Machine Learning Prediction of Surgical Intervention for Small Bowel Obstruction

Miles Turpin[1], Joshua Watson, MD*[2], Matthew Engelhard, MD, PhD[3], Ricardo Henao, PhD[4], David Thompson, MD[2], Lawrence Carin, PhD[1], Allan Kirk, MD, PhD[2]

*Corresponding Author
[1]Department of Electrical and Computer Engineering, Duke University
[2]Department of Surgery, Duke University School of Medicine
[3]Department of Psychiatry and Behavioral Sciences, Duke University School of Medicine
[4]Department of Biostatistics and Bioinformatics, Duke University School of Medicine

Corresponding Author:

Joshua Watson, MD
Department of Surgery
Duke University School of Medicine
Durham, NC, 27705
joshua.watson@duke.edu
Phone: 919-684-8111
Fax: 919-681-7905

Manuscript Details:

Title length: 80 characters
Abstract length: 324 words
Body length: 2997 words

4 Figures, 1 Table
2 eFigures, 6 eTables

ABSTRACT

Importance: Small bowel obstruction (SBO) results in over 350,000 operations and over $2 billion in health care expenditures each year in the US. Prompt, effective identification of patients at high and/or low risk of requiring surgery could improve survival, lower incidence of local systemic complications, and shorten the average length of hospitalization.

Objective: To develop a machine learning model that continuously and effectively predicts the risk of requiring surgery among patients admitted for SBO.

Design, Setting, and Participants: A prediction model was developed in 2018-2019 based on retrospective analysis of SBO-related encounters taking place in the Duke University Health System between 2013 and 2017.

Main Outcomes and Measures: Performance was assessed in each hour after admission when predicting whether the patient will (a) receive surgery in the next 24 hours, and (b) receive surgery at any time during the encounter. Measures of performance included the area under the receiver operating characteristic curve (AUROC) and concordance index (C-index). Measures of effectiveness for discharging patients included the incorrect discharge rate and average reduction in hospital stay.

Results: A total 3,910 encounters among 3,374 unique patients were identified and used in model development. Model-based discharge of low-risk patients was projected to reduce the average length of stay among patients not receiving surgery by over 60 hours while maintaining an incorrect discharge rate lower than the observed readmission rate (9.3%). AUROC for the 24-hour prediction task increased from 0.644 to 0.779 at 12 and 72 hours post-admission, respectively. Concordance index increased from 0.639 to 0.729 at 12 and 72 hours post-admission, respectively.

Conclusions and Relevance: A machine learning model can effectively and continuously stratify SBO patients by their risk of requiring surgery. This approach, which we show quantitatively to reduce the average length of stay, could be used to improve prioritization of operating room resources by discharging patients whose risk is low. Further study is needed to prospectively explore the benefits of model deployment in an inpatient setting.

KEY POINTS

Question: Can a machine learning model effectively stratify patients by their risk of requiring surgery for small bowel obstruction?

Findings: After learning from 3,910 hospital encounters, the model selected low-risk patients for discharge an average of 60 hours before their observed discharge time with a lower rate of incorrect discharge compared to clinicians. It also effectively stratified patients by risk and predicted whether surgery would be required in the next 24 hours (CI=0.729, AUROC=0.779 at 72 hours after admission).

Meaning: This model will assist in risk stratifying patients admitted for SBOs and help clinical teams identify patients who can safely and accurately be discharged without surgical management.

INTRODUCTION

Small bowel obstruction (SBO) continues to be one of the most common reasons for inpatient admissions in the US. More than 350,000 SBO operations are performed every year, resulting in 960,000 inpatient days and over $2 billion in health care expenditures.[12] A recent retrospective national study demonstrated that lysis of adhesions and small bowel resections were two of the seven most common emergency general surgery procedures, and these seven account for the highest admissions, deaths, complications, and inpatient costs of all emergency general surgical procedures.[3]

Three quarters of SBOs are caused by adhesive disease,[45] and in these patients, non-operative management results in resolution 74% of the time.[6] A number of decision-support tools have been developed to identify patients who can be managed non-operatively versus those requiring surgical intervention, including scoring systems based on radiologic findings, clinical findings, and combinations of the two. Early, accurate risk stratification is needed to mitigate negative outcomes associated with delayed clinical decision-making, including increased morbidity, increased cost, and increased post-operative length of stay.[7–9] The early identification of the need for surgical intervention and subsequent surgical treatment has demonstrated significant survival benefit, lower incidence of local systemic complications, and shorter hospitalizations.[8]

Although many of these findings have been known for decades, current practice continues to rely on ad hoc management strategies that differ between specialists. Some studies have suggested that management of adhesive-SBO by a primary medical (rather than surgical) team is associated with a higher healthcare utilization and worse perioperative outcomes.[10–12] An objective decision-making tool that updates patients' risk of requiring surgery as clinical and radiological findings are collected will help to streamline clinical practice and improve outcomes.

The aim of this study was to develop a predictive model capable of assisting clinicians as they assess the need for surgery in patients with SBO. Model development was based on admissions taking place within the Duke University Health System (DUHS) between 2013 and 2015, including readily available laboratory data, nursing collected information, and imaging when available. We hypothesized that this model would effectively stratify patients by their risk of requiring surgery, and that its performance would improve as data accumulated during the encounter. Importantly, we aimed to explore the cost/benefit of reducing the average length of stay for SBO by more promptly discharging patients whose model-predicted risk was low.

METHODS

Cohort Identification

Results were based on retrospective data analysis conducted at the Duke University School of Medicine and exempted from Institutional Review Board review. DUHS encounters in an observation window ranging January 1, 2013 – December 31, 2017 were initially identified. Inclusion criteria were (a) age ≥18, and (b) SBO diagnosis present on admission, as determined via retrospective billing data. SBO diagnoses were defined as codes 560.1 (paralytic ileus), 560.8X (other specified intestinal obstruction), or 560.9 (unspecified intestinal obstruction) from the International Classification of Diseases, 9th Revision, Clinical Modification (ICD9-CM). Diagnosis codes 560.2 (intussusception), 560.2 (volvulus), and 560.3X (impaction of intestine) were excluded. Encounters were labeled as 'Surgery' or 'Discharge' based on the presence or absence of surgery for SBO during the encounter, as defined by the following ICD9-CM procedure codes suggestive of laparoscopic or open intervention: 49320, 49000, 49321, 49002. Other exclusion criteria were (a) over 3 SBO-related encounters in the observation window, and (b) surgery for SBO over 3 weeks after admission.

Data Preprocessing

For each encounter, a time series of 24 clinical variables was extracted, including vital signs (*e.g.*, pulse, blood pressure), 11 laboratory results (*e.g.*, complete blood count, comprehensive metabolic panel), and 8 inputs and outputs (*e.g.*, intravenous fluid, stool and urine output). Variables were aggregated by hour by taking the mean or sum of all measurements, as appropriate. All variables and missingness rates are presented in eTables 1 and 2, respectively.

Missing values were imputed by carrying the last value forward or filling with the median across all encounters when no previous value was available. For each variable, the time since last measurement was included as an additional predictor. Three summary statistics were also calculated per raw variable to capture long-term and short-term trends. Exponential moving averages were calculated for vital signs (6-, 24-, and 72-hour half-life) and laboratory results (12-, 48-, and 144-hour half-life). Exponential moving sums were calculated for inputs and outputs (1-, 6-, and 24-hour half-life).

To incorporate radiological findings, frequently occurring text sequences (1, 2, and 3 words in length) were extracted from radiology notes (*e.g.*, CT and small bowel follow through studies) associated with both surgical and non-surgical encounters. Text sequences identified as clinically meaningful and interpretable (eTable 3), as well as indicators for CT and small bowel follow through studies, were used to create predictors indicating (a) appearance in the past 24 hours, and (b) appearance in the encounter up to the current time.

Age, hospital name, and the number of past SBO diagnoses were included as additional static predictors. Hours since admission and time of day were included as additional time-varying predictors. In total, 161 predictors were available for modeling.

Model Development

A Cox proportional hazards (Cox PH) framework was utilized to predict patients' relative risk of requiring surgery given the predictor variables available at each time point.[13] For any time point of interest, this approach allows the model to predict which patients are likely to receive surgery earlier versus later given that they have not yet received surgery by that time. If no SBO surgery took place during the encounter, the time to surgery was right censored for model training. The model was trained to maximize the Cox PH partial likelihood, which quantifies the probability of the observed order of surgery events, including censored events (*e.g.*, discharges), given the parameters of the model. An elastic net penalty was placed on model parameters to encourage sparsity and reduce overfitting.

The model was implemented and trained in Glmnet for Python with Python 3.5.2.[14] Training and evaluation took place using 5-fold nested cross-validation in which model performance is evaluated on each of 5 subsets of the data (*i.e.*, outer folds) after training on the other 4. In nested cross-validation, hyperparameters are selected through a similar procedure, wherein the training data are further partitioned into subsets.[1] Hyperparameters are then selected to maximize average performance across these subsets (*i.e.*, inner folds), preventing bias during model selection.[1] In the current model, hyperparameters included the two elastic net penalties corresponding to L1- and L2-regularization. Stratification was used to ensure the number of encounters resulting in SBO surgery was similar between partitions. All encounters for a given patient were assigned to the same partition(s).

Statistical Analysis and Performance Measures

Demographic factors and other descriptive statistics were compared between groups by two-tailed Mann-Whitney *U* test for numeric variables, and by chi-square test for categorical variables (see Table 1).

The model was evaluated based on its ability to predict (a) whether patients will require surgery within 24 hours, and (b) whether patients will require surgery at any time during the encounter. A 24-hour prediction window was chosen because it is long enough to allow surgeons to intervene, but short enough to allow the model to effectively identify patients whose condition is deteriorating.

Performance for these binary prediction tasks was quantified via the receiver operating characteristic (ROC) curve, which depicts the tradeoff between true positive rate (TPR; *i.e.*, sensitivity) and false positive rate (FPR; *i.e.*, 1 – specificity), as well as the area under the ROC curve (AUROC), which measures performance across all TPR-FPR pairs.[15] The AUROC may also be interpreted as the model's effectiveness in ranking encounters resulting in surgery as higher risk than encounters not resulting in surgery. A final evaluation measure, the concordance index (C-index), measures the model's effectiveness in (a) ranking encounters resulting in surgery as higher risk than those not resulting in surgery, and (b) correctly ranking encounters resulting in surgery by the time elapsed between admission and surgery.[16]

Aggregate performance across all time points would not be meaningful due to differences in length of stay between individual encounters and the average length of stay in 'Surgery' versus 'Discharge' groups. Further, very few laboratory results and inputs and outputs are available within the first few hours of admission. Instead, performance was calculated at specific time points after admission to provide information directly relevant to a prospective deployment. Performance has also been evaluated by the time remaining until surgery or discharge to examine how performance changes when approaching these endpoints.

An additional "early discharge task" was explored to evaluate the model's effectiveness in discharging low-risk patients. In this task, patients are discharged whenever their model-predicted risk falls below a fixed threshold. Early discharge performance may then be quantified across a range of thresholds using task-specific metrics. These include (a) the *discharge predictive value* (*i.e.*, positive predictive value for the discharge task), which quantifies the proportion of patients identified by the model for early discharge who did not later require surgery; (b) the *true discharge rate* (*i.e.*, sensitivity for the discharge task), which quantifies the proportion of patients eventually discharged who were correctly identified; (c) the *incorrect discharge rate* (*i.e.*, false positive rate for the discharge task), which quantifies the proportion of patients requiring surgery who were incorrectly identified for discharge; and (d) the *average reduction in hospital stay* among patients who did not receive surgery.

To compare model-based discharge with clinician performance, the incorrect discharge rate observed in our cohort was estimated by defining a false negative as a case where a patient requiring surgery had been readmitted for SBO fewer than three days after a previous discharge without surgery.

RESULTS

Description of Participants

A total of 7191 SBO-related encounters were identified between January 1, 2013 and December 31, 2017. After applying exclusion criteria, 3,910 encounters among 3,374 unique patients remained for analysis. Of these patients, 13.7% were readmitted at least once during the observation window (*i.e.*, 2013-2017) with an SBO diagnosis on admission. SBO-related surgery was identified in 606 of the encounters ('Surgery'; 15.5%), whereas no surgery was identified in 3304 ('Discharge'; 84.5%). Stepwise application of exclusion criteria is shown in Figure 1. Demographics, length of stay, and other descriptive statistics are presented in Table 1.

SBO surgery most commonly occurred in the first 24 hours after admission (median=25 hours). Patients not receiving surgery were most commonly discharged between 4 and 5 days after admission (median=4.7 days). Detailed encounter endpoints are presented in Figure 2.

Description of Data and Missing Values

Each vital and laboratory measurement (eTable 1) was observed in over 80% of encounters (eTable 2). CT studies were conducted in 69.8% of all encounters (mean=1.16/encounter). Abdominal radiographs were taken in 62.0% of all encounters (mean=2.37/encounter). Small bowel follow through studies were conducted in 6.6% of all encounters (mean=1.02/encounter).

Model Performance

Mean AUROC for the 24-hour prediction task increased from 0.644 at 12 hours after admission to 0.779 at 72 hours, with performance steadily improving between those times (see Figure 3, panels A-B). Similarly, performance for the within-encounter prediction task increased from 0.622 at 12 hours to 0.708 at 72 hours (see Figure 3, panels C-D), and the C-Index increased from 0.639 at 12 hours to 0.729 at 72 hours. Peak performance for the 24-hour task (0.824) was observed at 108 hours after admission, whereas peak performance for the within-encounter task (0.712) and C-Index (0.735) were observed at 96 hours after admission, respectively. Model performance is shown in Figure 3. AUROC and C-Index values for all folds are presented in eTables 4-8.

High variance across folds for the 24-hour task can be explained by small numbers of positive classes in the test set in each time window as time progresses (see Figure 2). At 72 hours, for example, the number of surgeries occurring in the next 24 hours in each test fold ranges from three to six. For this reason, performance could not be reliably assessed beyond 120 hours after admission.

Reduced Hospitalization Length

The observed incorrect discharge rate in our cohort was 9.3% (56 of 598 cases). When setting the discharge threshold to match this value, the model's true discharge rate was 50.4%, the discharge predictive value was 96.7%, and the average reduction in hospital stay was approximately 63 hours (see Figure 4). When setting the threshold to achieve an incorrect discharge rate of 2.0%, the average reduction in hospital stay was approximately 18 hours.

DISCUSSION

This study is the first to develop a machine learning model for determining the risk of requiring surgery among patients admitted for SBO. The model was developed using a large dataset comprising 3,910 SBO-related encounters at three hospitals within the Duke University Health System over a five-year period. In contrast to earlier scoring systems, the proposed model integrates information from multiple clinical data sources (*e.g.*, laboratory values, vital signs, inputs and outputs, radiological findings) to assign each patient a single score summarizing their relative risk. This risk score is time sensitive, changing as more clinical information is collected, to allow providers to dynamically assess patient status. Moreover, because the model determines relative risk rather than predicting surgery versus no surgery, it could be used to identify which cases are most urgent when prioritizing operating room resources, or to drive decisions about hospital resource allocation. It could also be used to determine which patients can safely be managed by a medical service versus a surgical service, or which patients should be transferred to a higher level of care or a center with surgical expertise.

Not surprisingly, model performance steadily improves as more data are collected. Peak performance on both the 24-hour and within-encounter prediction tasks was observed approximately three days after admission. Prediction performance is less robust in the first 24 hours, likely due to the limited amount of clinical data available at that time. Although many cases result in surgery within the first 24 hours, these cases are often immediately evaluated as urgent and therefore proceed to surgery before laboratory results, radiological findings, and other informative data have been collected. Due to the model's reliance on these findings, its greatest value may be in identifying whether patients not immediately requiring surgery remain at risk of deteriorating over time. In contrast, patients identified as low-risk by the model are unlikely to require surgery, and may be considered for earlier discharge.

Model-based discharge results confirm that identifying low-risk patients may be the most clinically actionable application of this work. On average, the model recommended discharge over 60 hours before discharge took place while maintaining an incorrect discharge rate lower than the true rate of SBO-related readmission in our cohort. It is impossible to determine precisely when the decision to discharge took place, but our experience suggests that this decision precedes discharge itself by several hours, not days. Thus, we conservatively estimate that model-guided discharge could reduce the average length of stay by over 2 days among patients not requiring surgery. Given the large number of SBO-related admissions per year, this could substantially reduce hospital costs, reduce costs to the patient, and improve patient satisfaction for hundreds of thousands of admissions in the US. Moreover, our simple decision rule for early discharge – namely, applying the same risk threshold across all time points -- should be fine-tuned to further improve performance in a real clinical deployment.

The next phase of this research is to prospectively deploy the model in a clinical setting. The model will be applied to all admitted patients with SBOs, providing near-real-time identification of patients predicted to require surgical intervention versus those who can be safely medically managed. An initial period of silent deployment, in which model predictions are not presented to clinicians, will be used to compare predictions with clinician decisions and patient outcomes to allow the model to be further validated and refined. If successful, the silent deployment will be followed by a prospective trial to evaluate the potential benefit of providing model-based recommendations to clinical teams. Model predictions would be used to inform provider decision-making, not replace it, and instances in which the team disagreed with the model's prediction would be logged and examined. Our long-term goal is to integrate this model into our electronic health record to be incorporated in the normal clinician workflow.

Limitations

This study has several limitations related to the retrospective nature of the analysis. Although thoughtful extraction and data cleaning procedures were applied, there was no mechanism to verify that clinical measurements were accurate or complete. Additionally, the model was developed using data from DUHS only, and may not generalize effectively to another hospital system. Prior to deployment at another location, the model should be prospectively tested and/or refined. It is also recognized that additional time spent in the hospital may influence the ultimate outcome of a discharge, and a more granular approach to prediction may be needed to guide overall health support (intravenous fluid, *etc.*) in order to optimize outcomes. Finally, this work assumes that patients deemed appropriate for surgical management were in fact appropriate for surgical management. We believe this is a reasonable assumption, however, our model has undoubtedly captured DUHS-specific clinical management strategies that may limit its generalizability. Further investigation is required, and a prospective randomized trial would be the optimal method to accomplish this.

CONCLUSION

This work demonstrates that a machine learning approach can be used to continuously, effectively stratify patients in a large health system by their risk of requiring SBO surgery. The model effectively determined whether patients would require surgery within the next 24 hours and within the current encounter. Performance improved as more data was collected, peaking approximately 3 days after admission. Given this trend in model performance and the high number of surgeries occurring within 24 hours of admission, the model is most promising as a means to identify low-risk patients for earlier discharge. Model predictions can also be used to prioritize operating room and hospital resources, or to determine which patients should be managed by medical versus surgical services or transferred to a higher level of care. Future study will quantify model performance when applied to other hospital systems and explore the benefit of a prospective deployment in an inpatient setting.

Acknowledgments

Conflicts of Interest

All authors have no conflicts of interest to report.

Funding

Authors reports funding from:

REFERENCES

1. Varma S, Simon R. [No title found]. *BMC Bioinformatics*. 2006;7(1):91. doi:10.1186/1471-2105-7-91

2. Arung W. Pathophysiology and prevention of postoperative peritoneal adhesions. *World J Gastroenterol*. 2011;17(41):4545. doi:10.3748/wjg.v17.i41.4545

3. Scott JW, Olufajo OA, Brat GA, et al. Use of National Burden to Define Operative Emergency General Surgery. *JAMA Surg*. 2016;151(6):e160480. doi:10.1001/jamasurg.2016.0480

4. Duron J-J, Silva NJ-D, du Montcel ST, et al. Adhesive Postoperative Small Bowel Obstruction: Incidence and Risk Factors of Recurrence After Surgical Treatment: A Multicenter Prospective Study. *Ann Surg*. 2006;244(5):750-757. doi:10.1097/01.sla.0000225097.60142.68

5. Miller G, Boman J, Shrier I, Gordon PH. Etiology of small bowel obstruction. *Am J Surg*. 2000;180(1):33-36. doi:10.1016/S0002-9610(00)00407-4

6. Foster NM, McGory ML, Zingmond DS, Ko CY. Small Bowel Obstruction: A Population-Based Appraisal. *J Am Coll Surg*. 2006;203(2):170-176. doi:10.1016/j.jamcollsurg.2006.04.020

7. Fevang BT, Fevang JM, Søreide O, Svanes K, Viste A. Delay in Operative Treatment among Patients with Small Bowel Obstruction. *Scand J Surg*. 2003;92(2):131-137. doi:10.1177/145749690309200204

8. Keenan JE, Turley RS, McCoy CC, Migaly J, Shapiro ML, Scarborough JE. Trials of nonoperative management exceeding 3 days are associated with increased morbidity in patients undergoing surgery for uncomplicated adhesive small bowel obstruction: *J Trauma Acute Care Surg*. 2014;76(6):1367-1372. doi:10.1097/TA.0000000000000246

9. Schraufnagel D, Rajaee S, Millham FH. How many sunsets? Timing of surgery in adhesive small bowel obstruction: A study of the Nationwide Inpatient Sample. *J Trauma Acute Care Surg*. 2013;74(1):181-189. doi:10.1097/TA.0b013e31827891a1

10. Bilderback PA, Massman JD, Smith RK, La Selva D, Helton WS. Small Bowel Obstruction Is a Surgical Disease: Patients with Adhesive Small Bowel Obstruction Requiring Operation Have More Cost-Effective Care When Admitted to a Surgical Service. *J Am Coll Surg*. 2015;221(1):7-13. doi:10.1016/j.jamcollsurg.2015.03.054

11. Malangoni MA, Times ML, Kozik D, Merlino JI. Admitting service influences the outcomes of patients with small bowel obstruction. *Surgery*. 2001;130(4):706-713. doi:10.1067/msy.2001.116918

12. Aquina CT, Becerra AZ, Probst CP, et al. Patients With Adhesive Small Bowel Obstruction Should Be Primarily Managed by a Surgical Team: *Ann Surg*. 2016;264(3):437-447. doi:10.1097/SLA.0000000000001861

13. Cox DR. Regression Models and Life-Tables. *J R Stat Soc Ser B Methodol*. 1972;34(2):187-220.

14. Balakumar, B.J., Hastie, T., Friedman, J., Tibshirani, Simon, N.R. *Glmnet for Python*.; 2016. http://www.stanford.edu/~hastie/glmnet_python/.

15. Bradley, A.P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition. *Pattern Recognit*. 1997;30:1145-1159. doi:10.1016/S0031-3203(96)00142-2

16. Harrell FE, Lee KL, Mark DB. MULTIVARIABLE PROGNOSTIC MODELS: ISSUES IN DEVELOPING MODELS, EVALUATING ASSUMPTIONS AND ADEQUACY, AND MEASURING AND REDUCING ERRORS. *Stat Med*. 1996;15(4):361-387. doi:10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4

Table 1: Demographics and other Descriptive Statistics

Variable	Value	Surgery	No surgery	Test Statistic	*p*-value
Encounters	N (%)	598 (15.3%)	3310 (84.7%)	-	-
Age	Mean ± SD (range)	61.6 ± 16.6 (18-97)	61.4 ± 16.4 (18-109)	U=9.9e5	0.68
Length of Stay	Mean ± SD (range)	13.3 ± 11.3 (0.76-89.2)	7.6 ± 9.5 (0.11-252.8)	U=1.4e6	0
Weight	Mean ± SD (range)	171.5 ± 46.9 (70-433.4)	174.0 ± 51.7 (48.3-456.3)	U=9.3e5	0.03
Sex	*Female*, N (%)	340 (56.9%)	1746 (52.7%)	χ^2=3.26	0.07
	Male, N (%)	258 (43.1%)	1564 (47.3%)		
Location	*Duke University Hospital*, N (%)	105 (17.6%)	531 (16.0%)	χ^2=3.75	0.15
	Duke Raleigh Hospital, N (%)	160 (26.8%)	796 (24.0%)		
	Duke Regional Hospital, N (%)	333 (55.7%)	1983 (59.9%)		
Ethnicity	*Hispanic or Latino*, N (%)	14 (2.3%)	69 (2.1%)	χ^2=0.16	0.99
	Not Hispanic/Latino, N (%)	573 (95.8%)	3180 (96.1%)		
	Not Reported/Declined, N (%)	11 (1.8%)	61 (1.8%)		
Race	*Two or more races*, N (%)	7 (1.2%)	29 (0.9%)	χ^2=12.82	0.08
	American Indian or Alaskan Native, N (%)	7 (1.2%)	22 (0.7%)		
	Asian, N (%)	15 (2.5%)	38 (1.1%)		
	Black or African American, N (%)	179 (29.9%)	1098 (33.2%)		
	Caucasian/White, N (%)	375 (62.7%)	2022 (61.1%)		
	Native Hawaiian or Other Pacific Islander, N (%)	0 (0.0%)	2 (0.1%)		
	Not Reported/Declined, N (%)	3 (0.5%)	34 (1.0%)		
	Other, N (%)	12 (2.0%)	65 (2.0)%		
Insurance	*Commercial*, N (%)	9 (1.6%)	38 (1.2%)	χ^2=15.8	0.11
	Managed Care, N (%)	60 (10.4%)	400 (12.4%)		
	Medicaid Pending, N (%)	0 (0.0%)	2 (0.1%)		
	Medicare, N (%)	253 (43.7%)	1431 (44.3%)		
	Medicare Advantage, N (%)	89 (15.4%)	409 (12.7%)		
	NC Blue Cross, N (%)	84 (14.5%)	397 (12.3%)		
	NC Medicaid, N (%)	36 (6.2%)	242 (7.5%)		
	OOS Blue Cross, N (%)	34 (5.9%)	213 (6.6%)		
	OOS Medicaid, N (%)	5 (0.9%)	9 (0.3%)		
	Other Government, N (%)	3 (0.5%)	25 (0.8%)		
	Special Programs, N (%)	6 (1.0%)	61 (1.9%)		

FIGURE LEGENDS

Figure 1: Stepwise Application of Exclusion Criteria.

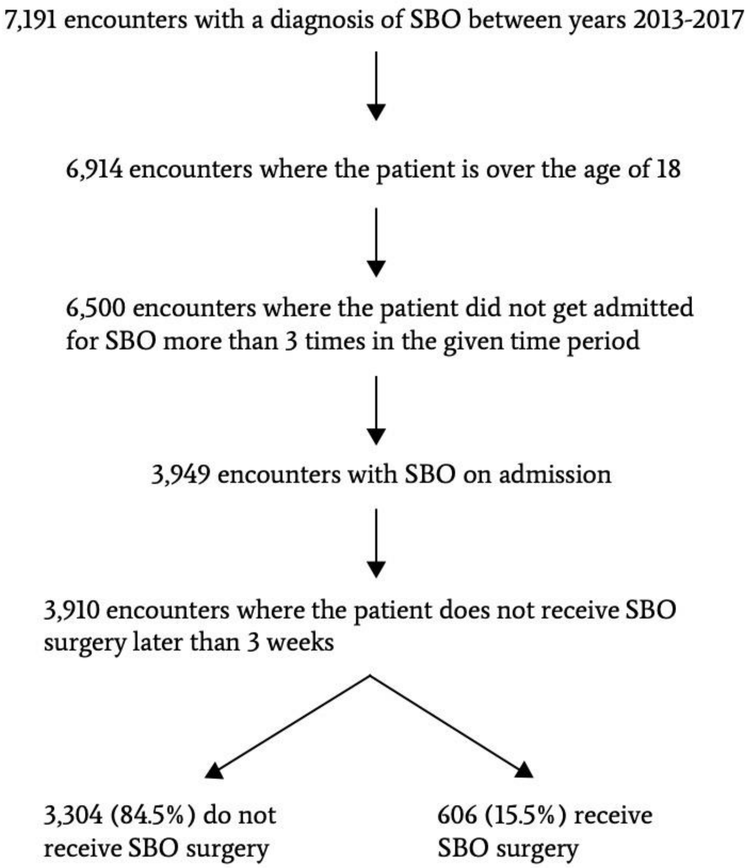


Figure 2: Encounter Endpoints. Panel (A) shows histograms of the time between admission and SBO-related surgery in the 'Surgery' group compared to the time between admission and discharge in the 'Discharge' group. Panel (B) shows the proportion of each group who have not yet received surgery or been discharged as a function of the time since admission.

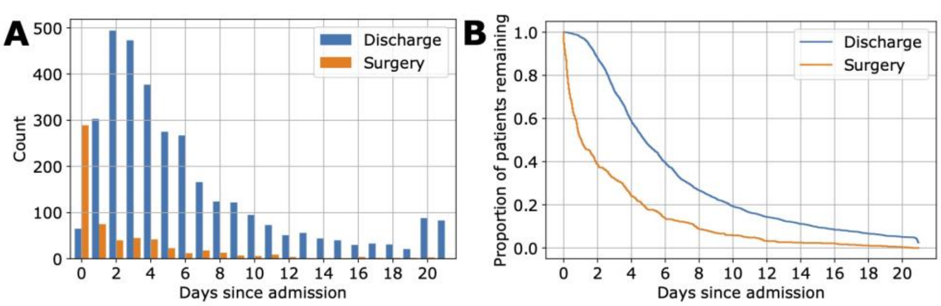


Figure 3: Prediction Performance over Time. Performance over time is quantified when predicting whether patients will receive surgery in the next 24 hours (*i.e.*, "24-hour prediction") and at any time during the current encounter (*i.e.*, "within-encounter prediction"), respectively. The left panels show the area under the receiver operating characteristic curve (AUROC) by time since admission for (A) the 24-hour task, and (C) the within-encounter task. The line plot indicates the mean AUROC at each time point and the boxplots show values across all folds, with the green center lines indicating median values, box edges indicating the interquartile range, and whiskers indicating the maximum and minimum values. The right panels show the receiver

operating characteristic (ROC) curve at 24, 72, and 120 hours since admission for (B) the 24-hour task, and (D) the within-encounter task.

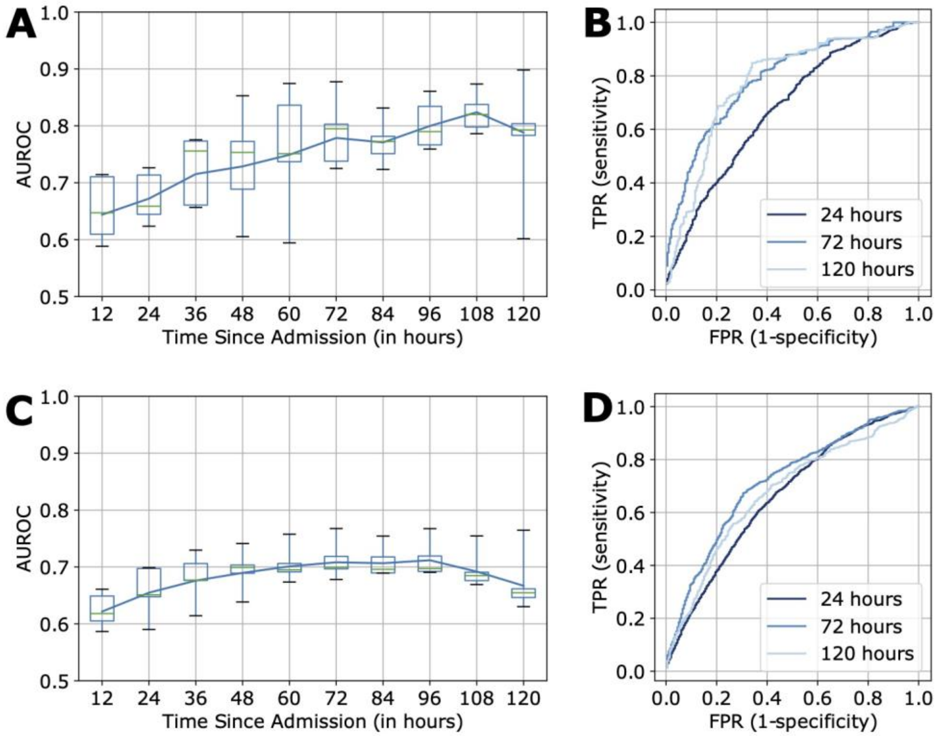
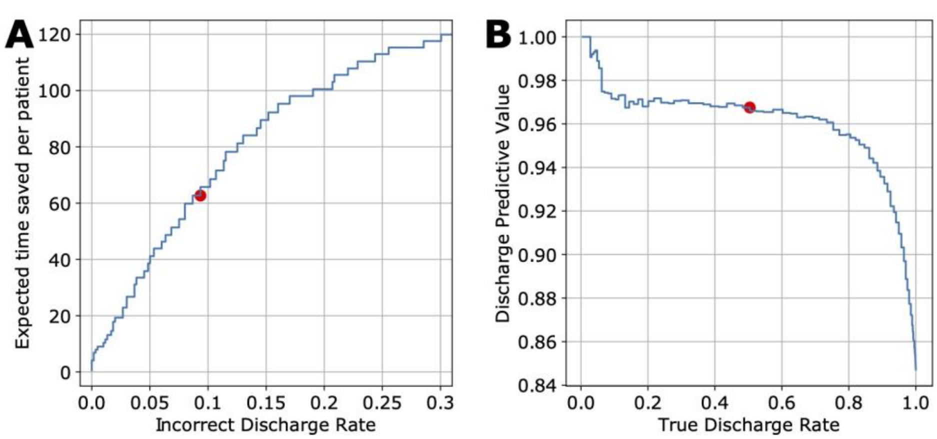


Figure 4: Discharge Prediction Performance. Panel (A) shows the average reduction in length of stay among patients eventually discharged versus the corresponding incorrect discharge rate. The red marker indicates the point on the curve that matches the incorrect discharge rate observed in our dataset as a consequence of clinician decision-making (9.4%). Panel (B) shows the tradeoff between discharge predictive value and the true discharge rate for the discharge prediction task. As before, the red marker indicates the point on the curve associated with clinician decision-making in the current cohort.



SUPPLEMENT: Machine Learning Prediction of Surgical Intervention for Small Bowel Obstruction

eTable 1: Overview of variables
eTable 2: Variable Descriptive Statistics and Observation Rates
eFigure 1: Concordance Index over Time
eFigure 2: Prediction Performance Relative to Time to Event
eTable 3: AUROC over Time for the 24-hour Prediction Task
eTable 4: AUROC over Time for Surgery at any Point
eTable 5: Concordance Index over Time
eTable 6: AUROC Relative to Time to Event
eTable 7: Concordance Index Relative to Time to Event
eTable 8: Words and phrases selected by expert panel

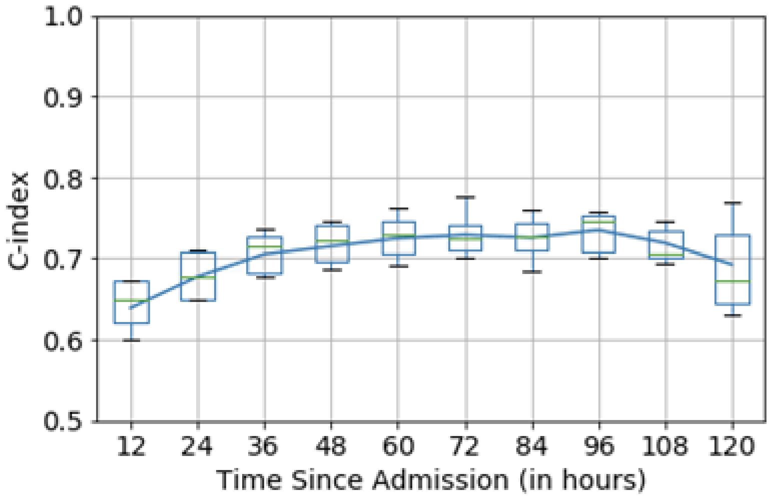
eTable 1: Overview of variables

Category	Variable	Description
	Age	Age at time of encounter
	Location	Duke University, Duke Regional, Duke Raleigh
	Number of Past SBOs	
	Hour Since Admission	Based on Admission Timestamp
	Time of Day	
Vitals	Diastolic Blood Pressure	Vital Sign Collected on Admission
	Systolic Blood Pressure	Vital Sign Collected on Admission
	Pulse	Vital Sign Collected on Admission
	Respiratory Rate	Vital Sign Collected on Admission
	SPO2	Vital Sign Collected on Admission
Labs	Albumin	Serum Level from Metabolic Panel
	Calcium	Serum Level
	Carbon Dioxide	Serum Level
	Chloride	Serum Level from Metabolic Panel
	Creatinine	Serum Level from Metabolic Panel
	Hematocrit	Serum Level
	Hemoglobin	Serum Level from CBC
	Platelet Count	Serum Level from CBC
	Potassium	Serum Level from Metabolic Panel
	Sodium	Serum Level from Metabolic Panel
	White Blood Cell Count	Serum Level from CBC
Inputs/outputs	Maintenance IV Volume	Intravenous Fluid Intake
	Piggyback IV Volume	Intravenous Fluid Intake
	Tube Intake	Nasogastric Tube Intake
	Tube Output	Nasogastric Tube Output
	Urine Output	Urine Output w/ and w/o catheter
	Emesis Occurrence	Emesis Present or not
	Stool Occurrence	Stool Output or not
	Urine Occurrence	Urine Output or not

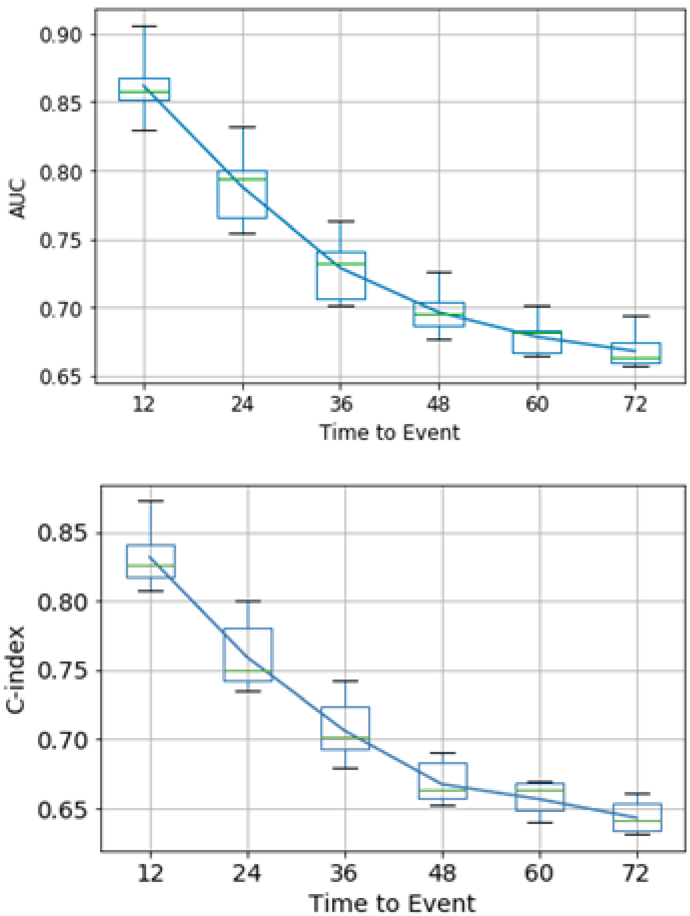
eTable 2: Variable Descriptive Statistics and Observation Rates

Variable	Mean ± SD (range)	Encounters with at least one measurement, N (%)
Diastolic Blood Pressure (mmHg)	70.2±13.6 (0-171)	3901 (99.8%)
Systolic Blood Pressure (mmHg)	127.8±23.4 (0-476)	3901 (99.8%)
Pulse (bpm)	87.9±18.5 (0-369)	3906 (99.9%)
Respiratory Rate (rpm)	19.2±4.8 (0-189)	3905 (99.9%)
SPO2 (%)	96.6±3.0 (0-100)	3904 (99.8%)
Platelet Count (K/mcL)	250.4±136.8 (1-2442)	3507 (89.7%)
White Blood Cell Count (K/mcL)	10.1±6.6 (0.1-117.6)	3508 (89.7%)
Hemoglobin (g/dL)	10.5±2.1 (2.9-20.3)	3508 (89.7%)
Hematocrit (%)	15.5±16.3 (9-58.9)	3188 (81.5%)
Sodium (mmol/L)	137.7±4.9 (110-177)	3818 (97.6%)
Potassium (mmol/L)	3.9±0.6 (1.5-9.1)	3814 (97.5%)
Chloride (mmol/L)	104.4±6.5 (54-155)	3818 (97.6%)
Albumin (g/dL)	2.9±0.9 (1-6)	3310 (84.7%)
Calcium (mg/dL)	8.5±0.7 (3.4-14.4)	3818 (97.6%)
Carbon Dioxide (mmol/L)	25.3±4.2 (5-50)	3818 (97.6%)
Creatinine (mg/dL)	1.4±1.4 (0.1-26.8)	3818 (97.6%)

eFigure 1: Concordance Index over Time. Boxplots depict the concordance index (c-index) across all folds by time since admission, with the green center lines indicating median values, box edges indicating the interquartile range, and whiskers indicating the maximum and minimum values. The line plot indicates the mean c-index at each time point.



eFigure 2: Prediction Performance Relative to Time to Event. Panels (A, B) show model performance by time until surgery or discharge when predicting whether the patient will receive surgery at any time during the current encounter. In panel (A), performance is quantified as the area under the receiver operating characteristic curve (AUROC), while in panel (B), performance is quantified as the concordance index (c-index). Boxplots depict model performance values across all folds, with the green center lines indicating median values, box edges indicating the interquartile range, and whiskers indicating the maximum and minimum values. The line plot indicates the mean performance value at each time point.



eTable 3: Words and phrases identified as clinically meaningful and interpretable. The table includes common words and phrases from CT and small bowel follow through studies, respectively, and whether they are more strongly associated with SBO surgery ('Surgery') or no SBO surgery ('Discharge') in our dataset.

CT		**Small bowel follow through**	
Surgery	**Discharge**	**Surgery**	**Discharge**
ischemia	ileus		partial
Closed loop	Early or partial		Normal transit
Internal hernia	No ascites		Bowel transit
Transition point	unchanged		Gaseous distention
Pneumotosis or free			Normal appearance
Impression high grade			Opacification of the

eTable 4: AUROC over Time for the 24-hour Prediction Task. Values for all folds presented in Figure 3, Panel (A) are provided in the table below.

	Time Since Admission (in hours)									
	12	24	36	48	60	72	84	96	108	120
Fold 1	0.647	0.714	0.756	0.773	0.836	0.803	0.751	0.790	0.873	0.783
Fold 2	0.610	0.624	0.657	0.689	0.751	0.877	0.831	0.861	0.838	0.602
Fold 3	0.710	0.726	0.773	0.853	0.874	0.795	0.723	0.759	0.798	0.804
Fold 4	0.714	0.659	0.775	0.753	0.737	0.738	0.782	0.834	0.820	0.898
Fold 5	0.588	0.644	0.661	0.605	0.594	0.725	0.772	0.767	0.786	0.793
Mean	0.644	0.672	0.715	0.729	0.749	0.779	0.771	0.800	0.824	0.788

eTable 5: AUROC over Time for Surgery at any Point. Values for all folds presented in Figure 3, Panel (B) are provided in the table below.

	Time Since Admission (in hours)									
	12	24	36	48	60	72	84	96	108	120
Fold 1	0.649	0.697	0.706	0.703	0.691	0.697	0.690	0.691	0.691	0.646
Fold 2	0.587	0.648	0.677	0.699	0.706	0.719	0.718	0.719	0.685	0.655
Fold 3	0.661	0.699	0.730	0.741	0.758	0.768	0.754	0.767	0.755	0.765
Fold 4	0.618	0.590	0.614	0.639	0.674	0.678	0.689	0.692	0.669	0.662
Fold 5	0.605	0.651	0.676	0.689	0.695	0.700	0.696	0.697	0.676	0.630
Mean	0.622	0.655	0.676	0.690	0.701	0.708	0.707	0.712	0.692	0.667

eTable 6: Concordance Index over Time

	Time Since Admission (in hours)									
	12	24	36	48	60	72	84	96	108	120
Fold 1	0.649	0.71	0.727	0.723	0.705	0.702	0.685	0.701	0.701	0.63
Fold 2	0.62	0.677	0.716	0.746	0.763	0.777	0.759	0.753	0.693	0.644
Fold 3	0.672	0.709	0.737	0.741	0.746	0.74	0.727	0.747	0.746	0.768
Fold 4	0.672	0.65	0.682	0.697	0.728	0.724	0.744	0.756	0.734	0.73
Fold 5	0.601	0.649	0.679	0.686	0.693	0.711	0.71	0.709	0.705	0.673
Mean	0.639	0.678	0.705	0.715	0.725	0.729	0.725	0.735	0.719	0.693

eTable 7: AUROC Relative to Time to Event. Values for all folds presented in eFigure 2, Panel (A) are provided in the table below.

	Time to Event (in hours)					
	12	24	36	48	60	72
Fold 1	0.871	0.805	0.733	0.697	0.677	0.672
Fold 2	0.851	0.780	0.717	0.672	0.651	0.646
Fold 3	0.909	0.828	0.753	0.709	0.676	0.667
Fold 4	0.854	0.764	0.698	0.672	0.660	0.648
Fold 5	0.835	0.763	0.710	0.683	0.673	0.647
Mean	0.863	0.786	0.720	0.685	0.666	0.654

eTable 8: Concordance Index Relative to Time to Event

	Time to Event (in hours)					
	12	24	36	48	60	72
Fold 1	0.841	0.78	0.724	0.683	0.668	0.661
Fold 2	0.817	0.75	0.702	0.657	0.64	0.631
Fold 3	0.873	0.8	0.742	0.69	0.669	0.654
Fold 4	0.826	0.743	0.679	0.652	0.648	0.641
Fold 5	0.808	0.735	0.693	0.664	0.663	0.634
Mean	0.832	0.759	0.706	0.667	0.656	0.643