AUC: 0.725/0.735, when training on one site and testing on the other). To synthesize a human-oriented conceptualization of pediatric surgery cancellation, an iterative step-forward approach was applied to identify key predictors which may inform the design of future preventive interventions.

Conclusions: Our study demonstrated the capacity of machine learning models for predicting pediatric patients at risk of last-minute surgery cancellation and providing useful insight into root causes of cancellation. The approach offers the promise of targeted interventions to significantly decrease both healthcare costs and also families' negative experiences.

1. Introduction

When surgery is cancelled at the last minute, families frequently suffer psychological stress and financial hardships. Moreover, from the perspective of health care providers, cancellation leaves unutilized resources valued as high as $1 per second (approximately $100,000 per week at our large, tertiary children's hospital) [1].

Although last-minute cancellation is infrequent, with rates from 2% to 20% reported, over 50 million procedures are performed annually in American hospitals [2], so the absolute number of cancelled cases is high. As such, surgical cancellation figures as a leading source of perioperative wastage. The cost of unused staff and facilities due to last-minute surgery cancellations forms a substantial contribution to the cost of surgical care. Furthermore, the negative impact on patients and families of last-minute cancellation is substantial. For example, Tait et al. found that the average round-trip for cancelled children was 160 miles to their children's hospital in Michigan [3]. In their study, more than one-third of family members missed a day of work, which was unpaid in half of cases. Many parents and children expressed disappointment, frustration and even anger. Avoidable last-minute

[*] Corresponding author at: Cincinnati Children's Hospital Medical Center, Department of Anesthesia, 3333 Burnet Avenue, MLC 7024, Cincinnati, OH 45229-3039, USA.

E-mail address: jayant.pratap@cchmc.org (J. "Nick" Pratap).

https://doi.org/10.1016/j.ijmedinf.2019.06.007
Received 31 August 2018; Received in revised form 13 May 2019; Accepted 6 June 2019

cancellation, therefore, leads to a poor patient and family experience.

Reasons for cancellation include acute patient illness, inadequate work-up of co-existing chronic medical conditions, failure to comply with eating/drinking instructions (“NPO violation”), and failure to attend for surgery (“no show”). Our institution has already reduced cancellation by one-sixth using cheap “across-the-board” interventions, including clearer pre-operative instructions and text-message reminders [1].

Our earlier work has elucidated a range of psychosocial factors underlying last-minute surgery cancellation and raises the possibility of developing interventions to support families struggling to prepare their children for surgery [4]. To ensure efficient use of resources, we aim to target additional support to patients and families at greatest risk of cancellation. However, we have been hampered by our inability to predict last-minute surgery cancellations. To date, the literature on last-minute surgical cancellation has been entirely descriptive with classical statistical techniques applied to demonstrate association. For example, Schuster et al. [5] (including review of previous studies) found higher cancellation rates in German university hospitals than in community hospitals, and also more cancellations in general surgery than other services. Despite these findings, no consistent single factor or group of factors has been identified that would allow a definite determination of surgery cancellation, which is perhaps unsurprising given the wide variety of patients, types of surgery and of institutions offering surgical care [6–11]. Consequently, development of a more sophisticated approach to predicting surgery cancellation promises great potential benefits for clinical practice.

Machine learning, as a field of computer science, utilizes computerized algorithms to identify hidden patterns within datasets that are useful for prediction. By learning from a set of training data, machine learning algorithms construct a predictive model to make data-driven predictions on unseen examples (test data). Machine learning has been widely utilized in clinical decision support, such as detecting patient clinical status and identifying signs and symptoms of specific diseases [12–14].

1.1. Objective

Our overall objective was to develop a system to predict last-minute cancellation of children’s surgery, with a long-term goal of implementing such a system to facilitate targeting of future quality improvement efforts. The rationale for the current study is that machine learning uncovers patterns in historical data to identify subtle predictors, and captures relationships among many factors to allow assessment of risk associated with a particular cause of surgery cancellation. As such, machine learning-based models offer the potential both to predict cancellation and to understand its antecedents. Such knowledge offers the promise of interventions to significantly decrease both healthcare costs and also families’ negative experiences associated with last-minute cancellation of surgical procedures.

This study aimed to develop machine learning models for predicting surgery cancellation from patient-specific and contextual data from two distinct pediatric surgical sites of a single institution, together providing the surgical capacity for children in a large Midwest conurbation in the United States, and to identify specific key predictors. Our study is the first, known to us, to investigate automated prediction of surgery cancellation based on large and detailed surgery cancellation datasets drawing from the electronic health record and publicly available contextual data.

2. Materials and methods

2.1. Data

Five-year datasets (2012–2017) were extracted from the Electronic Health Record (EHR) at Cincinnati Children’s Hospital Medical Center (CCHMC) to reflect surgical activities at the institution’s two distinct campuses (Burnet and Liberty). The study was approved by the CCHMC institutional review boards and a waiver of individual consent was authorized.

Burnet campus is the main campus of CCHMC and is located close to downtown Cincinnati. Approximately 17,000 scheduled surgeries are performed each year at the Burnet campus. Most complex surgeries are performed at the Burnet site, and medically complex children are generally booked for procedures here. Liberty campus is located in Liberty Township, 15 miles to the north of Cincinnati. Approximately 9,000 surgeries are performed annually at the Liberty campus, and most of the procedures are short and less complex surgeries on children who are typically not medically complex. Most patients at the Liberty campus live in the Cincinnati area, but many patients at the Burnet campus travel from across the Midwest, further afield in the United States or beyond.

For all surgical activities at CCHMC, cancellations have been comprehensively adjudicated to one of ten codes, thus allowing prediction for specific causes. To capture patient-specific information for a surgical activity, we extracted variables including patient demographic data, recent health care use, patient insurance information, schedule-related factors, prior cancellation behaviors, and information gleaned from a routine pre-operative phone call with a nurse. From an institutional database of more than 330,000 dated laboratory results, we developed a measure of locally circulating pathogens for febrile, respiratory and gastrointestinal diseases, which served as a potential predictor of patient illness. Finally, we extracted complete daily weather records for the nearby Cincinnati/Northern Kentucky International Airport [15], to investigate any effect of weather conditions on last-minute surgery cancellation. A summary of variables utilized in this study is presented in Table 1. A description of each variable is presented in Table S.1 (Appendix).

Table 1
Summary of the variables.

Category	Number of variables	Description	Data source
Demographics	5	Patient age, sex, race, ethnicity, distance to CCHMC from home	Institutional EHR
Insurance info	2	Payer, payer type	Institutional EHR
Pre-op phone call	5	Number of call attempts,’ live’ contact reached, first and final contacts, history & physical completed	Institutional EHR
Recent health care use	7	Number of medications taken as outpatient before surgery, recent ER attendance (4 timepoints), office visits, hospitalizations in 6 months	Institutional EHR
Prior cancellation behaviors	5	Numbers of previous cancellations, previous” no shows,” previous other cancellations, clinic “no shows,” previous surgeries	Institutional EHR
Surgery related factors	9	Hour, day of week and month of surgery, lead time, “work in” case, surgical specialty, estimated case length, post-op disposition, time since original QI project	Institutional EHR
Infection risk	1	Local circulating load of respiratory, gastrointestinal and other febrile pathogens	Institutional EHR
Weather	24	Detailed daily weather records from nearby airport (NOAA)	Public Resource

2.2. Feature extraction from the data

All categorical variables (e.g., sex and insurance payer type) were converted to binary features using zero and one to indicate absence and presence, respectively [16]. For example, if a variable had five categories, it was expanded to four features to avoid linear dependencies induced between the features.

Age was categorized into seven distinct features (0–27 days: neonatal; 28 days to 12 months: infancy; 13–24 months: toddler; 2–5 years: early childhood; 6–11 years: middle childhood; 12–18 years: early adolescence; 19–21 years: late adolescence) [17]. All home locations were geocoded with an in-house geographic information system to ensure that no protected health information was sent outside the institution (90.2% to city block level, and a further 6.8% to street level). Given that missing data for certain variables represented an important feature in this dataset (e.g., responses to the nurse's questions, in the case that the nurse was unable to reach the family for the routine preoperative call), a unique category was created to represent "unknown" or "NA" for categorical variables, while all missing values for numeric variables were replaced by appropriate values (e.g., 0). Finally, all rescheduled cases (1,354 surgeries, 1.6%) were identified and excluded from the training data to avoid diluting the effects of cancellation predictors by subsequently completed surgeries.

2.3. Machine learning classifiers

In this study we sought to predict *all patient-related cancellations* (denoted by "all causes") and the most frequent four cancellation causes *individually*: patient illness, "no show," NPO violation and refusal to undergo surgery by either patient or family. We modeled prediction of last-minute surgery cancellation as a supervised classification problem and utilized a representative set of machine learning classifiers, including naïve Bayes, multivariate logistic regression (LR) with L1 and L2 normalization, support vector machines with polynomial (SVM-P) and radial basis function (SVM-R) kernels, decision trees (both CART and C5.0), random forests (RF), gradient boosted LR (GBL), artificial neural networks (aNNs) and a deep learning algorithm (TensorFlow) [18–26]. Classifiers such as RF, SVM, GBL, aNNs and deep learning were known for their high predictive performance for nonlinear problems and ability to find complex interactions among features, while the other classifiers were chosen for their better interpretability that is important for model understanding and interpretation. All machine learning classifiers used were implemented within the R programming language or MATLAB [27,28].

2.4. Feature selection

To facilitate a human-oriented conceptualization of surgery cancellation, we applied an iterative step-forward approach with "best first" search on the training sets to identify key predictors [29]. In each iteration, the feature generating the greatest increase in cross-validation performance was added to an L1-normalized LR model. We determined the optimal feature set as the point at which additional features did not increase the performance. The top five most important features were extracted for presentation.

2.5. Dealing with imbalanced data

In view of the relative rarity of the outcome of interest (cancellation), which could negatively impact the performance of machine learning algorithms, we tested both up- and down-sampling and also the synthetic minority oversampling technique (SMOTE) to address this issue [30,31]. The SMOTE algorithm oversamples the minority class (i.e., cancelled cases) by creating "synthetic" examples in the training data. The balanced data were then used to train the machine learning algorithms. To identify the best sampling approach for each machine learning classifier independently, the original and all resampled variants were tested for each classifier and selected for subsequent use according to cross-validation performance.

2.6. Experiment setup

We performed stratified random sampling to divide each dataset into two: 70% for training and 30% for performance evaluation and error analysis. Ten-fold cross-validation was applied on the training set to tune hyper-parameters of the machine learning classifiers with grid search parameterization, including: (1) cost parameters for LR, SVM-P, SVM-R and aNNs (screened from 10^{-6} to 10^{6}); (2) optimal degree for SVM-P (screened from 1 to 6); (3) parameter γ for SVM-R (screened from 2^{-15} to 2^{5}); (4) minimum number of observations in a node (3, 5, 10, 15 and 20) and the complexity parameters (screened from 10^{-6} to 10^{-1}, 0.3, 0.5 and 0.8) for CART; (5) the number of boosting iterations for C5.0 (screened at 5 increments from 1 to 20); (6) number of trees for RF (screened from 2^{2} to 2^{11}); (7) total number of trees (300, 500, 1000, 1500, 2000 and 2500), the maximum depth of variable interactions (1, 3, 5 and 10) and the minimum number of observations in the trees' terminal nodes (5 and 10) for GBL; (8) number of neurons for aNNs (screened at 20 increments from 10 to 100); and (9) dimensionality of output space (10, 30, 50, 100) and batch size (screened from 2^{3} to 2^{7}) for deep learning. The machine learning classifiers with optimal parameters were then applied on the test data for performance evaluation and comparison.

To assess the generalizability of our models, optimized L1- and L2-normalized LR and GBL models generated from the Burnet campus dataset were applied to make predictions on the full Liberty campus dataset and vice versa.

2.7. Evaluation metrics

We evaluated model performance by six evaluation metrics: 1) Accuracy = (True positives + True negatives)/Total cancellations; 2) Precision = True positives/(True positives + False positives); 3) Recall = True positives/(True positives + False negatives); 4) Specificity = True negatives/(True negatives + False positives); 5) Negative predictive value = True negatives/(True negatives + False negatives); and 6) AUC that measures the balance between recall and specificity [32–34]. AUC was used as the primary measure for selecting the best-performing classifiers.

2.8. Error analysis

The Local Interpretable Model-Agnostic Explanations (LIME) approach was applied to the best-performing machine learning classifier, with optimal parameters, for error analysis [35]. The LIME algorithm explains each classifier prediction by developing a linear model locally around the prediction to identify interpretable features. To understand why the predictive model would predict some cancelled observations as completed and vice versa, the LIME algorithm was applied on false positive and false negative cases generated by the optimized machine learning classifier (GBL model). For each false positive/false negative case, the top 10 features selected by the highest weights method were presented for error interpretation.

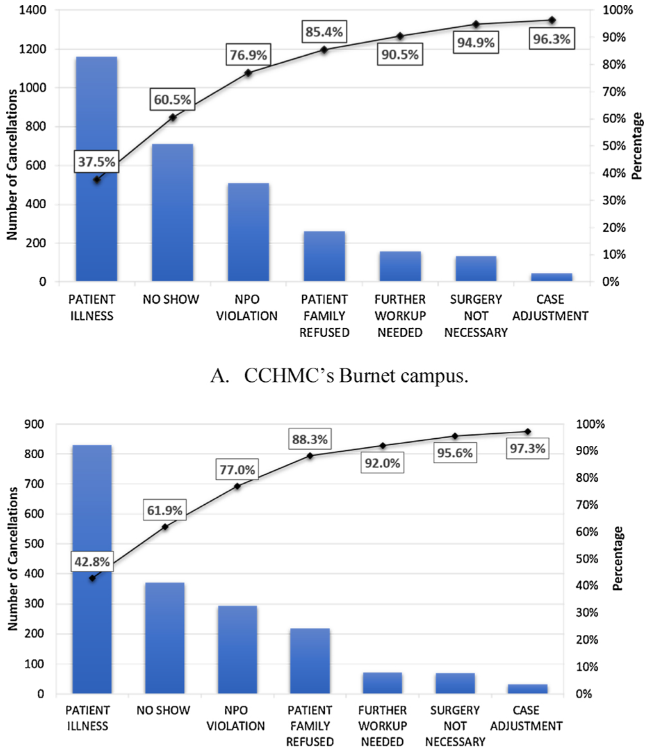
3. Results

3.1. Details of the datasets

Summary statistics of the two datasets utilized in our study are presented in Table 2. To avoid the negative impact of rescheduled activities on model training, 1,354 and 730 examples were excluded from Burnet and Liberty datasets, respectively, as described above. After performing pre-processing and stratified random sampling, 58,301

Table 2
Statistics of two datasets used in the study.

	Number of surgeries	Number of cancellations	Number of rescheduled activities
Burnet campus	84615	3088 (3.6%)	1354 (1.6%)
Liberty campus	43162	1940 (4.5%)	730 (1.69%)



A. CCHMC's Burnet campus.

B. CCHMC's Liberty campus.

Fig. 1. Cumulative distribution of cancellation reasons. A) CCHMC's Burnet campus; B) CCHMC's Liberty campus.

examples with 58 variables were present in the Burnet training set. These examples were utilized in experiments with ten-fold cross-validation to tune hyper-parameters of different machine learning classifiers. The Burnet test set consists of 24,960 cases. As for the Liberty dataset, 29,729 examples were in the training set, while 12,703 examples were used for performance evaluation and error analysis. Fig. 1 illustrates the cumulative distribution of cancellation reasons. The top four most frequent cancellation reasons were patient illness, "no show", NPO violation, and patient/family refusal for both campuses. These cancellation causes accounted for over 85% of all last-minute cancelled surgeries.

3.2. Model comparison

The performance of different classifiers for predicting surgery cancellation is presented in Tables 3a and 3b. The highest AUCs were generated by the GBL models, with 0.781 (95% CI: [0.764,0.797]) and 0.740 (95% CI: [0.726,0.771]) on the training sets for Burnet and Liberty campuses respectively. L1-normalized LR was the second best-performing classifier, yielding AUCs of 0.770 (Burnet campus, 95% CI: [0.755,0.785]) and 0.742 (Liberty campus, 95% CI: [0.721,0.763]) on the training samples. RF achieved comparable performance with AUCs of 0.783 for Burnet and 0.745 for Liberty on test data but the AUCs were lower for individual causes of cancellation. All classifiers achieved higher AUCs for prediction of "no show" and NPO violation cancellations compared with the other two specific causes.

3.3. Feature selection

Fig. 2 shows the change in classifier performance, for all patient-related cancellation reasons and individual causes, as features generating the greatest increment were added iteratively. For all models and both datasets, the top five variables, highlighted in the figure, yielded more than 95% of performance gain in feature selection, supporting high relative importance in predicting last-minute cancellations.

3.4. Generalizability

To assess their generalizability, the L1- and L2-normalized LR and GBL models optimized on the Burnet campus data were applied and evaluated on the Liberty dataset, and vice versa. The evaluation performances are presented in Table 4.

3.5. Error analysis

Using the standard probability threshold of 0.5, the optimized all-cause GBL classifier made 32 false-positive and 764 false-negative predictions of cancellation on the Burnet campus dataset. For the purposes of error interpretation and visualization, 50 false-negative cases were randomly selected. An overview of the error interpretation generated by the LIME algorithm is displayed in Fig. 3 with the horizontal-axis representing the false-positive/false-negative cases and the vertical-axis the top features selected, with predicted labels shown at the top of the figure. The color of each cell in the figure indicates local importance of the selected features, with green representing positive weights supporting the predicted label and red representing negative weights. As shown in Fig. 3, "payer name", "number of medications" and "call attempts more than two" are important features across almost all false-positive cases. For the false-negative cases, clinic "no shows" in last six months and "number of call attempts" are the most important features to explain why the model predicted these cancelled cases as completed.

4. Discussion

Our results indicate that machine learning techniques, using primarily EHR-derived data, predict all-cause surgery cancellation at both campuses with AUC up to 0.78. Logistic regression models, particularly a gradient-boosted variant, proved most powerful. Despite differences in clinical workload and local population characteristics between the two campuses, cross-trained models performed almost as well as models both trained and evaluated on data from the same campus. Of the four most frequent individual causes of cancellation, "no show" and NPO violation were predicted better than patient illness and patient/family refusal by the machine learning models.

In addition, machine learning was helpful in identifying predictors of all-cause cancellations and in differentiating its most frequent individual causes. Patient age and the identity of the healthcare payer predict patient-illness cancellation most strongly; the time from the start of the intervention program is also important, which may reflect effectiveness of prior quality improvement efforts [1]. Time of year and the circulating pathogen load are also influential. At the Liberty satellite campus, where mostly shorter and most straightforward surgeries on generally healthy children were performed, individual surgical services differ importantly in terms of cancellation risk. For "no show" cancellation, payer is of prime importance, likely reflecting families' socioeconomic status. Also, if, at the time of the pre-operative telephone consultation (two working days before the scheduled procedure), the state-mandated pre-operative history and physical examination has not been completed by the patient's primary care physician, or if the

Table 3a

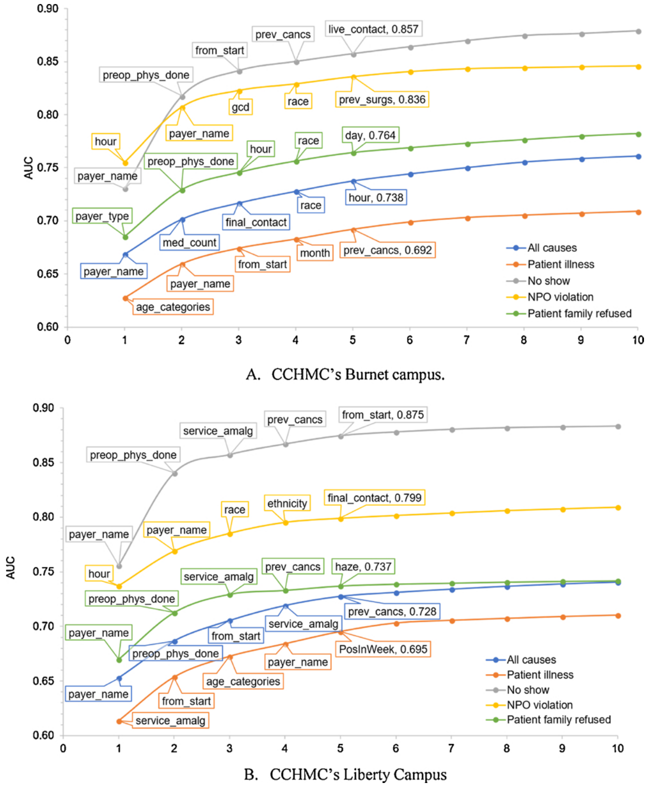
Performance of different machine learning classifiers for the Burnet campus dataset.

	Ten-fold Cross Validation Performance					Test Set Performance				
Classifier	All-cause	Patient illness	No show	NPO violation	Patient family refused	All-cause	Patient illness	No show	NPO violation	Patient family refused
Naïve Bayes	0.691	0.640	0.755	0.714	0.655	0.711	0.645	0.761	0.744	0.637
LR + L1	0.770	0.715	0.876	0.840	0.751	0.787	0.725	0.898	0.842	0.732
LR + L2	0.770	0.712	0.874	0.831	0.751	0.787	0.724	0.891	0.832	0.724
SVM-P	0.735	0.673	0.847	0.817	0.732	0.730	0.644	0.838	0.751	0.704
SVM-R	0.731	0.672	0.840	0.806	0.730	0.685	0.607	0.818	0.706	0.686
Decision Tree	0.699	0.627	0.805	0.708	0.686	0.719	0.661	0.820	0.739	0.584
C5.0	0.706	0.625	0.805	0.758	0.641	0.721	0.618	0.839	0.757	0.712
RF	0.769	0.713	0.876	0.826	0.760	0.783	0.712	0.893	0.815	0.736
GBL	0.781	0.725	0.880	0.826	0.775	0.793	0.725	0.898	0.828	0.726
aNN	0.710	0.650	0.833	0.805	0.725	0.655	0.562	0.758	0.716	0.658
DNN	0.760	0.702	0.844	0.789	0.697	0.771	0.706	0.866	0.797	0.702

family cannot be contacted to ascertain this, “no show” cancellation is much more likely. The time from the start of our quality improvement work [1] and prior surgery cancellation behaviors are also salient to “no show.” NPO violation is more likely when surgery is scheduled later in the day. Specific payers and patient race are also important to NPO violation, suggesting socioeconomic disadvantage. Differences in key predictors between the two campuses (Fig. 2A vs. B) are also noteworthy. We speculate that they reflect differences in the patient mix. For example, the importance of the number of regular medications taken by the patient may reflect the more medically complex children managed at CCHMC’s main Burnet campus. Likewise, patient race perhaps represents a mixed group of patients including a substantial socioeconomically deprived African-American community located near the main campus, as compared to a more homogenously affluent, predominantly Caucasian population in the suburbs surrounding the Liberty satellite campus [36]. Such distinctions suggest that approaches to reducing cancellations would need to be tailored to the patient mix at each campus.

In this study we describe, for the first time, the application of machine learning techniques to predict surgery cancellation. The most comprehensive previous study in the literature comprised only around 6,000 cases, and was limited to comparing gross cancellation rates between 25 different hospitals [5]. Our study is differentiated by analyzing a large-scale dataset as well as by offering insight into predictors of cancellation. A particular feature of surgery cancellation prediction, as a machine learning problem, is the marked class imbalance generated by the 3–5% cancellation rate, thereby creating a low-frequency class of interest. To deal with this challenge, we therefore utilized up- and down-sampling techniques, as well as SMOTE.

The promising performance achieved in this study suggests that our machine learning models offer potential for use in targeting interventions towards children and their families at elevated risk of surgery cancellation. In this way, more costly support can be focused efficiently



A. CCHMC’s Burnet campus.

B. CCHMC’s Liberty Campus

Fig. 2. Cross-validation performance of iterative step-forward feature selection (L1-normalized LR). A) CCHMC’s Burnet campus; B) CCHMC’s Liberty campus.

Table 3b

Performance of different machine learning classifiers for the Liberty campus dataset.

	Ten-fold Cross Validation Performance					Test Set Performance				
Classifier	All-cause	Patient illness	No show	NPO violation	Patient family refused	All-cause	Patient illness	No show	NPO violation	Patient family refused
Naïve Bayes	0.680	0.632	0.752	0.668	0.615	0.666	0.628	0.776	0.657	0.712
LR + L1	0.742	0.705	0.876	0.788	0.715	0.743	0.715	0.862	0.787	0.815
LR + L2	0.741	0.696	0.862	0.785	0.725	0.739	0.706	0.871	0.779	0.784
SVM-P	0.700	0.655	0.834	0.753	0.687	0.688	0.615	0.843	0.738	0.726
SVM-R	0.691	0.653	0.819	0.732	0.687	0.674	0.606	0.789	0.747	0.761
Decision Tree	0.661	0.646	0.796	0.653	0.652	0.675	0.673	0.785	0.706	0.732
C5.0	0.669	0.640	0.813	0.689	0.643	0.692	0.596	0.800	0.690	0.691
RF	0.742	0.704	0.874	0.753	0.736	0.745	0.686	0.850	0.757	0.823
GBL	0.749	0.711	0.877	0.783	0.737	0.754	0.707	0.860	0.740	0.822
aNN	0.682	0.591	0.584	0.528	0.675	0.692	0.547	0.575	0.520	0.664
DNN	0.721	0.664	0.826	0.705	0.667	0.729	0.671	0.829	0.715	0.740

Table 4
Performance of the cross-trained classifiers.

	Train on Liberty Data and Test on Burnet Data					Train on Burnet Data and Test on Liberty Data				
Classifier	All-cause	Patient illness	No show	NPO violation	Patient family refused	All-cause	Patient illness	No show	NPO violation	Patient family refused
LR + L1	0.725	0.688	0.842	0.797	0.755	0.735	0.684	0.872	0.830	0.725
LR + L2	0.724	0.685	0.838	0.786	0.758	0.728	0.673	0.848	0.820	0.714
GBL	0.723	0.653	0.848	0.790	0.729	0.728	0.663	0.865	0.810	0.707

towards those who are both in need and also most likely to benefit. Moreover, the specific predictors identified for individual cancellation causes may inform the design of interventions to prevent the appropriate failure modes, in conjunction with findings from our psychosocial research [4].

In view of the differences between local communities and institutional policies and cultures, a model trained on data from one hospital may show poor ability to predict cancellations at another. The similarities of AUCs for same-site and cross-trained models, however, support adequate generalizability between CCHMC's two surgical sites. Although both sites form part of the same institution, with many similarities in policies and culture, plus a proportion of health care professionals in common, differences in the patient population and procedures performed raise the possibility that certain predictors may have more widespread applicability. In any case, machine learning methodology is likely practicable at any hospital with an EHR system, using classifiers trained from a centralized dataset.

Our findings support the utility of machine learning approaches to investigating surgery cancellation. Moreover, related techniques may be relevant to the study of other problems in health care utilization, such as physician office visit cancellation or unscheduled re-admission after in-patient stays.

4.1. Error analysis, limitations and future work

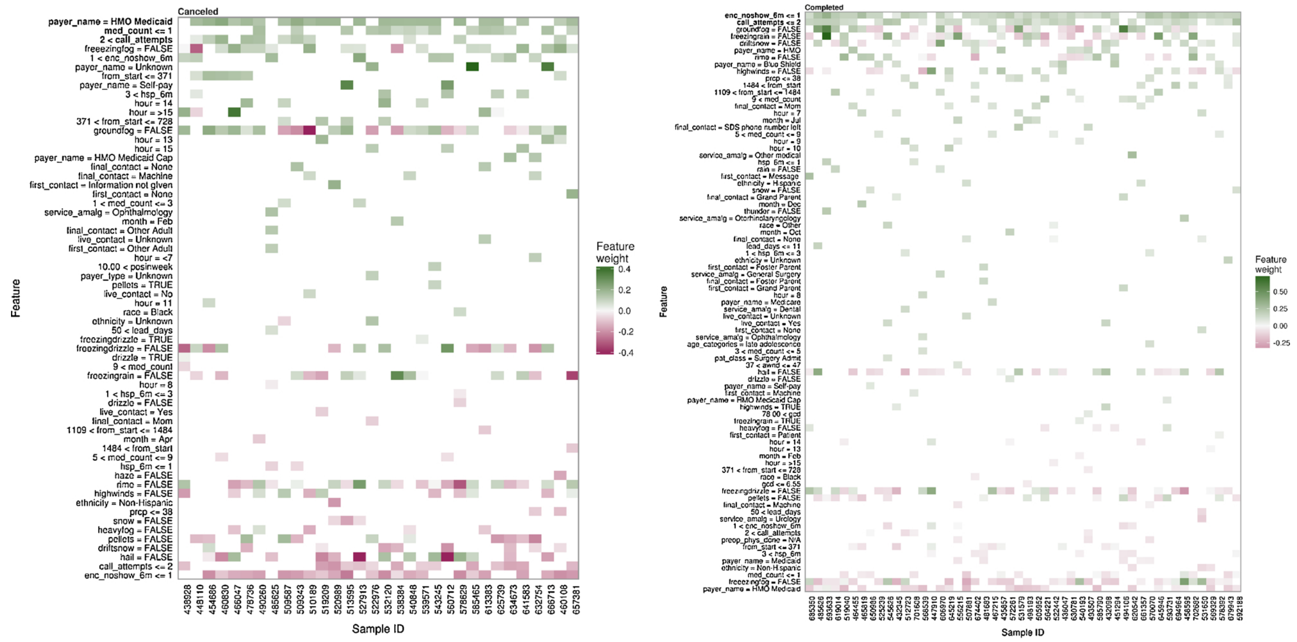
An error analysis was performed on predictions of all-cause cancellation made by the optimized GBL model for the Burnet campus. The results of error analysis and feature selection suggested that some key features in predicting surgery cancellation, including insurance payer, number of call attempts, number of outpatient medications, and clinic "no shows" in last six months, led to misclassifications. To alleviate this problem, in our future work we will develop advanced multi-layer classifiers to balance weights between different variable sets before aggregating them for predicting surgery cancellation [37].

In common with other database research, our results could potentially have been affected by errors or inadvertent omissions in the data. All cases had a valid patient identifier for tracking with just 35 (0.068%) excluded for missing data (18 invalid zip code, 14 duplicated case identifiers, 3 missing admission class). Moreover, the overwhelming majority (97%) were successfully geocoded to street level, at least. Therefore, likely reflecting its clinical and operational importance, our dataset is of very high quality.

Specific predictors of cancellation may change over time, particularly if quality improvement projects are effective in reducing the rate of cancellation. This is supported by our finding that cancellation becomes less likely with time in our dataset, which may coincide with our previously reported quality improvement work [1]. The machine learning approach to predict cancellation will likely however remain valid, and specific predictors and coefficients may be calibrated periodically.

As a final limitation, the work was limited to reporting system performance on a population collected in a single institution. Similar machine learning-based studies of surgery cancellation at other institutions, both adult and pediatric, would further establish the feasibility and utility of the approach.

In addition to an improved understanding of the etiology of surgery



A. False positive

B. False negative

Fig. 3. Heatmap visualization of error analysis for all-causes cancellation for the Burnet campus dataset. The LIME algorithm was applied on A) the false positives and B) 50 randomly selected false negatives generated by the optimized GBL classifiers.

cancellations gleaned from this study, we plan also to apply the best trained model to reduce both the number of surgery cancellations and their impact. The tool will enable us to pinpoint families in most need of support, in order to target resources to them efficiently. Also, by providing advance notice to operating room operations staff of slots most likely to be opened up by cancellation, we hope to facilitate better use of freed slots for add-on cases.

5. Conclusions

Our study demonstrated that machine learning models had capacity for predicting patients at risk of last-minute surgery cancellation, particularly "no show" and NPO violation. The models performed equally well at both campuses in our institution with the highest AUC for all-cause cancellation of 0.78. The feature selection process identified multiple predictors that uncovered useful insight into root causes of surgery cancellation. Performance of classifiers for all and specific causes supports the feasibility of operationally useful prediction of last-minute surgery cancellation. As such, work is in progress to integrate our predictive models into the institutional EHR system to facilitate rational targeting of quality improvement efforts towards patients and families at highest risk of cancellation. Through the early identification of surgery cancellation, timely interventions could be delivered to prevent cancellation in advance and to mitigate its effects, which has great potential to significantly decrease healthcare costs and cancellation-related negative patient and family experiences.

Author contributions

LL developed the algorithms, ran the experiments, analyzed the results, created the tables and figures, and wrote the manuscript. YN assisted with design of the study, coordinated the data extraction, provided suggestions in algorithm development, ran the experiments, analyzed the results and contributed to the manuscript. NZ provided suggestions in result analysis and contributed to the manuscript. JNP conceptualized the study, coordinated the data extraction, pre-processed the data, analyzed the results, provided suggestions in error analysis and wrote the manuscript. All authors read and approved the final manuscript.

Author declaration

We confirm that the manuscript has been read and approved by all named authors and that there are no other persons who satisfied the criteria for authorship but are not listed. We further confirm that the order of authors listed in the manuscript has been approved by all of us.

We confirm that we have given due consideration to the protection of intellectual property associated with this work and that there are no impediments to publication, including the timing of publication, with respect to intellectual property. In so doing we confirm that we have followed the regulations of our institutions concerning intellectual property.

We understand that the Corresponding Author is the sole contact for the Editorial process (including Editorial Manager and direct communications with the office). He/she is responsible for communicating with the other authors about progress, submissions of revisions and final approval of proofs. We confirm that we have provided a current, correct email address which is accessible by the Corresponding Author and which has been configured to accept email from ijmi@elsevier.com.

Summary table

What was already known on the topic:

- Last-minute surgery cancellation wastes resources and can lead to poor patient and family experience.
- No single factor or group of factors to predict surgery cancellation has been identified by existing studies.
- Machine learning technologies identify hidden patterns within datasets to construct predictive models and have been widely utilized in clinical decision support.

What this study added to our knowledge:

- Machine learning models have capacity for predicting patients at risk of last-minute surgery cancellation.
- By utilizing feature selection, this study identifies key predictors of cancellation which could inform the design of future preventive interventions.
- The machine learning models offer potential for timely interventions to significantly decrease healthcare costs and cancellation-related negative patient and family experiences.

;1;

Acknowledgments

This work was supported by the Agency for Healthcare Research and Quality [grant number: 1R21HS024983-01]. JNP, LL and YN were also supported by internal funds from Cincinnati Children's Hospital Medical Center. The authors thank John Rugg, William "Ray" Stone, Anthony Coleman, Wayne Geers and Vincent Evans for EHR data extraction.

References

[1] J.N. Pratap, A.M. Varughese, P. Mercurio, T. Lynch, T. Lonnemann, A. Ellis, et al., Reducing cancelations on the day of scheduled surgery at a children's hospital, Pediatrics 135 (5) (2015) e1292–9.

[2] Centers for Disease Control and Prevention, FastStats, 2013 (Accessed 20 July 2016) Available from: http://www.cdc.gov/nchs/fastats/inpatient-surgery.htm.

[3] A.R. Tait, T. Voepel-Lewis, H.M. Munro, H.B. Gutstein, P.I. Reynolds, Cancellation of pediatric outpatient surgery: economic and emotional implications for patients and their families, J. Clin. Anesth. 9 (3) (1997) 213–219.

[4] L.M. Vaughn, M. DeJonckheere, J.N. Pratap, Putting a face and context on pediatric surgery cancelations: the development of parent personas to guide equitable surgical care, J. Child Health Care 21 (1) (2017) 14–24.

[5] M. Schuster, C. Neumann, K. Neumann, J. Braun, G. Geldner, J. Martin, et al., The effect of hospital size and surgical service on case cancellation in elective surgery: results from a prospective multicenter study, Anesth. Analg. 113 (3) (2011) 578–585.

[6] R. Hand, P. Levin, A. Stanziola, The causes of cancelled elective surgery, Qual. Assur. Util. Rev. 5 (1) (1990) 2–6.

[7] M.J. Lacqua, J.T. Evans, Cancelled elective surgery: an evaluation, Am. Surg. 60 (11) (1994) 809–811.

[8] J.L. Argo, C.C. Vick, L.A. Graham, K.M. Itani, M.J. Bishop, M.T. Hawn, Elective surgical case cancellation in the Veterans Health Administration system: identifying areas for improvement, Am. J. Surg. 198 (5) (2009) 600–606.

[9] A.R. Seim, T. Fagerhaug, S.M. Ryen, P. Curran, O.D. Saether, H.O. Myhre, et al., Causes of cancellations on the day of surgery at two major university hospitals, Surg. Innov. 16 (2) (2009) 173–180.

[10] N. Al Talalwah, K.H. McIltrot, Cancellation of surgeries: integrative review, J. Perianesth. Nurs. 34 (1) (2019) 86–96.

[11] U. Caesar, J. Karlsson, L.E. Olsson, K. Samuelsson, E. Hansson-Olofsson, Incidence and root causes of cancellations for elective orthopaedic procedures: a single center experience of 17,625 consecutive cases, Patient Saf. Surg. 8 (1) (2014) 24.

[12] L. Deleger, H. Brodzinski, H. Zhai, Q. Li, T. Lingren, E.S. Kirkendall, et al., Developing and evaluating an automated appendicitis risk stratification algorithm for pediatric patients in the emergency department, J. Am. Med. Inform. Assoc. 20 (e2) (2013) e212-e20.

[13] T. Lingren, P. Chen, J. Bochenek, F. Doshi-Velez, P. Manning-Courtney, J. Bickel, et al., Electronic health record based algorithm to identify patients with autism spectrum disorder, PLoS One 11 (7) (2016) e0159621.

[14] H. Zhai, P. Brady, Q. Li, T. Lingren, Y. Ni, D.S. Wheeler, et al., Developing and evaluating a machine learning based algorithm to predict the need of pediatric intensive care unit transfer for newly hospitalized children, Resuscitation 85 (8) (2014) 1065–1071.

[15] National Oceanic and Atmospheric Administration, (2019) Available from: http://www.noaa.gov.

[16] D.B. Suits, Use of dummy variables in regression equations, J. Am. Stat. Assoc. 52 (280) (1957) 548–551.

[17] K. Williams, D. Thomson, I. Seto, D.G. Contopoulos-Ioannidis, J.P. Ioannidis, S. Curtis, et al., Standard 6: age groups for pediatric trials, Pediatrics 129 (Suppl. 3) (2012) S153-S60.

[18] I. Rish (Ed.), An Empirical Study of the Naive Bayes Classifier. IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence, IBM, 2001.
[19] J. Friedman, T. Hastie, R. Tibshirani, Regularization paths for generalized linear models via coordinate descent, J. Stat. Softw. 33 (1) (2010) 1.
[20] J. Shawe-Taylor, N. Cristianini, Kernel Methods for Pattern Analysis, Cambridge university press, 2004.
[21] J.R. Quinlan, Induction of decision trees, Mach. Learn. 1 (1) (1986) 81–106.
[22] J.R. Quinlan, C4. 5: Programs for Machine Learning, Elsevier, 2014.
[23] M. Kuhn, K. Johnson, Applied Predictive Modeling, Springer, 2013.
[24] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[25] T.M. Mitchell, Artificial neural networks, Mach. Learn. 45 (1997) 81–127.
[26] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436.
[27] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2013.
[28] MathWorks, Matlab – The language of technical computing, Available from: (2019) https://www.mathworks.com/products/matlab.html.
[29] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (March) (2003) 1157–1182.
[30] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2002) 321–357.
[31] Addressing the curse of imbalanced data sets: one sided sampling, in: M. Kubat, S. Matwin (Eds.), Proceedings of the Fourteenth International Conference on Machine Learning, 1997.
[32] D.G. Altman, J.M. Bland, Diagnostic tests. 1: sensitivity and specificity, BMJ 308 (6943) (1994) 1552.
[33] D.G. Altman, J.M. Bland, Statistics Notes: diagnostic tests 2: predictive values, BMJ 309 (6947) (1994) 102.
[34] A.P. Bradley, The use of the area under the ROC curve in the evaluation of machine learning algorithms, Pattern Recognit. 30 (7) (1997) 1145–1159.
[35] Why should I trust you? Explaining the predictions of any classifier, in: M.T. Ribeiro, S. Singh, C. Guestrin (Eds.), Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2016).
[36] A.F. Beck, T. Moncrief, B. Huang, J.M. Simmons, H. Sauers, C. Chen, et al., Inequalities in neighborhood child asthma admission rates and underlying community characteristics in one US county, J. Pediatr. 163 (2) (2013) 574–580.
[37] Mining a large-scale EHR with machine learning methods to predict all-cause 30-day unplanned readmissions, in: H. Zhai, S. Iyer, Y. Ni, T. Lingren, E. Kirkendall, Q. Li (Eds.), Proc of the 2nd ASE International Conference on Big Data Science and Computing, 2014.