Check for updates

Review Article

Performance and usability of pre-operative prediction models for 30-day peri-operative mortality risk: a systematic review

J. E. M. Vernooij,[1] **N. J. Koning,**[1] **J. W. Geurts,**[2] **S. Holewijn,**[3] **B. Preckel,**[4] **C. J. Kalkman**[5] and **L. M. Vernooij**[6]

1 Anaesthetist, 2 Research Co-ordinator, Department of Anaesthesia, Rijnstate Hospital, the Netherlands
3 Senior Researcher, Department of Vascular Surgery, Rijnstate Hospital, the Netherlands
4 Professor, Department of Anaesthesia, Amsterdam UMC, Amsterdam, the Netherlands
5 Emeritus Professor, University Medical Centre, Utrecht, the Netherlands
6 Clinical Epidemiologist, Department of Anaesthesia, University Medical Centre Utrecht, the Netherlands

Summary

Estimating pre-operative mortality risk may inform clinical decision-making for peri-operative care. However, pre-operative mortality risk prediction models are rarely implemented in routine clinical practice. High predictive accuracy and clinical usability are essential for acceptance and clinical implementation. In this systematic review, we identified and appraised prediction models for 30-day postoperative mortality in non-cardiac surgical cohorts. PubMed and Embase were searched up to December 2022 for studies investigating pre-operative prediction models for 30-day mortality. We assessed predictive performance in terms of discrimination and calibration. Risk of bias was evaluated using a tool to assess the risk of bias and applicability of prediction model studies. To further inform potential adoption, we also assessed clinical usability for selected models. In all, 15 studies evaluating 10 prediction models were included. Discrimination ranged from a c-statistic of 0.82 (MySurgeryRisk) to 0.96 (extreme gradient boosting machine learning model). Calibration was reported in only six studies. Model performance was highest for the surgical outcome risk tool (SORT) and its external validations. Clinical usability was highest for the surgical risk pre-operative assessment system. The SORT and risk quantification index also scored high on clinical usability. We found unclear or high risk of bias in the development of all models. The SORT showed the best combination of predictive performance and clinical usability and has been externally validated in several heterogeneous cohorts. To improve clinical uptake, full integration of reliable models with sufficient face validity within the electronic health record is imperative.

Correspondence to: J. E. M. Vernooij
Email: jvernooij@rijnstate.nl

Accepted: 6 January 2023
Keywords: 30-day peri-operative mortality; clinical usability; peri-operative mortality risk; prediction; pre-operative risk assessment; risk calculator
Presented in part at the Netherlands Society of Anesthesiologists Annual Science Symposium, Utrecht, the Netherlands, September 2022.
This article is accompanied by an editorial by A. Lee and S. R. Moonesinghe, *Anaesthesia* 2023; **78**: 547–550.
Twitter: @sjaaksaam

© 2023 The Authors. *Anaesthesia* published by John Wiley & Sons Ltd on behalf of Association of Anaesthetists.
This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

Introduction

Globally, over 300 million surgical procedures are performed annually [1]. Early postoperative mortality rates vary from 1% to 22% in patients undergoing a wide range of non-cardiac surgical procedures in Europe [2, 3]. High-risk patients who require non-cardiac surgical procedures account for 84% of postoperative deaths [4]. Reliable pre-operative risk prediction for high-risk surgical patients is needed to promote pre-operative optimisation and appropriate resource allocation, including elective postoperative admission to critical care [5, 6]. Furthermore, pre-operative risk prediction may improve peri-operative clinical decision-making including informed consent discussions and shared decision-making with the multidisciplinary team [7]. The European Society for Cardiology, the American College of Cardiologists, the American Heart Association and the Canadian Society of Anesthesiologists recommend using pre-operative prediction models to estimate peri-operative mortality risk [8–11]. However, the use of pre-operative prediction models in clinical practice remains limited [7, 12–15]. Before implementation of pre-operative prediction models, internal and external validation are required to demonstrate acceptable predictive performance in relevant populations [16, 17]. Furthermore, good clinical usability is essential for clinical uptake, including low burden of data collection; ease of use and non-proprietary, reliable models [17]. The objective of this systematic review was to identify, describe and appraise reliability and clinical usability of pre-operative prediction models for 30-day postoperative all-cause mortality in adult non-cardiac surgery patients.

Methods

We used the preferred reporting items for systematic reviews and meta-analyses [18, 19] and the critical appraisal and data extraction for systematic reviews of prediction modelling studies (CHARMS) checklist to report the systematic review, respectively[20]. This review was registered with PROSPERO (CRD42020155049).

All original research reports on the development, updates to, or external validation of a pre-operative risk model to predict 30-day mortality for surgical cohorts with more than two surgical subspecialties other than cardiac surgery were included. Studies that only reported on the prediction of in-hospital mortality, conference abstracts and studies published in languages other than English, German or Dutch were excluded. Whilst in-hospital mortality is easier to assess, we used 30-day mortality as this metric is more accurate and comparable, facilitating both early at-home and institutional mortality [21].

PubMed and Embase were searched from inception to 14 December 2022, using search terms to identify articles reporting on the development, update or external validation of prediction models to predict 30-day peri-operative mortality. The search terms consisted of the Ingui filter (with additional string) to identify prognostic and diagnostic models [22], combined with terms related to surgery and postoperative complications (online Supporting Information Table S1). For Embase, the PubMed search was adapted according to the rules for adaptation.

Two reviewers independently assessed eligibility based on the title and abstract (SG and JV). Disagreements were resolved by a third investigator (JG). Two reviewers (JG and JV) than screened publication for potential inclusion in the review. Again, disagreements were resolved by a third investigator (NK). Subsequently, we searched SCOPUS for manuscripts citing the retrieved models and hand-searched the reference lists of included studies for potentially missed publications.

According to recommendations in the CHARMS checklist, one author (JV) extracted the data following a preconstructed data extraction form [20]. Items extracted were as follows: patient characteristics; the number and type of candidate predictors; the predictors in the prediction model described; the sample size of the development or external validation cohort; the number of patients with the outcome of interest (i.e. 30-day postoperative mortality); the number of hospitals involved in the study; the number of missing data and handling of missing values; the method of modelling, including shrinkage methods; performance measures regarding discrimination (e.g. c-statistic), calibration (e.g. calibration plot and Hosmer–Lemeshow test); and overall performance (Brier score, net reclassification index). Risk of bias and concerns for applicability were assessed using the prediction model risk of bias assessment tool [23, 24]. Risk of bias assessment was executed per model (development, validation or update). Assessment of risk of bias and concern of applicability for the retrieved studies were performed independently by two researchers (JG/LV and JV) [23]. Conflicts were resolved by a third reviewer (NK).

A pre-operative mortality risk prediction model is designed to guide clinical decision-making. Good clinical usability is necessary to improve clinical implementation of the model. Since guidance on scoring the clinical usability of prediction models does not exist, we followed recommendations as previously described [7, 15–17, 25]. Items assessed include the burden of data collection; integration in electronic health records; objectivity in predictor definitions; whether the predictive model has been externally validated and whether the model is periodically updated [7, 15–17, 25]. We used all items to

13652044, 2023, 5, Downloaded from https://associationofanaesthetists-publications.onlinelibrary.wiley.com/doi/10.1111/anae.15988 by Cochrane Netherlands, Wiley Online Library on [02/02/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

develop a scoring system for clinical usability. Definitions and grading of these items are presented in Table 1. The definitions of the items on clinical usability are explained in further detail in online Supporting Information Appendix S1.

Results were summarised using descriptive statistics. We assessed the models on discrimination and calibration. For discriminative predictive performance, c-statistics were collected. We considered a c-statistic of ≤ 0.7 as showing poor predictive performance, 0.7–0.9 as moderate predictive performance and a c-statistic > 0.9 as showing high predictive performance as defined before [29]. For the calibration measures, including the Hosmer–Lemeshow test, a p value of > 0.05 was considered to indicate that there was no evidence of a lack of model fit. Overall performance was reported with a Brier score, a combination of discrimination and calibration properties of a model. A Brier score of 0 means perfect accuracy, and a Brier score of 1 means total inaccuracy. Reclassification was assessed using the net reclassification index [30].

Results

In total, 31,436 records were identified through database and hand-searching. After removal of duplicates, 18,090 records were screened on title and abstract, from which 106 full-text articles were retrieved. After the full-text articles were screened, 15 were included in this review reporting on 10 prediction models (Fig. 1 and online Supporting Information Table S2). Included articles describe the development [25, 31–36] (seven studies); a combination of a new model and its external validation [37] (one study); the external validation [38–40] (three studies); a combination of an external validation of a current model combined with the development of a new model [12, 41] (two studies); or an update of a current prediction model [42, 43] (two studies). Prediction models identified were the surgical outcome risk tool (SORT) [31]; NewZealandRISK (NZRISK) [41]; SORT clinical judgement [12]; surgical risk pre-operative assessment system (SURPAS) [25]; surgical risk calculator (SRC) [32]; risk quantification index (RQI) [34]; surgical mortality probability model (S-MPM) [33]; MySurgeryRisk [36]; Pythia [35]; and the extreme gradient boosting (XGB) machine learning model by Choi et al. [37]. Five studies reported according to the TRIPOD guidelines [12, 35, 36, 40, 41, 44]. More detailed information on the prediction models is presented in online Supporting Information Appendix S2.

In all, 10 of the 15 identified studies were multicentre (Table 2). The cohorts varied in sample size, with a median (IWR [range]) of 168,442 36,451–792,450 [11,129–4,600,000]) patients. The patient inclusion period varied from 1 week [12, 31] to 10 years [42]. Data were collected between 2005 and 2021. Seven studies used data from the American College of Surgeons National Surgical Quality Improvement Program (ACS NSQIP) (online Supporting Information Table S2); one used the New Zealand National Minimal Dataset [41]; another used the second Sprint National Anaesthesia Project: epidemiology of critical care provision after surgery (SNAP-2; EPICCS) [12]; and seven used a mix of administrative and hospital data [35–38, 40].

Table 1 Grading of clinical usability qualities of 30-day mortality risk prediction models.

Qualities	Definition and grading
Low burden of data collection [7, 25]	≤11 predictors = 2 points >11 predictors = 0 points (except machine learning models)

Automated prediction model built into electronic health record [26]	At least one example = 2 points
	Partially = 1 point
	No = 0 points

Uses objective data Objective data were defined as data based on facts (e.g. age, laboratory measurements), unlikely to be influenced by personal interpretation; subjective data: data prone to interpretation, such as ASA physical status, dependency, surgical complexity	Only objective data = 2 points Mix of subjective (based on interpretation) and objective data = 1 point Subjective data only = 0 points
Be updated periodically [25] Since healthcare performance and patient outcome change over time, regression coefficients should be adapted every 5 years	Yes = 2 points; No = 0 points
Transparency of risk equation [27] The risk equation is available in the public domain	Yes = 2 point; No = 0 points
External validation on heterogeneous noncardiac cohorts [17, 28]	Yes = 2 points; No = 0 points

For an explanation of awarding of points, see online Supporting Information Appendix S1. Maximum score is 12 points.

© 2023 The Authors. *Anaesthesia* published by John Wiley & Sons Ltd on behalf of Association of Anaesthetists.

13652044, 2023, 5, Downloaded from https://associationofanaesthetists-publications.onlinelibrary.wiley.com/doi/10.1111/anae.15988 by Cochrane Netherlands, Wiley Online Library on [02/02/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

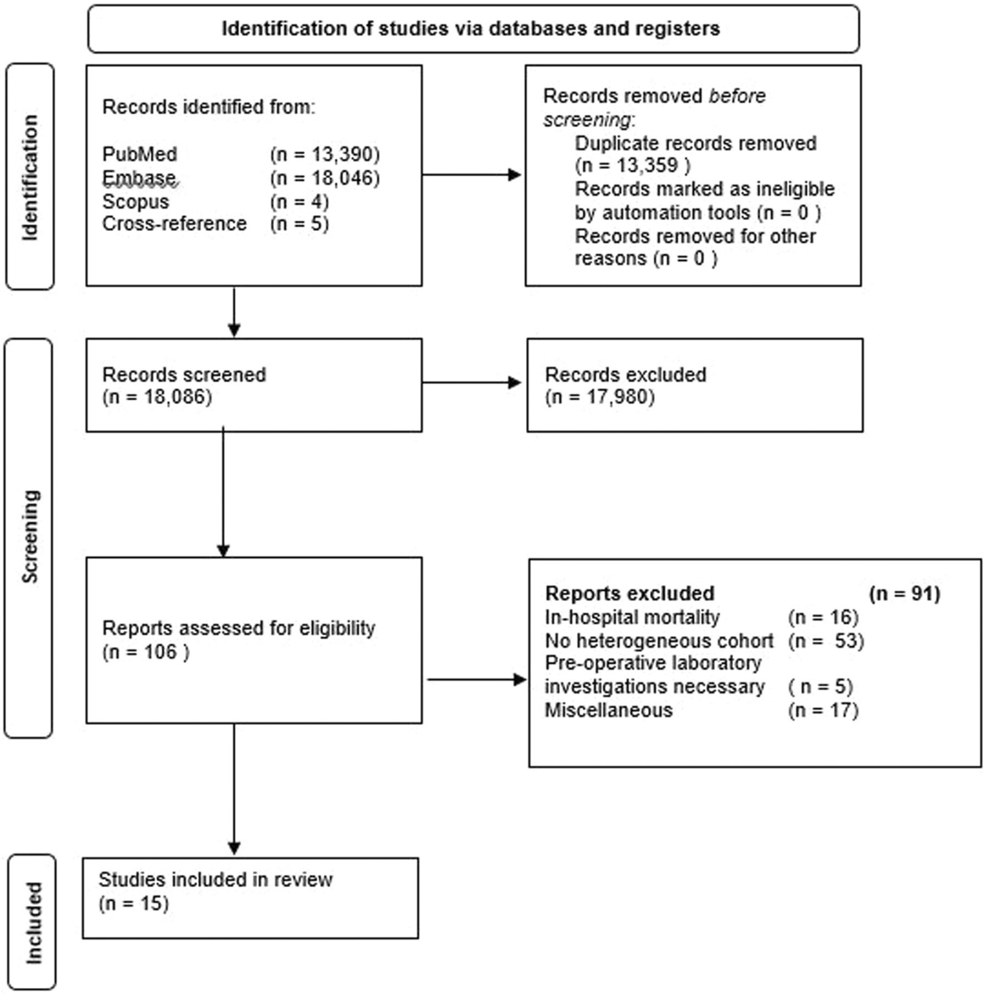

Identification of studies via databases and registers
Identification
Records identified from:
PubMed (n = 13,390)
Embase (n = 18,046)
Scopus (n = 4)
Cross-reference (n = 5)
Records removed *before screening*:
Duplicate records removed (n = 13,359)
Records marked as ineligible by automation tools (n = 0)
Records removed for other reasons (n = 0)
Screening
Records screened (n = 18,086)
Records excluded (n = 17,980)
Reports assessed for eligibility (n = 106)
Reports excluded (n = 91)
In-hospital mortality (n = 16)
No heterogeneous cohort (n = 53)
Pre-operative laboratory investigations necessary (n = 5)
Miscellaneous (n = 17)
Included
Studies included in review (n = 15)

Figure 1 Study flow diagram of database searches.

The patient characteristics from the different studies are presented in online Supporting Information Table S3. The studies that reported on sex included more women than men (median (IQr [range]) female patients 52.8% 49.4–56.8% [41.9–65%])). The mean (SD) reported age was 56.5 (4.2) y. Most of the studies included patients that underwent surgery from one of the following non-cardiac surgery subspecialties: vascular surgery; abdominal surgery; thoracic surgery; neurosurgery; musculoskeletal surgery; plastic surgery; urology; gynaecology; orthopaedics; otolaryngology and other surgery (online Supporting Information Table S2). Three studies additionally included patients that underwent cardiac surgery [35, 36, 43] (online Supporting Information Table S2). Six studies did not report on urgency of surgery [34, 35, 37, 38, 42, 43]. Among the studies that did report on urgency, reports ranged from 88.3% elective surgery procedures [32] to emergency surgery procedures only [39] (online Supporting Information Table S2). Four studies included elective and emergency surgery; in those studies, patients were more often classified as ASA physical status 3 and 4 compared with studies where patients only underwent elective surgery [25, 32, 35, 42]. One study included only emergency surgery [39]. Day-case surgery was not included in six studies [12, 31, 35, 36, 38, 40, 41] (online Supporting Information Table S2). Some of the publications did not report on the included subspecialties (Online Supporting Information Table S2). The studies that did not report on surgical severity used the current procedure terminology code; the work relative value unit; the procedure specific score for severity of the surgical intervention [25, 33, 34, 38, 39, 42, 43] or reported no surgical procedure (but non-cardiac) at all [37]. None of the studies conducted subgroup validations per subspecialty (online Supporting Information Table S2).

Most studies used multivariable logistic regression to develop the prediction model. However, four used machine learning techniques [35–37, 40] (online Supporting

© 2023 The Authors. *Anaesthesia* published by John Wiley & Sons Ltd on behalf of Association of Anaesthetists.

13652044, 2023, 5, Downloaded from https://associationofanaesthetists-publications.onlinelibrary.wiley.com/doi/10.1111/anae.15988 by Cochrane Netherlands, Wiley Online Library on [02/02/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 2 Characteristics and performance measures of pre-operative 30-day mortality risk prediction models.

Model	Study	Mortality	EPV	Size development set (no. of events)	No of variables (candidate variables)	Discrimination [CI]	Calibration	Performance overall
SORT	Protopapa [31]	1.4%	3.5	11,219 (158)	6 (45)	0.91 [0.88–0.94]	HL = 12.16, p = 0.204	
SORT external validation	Campbell [41]	0.7%	210	270,105 (2053)	6 (6)	0.91 [0.90–0.92]	Intercept = −0.007 Slope = 5.32	
NZRISK SORT update	Campbell [41]	0.7%	236	270,105 (2053) Internal validation: 90,035 (684)	8 (6)	0.92 [0.91–0.93]	Intercept = −0.001 Slope = 1.12	
SORT external validation	Wong [12]	1.4%	53	22,361 (317)	6	0.90 [0.88–0.92]	HL > p < 0.001	NRI; 0.073 (p < 0.309); decision curve analysis and net benefit calculated
SORT update clinical judgement	Wong [12]	1.05%	27	17,845 (188)	7	0.92 [0.90–0.94]	HL– > p < 0.001	NRI: 0.130 (p < 0.001); Decision Curve Analysis and Net benefit calculated.
SURPAS update SRC	Meguid [25]	1.4%	631	2,275,240 (31,853)	8 (28)	0.93 [0.93–0.93]		Brier = 0.012
SURPAS update	Henderson [42]	1.2%	2187	4,600,000 (55,300)	8 (8)	0.93 [0.93–0.93]		Brier = 0.010
SURPAS external validation	Rozeboom [39]	8.8%	2266	66,720 (18,133)	8	0.86 [0.85–0.86]		Brier = 0.068
SRC	Bilimoria [32]	1.3%	875	1,414,006 (18,909)	21 (24)	0.94 [0.94–0.94]		Brier = 0.011
SRC update	Liu [43]	1.3%	1019	987,744 (12,840)	21 (21)	0.94 [0.94–0.94]	HL p-value = 0	
RQI	Dalton [34]	1.6%	390	585,265 (9363)	3 (24)	0.92 [0.91–0.92]		
RQI external validation	Sigakis [38]	1.9%	386	62,640 (1190)	3 (3)	0.89 [0.88–0.90]		Brier = 0.017
S-MPM	Glance [33]	1.3%	1334	298,772 (4004)	3 (3)	0.90 [0.90–0.90]	HL = 11.8, p value = 0.04	
MySurgeryRisk machine learning	Bihorac [36]	3.4%	6	41,148 (1750)	285 (285)	0.83 [0.81–0.85]		Sensitivity = 0.39 Specificity = 0.93 PPV = 0.18 NPV = 0.98 Accuracy = 0.92
MySurgeryRisk update	Ren [40]	1.9%	3	19,132 (429)	135	0.82 [0.80–0.84]		Sensitivity = 0.76 Specificity = 0.8 PPV = 0.06 NPV = 1.0
XGB model machine learning development	Choi [37]	0.16%	13	276,341 (442)	31	0.96 [0.94–0.98]	ICI = 0.0044	Sensitivity = 0.89 Specificity = 0.91 PPV = 0.08 NPV = 0.99 Brier = 0.0015
XGB Machine learning External validation	Choi [37]	0.34%	6	63,384 (101)	31	0.93 [0.92–0.95]	ICI = 0.0017	Sensitivity = 0.87 Specificity = 0.85 PPV = 0.17 NPV = 0.99 Brier = 0.0036
Pythia machine learning	Corey [35]	0.51%	2	66,370 (338)	194 (194)	0.92 [0.88–0.95]		Sensitivity = 0.92 Specificity = 0.59 PPV = 0.3

EPV, Events per variable; SORT, surgical outcome risk tool; HL, Hosmer–Lemeshow test; NZRISK, New Zealand Risk Calculator; NRI, net reclassification index; SURPAS, surgical preoperative assessment system; SRC, surgical risk calculator; RQI, risk quantification index; S-MPM, surgical mortality probability model; PPV, positive predictive value; NPV, negative predictive value; XGB, extreme gradient boosting; ICI, integrated calibration index.

© 2023 The Authors. *Anaesthesia* published by John Wiley & Sons Ltd on behalf of Association of Anaesthetists.

13652044, 2023, 5, Downloaded from https://associationofanaesthetists-publications.onlinelibrary.wiley.com/doi/10.1111/anae.15988 by Cochrane Netherlands, Wiley Online Library on [02/02/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Information Table S2). Updating (applying the model to a new population and adjusting the regression coefficients to obtain a new model) was performed for SRC and SORT, and recalibration for SURPAS [42]. Ren et al. updated MySurgeryRisk with an application on a mobile phone [40]. Except for the SRC and its update, all development studies conducted internal validation [32, 43]. Table 2 shows the number of candidate variables considered for inclusion in the prediction models, median (IQR [range]) 8 (6–31 [3–286]). The median (IQR [range]) number of variables while excluding machine learning models [35–37, 40] was 7 (5–8 [3–21]).

All models, except the machine-learning-derived MySurgeryRisk and Pythia, included ASA physical status as a predictor. Instead, MySurgeryRisk and Pythia used individual comorbidities as predictors for a measure of physical status. The SRC combined comorbidities and ASA physical status. All models except S-MPM included age as a predictor. Surgical complexity was a predictor in all models except for the study by Choi et al., and urgency was used in all models except for RQI and the XGB model. Thirty-day mortality was highest in the study by Rozeboom et al. [37] (8.8%) and lowest in the study by Choi et al. (0.16%) [37] (Table 2). The number of outcome events (30-day mortality) reported, ranged from 158 [31] to 55,300 [42]. Only one of the studies developing a prediction model, presented an external validation of the model in the same publication [37]. External validation in different geographical or temporal heterogeneous surgical cohorts was performed for the SORT [12, 41], RQI [38] and SURPAS [39, 42] prediction models (online Supporting Information Table S2).

C-statistic was reported as a measure of discrimination in all studies, ranging from 0.82 to 0.96 (Fig. 2 and Table 2). Discrimination was moderate in both studies on MySurgeryRisk (c-statistic = 0.83, 95%CI 0.81–0.85) [36] and (c-statistic = 0.82, 95%CI 0.80–0.84) [40] and for the external validation of SURPAS (c-statistic = 0.86, 95%CI 0.85–0.86) and RQI (c-statistic = 0.89, 95%CI 0.88–0.90) [38, 42]. The other models all scored high on discrimination (Fig. 2). The external

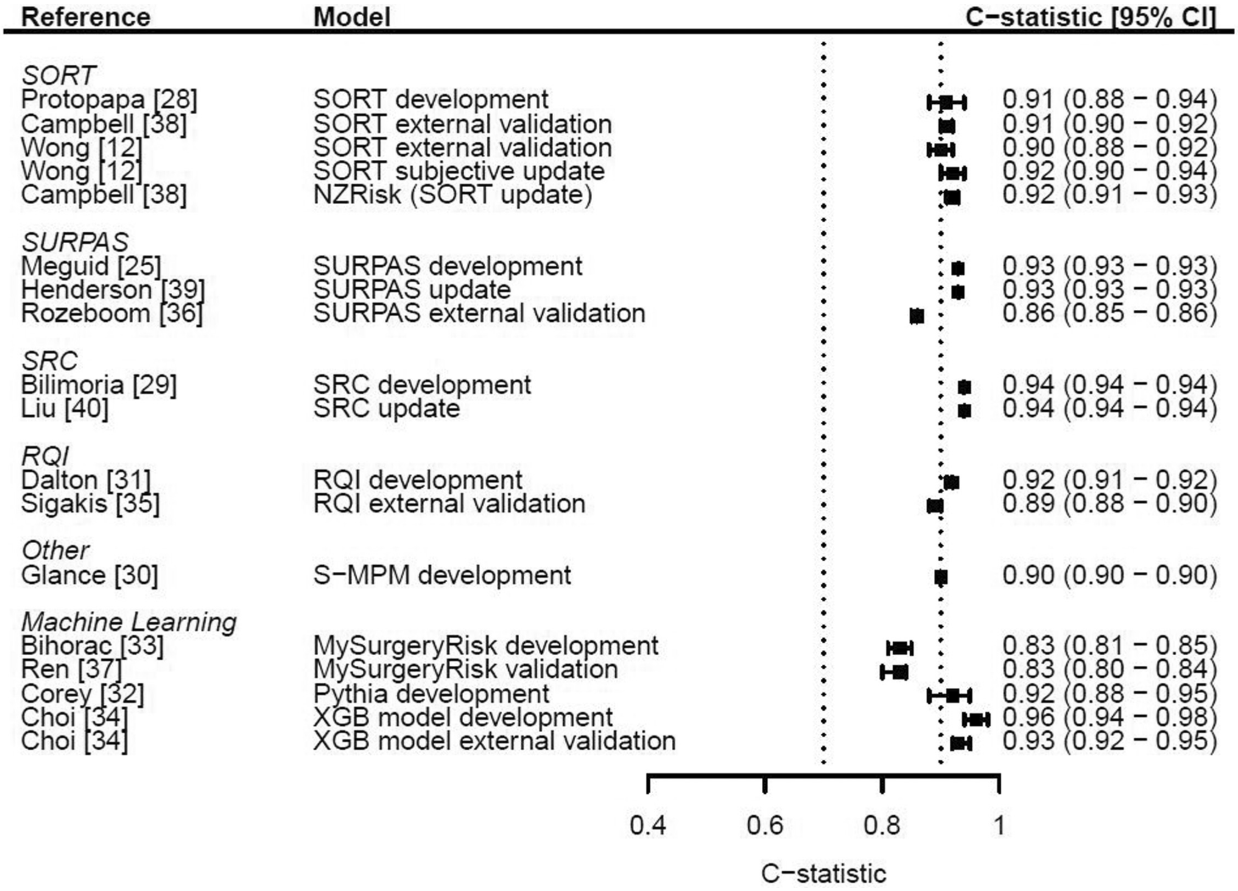


Figure 2 Forest plot of c-statistic for discussed prediction models. SORT, surgical outcome risk tool; SURPAS, surgical preoperative assessment system; SRC, surgical risk calculator; RQI, risk quantification index.

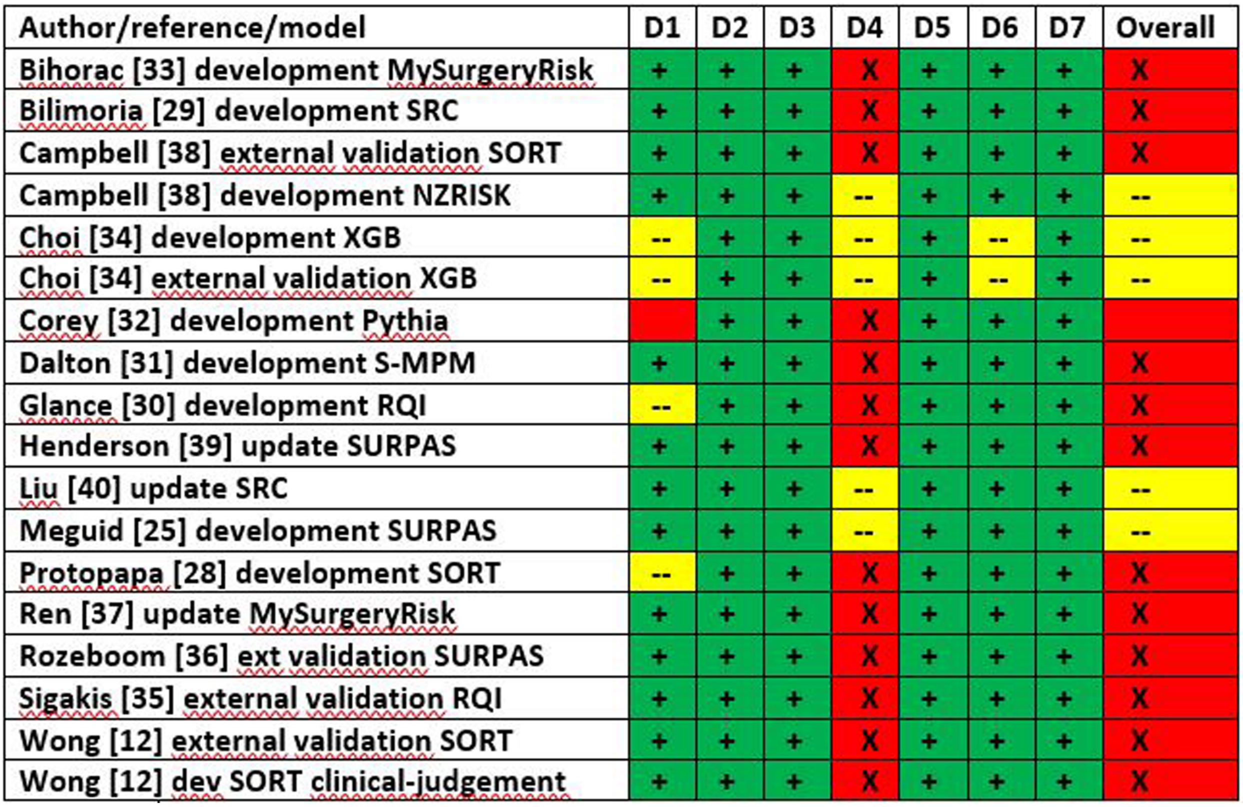
© 2023 The Authors. *Anaesthesia* published by John Wiley & Sons Ltd on behalf of Association of Anaesthetists.

13652044, 2023, 5, Downloaded from https://associationofanaesthetists-publications.onlinelibrary.wiley.com/doi/10.1111/anae.15988 by Cochrane Netherlands, Wiley Online Library on [02/02/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

validations of SORT and the XGB model (as published by Choi et al.) also showed high discrimination [12, 37, 41]. Calibration was less often described (6 out of 15 studies, Table 2). Good calibration (based on the Hosmer-Lemeshow test) was only reported for SORT [12, 41]. Choi et al. reported calibration with the integrated calibration index [37]. Wong et al. reported the net reclassification index for the external validation of SORT compared to clinical judgement alone: net reclassification index: 0.073 (95%CI 0.062–0.208); and the improvement for the combination of SORT with the clinical judgement of the team: net reclassification index: 0.130 (95%CI 0.057–0.202, p < 0.001) [12].

The overall risk of bias was judged as either unclear [37, 41–43] or high [12, 25, 31–36, 38–41], primarily because of the risk of bias in the analysis domain (Figs. 3–5). Reasons for the high risk of bias were one or more of the following aspects: not reporting any missing data or inappropriate handling of missing data; no shrinkage techniques applied in model development studies; no accounting for complexities in the data; a low number of events per variable or no calibration assessed at all [12, 31–36, 38–40, 42]. There were concerns of applicability for the participants in one study [37]. The clinical usability scoring showed that SRC, MySurgeryRisk, Pythia and the XGB model include a large number of predictors: 21, 285, 194 and 135, respectively (Tables 1 and 3). As a result, SRC has a high burden for data collection. MySurgeryRisk, Pythia and the XGB model are machine learning models and therefore have a low data collection burden, provided they are built into electronic health records and have good data validity/accuracy. The other models show a low burden for data collection. The SURPAS model has been partially integrated into an electronic health record [42] and for MySurgeryRisk, a mobile phone application was designed for clinical use [40].

All models use a mix of objective and subjective variables for mortality risk prediction. We identified two updates of SORT [12, 41], one of SURPAS [42], one of SRC [43] and one of MySurgeryRisk [40]. We did not find a model that had been structurally updated. For SRC and MySurgeryRisk, the regression formula is not publicly available for use, making external validation difficult [32, 36]. We found five external validations (SORT twice; SURPAS, RQI and XGB once) on heterogeneous cohorts [12,



Author/reference/model	D1	D2	D3	D4	D5	D6	D7	Overall
Bihorac [33] development MySurgeryRisk	+	+	+	X	+	+	+	X
Bilimoria [29] development SRC	+	+	+	X	+	+	+	X
Campbell [38] external validation SORT	+	+	+	X	+	+	+	X
Campbell [38] development NZRISK	+	+	+	--	+	+	+	--
Choi [34] development XGB	--	+	+	--	+	--	+	--
Choi [34] external validation XGB	--	+	+	--	+	--	+	--
Corey [32] development Pythia	🟥	+	+	X	+	+	+	🟥
Dalton [31] development S-MPM	+	+	+	X	+	+	+	X
Glance [30] development RQI	--	+	+	X	+	+	+	X
Henderson [39] update SURPAS	+	+	+	X	+	+	+	X
Liu [40] update SRC	+	+	+	--	+	+	+	--
Meguid [25] development SURPAS	+	+	+	--	+	+	+	--
Protopapa [28] development SORT	--	+	+	X	+	+	+	X
Ren [37] update MySurgeryRisk	+	+	+	X	+	+	+	X
Rozeboom [36] ext validation SURPAS	+	+	+	X	+	+	+	X
Sigakis [35] external validation RQI	+	+	+	X	+	+	+	X
Wong [12] external validation SORT	+	+	+	X	+	+	+	X
Wong [12] dev SORT clinical-judgement	+	+	+	X	+	+	+	X

Figure 3 Risk of bias and applicability of pre-operative 30-day mortality risk models with PROBAST [24, 70]. Red, high risk; Yellow, unclear; Green, low risk. Concerns of risk of bias: D1, participants; D2, predictors; D3, outcome; D4, analysis. Concerns of applicability for the systematic review: D5, participants; D6, predictors; D7, outcome; Overall: overall risk of bias.

© 2023 The Authors. *Anaesthesia* published by John Wiley & Sons Ltd on behalf of Association of Anaesthetists.

13652044, 2023, 5, Downloaded from https://associationofanaesthetists-publications.onlinelibrary.wiley.com/doi/10.1111/anae.15988 by Cochrane Netherlands, Wiley Online Library on [02/02/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

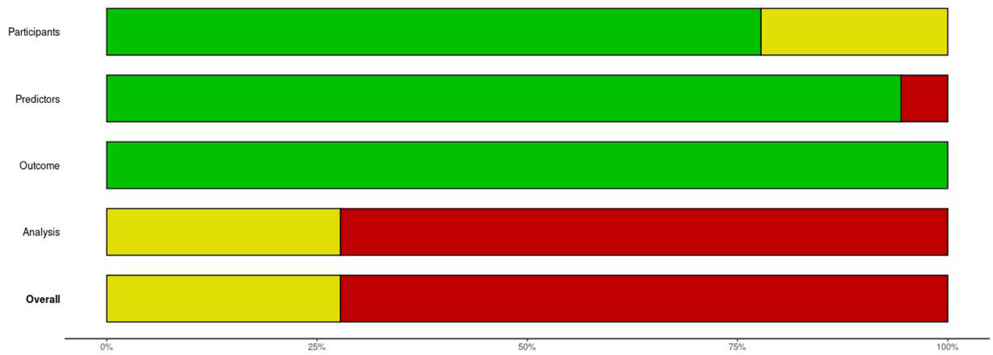


Figure 4 Summary analysis of risk of bias of discussed pre-operative 30-day risk mortality prediction models [70]. Red, high risk; yellow, unclear; green, low risk.

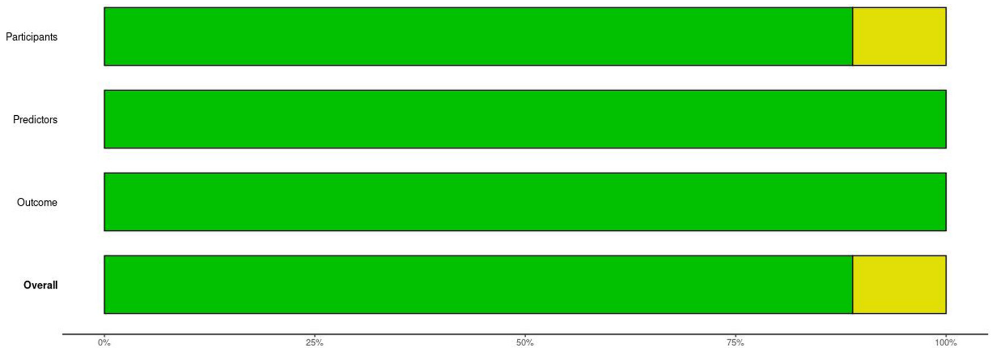


Figure 5 Summary analysis of concerns of applicability of discussed pre-operative 30-day risk mortality prediction models for systematic review. Red, high risk; yellow, unclear; green, low risk.

37–39, 41]. The SURPAS model scored highest on clinical usability with 9 out of 12 possible points, followed by SORT and RQI (8 and 7 points; Table 3).

Discussion

We found 15 studies discussing 10 pre-operative prediction models to predict 30-day mortality risk in adult patients undergoing non-cardiac surgery. Although none of the models combined high predictive accuracy with good clinical usability, SORT performed best of the identified models in the combination of predictive performance and clinical usability.

We assessed the risk of bias of studies included in this review with the prediction model risk of bias assessment tool [23]. In four studies, update SRC [43], update SURPAS [42], NZRISK [41] and the XGB model (development and external validation) [37], we found an unclear risk of bias in the analysis domain, whereas for the other models, a high risk of bias was found. This is important knowledge because some of the models are freely available on the internet and can be used for mortality risk prediction in clinical practice. Mortality risks calculated with models that are not yet made fit for the population it is used on (external validation with update if necessary) may deliver unreliable risk calculations for safe use in high-risk surgical patients. The SORT model seems the most promising model for use, but it also needs external validation on new populations before physicians can safely use it in clinical practice. Another systematic review on pre-operative mortality risk models by Reilly et al. [15] identified four prediction models as candidates with a low risk of bias for adapting in the Australian context, including S-MPM, SORT, NZRISK and the preoperative score to predict postoperative outcome (POSPOM) [15, 45]. We could not reproduce this low risk of bias, while assessing the development of the same models (except POSPOM) in the current study. High risk of bias in the development procedure of a model can induce over- or underestimation of predicted risks, which impacts on clinical decision-making [46]. Obviously, inadequate model performance can lead to erroneous estimates of predicted risk [47]. We suggest that future research focuses on external validations and clinical usability.

© 2023 The Authors. *Anaesthesia* published by John Wiley & Sons Ltd on behalf of Association of Anaesthetists.

13652044, 2023, 5, Downloaded from https://associationofanaesthetists-publications.onlinelibrary.wiley.com/doi/10.1111/anae.15988 by Cochrane Netherlands, Wiley Online Library on [02/02/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 3 Clinical usability matrix and grading of reviewed pre-operative mortality risk prediction models.

	SORT	NZRISK	SORT clinical judgement	SURPAS	SRC	RQI	S-MPM	MySurgeryRisk	Pythia	XGB model
Qualities										
Low burden of data collection	2	2	2	2	0	2	2	2	2	2
Integrated in EHR?	0	0	0	1	0	0	0	2	0	1
Objective data?	1	1	1	1	1	1	1	1	1	1
Is updated periodically	1	0	0	1	1	0	0	0	0	0
Transparency	2	2	2	2	0	2	2	0	2	0
Externally validated in heterogeneous cohorts	2	0	0	2	0	2	0	0	0	1
Total score	8	5	5	9	2	7	5	5	5	5

SORT, surgical outcome risk tool; NZRISK, New Zealand Risk Calculator; SURPAS, surgical preoperative assessment system; SRC, surgical risk calculator; RQI, risk quantification index; S-MPM, surgical mortality probability model; EHR, electronic health record; XGB, extreme gradient boosting.
Points are awarded following online Supporting Information Appendix S1. Maximum total score is 12 points. The three highest scoring models were regarded as promising.

Although the number of published pre-operative risk prediction models is increasing, thus far their use in clinical practice has been limited [7, 12–16, 48]. Possible reasons for this lack of implementation include lack of face validity, limited integration in electronic health records and increased burden of data collection. Since no published guidance exists for clinical prediction model clinical usability, we developed a framework to assess clinical usability, based on previously suggested desirable characteristics of pre-operative risk prediction models [7, 15–17, 25]. Our assessment showed that SURPAS, SORT and RQI had the highest potential for being used in daily practice. The SURPAS model uses current procedure terminology codes for surgical severity, which makes it less suitable for use outside the United States.

All studies reported on discrimination using c-statistics. Discrimination quantifies the model's ability to distinguish between patients who do, or do not, experience the event of interest [49]. The SORT, SURPAS, SRC, RQI, Pythia and XGB models showed good discrimination in their development study. Calibration was reported with a test only in six studies [12, 31, 33, 37, 41, 43].

Calibration refers to the agreement between the predicted and observed number of events and is essential in this era of precision medicine [50]. Reliable, well-calibrated predictions are necessary for informed decision-making, and to optimally allocate scarce resources such as ICU capacity [50]. The integrated calibration index was reported in one study [37] but this is only usable in comparison with other models [51]. Unfortunately, calibration is vastly under-reported. Wessler et al. noted in their review on cardiovascular prediction models that only 36% of models provided a measure of calibration [52]. In this review, the reporting rate of calibration measures was similarly low at 40%. Guidance on uniform reporting of calibration measures would make interpretation of, and comparisons between, models easier for clinicians. However, disagreement exists on the best way to calculate calibration [50, 53, 54].

External validation of prediction models is required to assess predictive performance on the targeted population and, if necessary, to update the prediction model [16, 55]. The current systematic review revealed that most pre-operative mortality risk prediction models lack external validation. In all areas of medicine, the number of publications reporting on developing new prediction models far outweighs the number of external validation studies [26]. We found that only SORT, SURPAS, RQI and the XGB models had been externally validated in heterogeneous non-cardiac surgical cohorts [12, 37–39, 41]. External validation of SURPAS showed moderate performance, although it should be noted that this validation was performed in an emergency, heterogeneous surgical cohort [39]. In contrast, SORT and XGB performed well in external validations. Importantly, the weights of the predictors in some models are proprietary and thus inaccessible to researchers. In that case, external validation can only be performed by the original developers of the prediction model [27]. For example, the coefficients of predictors of the SRC are proprietary, which hampers

© 2023 The Authors. *Anaesthesia* published by John Wiley & Sons Ltd on behalf of Association of Anaesthetists.

13652044, 2023, 5, Downloaded from https://associationofanaesthetists-publications.onlinelibrary.wiley.com/doi/10.1111/anae.15988 by Cochrane Netherlands, Wiley Online Library on [02/02/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

external validation and precludes integration of the rule into an electronic health record. The SRC model has only been externally validated on small and single-specialty cohorts, with variable performance [56–63].

Implementation of prediction models in clinical practice is likely to increase when predictive performance (especially calibration) is high in combination with clinical usability. A significant but underappreciated barrier to adopting prediction models in clinical practice is the lack of integration within electronic health records. This limitation adds considerable administrative burden to healthcare workers [26]. Stakeholders from electronic health record vendors should be involved, for faster implementation of a useful prediction score in their systems. For MySurgeryRisk, the authors developed a platform-based application [40] with integration in the electronic health record, which may prove a worthy asset to diminish the burden of data collection. However, for the other models, complete integration still needs to be completed. In general, machine learning models in clinical practice require validated and reliable data to be able to provide accurate predictions. Most predictors related to clinical care are prone to bias because clinical information is subjective or only available in a selected group of patients. Another barrier to adoption of risk models in clinical practice is the lack of face validity. Because assigning values to some categorical variables (e.g. ASA physical status 2 vs. 3) is prone to subjectivity, inter- and intra-rater variability may cause under- or overestimation of mortality risk [64–66]. As many physicians are aware of the variability problem, they may not believe the presented risks and – as a result – decide not to use the risk models in their clinical practice. Nonetheless, several anaesthesia and cardiologic societies advise using pre-operative risk prediction models [8, 9, 11, 67]. For the above reasons, prediction models should be considered valuable adjuncts during the pre-operative consultation.

We cannot overestimate the importance of adequate reporting on discrimination and calibration to assess the usability of a model [47]. In addition to predictive performance, clinical usability and adequate external validation are required measures that one should take into account to decide whether a clinical prediction model suffices for implementation [68]. Formal `impact studies´ are needed to further evaluate the clinical usability and impact of routinely using these prediction models. Impact studies are also mandatory to determine if the use of the models will improve quality of life and cost-effectiveness.

Future research should focus on external validation and updating of existing models in respective patient populations [16]. Nationwide auditing initiatives like the Peri-operative Quality Improvement Program may be used to externally validate pre-operative mortality risk prediction models on current real-world data [69]. In addition, efforts should be made to increase both the clinical uptake and usability of pre-operative mortality risk prediction models. Finally, it remains unknown how identifying high-risk non-cardiac surgical patients leads to improved care. The added value of multidisciplinary team discussions for balancing the harm–benefit ratio of the planned surgery or peri-operative management alterations in the high-risk surgical population should be further elucidated.

This study had some limitations. To increase clinical relevance, we focused on heterogeneous non-cardiac surgery adult patient cohorts, and therefore numerous external validations on single surgical specialties or even single surgical procedure studies were not included. Our study included only publications from heterogeneous patient populations for which the degree of heterogeneity varied among studies, including the urgency and subtype of surgical specialties. Both factors may have affected the predictive accuracy of models in different studies. However, we believe that the discussed prediction models should be applicable for a broad range of surgical patients to be considered for clinical use.

We aggregated and reported on several elements of clinical usability but must acknowledge that there is currently no accepted standard to gauge usability. Future research is needed to validate the clinical usability score. Research shows that models with low predictive performance on development or during external validation are often not submitted or accepted for publication. Currently, there is no established standard for assessing the likelihood of publication bias in research on predictive models.

The current systematic review of models to predict 30-day peri-operative mortality found that SORT combines good predictive model performance with clinical usability. In addition, SORT has been externally validated in heterogeneous cohorts and can be used on the population where validation was executed. External validation and updating of existing prediction models to specific patient populations have scarcely been performed in pre-operative mortality risk prediction models. Still, this is a necessary step to improve clinical uptake. Adequate reporting of calibration is required to make it easier for clinicians to understand which models provide accurate predictions across the entire risk spectrum. Furthermore, integrating reliable models with face validity in the electronic health record is indispensable for improving clinical uptake.

© 2023 The Authors. *Anaesthesia* published by John Wiley & Sons Ltd on behalf of Association of Anaesthetists.

13652044, 2023, 5, Downloaded from https://associationofanaesthetists-publications.onlinelibrary.wiley.com/doi/10.1111/anae.15988 by Cochrane Netherlands, Wiley Online Library on [02/02/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Acknowledgements

We thank S. Gadgil and P. Wiersma, University Medical Centre Utrecht, for their help with the PubMed search. BP was a member of the Advisory Board of Sensium Healthcare. All data will be published in the Harvard data-verse after publication of the manuscript. No external funding or other competing interests declared.

References

1. Weiser TG, Haynes AB, Molina G, et al. Size and distribution of the global volume of surgery in 2012. *Bulletin of the World Health Organization* 2016; **94**: 201–209F.
2. Abbott TEF, Pearse RM, Archbold RA, et al. A prospective international multicentre cohort study of intraoperative heart rate and systolic blood pressure and myocardial injury after noncardiac surgery: results of the VISION study. *Anesthesia and Analgesia* 2018; **126**: 1936–45.
3. Pearse RM, Moreno RP, Bauer P, et al. Mortality after surgery in Europe: a 7 day cohort study. *Lancet* 2012; **380**: 1059–65.
4. Pearse RM, Harrison DA, James P, et al. Identification and characterisation of the high-risk surgical population in the United Kingdom. *Critical Care* 2006; **10**: R81.
5. Tjeertes EK, Ultee KH, Stolker RJ, et al. Perioperative complications are associated with adverse long-term prognosis and affect the cause of death after general surgery. *World Journal of Surgery* 2016; **40**: 2581–90.
6. Toner A, Hamilton M. The long-term effects of postoperative complications. *Current Opinion in Critical Care* 2013; **19**: 364–8.
7. Moonesinghe SR, Mythen MG, Das P, Rowan KM, Grocott MP. Risk stratification tools for predicting morbidity and mortality in adult patients undergoing major surgery: qualitative systematic review. *Anesthesiology* 2013; **119**: 959–81.
8. Kristensen SD, Knuuti J, Saraste A, et al. 2014 ESC/ESA Guidelines on non-cardiac surgery: cardiovascular assessment and management: The Joint Task Force on non-cardiac surgery: cardiovascular assessment and management of the European Society of Cardiology (ESC) and the European Society of Anaesthesiology (ESA). *European Journal of Anaesthesiology* 2014; **31**: 517–73.
9. De Hert S, Staender S, Fritsch G, et al. Pre-operative evaluation of adults undergoing elective noncardiac surgery: updated guideline from the European Society of Anaesthesiology. *European Journal of Anaesthesiology* 2018; **35**: 407–65.
10. Duceppe E, Parlow J, MacDonald P, et al. Canadian Cardiovascular Society guidelines on perioperative cardiac risk assessment and management for patients who undergo noncardiac surgery. *Canadian Journal of Cardiology* 2017; **33**: 17–32.
11. Halvorsen S, Mehilli J, Cassese S, et al. 2022 ESC Guidelines on cardiovascular assessment and management of patients undergoing non-cardiac surgery. *European Heart Journal* 2022; **43**: 3826–924.
12. Wong DJN, Harris S, Sahni A, et al. Developing and validating subjective and objective risk-assessment measures for predicting mortality after major surgery: an international prospective cohort study. *PLoS Medicine* 2020; **17**: e1003253.
13. Kappen TH, van Klei WA, van Wolfswinkel L, Kalkman CJ, Vergouwe Y, Moons KGM. Evaluating the impact of prediction models: lessons learned, challenges, and recommendations. *Diagnostic and Prognostic Research* 2018; **2**: 11.
14. Moons KG, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? *British Medical Journal* 2009; **338**: b375.
15. Reilly JR, Gabbe BJ, Brown WA, Hodgson CL, Myles PS. Systematic review of perioperative mortality risk prediction models for adults undergoing inpatient non-cardiac surgery. *ANZ Journal of Surgery* 2021; **91**: 860–70.
16. Steyerberg EW, Moons KG, van der Windt DA, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Medicine* 2013; **10**: e1001381.
17. Counsell C, Dennis M. Systematic review of prognostic models in patients with acute stroke. *Cerebrovascular Diseases* 2001; **12**: 159–70.
18. Moher D, Shamseer L, Clarke M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic Reviews* 2015; **4**: 1.
19. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *British Medical Journal* 2021; **372**: n71.
20. Moons KG, de Groot JA, Bouwmeester W, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. *PLoS Medicine* 2014; **11**: e1001744.
21. Borzecki AM, Christiansen CL, Chew P, Loveland S, Rosen AK. Comparison of in-hospital versus 30-day mortality assessments for selected medical conditions. *Medical Care* 2010; **48**: 1117–21.
22. Geersing GJ, Bouwmeester W, Zuithoff P, Spijker R, Leeflang M, Moons KG. Search filters for finding prognostic and diagnostic prediction studies in Medline to enhance systematic reviews. *PLoS One* 2012; **7**: e32844.
23. Moons KGM, Wolff RF, Riley RD, et al. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. *Annals of Internal Medicine* 2019; **170**: W1–W33.
24. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Annals of Internal Medicine* 2019; **170**: 51–8.
25. 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. *Annals of Surgery* 2016; **264**: 23–31.
26. Sharma V, Ali I, van der Veer S, Martin G, Ainsworth J, Augustine T. Adoption of clinical risk prediction tools is limited by a lack of integration with electronic health records. *British Medical Journal Health and Care Informatics* 2021; **28**: e100253.
27. Wanderer JP, Ehrenfeld JM. Toward external validation and routine clinical use of the American College of Surgeons NSQIP surgical risk calculator. *Journal of the American College of Surgeons* 2016; **223**: 674.
28. Collins GS, de Groot JA, Dutton S, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC Medical Research Methodology* 2014; **14**: 40.
29. Swets JA. Measuring the accuracy of diagnostic systems. *Science* 1988; **240**: 1285–93.
30. Cook NR. Quantifying the added value of new biomarkers: how and how not. *Diagnostic and Prognostic Research* 2018; **2**: 14.
31. Protopapa KL, Simpson JC, Smith NC, Moonesinghe SR. Development and validation of the Surgical Outcome Risk Tool (SORT). *British Journal of Surgery* 2014; **101**: 1774–83.
32. Bilimoria KY, Liu Y, Paruch JL, et al. Development and evaluation of the universal ACS NSQIP surgical risk calculator: a decision aid and informed consent tool for patients and surgeons. *Journal of the American College of Surgeons* 2013; **217**: 833–42.
33. Glance LG, Lustik SJ, Hannan EL, Osler TM, Mukamel DB, Qian F, Dick AW. The Surgical Mortality Probability Model: derivation and validation of a simple risk prediction rule for noncardiac surgery. *Annals of Surgery* 2012; **255**: 696–702.
34. Dalton JE, Kurz A, Turan A, Mascha EJ, Sessler DI, Saager L. Development and validation of a risk quantification index for

© 2023 The Authors. *Anaesthesia* published by John Wiley & Sons Ltd on behalf of Association of Anaesthetists.

13652044, 2023, 5, Downloaded from https://associationofanaesthetists-publications.onlinelibrary.wiley.com/doi/10.1111/anae.15988 by Cochrane Netherlands, Wiley Online Library on [02/02/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

30-day postoperative mortality and morbidity in noncardiac surgical patients. *Anesthesiology* 2011; **114**: 1336–44.

35. 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 Medicine* 2018; **15**: e1002701.
36. Bihorac A, Ozrazgat-Baslanti T, Ebadi A, et al. MySurgeryRisk: development and validation of a machine-learning risk algorithm for major complications and death after surgery. *Annals of Surgery* 2019; **269**: 652–62.
37. Choi B, Oh AR, Lee SH, et al. Prediction model for 30-day mortality after non-cardiac surgery using machine-learning techniques based on preoperative evaluation of electronic medical records. *Journal of Clinical Medicine* 2022; **11**: 6487.
38. Sigakis MJ, Bittner EA, Wanderer JP. Validation of a risk stratification index and risk quantification index for predicting patient outcomes: in-hospital mortality, 30-day mortality, 1-year mortality, and length-of-stay. *Anesthesiology* 2013; **119**: 525–40.
39. Rozeboom PD, Bronsert MR, Velopulos CG, et al. A comparison of the new, parsimonious tool Surgical Risk Preoperative Assessment System (SURPAS) to the American College of Surgeons (ACS) risk calculator in emergency surgery. *Surgery* 2020; **168**: 1152–9.
40. Ren Y, Loftus TJ, Datta S, et al. Performance of a machine learning algorithm using electronic health record data to predict postoperative complications and report on a Mobile platform. *Journal of the American Medical Association Network Open* 2022; **5**: e2211973.
41. Campbell D, Boyle L, Soakell-Ho M, et al. National risk prediction model for perioperative mortality in non-cardiac surgery. *British Journal of Surgery* 2019; **106**: 1549–57.
42. Henderson WG, Bronsert MR, Hammermeister KE, Lambert-Kerzner A, Meguid RA. Refining the predictive variables in the ``Surgical Risk Preoperative Assessment System´´ (SURPAS): a descriptive analysis. *Patient Safety in Surgery* 2019; **13**: 28.
43. Liu Y, Cohen ME, Hall BL, Ko CY, Bilimoria KY. Evaluation and enhancement of calibration in the American College of Surgeons NSQIP surgical risk calculator. *Journal of the American College of Surgeons* 2016; **223**: 231–9.
44. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *British Medical Journal* 2015; **350**: g7594.
45. Le Manach Y, Collins G, Rodseth R, et al. Preoperative Score to Predict Postoperative Mortality (POSPOM): derivation and validation. *Anesthesiology* 2016; **124**: 570–9.
46. Venema E, Wessler BS, Paulus JK, et al. Large-scale validation of the prediction model risk of bias assessment tool (PROBAST) using a short form: high risk of bias models show poorer discrimination. *Journal of Clinical Epidemiology* 2021; **138**: 32–9.
47. Hoesseini A, van Leeuwen N, Sewnaik A, Steyerberg EW, Baatenburg de Jong RJ, Lingsma HF, Offerman MPJ. Key aspects of prognostic model development and interpretation from a clinical perspective. *Journal of the American Medical Association Otolaryngology. Head and Neck Surgery* 2022; **148**: 180–6.
48. Mureddu GF. Current multivariate risk scores in patients undergoing non-cardiac surgery. *Monaldi Archives for Chest Disease* 2017; **87**: 848.
49. Altman DG, Vergouwe Y, Royston P, Moons KG. Prognosis and prognostic research: validating a prognostic model. *British Medical Journal* 2009; **338**: b605.
50. Huang Y, Li W, Macheret F, Gabriel RA, Ohno-Machado L. A tutorial on calibration measurements and calibration models for clinical prediction models. *Journal of the American Medical Informatics Association* 2020; **27**: 621–33.
51. Austin PC, Steyerberg EW. The Integrated Calibration Index (ICI) and related metrics for quantifying the calibration of logistic regression models. *Statistics in Medicine* 2019; **38**: 4051–65.
52. Wessler BS, Lai YL, Kramer W, et al. Clinical prediction models for cardiovascular disease: tufts predictive analytics and comparative effectiveness clinical prediction model database. *Circulation. Cardiovascular Quality and Outcomes* 2015; **8**: 368–75.
53. Van Calster B, McLernon DJ, van Smeden M, et al. Calibration: the Achilles heel of predictive analytics. *BMC Medicine* 2019; **17**: 230.
54. Moons KG, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Annals of Internal Medicine* 2015; **162**: W1–W73.
55. Toll DB, Janssen KJ, Vergouwe Y, Moons KG. Validation, updating and impact of clinical prediction rules: a review. *Journal of Clinical Epidemiology* 2008; **61**: 1085–94.
56. Scotton G, Del Zotto G, Bernardi L, et al. Is the ACS-NSQIP risk calculator accurate in predicting adverse postoperative outcomes in the emergency setting? An Italian single-center preliminary study. *World Journal of Surgery* 2020; **44**: 3710–9.
57. Ma M, Liu Y, Gotoh M, et al. Validation study of the ACS NSQIP surgical risk calculator for two procedures in Japan. *American Journal of Surgery* 2021; **222**: 877–81.
58. Arce K, Moore EJ, Lohse CM, Reiland MD, Yetzer JG, Ettinger KS. The American College of Surgeons National Surgical Quality Improvement Program surgical risk calculator does not accurately predict risk of 30-day complications among patients undergoing microvascular head and neck reconstruction. *Journal of Oral and Maxillofacial Surgery* 2016; **74**: 1850–8.
59. Johnson C, Campwala I, Gupta S. Examining the validity of the ACS-NSQIP Risk Calculator in plastic surgery: lack of input specificity, outcome variability and imprecise risk calculations. *Journal of Investigative Medicine* 2017; **65**: 722–5.
60. Wang X, Hu Y, Zhao B, Su Y. Predictive validity of the ACS-NSQIP surgical risk calculator in geriatric patients undergoing lumbar surgery. *Medicine (Baltimore)* 2017; **96**: e8416.
61. Chudgar NP, Yan S, Hsu M, et al. External validation of surgical risk preoperative assessment system in pulmonary resection. *Annals of Thoracic Surgery* 2021; **112**: 228–37.
62. van der Hulst HC, Dekker JWT, Bastiaannet E, et al. Validation of the ACS NSQIP surgical risk calculator in older patients with colorectal cancer undergoing elective surgery. *Journal of Geriatric Oncology* 2022; **13**: 788–95.
63. Hamade S, Alshiek J, Javadian P, Ahmed S, McLeod FN, Shobeiri SA. Evaluation of the American College of Surgeons National Surgical Quality Improvement Program Risk Calculator to predict outcomes after hysterectomies. *International Journal of Gynaecology and Obstetrics* 2022; **158**: 714–21.
64. Mansmann U, Rieger A, Strahwald B, Crispin A. Risk calculators-methods, development, implementation, and validation. *International Journal of Colorectal Disease* 2016; **31**: 1111–6.
65. Stones J, Yates D. Clinical risk assessment tools in anaesthesia. *British Journal of Anaesthesia Education* 2019; **19**: 47–53.
66. Aakre C, Dziadzko M, Keegan MT, Herasevich V. Automating clinical score calculation within the electronic health record. A feasibility assessment. *Applied Clinical Informatics* 2017; **8**: 369–80.

© 2023 The Authors. *Anaesthesia* published by John Wiley & Sons Ltd on behalf of Association of Anaesthetists.

13652044, 2023, 5, Downloaded from https://associationofanaesthetists-publications.onlinelibrary.wiley.com/doi/10.1111/anae.15988 by Cochrane Netherlands, Wiley Online Library on [02/02/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

67. Fleisher LA, Fleischmann KE, Auerbach AD, et al. 2014 ACC/AHA guideline on perioperative cardiovascular evaluation and management of patients undergoing noncardiac surgery: executive summary: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *Circulation* 2014; **130**: 2215–45.
68. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *European Heart Journal* 2014; **35**: 1925–31.
69. Bedford J, Martin P, Crowe S, et al. Development and internal validation of a model for postoperative morbidity in adults undergoing major elective colorectal surgery: the peri-operative quality improvement programme (PQIP) colorectal risk model. *Anaesthesia* 2022; **77**: 1356–67.
70. McGuinness LA, Higgins JPT. Risk-of-bias VISualization (robvis): an R package and shiny web app for visualizing risk-of-bias assessments. *Research Synthesis Methods* 2021; **12**: 55–61.

Supporting Information

Additional supporting information may be found online via the journal website.

Appendix S1. Clinical usability assessment scores.

Appendix S2. Pre-operative mortality risk prediction models.

Table S1. Search terms, databases and search strategies.

Table S2. Study characteristics of the included studies.

Table S3. Patient characteristics from study cohorts.

© 2023 The Authors. *Anaesthesia* published by John Wiley & Sons Ltd on behalf of Association of Anaesthetists.

13652044, 2023, 5, Downloaded from https://associationofanaesthetists-publications.onlinelibrary.wiley.com/doi/10.1111/anae.15988 by Cochrane Netherlands, Wiley Online Library on [02/02/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License