Surgery 172 (2022) 663–669

Contents lists available at ScienceDirect

Surgery

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

Optimizing discharge after major surgery using an artificial intelligence–based decision support tool (DESIRE): An external validation study

Davy van de Sande, BSc[a], Michel E. van Genderen, MD, PhD[a,*], Cornelis Verhoef, MD, PhD[b], Joost Huiskens, MD, PhD[c], Diederik Gommers, MD, PhD[a], Edwin van Unen, Ir.[c], Renske A. Schasfoort, MD[d], Judith Schepers, MSc[e], Jasper van Bommel, MD, PhD[a], Dirk J. Grünhagen, MD, PhD[b]

[a] Department of Adult Intensive Care, Erasmus University Medical Center, Rotterdam, The Netherlands
[b] Department of Surgical Oncology, Erasmus MC Cancer Institute University Medical Center, Rotterdam, The Netherlands
[c] SAS Institute, Health, Huizen, The Netherlands
[d] Department of Surgery, Treant Care Group, Emmen, The Netherlands
[e] Department of Business Intelligence, Treant Care Group, Emmen, The Netherlands

ARTICLE INFO

Article history:
Accepted 21 March 2022
Available online 4 May 2022

ABSTRACT

Background: In the DESIRE study (Discharge aftEr Surgery usIng aRtificial intElligence), we have previously developed and validated a machine learning concept in 1,677 gastrointestinal and oncology surgery patients that can predict safe hospital discharge after the second postoperative day. Despite strong model performance (area under the receiver operating characteristics curve of 0.88) in an academic surgical population, it remains unknown whether these findings can be translated to other hospitals and surgical populations. We therefore aimed to determine the generalizability of the previously developed machine learning concept.

Methods: We externally validated the machine learning concept in gastrointestinal and oncology surgery patients admitted to 3 nonacademic hospitals in The Netherlands between January 2017 and June 2021, who remained admitted 2 days after surgery. Primary outcome was the ability to predict hospital interventions after the second postoperative day, which were defined as unplanned reoperations, radiological interventions, and/or intravenous antibiotics administration. Four forest models were locally trained and evaluated with respect to area under the receiver operating characteristics curve, sensitivity, specificity, positive predictive value, and negative predictive value.

Results: All models were trained on 1,693 epsiodes, of which 731 (29.9%) required a hospital intervention and demonstrated strong performance (area under the receiver operating characteristics curve only varied 4%). The best model achieved an area under the receiver operating characteristics curve of 0.83 (95% confidence interval [0.81–0.85]), sensitivity of 77.9% (0.67–0.87), specificity of 79.2% (0.72–0.85), positive predictive value of 61.6% (0.54–0.69), and negative predictive value of 89.3% (0.85–0.93).

Conclusion: This study showed that a previously developed machine learning concept can predict safe discharge in different surgical populations and hospital settings (academic versus nonacademic) by training a model on local patient data. Given its high accuracy, integration of the machine learning concept into the clinical workflow could expedite surgical discharge and aid hospitals in addressing capacity challenges by reducing avoidable bed-days.

Introduction

Because of the high demand for hospital services, efficient capacity management is critical for the continuous availability of postoperative care beds, especially during pressing times.[1] Also,

* Reprint requests: Michel E. van Genderen, Erasmus Medical Center, Department of Adult Intensive Care, Room Ne-403, Doctor Molewaterplein 40, 3015 GD Rotterdam, The Netherlands.
E-mail address: m.vangenderen@erasmusmc.nl (M.E. van Genderen);
Twitter: @davy_sande, @ErasmusMC

https://doi.org/10.1016/j.surg.2022.03.031

unnecessary prolonged hospital stay can be harmful because patients are exposed to an increased risk of iatrogenic complications.[2] Conversely, too early discharge can lead to delayed recognition and treatment of complications. Both situations contribute to postoperative morbidity and mortality with a negative impact on quality of life.[3,4] Thus, the decision for a timely and safe time of discharge is key.[5] Despite initiatives to expedite postoperative recovery such as the Enhanced Recovery After Surgery pathways and the coordinated efforts of physicians, nurses, and policy makers, there is still room for improvement.[6,7]

Clinical artificial intelligence (AI) or machine learning (ML)–based prediction models are increasingly reported in medicine and have the potential to improve patient care as well as the surgeon workflow.[8–10] Some examples include predicting in-hospital mortality after aneurysm repair surgery, analyzing operative reports to determine anastomotic leak after colorectal surgery, and automatic surgical phase recognition from intraoperative imaging.[11,12] Although developing such AI models can be challenging in itself, the real challenge is to make it to the bedside and to use them at large scale. To illustrate, 90% to 94% of the AI models in the intensive care unit and radiology department remain in the development and prototyping phase (ie, a phase where only retrospective data are analyzed), respectively.[9,13] Although these were nonsurgical studies, these findings could also be extrapolated to surgery. Such models may work perfectly in one hospital but may be poorly generalizable in other clinical settings; therefore, an external validation is a crucial step toward safe clinical implementation.[14,15] While some surgical AI models underwent external validation, it is not yet standard practice, and thus there is little evidence if such models would alter clinical practice or improve clinical outcomes.[16–19]

In the DESIRE (Discharge aftEr Surgery usIng aRtificial intElligence) study, we have previously developed and validated a ML concept in 1,677 gastrointestinal and oncology surgery patients in a tertiary referral hospital that can predict safe hospital discharge after the second postoperative day.[20] Despite strong model performance (area under the receiver operating characteristics curve [AUROC] of 0.88), it remains unknown whether these findings can be translated to other hospitals and surgical populations. In this study we therefore aimed to determine generalizability of the previously developed ML concept, which is a crucial step before clinical implementation, and, as such, we externally validated the concept and assessed its performance in gastrointestinal and oncology surgery patients in a nonacademic hospital.

Methods

We conducted a retrospective cohort study to externally validate the ML concept following the TRIPOD (transparent reporting of a multivariable prediction model for individual prognosis or diagnosis) guideline.[21] The study protocol was approved by the Ethics Committee of the Erasmus MC University Medical Center (protocol no. MEC-2021-0625), and the need for informed consent was waived.

Participants and outcome

Adult gastrointestinal and oncology surgery patients (≥18 years), admitted to 3 different hospitals (Bethesda, Refaja, and Scheper Hospitals) that are part of the Treant Care Group in The Netherlands were selected based on surgical procedure descriptions to match the development cohort.[20] The Treant Care Group provides secondary care to approximately 300,000 people in the northeastern area of The Netherlands. Only patients admitted more than 2 days after initial surgery and who were discharged between January 2017 and June 2021 were included.

The primary outcome was the performance in terms of AUROC and negative predictive value (NPV) to predict hospital interventions (ie, care that is strictly provided by hospitals) after the second postoperative day. Hospital interventions were defined as reoperations, radiological interventions, and intravenous antibiotics. As a secondary outcome, we calculated the predicted number of avoidable bed-days for 2017-2019 to determine potential clinical impact. Avoidable bed-days were defined as the number of admission days after the second postoperative day for patients for whom safe discharge to a lower level of care (eg, to a postoperative nursing home or home if home-care can be arranged) was predicted.

Data collection

For each patient, a minimum list of 17 perioperative variables needed to be collected, which were previously used to develop the ML concept.[20] Additionally, the corresponding admission diagnosis was collected, totaling to 18 variables per patient (Table I). Data were retrieved and linked from multiple medical information systems including Xcare, Metavision, and Zamicom. All variables were unaggregated, except for medication history, for which the total number of unique administered medications until the second postoperative day was calculated.

Data preparation

All identifiable patient data were removed, and unique episodes were created for each patient encounter. If a patient underwent multiple surgeries within 1 episode, the surgery date closest to hospital admission was marked as the initial surgery. All surgeries that occurred within the same episode and more than 2 days after the initial surgery were marked as reoperations. In addition, all intravenous antibiotic administrations and radiological interventions (eg, percutaneous abscess drainage, computed tomography scan-guided

Table I
Machine learning model input variables

Input variable description (unit)	Variable derivation
Length of time in the operating room (OR) (minutes)	OR exit time - OR entrance time
Surgical procedure description	NA
Expected duration of surgery (minutes)	NA
Number of unique administered medications	Count of distinct medication descriptions where date of administration ≤2 days after initial surgery
Length of stay until surgery (days)	Surgery date−hospital admission date
BMI on admission (kg/m^2)	Weight on admission (in kg)/(length on admission (in m)^2)
Department of admission after surgery	NA
Age on admission (years)	(Hospital admission date−date of birth)/365.25)
Location of origin	NA
ASA score	NA
Responsible specialty	NA
Surgical urgency	NA
Specialty of admission	NA
Anesthesia type	NA
Sex	NA
Hospital location	NA
Surgical indicator	NA
Admission diagnosis	NA

ASA, American Society of Anesthesiologists; BMI, body mass index; NA, not applicable; OR, operating room.

puncture) that occurred between the second postoperative day and hospital discharge were marked as hospital interventions. The outcome to be predicted was the occurrence of 1 of the 3 interventions (reoperations, radiological interventions, and intravenous antibiotics).

Statistical analysis and model training

A minimum sample size of 1,374 episodes was required to validate the ML concept on local patient data.[20] Patient characteristics and clinical outcomes were reported as median and interquartile range (IQR) for numerical variables and count and percentage (%) for categorical variables. Correlation of continuous variables was calculated by using Pearson's correlation and was classified as weak (<0.3), moderate (>0.3 and ≤0.6), and strong (>0.6).

Four random forest models were trained to validate the ML concept. Model parameters were drawn from the previous study and were unmodified (settings: 100 trees, 5 variables per split, 50 interval bins, and 2 branches).[20] Missing values were used as splitting criteria in the models. The first model was trained using a total of 17 variables (hereafter referred to as the full model), presented in Table I. The second model was trained on a reduced number of variables (hereafter referred to as the reduced model); strongly correlated variables were excluded and a reduced list of clinically relevant variables was selected based on clinical expert knowledge (ie, informative variables were selected).[22] Additionally, the reduced and the full models were evaluated by adding the admission diagnosis as input variable.

Evaluation of model performance and impact analysis

The data set was randomly divided into 3 nonoverlapping data sets: train (70%), validation (20%), and test data set (10%).[23] Multiple metrics were calculated on the test data set (ie, unseen data) for each model following the guideline of Park et al[24]: AUROC curve, misclassification rate, sensitivity (%), specificity (%), positive predictive value (PPV) (%), and negative predictive value NPV (%). Youden's J statistic was used to calculate the optimal statistical classification threshold on the validation data set.[25] A nomogram, reflecting variables' relative importance, was also constructed.

The best model (in terms of AUROC and NPV [%]) was used to predict the total number of avoidable bed-days for 2017, 2018, and 2019. Potential misclassification was penalized by subtracting half the predicted avoidable bed-days, multiplied by the hospital interventions probability, from the predicted avoidable bed-days. For each episode these days were calculated as follows:

predicted avoidable bed-days; ((length of stay after initial surgery (in days)−2)(1-hospital interventions probability (%)))−penalty; (((length of stay after initial surgery (in days)−2)*0.5)*hospital interventions probability (%)).*

Statistical Analysis System (SAS) Viya version 3.5 was used for statistical analysis. The SAS Visual Data Mining and Machine Learning package of this software was used to externally validate the ML concept.

Results

Patient characteristics for the train, validate, and test data set are presented in Table II. After data preparation, a total of 2,035 patients with 2,447 unique epsiodes were identified and randomly divided over the different data sets. The train data set included 1,693 epsiodes (median [IQR] age, 69 [56–77] years; 888 [52.5%] men), the validation set 505 episodes (age, 69 [57–77] years; 267 [52.9%] men), and the test set 249 episodes (age, 69 [58–76] years; 140 [56.2%] men).

Table II
Characteristics of the training, validation, and test data set

	Patients, no. (%)		
Characteristic	Training	Validation	Test
No. of unique episodes	1,693 (69.2)	505 (20.6)	249 (10.2)
Sex			
Men	888 (52.5)	267 (52.9)	140 (56.2)
Women	805 (47.5)	238 (47.1)	109 (43.8)
Age, years, median (IQR)	69 (56–77)	69 (57–77)	69 (58–76)
BMI, kg/m^2, median (IQR)	23.6 (20.8–27.1)	23.9 (20.6–27.1)	23.7 (20.9–28.1)
ASA score, median (IQR)	2 (2–3)	2 (2–3)	2 (2–3)
Length of stay after surgery, median (IQR)	5 (4–10)	5 (4–10)	5 (4–9)
Surgical urgency			
Elective	880 (52)	241 (47.8)	113 (45.4)
Emergency	813 (48)	264 (52.2)	136 (54.6)
Surgery type			
Breast	28 (1.7)	6 (1.2)	-
Colon and rectum	626 (37)	171 (33.9)	80 (32.1)
Diagnostic laparoscopy	68 (4)	24 (4.8)	11 (4.4)
Hernia	64 (3.8)	28 (5.5)	13 (5.2)
Lymph node dissection	5 (0.3)	3 (0.6)	-
Melanoma	2 (0.1)	-	-
Ostomy	160 (9.5)	40 (8)	19 (7.6)
Stomach	24 (1.4)	10 (2)	5 (2)
Thyroid gland	5 (0.3)	1 (0.2)	2 (0.8)
Other	560 (33.1)	178 (35.2)	89 (35.7)

ASA, American Society of Anesthesiologists; *BMI*, body mass index; *IQR*, interquartile range.

In 731 (29.9%) out of 2,447 episodes, a hospital intervention occured beyond the second postoperative day of which 620 [84.8%] consisted of at least the administration of intravenous antibiotics (Table III). Episodes without a hospital intervention had a median stay of 4 days (IQR; 3–6) after surgery, compared to 13 days for episodes with a hospital intervention.

Variable reduction

The total number of 17 variables ± admission diagnosis was used by the full models, and a reduced number was selected for the reduced models. Expected duration of surgery was excluded due to high correlation with length of time in the operating room (Pearson's $r = 0.77$); correlations are presented in the heatmap in Figure 1. All other variables were weakly correlated (Pearson's $r < 0.3$). Out of the remaining 16 variables, 11 were selected based on their clinical relevance; anesthesia type, location of origin, responsible specialty, specialty of admission, and surgical indicator

Table III
Frequency of hospital interventions

Reoperations	Intravenous antibiotics	Radiological interventions	Number of episodes
Yes	Yes	Yes	64
		No	55
	No	Yes	15
		No	43
No	Yes	Yes	100
		No	401
	No	Yes	53
		No	1,716
Total 2,447			

In 731 episodes (29.9%) at least 1 intervention; in 43 episodes, only 1 reoperation; in 53 episodes only 1 radiological intervention, and in 401 episodes only intravenous antibiotics.

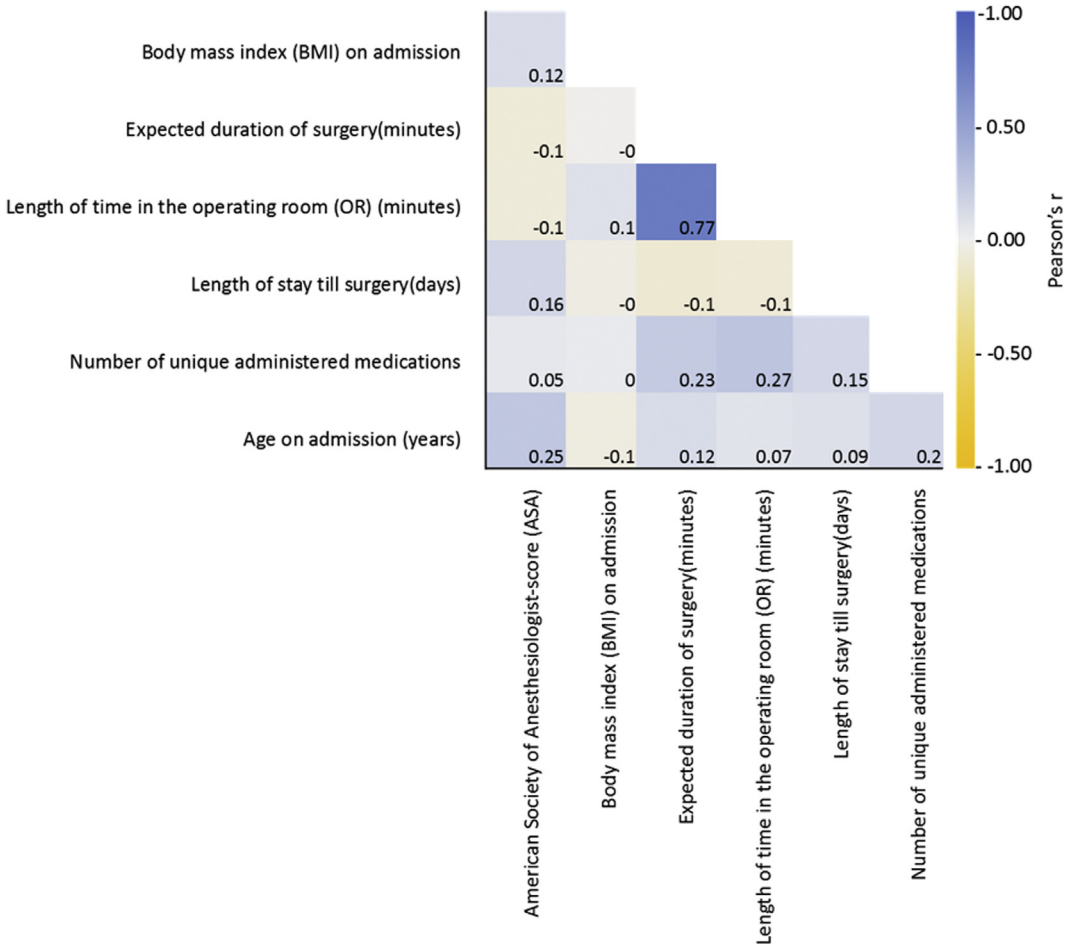

Figure 1. Variable concurvity. Pearson's *r* was used to evaluate correlation between each combination of continuous variables.

were excluded. Finally, the reduced models were trained on 11 variables ± admission diagnosis.

Model performance

All 4 models demonstrated good discriminative performance with at least an AUROC of 0.79 (95% CI [0.77–0.81]) and a maximum difference of 4% between models (Figure 2). The reduced models (± admission diagnosis) outperformed the full models (± admission diagnosis) in terms of AUROC and NPV; the reduced models had an AUROC of 0.83 (0.81–0.85) and 0.81 (0.79–0.83) compared to 0.81 (0.79–0.83) and 0.79 (0.77–0.81) for the full models. Furthermore, the reduced models had a NPV of 89.3% (0.85–0.93) and 89.0% (0.84–0.93) compared to 88.2% (0.83–0.92) and 85.9% (0.81–0.90) for the full models (Table IV).

The reduced model that additionally used the admission diagnosis ($n = 12$ variables) performed best. On the test data set this model had an AUROC of 0.83 (0.81–0.85) and a misclassification rate of 0.188. The optimal Youden's J statistic was 0.57 at a probability threshold of 0.35. By using this threshold, the model had a sensitivity of 77.9% (0.67–0.87), a specificity of 79.2% (0.72–0.85), a PPV of 61.6% (0.54–0.69), and a NPV of 89.3% (0.85–0.93). Variable importance is presented in Figure 3, which demonstrated that surgical procedure description and admission diagnosis were relatively the most important variables, whereas sex and hospital location were relatively the least important variables.

Impact on avoidable bed-days

An impact analysis was performed to obtain an estimate of the unnecessary prolonged stays (ie, the potential avoidable bed-days). For this purpose, we used a hospital interventions probability threshold of 15%—that is, the probability of no hospital interventions (=1-hospital interventions probability [%]) must be larger than 85%. Safe discharge was predicted in 92 out of 579 episodes in 2019 (ie, the probability did not exceed the 15% threshold while they actually remained admitted). Application of the previously described formula resulted in 238 avoidable bed-days that could have been saved in the 3 hospital locations, while only patients in 7 episodes (7.6% [7/92]) would need to be readmitted to the hospital from the nursing home. The avoidable bed-days were 366 (based on safe discharge predicted in 147 out of 677 episodes) and 309 (based on safe discharge predicted in 132 out of 656 episodes) for 2017 and 2018, respectively.

Discussion

This study demonstrated that a previously developed and validated ML concept, able to predict safe hospital discharge after the second postoperative day, can also be translated to nonacademic hospitals' surgical populations.[20] We tested 4 models and found that the reduced model that additionally used the admission diagnosis ($n = 12$ variables) performed best. Further, the results have nicely shown that the ML concept may be used to save on

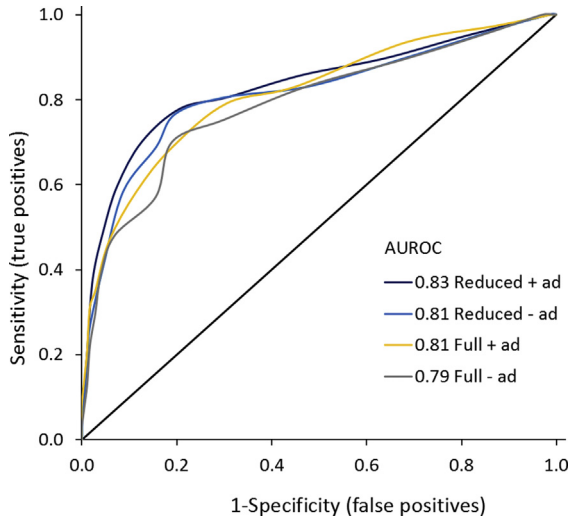

Figure 2. Receiver operating characteristic (ROC) curves. Models' discriminative performance on the test data set for the 4 models ($n = 249$ patients). The full models ("Full") used 17 variables ± admission diagnosis. The reduced models ("Reduced") used 11 variables ± admission diagnosis. *AD*, admission diagnosis; *AUROC*, area under the receiver operating characteristic curve.

avoidable bed-days; 913 bed-days could have been avoided in this particular surgical population between 2017 and 2019. We specifically reported this period, just before the global COVID-19 pandemic put a strain on operating room capacity starting from early 2020 (which probably could lead to an unrepresentative number of surgical admissions).[26]

Whereas most studies remain in levels 3 and 4 (ie, development and prototype phase) on the AI levels of clinical readiness scale, we conducted a level 5 study (ie, external validation), which is an important step to warrant safe clinical implementation.[9,27–29] Despite the increasing interest in AI and ML, few models have undergone external validation, ranging from 7% in the ICU to 6% and 30% in radiology/imaging.[9,13,30] The generalizability of such models is the subject of debate because it is threatened by multiple factors.[31] AI models are typically context-specific and susceptible to variations in care practices (local, surgical), patient populations, and information systems, which can affect model performance; it has even been argued that entire generalizability may be a utopia.[31] To rule out such potential differences, it has been suggested that a new model should be trained on local patient data (which is also known as "site-specific training"), as in our study.[32] Since we trained a model on local patient data and did not validate the exact model (including its underlying variable distributions), variables did not need to be mapped to a common ontology (ie, transforming variable units and descriptions such as surgical procedure description to a uniform data format that can be used by multiple hospitals), which is time and resource consuming but often required due to variations in local terminologies.[33] Furthermore, we only used routinely collected data, available until the second postoperative day, which are often widely available in hospital information systems as standard of care data collection; the same variables were available during development and external validation. This suggests that the ML concept could be seamlessly integrated with other hospital information systems while incorporating differences in site-specific care practices and populations and thereby optimize surgical discharge from cure to care facilities.

Although the ML concept was developed in the largest academic tertiary hospital of The Netherlands where the most complex cancer surgeries are performed (eg, esophageal cancer, liver surgery, sarcomas), this external validation took place in hospitals providing secondary care and still achieved an AUROC of 0.83 (95% CI [0.81–0.85]), which is comparable to 0.88 (0.83–0.93) previously.[20] In addition, with respect to other metrics such as sensitivity and specificity, external validation achieved similar results with an even higher PPV of 61.6% (95% CI [0.54–0.69]) compared to 57.6% (95% CI [0.45–0.70]). However, the challenge in selecting the appropriate prediction probability threshold, which is used to calculate these metrics, is to weigh hospital needs (eg, improved patient flow) with the number of readmissions (ie, potential harm) considered clinically acceptable, as we have previously extensively described.[20] In addition, in this study, a reduced model outperformed the full model (Table IV). An explanation could be that redundant variables such as the "expected duration in operating room" (which is highly correlated to "total length in operating room") contribute little independent information and may even negatively affect model performance (ie, they can make the model unnecessarily complex).[22,34] Furthermore, admission diagnosis improved model performance (AUROC improved 2% in the full as well as the reduced models), which may be due to a better reflection of patients' preoperative state.[20]

Over the last decades, efforts have been made to expedite postoperative discharge—for example, by the introduction of the Enhanced Recovery After Surgery pathways.[6,7] However, unnecessary prolonged stays after major surgery are still a common problem and pose a challenge to capacity management due to the restricted number of hospital beds.[5] Over the years, several ML models have been developed to predict adverse postoperative outcomes such as sepsis, mortality, or unplanned readmissions.[35–38] Such models have high predictive value, are useful during the preoperative consent procedure, and improve patients' satisfaction.[39–42] These models can also benefit the logistics of the operating room and postoperative care planning.[8] However, these predictions are made before the actual surgery, do not involve individual perioperative factors, and, importantly, are not yet clinically implemented; our study is a crucial step toward clinical implementation. Furthermore, such models often arise from a wealth of data and technical opportunities rather than that they are developed to generate actionable clinical output, which makes it difficult for clinicians to interpret.[32] In contrast, our ML concept

Table IV
Model performance results

Model	Number Variables	AUROC (95% CI)	Sensitivity, % (95% CI)	Specificity, % (95% CI)	PPV, % (95% CI)	NPV, % (95% CI)	Optimal Threshold*
Reduced model + AD	12	0.83 (0.81–0.85)	77.9% (0.67–0.87)	79.2% (0.72–0.85)	61.6% (0.54–0.69)	89.3% (0.85–0.93)	0.35
Reduced model - AD	11	0.81 (0.79–0.83)	80.5% (0.70–0.89)	69.9% (0.63–0.77)	54.4% (0.48–0.61)	89.0% (0.84–0.93)	0.30
Full model + AD	18	0.81 (0.79–0.83)	79.2% (0.68–0.88)	69.4% (0.62–0.76)	53.5% (0.47–0.60)	88.2% (0.83–0.92)	0.30
Full model - AD	17	0.79 (0.77–0.81)	70.1% (0.59–0.80)	80.9% (0.74–0.86)	62.1% (0.54–0.70)	85.9% (0.81–0.90)	0.35

Model performance was assessed on the test data set ($n = 249$). The full models used 17 variables ± admission diagnosis (AD). The reduced models used 11 variables ± admission diagnosis (AD).
AD, admission diagnosis; *AUROC*, area under the receiver operating characteristics; *CI*, confidence interval; *NPV*, negative predictive value; *PPV*, positive predictive value.
* The optimal classification threshold was calculated on the validation data set using the Youden's J statistic.

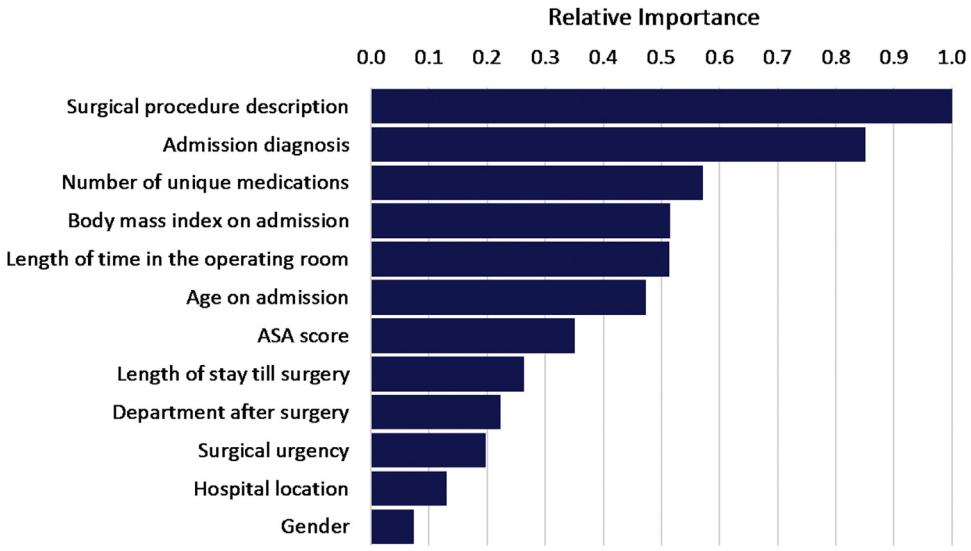

Figure 3. Nomogram of the reduced model with admission diagnosis. This model used a total of 12 variables. *ASA*, American Society of Anesthesiologists; *BMI*, body mass index.

predicts a one-time discharge safety score instead of a particular clinical condition or complication such as sepsis (ie, a patient can either safely recover outside the hospital or not, which can be presented in more detail for different risk groups). Of note, while patients might not need care that is strictly provided by hospitals, some may still require care that cannot be arranged at home (eg, pain management) and, as such, may need to be discharged to a postoperative nursing home before being discharged home.

Although the ML concept currently generates a one-time discharge safety score after the second postoperative day, it may be extended to other postoperative days using similar aims and variables. This is particularly interesting since multiple ML studies that analyzed longitudinal data (eg, prediction of hypotension or intracranial hypertension) demonstrated that predictive performance increases as time to outcome decreases (ie, it may be easier to predict safe discharge later after surgery as more data become available).[43–45] Nevertheless, the clinical impact of such models will decrease as postoperative days pass and the avoidable bed-days decrease. Furthermore, the current concept can be extended to other surgical populations such as cardiothoracic, transplant, vascular, plastic, and reconstructive.

At this moment, the ML concept performed well in both the development and validation populations. Because clinical utility, usability, and health benefits still need to be determined, a clinical study is warranted. Since AI model output cannot inform clinical decision-making on its own but needs to be integrated in an end-to-end solution to appropriately convey information to the clinicians (end-users), a next step is to integrate the current ML concept with such a display and test the end-to-end solution in a clinical implementation study. For example, Barda et al[46] proposed a framework to design user-centered displays to enhance communication and provide explanations of ML model output toward the end-users.

Limitations

Some limitations of this study must be addressed. First, we validated the ML concept instead of the exact model that was previously developed. This is called site-specific training and requires sufficiently large data sets to locally train a model, which may not always be feasible in smaller hospitals. Nonetheless, by adopting this approach we ensured that the ML concept adapts to local care practices and patient populations, which are often highly predictive and may therefore yield better outcomes for this particular surgical population.[31]

Second, the predictor variables consisted of routinely collected clinical data variables and were thus limited to those entered in the electronic health record. Some of these variables rely on manual input and are hence prone to human error. Thus, a future clinical implementation study should be accompanied by efforts to enhance and ensure data quality.

In conclusion, this study demonstrated that an earlier developed ML concept can be used to predict safe discharge in different surgical populations and hospital settings (academic versus nonacademic) by training a model on local patient data. The consistently strong performance suggests that the ML concept can be used to guide capacity challenges by reducing the number of avoidable bed-days. As a next step, a clinical implementation study is needed to assess DESIRE's clinical utility and usability.

Funding/Support

No funding declared.

Conflicts of interest/Disclosure

Diederik Gommers has received speakers fees and travel expenses from Dräger, GE Healthcare (medical advisory board 2009–12), Maquet, and Novalung (medical advisory board 2015–18). All other authors declare no competing interests. Joost Huiskens currently works as industry expert healthcare at SAS Institute. Edwin van Unen currently works as principal analytics consultant at SAS Institute. No financial relationships exists that could be construed as a potential conflict of interest. All other authors declare no conflicts of interests.

References

1. Wick EC, Pierce L, Conte MC, Sosa JA. Operationalizing the operating room: ensuring appropriate surgical care in the era of COVID-19. *Ann Surg*. 2020;272:e165–e167.
2. Rojas-Garcia A, Turner S, Pizzo E, Hudson E, Thomas J, Raine R. Impact and experiences of delayed discharge: a mixed-studies systematic review. *Health Expect*. 2018;21:41–56.

3. Shojania KG, Duncan BW, McDonald KM, Wachter RM. Safe but sound: patient safety meets evidence-based medicine. *JAMA*. 2002;288:508–513.
4. Covinsky KE, Palmer RM, Fortinsky RH, et al. Loss of independence in activities of daily living in older adults hospitalized with medical illnesses: increased vulnerability with age. *J Am Geriatr Soc*. 2003;51:451–458.
5. Jerath A, Sutherland J, Austin PC, et al. Delayed discharge after major surgical procedures in Ontario, Canada: a population-based cohort study. *CMAJ*. 2020;192:E1440–E1452.
6. Lassen K, Soop M, Nygren J, et al. Consensus review of optimal perioperative care in colorectal surgery Enhanced Recovery After Surgery (ERAS) group recommendations. *Arch Surg-Chicago*. 2009;144:961–969.
7. Varadhan KK, Neal KR, Dejong CHC, Fearon KCH, Ljungqvist O, Lobo DN. The enhanced recovery after surgery (ERAS) pathway for patients undergoing major elective open colorectal surgery: a meta-analysis of randomized controlled trials. *Clin Nutr*. 2010;29:434–440.
8. Hashimoto DA, Rosman G, Rus D, Meireles OR. Artificial intelligence in surgery: promises and perils. *Ann Surg*. 2018;268:70–76.
9. van de Sande D, van Genderen ME, Huiskens J, Gommers D, van Bommel J. Moving from bytes to bedside: a systematic review on the use of artificial intelligence in the intensive care unit. *Intens Care Med*. 2021;47:750–760.
10. Secinaro S, Calandra D, Secinaro A, Muthurangu V, Biancone P. The role of artificial intelligence in healthcare: a structured literature review. *BMC Med Inform Decis*. 2021:21.
11. Monsalve-Torra A, Ruiz-Fernandez D, Marin-Alonso O, Soriano-Paya A, Camacho-Mackenzie J, Carreno-Jaimes M. Using machine learning methods for predicting inhospital mortality in patients undergoing open repair of abdominal aortic aneurysm. *J Biomed Inform*. 2016;62:195–201.
12. Soguero-Ruiz C, Hindberg K, Rojo-Alvarez JL, et al. Support vector feature selection for early detection of anastomosis leakage from bag-of-words in electronic health records. *IEEE J Biomed Health Inform*. 2016;20:1404–1415.
13. Kim DW, Jang HY, Kim KW, Shin Y, Park SH. Design characteristics of studies reporting the performance of artificial intelligence algorithms for diagnostic analysis of medical images: results from recently published papers. *Korean J Radiol*. 2019;20:405–410.
14. Wong A, Otles E, Donnelly JP, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Intern Med*. 2021.
15. Wiens J, Saria S, Sendak M, et al. Do no harm: a roadmap for responsible machine learning for health care (vol 25, pg 1337, 2019). *Nature Med*. 2019;25:1627.
16. Steyerberg EW, Moons KG, van der Windt DA, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Med*. 2013;10:e1001381.
17. Safavi KC, Khaniyev T, Copenhaver M, et al. Development and validation of a machine learning model to aid discharge processes for inpatient surgical care. *JAMA Netw Open*. 2019;2:e1917221.
18. Lazar DJ, Kia A, Freeman R, Divino CM. A machine learning model enhances prediction of discharge for surgical patients. *J Am Coll Surgeons*. 2020;231:S132.
19. Levin SB, Toerper M, Debraine A, et al. Machine-learning-based hospital discharge predictions can support multidisciplinary rounds and decrease hospital length-of stay. *BMJ Innovations*. 2021;7:414–421.
20. van de Sande D, van Genderen ME, Verhoef C, et al. Predicting need for hospital-specific interventional care after surgery using electronic health record data. *Surgery*. 2021.
21. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Br J Surg*. 2015;102:148–158.
22. Frank E, Harrell J. *Regression Modeling Strategies With Applications to Linear Models, Logistic Regression, and Survival Analysis*. New York: Springer; 2015.
23. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning, Data Mining, Inference and Prediction*. New York: Springer; 2017.
24. Park SH, Han K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. *Radiology*. 2018;286:800–809.
25. Youden WJ. Index for rating diagnostic tests. *Cancer*. 1950;3:32–35.
26. Stoss C, Steffani M, Kohlhaw K, Rudroff C, Staib L, Hartmann D, et al. The COVID-19 pandemic: impact on surgical departments of non-university hospitals. *BMC Surg*. 2020;20:313.
27. Riley. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges (vol 353, i3140, 2016). *BMJ*. 2019:365.
28. Ramspek CL, Jager KJ, Dekker FW, Zoccali C, van Diepen M. External validation of prognostic models: what, why, how, when and where? *Clin Kidney J*. 2021;14:49–58.
29. Fleuren LM, Thoral P, Shillan D, Ercole A, Elbers PWG. Right Data Right Now Collaborators. Machine learning in intensive care medicine: ready for take-off? *Intensive Care Med*. 2020;46:1486–1488.
30. Liu X, Faes L, Kale AU, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit Health*. 2019;1:e271–e297.
31. Futoma J, Simons M, Panch T, Doshi-Velez F, Celi LA. The myth of generalisability in clinical research and machine learning in health care. *Lancet Digit Health*. 2020;2:e489–e492.
32. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med*. 2019;17:195.
33. Fleuren LM, de Bruin DP, Tonutti M, et al. Large-scale ICU data sharing for global collaboration: the first 1633 critically ill COVID-19 patients in the Dutch Data Warehouse. *Intens Care Med*. 2021;47:478–481.
34. Nicodemus KK, Malley JD. Predictor correlation impacts machine learning algorithms: implications for genomic studies. *Bioinformatics*. 2009;25:1884–1890.
35. Hicks CW, Bronsert M, Hammermeister KE, et al. Operative variables are better predictors of postdischarge infections and unplanned readmissions in vascular surgery patients than patient characteristics. *J Vasc Surg*. 2017;65:1130–1141e9.
36. Corey KM, Kashyap S, Lorenzi E, et al. Development and validation of machine learning models to identify high-risk surgical patients using automatically curated electronic health record data (Pythia): a retrospective, single-site study. *PLoS Med*. 2018;15:e1002701.
37. Xue B, Li DW, Lu CY, et al. Use of machine learning to develop and evaluate models using preoperative and intraoperative data to identify risks of postoperative complications. *JAMA Netw Open*. 2021;4.
38. Bonde A, Varadarajan KM, Bonde N, et al. Assessing the utility of deep neural networks in predicting postoperative surgical complications: a retrospective study. *Lancet Digit Health*. 2021;3:e471–e485.
39. Meguid RA, Bronsert MR, Juarez-Colunga E, Hammermeister KE, Henderson WG. Surgical Risk Preoperative Assessment System (SURPAS) I: parsimonious, clinically meaningful groups of postoperative complications by factor analysis. *Ann Surg*. 2016;263:1042–1048.
40. Meguid RA, Bronsert MR, Juarez-Colunga E, Hammermeister KE, Henderson WG. Surgical Risk Preoperative Assessment System (SURPAS) II: parsimonious risk models for postoperative adverse outcomes addressing need for laboratory variables and surgeon specialty-specific models. *Ann Surg*. 2016;264:10–22.
41. Meguid RA, Bronsert MR, Juarez-Colunga E, Hammermeister KE, Henderson WG. Surgical Risk Preoperative Assessment System (SURPAS) III: accurate preoperative prediction of 8 adverse outcomes using 8 predictor variables. *Ann Surg*. 2016;264:23–31.
42. Wiesen BM, Bronsert MR, Aasen DM, et al. Use of Surgical Risk Preoperative Assessment System (SURPAS) and patient satisfaction during informed consent for surgery. *J Am Coll Surgeons*. 2020;230:1025.
43. Hatib F, Jian Z, Buddi S, et al. Machine-learning algorithm to predict hypotension based on high-fidelity arterial pressure waveform analysis. *Anesthesiology*. 2018;129:663–674.
44. Lee S, Lee HC, Chu YS, Song SW, Ahn GJ, Lee H, et al. Deep learning models for the prediction of intraoperative hypotension. *Br J Anaesth*. 2021;126:808–817.
45. Raj R, Luostarinen T, Pursiainen E, et al. Machine learning-based dynamic mortality prediction after traumatic brain injury. *Sci Rep*. 2019;9:17672.
46. Barda AJ, Horvat CM, Hochheiser H. A qualitative research framework for the design of user-centered displays of explanations for machine learning model predictions in healthcare. *BMC Med Inform Decis*. 2020;20.