Anaesthesia 2024, 79, 389–398 doi:10.1111/anae.16248

Original Article

Systematic development and validation of a predictive model for major postoperative complications in the Peri-operative Quality Improvement Project (PQIP) dataset

C. M. Oliver,[1,2] **D. Wagstaff,**[2,3] **J. Bedford**[2,3] and **S. R. Moonesinghe**[2,4]
for the Peri-operative Quality Improvement Project delivery team and collaborative

1 Associate Professor, 3 Honorary Clinical Fellow, 4 Professor and Head, Centre for Peri-operative Medicine, University College London, UK
2 Consultant, Department of Anaesthesia and Peri-operative Medicine, UCL Hospitals, London, UK

Summary

Complications are common following major surgery and are associated with increased use of healthcare resources, disability and mortality. Continued reliance on mortality estimates risks harming patients and health systems, but existing tools for predicting complications are unwieldy and inaccurate. We aimed to systematically construct an accurate pre-operative model for predicting major postoperative complications; compare its performance against existing tools; and identify sources of inaccuracy in predictive models more generally. Complete patient records from the UK Peri-operative Quality Improvement Programme dataset were analysed. Major complications were defined as Clavien–Dindo grade ≥ 2 for novel models. In a 75% train:25% test split cohort, we developed a pipeline of increasingly complex models, prioritising pre-operative predictors using Least Absolute Shrinkage and Selection Operators (LASSO). We defined the best model in the training cohort by the lowest Akaike's information criterion, balancing accuracy and simplicity. Of the 24,983 included cases, 6389 (25.6%) patients developed major complications. Potentially modifiable risk factors (pain, reduced mobility and smoking) were retained. The best-performing model was highly complex, specifying individual hospital complication rates and 11 patient covariates. This novel model showed substantially superior performance over generic and specific prediction models and scores. We have developed a novel complications model with good internal accuracy, re-prioritised predictor variables and identified hospital-level variation as an important, but overlooked, source of inaccuracy in existing tools. The complexity of the best-performing model does, however, highlight the need for a step-change in clinical risk prediction to automate the delivery of informative risk estimates in clinical systems.

Correspondence to: C. M. Oliver
Email: charles.oliver@ucl.ac.uk
Accepted: 4 January 2024
Keywords: feature selection; major surgery; peri-operative medicine; postoperative complications; predictive modelling
Twitter/X: @CMOliver_; @duncanwagstaff; @jbedford84; @rmoonesinghe; @PQIPNews

Introduction

Major surgery is associated with a high incidence of short-term mortality (> 4%) and postoperative complications (incidence 15–35%) across populations [1, 2]. For individuals, as `unintended and undesirable consequences of surgery´, complications have the potential to de-rail the trajectory of postoperative recovery, leading to disability, psychological distress, increased use of healthcare resources and reduced survival in both the short-term and over subsequent decades [3–5]. The consequences of complications vary according to the nature, sequence and severity of complications; initiated treatment; and baseline physiological fitness of the patient.

Even accounting for increases in the age and complexity of surgical patient populations, advances in healthcare and efforts to improve the recognition and management of complications have failed to deliver material reductions in complication rates over recent decades. Contrasting, preventative approaches require targeted timely interventions in at-risk individuals to optimise comorbidities and prevent avoidable harm. But, despite the development of multiple candidate tools, accurate early prediction remains elusive and the success of preventative strategies has also been limited.

Predictive tools for complications may be classed as generic (designed to predict multiple outcomes across diverse populations) or specific (defined by population, procedure or complication type). Generic tools have the advantage of being widely applicable, but validation studies typically show them to be too inaccurate for clinical practice. In contrast, specific tools may be highly accurate in narrowly defined groups, but they become impractical when multiple tools are required for every patient. Complication tools are, therefore, rarely used. A notable exception is the American College of Surgeons National surgical quality improvement programme (ACS-NSQIP) calculator [6], which estimates multiple outcomes. However, more than 20 variables must be inputted, and its performance in almost exclusively American patient populations has been modest [7, 8].

Discussions of peri-operative risk are therefore limited to the likelihood of death within (or survival to) 30 days following surgery, because mortality prediction models tend to be more accurate in unselected populations. However, this fails to describe the trajectory and destination of postoperative recovery required by patients for informed decision-making. Moreover, mortality-based clinical decisions risk perversely prioritising preventative interventions to individuals most at risk of failure to rescue, missing opportunities to prevent avoidable complications [9]. Accurate methods to predict complications are therefore required.

We hypothesised that the inaccuracy of existing generic tools for the prediction of postoperative complications results from a failure to model important sources of variation, necessitating advanced modelling approaches and deployment solutions [10, 11]. Our three-fold aims were: to develop an accurate pre-operative model for the predictions of major postoperative complications in patients undergoing major surgery; to characterise sources of inaccuracies and limitations of developing regression-based predictive models; and to compare the accuracy of our novel model against existing tools.

Methods

We followed TRIPOD guidelines for this report [12]. The UK Health Research Authority approved analyses by the Peri-operative Quality Improvement Programme (PQIP) of adults who had scheduled surgery from December 2016 to June 2020 [13]. The PQIP is a prospective observational cohort study of a sample of adults (age ≥ 18 y on the date of surgery) undergoing one of a list of major, planned surgical procedures in UK NHS hospitals. The full list of included operations is available on the PQIP website. Casemix, process, outcome and patient questionnaire data are collected at six timepoints relative to surgery: immediately pre-operatively; during or immediately following surgery; 7 days postoperatively; on hospital discharge; and 6 and 12 months postoperatively. These data are then submitted electronically. Due to PQIP inclusion criteria, we did not study children (age < 18 y) or patients who had emergency surgery, minor surgery, cardiac surgery, neurosurgery or an obstetric procedure. We have published descriptive analyses of this cohort previously [14, 15]. The sample size was determined by the number of submitted records at the time of data extraction. Cases were eligible for inclusion if the primary outcome was recorded.

The primary outcome was major complications, defined as any Clavien–Dindo postoperative complication ≥ grade 2 that occurred at any time before hospital discharge [16]. This composite outcome represents a material deviation from an `uncomplicated´ recovery, ranging from requirement for a new drug treatment to death. Secondary outcomes were defined to match definitions of complications or morbidity used in derivation studies of eligible existing tools (as closely as possible within the confines of the dataset). Patient characteristics were assessed. To quantify inter-hospital variation in

comparable patients, a funnel plot was created of hospital-level incidence of the primary outcome in patients undergoing major colorectal surgery (anterior resection; right hemicolectomy with anastomosis; excision of sigmoid colon; or left hemicolectomy with anastomosis) for hospitals that submitted at least five patient records.

Records were divided at random to create 75% model training: 25% model testing cohorts. Candidate predictor variables were processed if recorded in > 90% of the training cohort. For categorical variables, we changed thresholds to balance case numbers between classes, as necessary. Surgical specialties were categorised as one of four classes, defined by univariate coefficient ranges (95% CIs < -1, < 0, including 0, > 0), with colorectal surgery as the reference category. For continuous variables, we Winsorised the extreme 2% values, and for those that were non-linearly associated with the outcome, restricted cubic splines were used to identify nodal points.

We inspected 32 candidate predictor variables (online Supporting Information Table S1) and their interactions for strength of association with the primary outcome using bootstrap sampling (× 200) multiple logistic regression in the training dataset. Candidate variables were: age; sex; BMI; ASA physical status; New York Heart Association (NYHA) heart failure class; cardiac or respiratory signs and history (as per definitions in the Physiological and Operative Severity Score for the enUmeration of Mortality and Morbidity (POSSUM)); ECG findings; smoking status; history of cerebrovascular disease, dementia, diabetes mellitus or liver disease; chest infection within the preceding month; EuroQol five-dimension questionnaire; pre-operative oxygen saturation; and pre-operative concentrations of sodium, potassium, creatinine, haemoglobin and white blood cells. We also tested surgical variables that were available pre-operatively: urgency (elective or expedited); surgery for cancer; minimally invasive surgery; surgical specialty; invasiveness of surgery; duration of surgery; and number of surgical procedures in the preceding 30 days.

The final selection of variables and sequence of forward model construction was guided by least absolute shrinkage and selection operator (LASSO) estimates, comparing cross-validation, adaptive and plugin tuning parameters. We built six models in stepwise fashion using forward variable selection, informed by LASSO prioritisation and forcing no variables, and recording incremental change in Bayesian information criterion (BIC), a measure of predictive value. A baseline model (model 1) comprised only patient-level variables and interaction terms. The first multi-level model (model 3) included statistical adjustment for the hospital that reported the operation. Hospital-level incidences were specified in the second multi-level model (model 5). We simplified each model by removing variables that added only minimal predictive value, identified by the inflection point in BIC reduction, to create three parsimonious models (model 2, model 4 and model 6 respectively). Coefficients and intercepts were then estimated, using bootstrapped sampling (× 1000), for each model, and individual estimates were computed. We selected the most accurate model, defined as the lowest Akaike's information criterion (AIC) value in the test cohort, to compare with existing tools.

We performed two separate exploratory analyses on the baseline model (model 1) to assess the importance of potential predictor variables. First, to assess the importance of frailty (which was recorded for only 40% of records) as a predictor variable, we forced the Rockwood clinical frailty scale (dichotomised < 4 vs. ≥ 4) into the model. Second, to explore the validity of the model in the wake of the first novel coronavirus 2019 wave, we forced a date (dichotomised as 29 February 2020).

From the existing literature, we identified existing prediction tools that had been developed for the purpose of, or that are commonly used as, surrogates for the prediction of postoperative complications and/or morbidity. Tools were eligible for testing if every component variable was recorded for > 90% of the cohort and if the calculation for risk estimation is publicly available. To enable fair comparison for existing prediction models that had been fitted on different incidences, log odds were estimated for each patient using these published formulae, outcome-specific intercepts (for the relevant outcomes) fitted and individual probabilities calculated.

We assessed and compared the performance of the novel model against existing tools, calculating the Brier score and Pearson's χ^2 goodness of fit for accuracy and calibration of predictive models and the area under the receiver operator characteristic curve (AUROC) for discrimination. We used a chi-squared test for comparison of AUROC and AIC for all tools. Calibration belts and decision curves were plotted for predictive models. Decision curves provide clinically meaningful information, balancing the value and consequences of an intervention based on the prediction. We used Stata®15 (StataCorp LP, College Station, TX, USA) for all analyses.

Results

Of the 25,523 extracted records, 24,983 were eligible for analysis of Clavien–Dindo complications grade: 18,735 (75%) were used to generate models for the primary outcome and 6248 (25%) were used to test the models. In

total, 4810 (26%) patients had Clavien–Dindo grade ≥ 2 complications. Patient characteristics are reported in online Supporting Information Table S2. A breakdown of Clavien–Dindo-grade complications and associated diagnoses are shown in online Supporting Information Table S3. Complication rates varied substantially between hospitals within a relatively homogenous cohort of surgical procedures (online Supporting Information Figure S1).

Table 1 details the coefficients (95%CI) that associated complications with the 11/32 assessed variables that were selected for the full model and the 6/32 variables selected for the simplified model. Surgical specialty classes are reported in online Supporting Information Table S4.

The full (11-variable) model, specifying hospital-level intercept for each patient, showed the best balance of accuracy and simplicity (the lowest AIC), as reported in Table 2. Stepwise increases in complexity (model 1 to model 3 to model 5) resulted in increasing net benefit (decision curve analysis, Fig. 1) and accuracy (calibration, online Supporting Information Figure S2). But interestingly, simplification (from 11 to 6 variables) resulted in only modest reductions in performance. The method, hospital-specific intercepts and predictor variable coefficients to calculate the predicted likelihood of Clavien–Dindo grade ≥ 2 complications are provided in online Supporting Information Appendix S1.

We identified 10 scores or models for the prediction of postoperative outcomes from the literature [6, 17–26]. Four fulfilled our criteria for assessment: the surgical outcomes risk tool (SORT) for morbidity (measured on postoperative day 7) [20]; the assess respiratory risk in surgical patients in Catalonia (ARISCAT) score for postoperative pulmonary complications (24,755 available operations) [27]; the revised cardiac risk index (RCRI) (24,718 available operations) [28]; and the surgical outcomes risk tool (SORT) for mortality [26]. Predictor variables for each of these tools are reported in online Supporting Information Table S5.

The SORT morbidity score was developed for any of 30 diagnoses or interventions, 14 of which would qualify as a Clavien–Dindo complication grade ≥ 2. The ARISCAT score was developed to predict seven pulmonary complications (five of which would qualify as a Clavien–Dindo complication grade ≥ 2) occurring within 90 days of elective or emergency surgery but excluded patients who had more than one operation. The RCRI was developed for five outcomes (four of which would qualify as a Clavien–Dindo complication grade ≥ 2) during an undefined period after scheduled major non-cardiac surgery in patients aged > 49 y. The SORT mortality score was developed for death within 30 days of a selection of major elective operations. Definitions of cardiovascular and respiratory complications used in this study are reported in online Supporting Information Table S6.

Model 5 showed superior accuracy, calibration and discrimination for the primary outcome compared with SORT morbidity and SORT mortality in the test cohort (Table 3), and substantial benefit over these tools in decision curve analysis (Fig. 2). Model 5 was also superior to the SORT morbidity and RCRI models for predicting major adverse cardiac events. For predicting postoperative pulmonary complications, the performance of Model 5 was equivalent to ARISCAT.

Supplementary analyses identified that frailty was statistically significantly associated with major complications, but that it was lower priority than the 11 modelled variables using LASSO (data not reported) and would therefore not have been included in the final model. Because the incidence of major complications was substantially lower (15.3%, unadjusted) in the 478 patients who had surgery in March–June 2020, Model 5 substantially over-predicted the risk of complications during this period (data not reported).

Discussion

We have developed an accurate model for predicting postoperative complications, using modern variable selection and modelling techniques, in a contemporary multicentre and multispecialty prospective cohort of patients undergoing major surgery. Our novel model showed superior performance and substantial benefit over existing tools, both in identifying the primary outcome and alternative generic and organ-specific types of complication. The accuracy of prediction improved with increasingly complex models, notably the inclusion of hospital complication rates, but reduced only modestly on removing predictor variables. Our approach to model building highlights the importance of overlooked sources of inaccuracy in existing prediction models.

Our reliance on mortality prediction risks harming patients and health systems. The accurate identification of high-risk patients is used to focus targeted interventions peri-operatively and better inform patient choice and expectations. But while the aim of these interventions is to prevent or mitigate postoperative complications (or morbidity), risk assessment almost exclusively relies on mortality prediction. Risk factors for death and complications differ and, by focusing on the downstream consequences of complications, opportunities to

Table 1 Coefficients (95%CI) for variables associated with postoperative complications of Clavien–Dindo grade ≥ 2, in six models with 11 or 6 variables, adjusted for hospital or not and, if so, whether the intercept was hospital-specific. Variables are listed in descending order of LASSO-assigned priority. Hospital-specific intercepts are listed in online Supporting Information Appendix S1.

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Variables; n	11	6	11	6	11	6
Adjusted for hospital	No	No	Yes	Yes	Yes	Yes
Intercept specific to hospital	n/a	n/a	No	No	Yes	Yes
Intercept	-3.6 (-4.4 to -2.8)	-3.8 (-4.5 to -3.0)	n/a	n/a	Various	Various
Specialty						
Class 1	Reference					
Class 2	1.3 (0.5–2.2)	1.4 (0.6–2.3)	1.0 (0.2–1.8)	1.1 (0.3–1.9)	1.0 (0.2–1.8)	1.1 (0.3–1.9)
Class 3	1.6 (0.8–2.4)	1.6 (0.8–2.4)	1.3 (0.6–2.1)	1.4 (0.6–2.2)	1.3 (0.6–2.1)	1.4 (0.6–2.2)
Class 4	1.3 (0.2–2.4)	1.3 (0.2–2.4)	1.0 (0.0–2.0)	1.0 (-0.1–2.0)	1.0 (0.0–2.0)	1.0 (-0.1–2.0)
Duration of surgery; h						
< 2 h	Reference					
2–3	1.0 (0.1–1.8)	0.9 (0.1–1.8)	1.0 (0.2–1.7)	0.9 (0.2–1.7)	1.0 (0.2–1.7)	0.9 (0.2–1.7)
> 3	1.7 (0.9–2.5)	1.7 (0.9–2.4)	1.5 (0.8–2.3)	1.5 (0.7–2.2)	1.5 (0.8–2.3)	1.5 (0.7–2.2)
EQ-5D-5L walking difficulty						
None/mild/moderate	Reference					
Severe	0.2 (0.1–0.3)	0.3 (0.2–0.4)	0.2 (0.1–0.3)	0.3 (0.2–0.4)	0.2 (0.1–0.3)	0.3 (0.2–0.4)
Impossible	0.8 (0.4–1.2)	0.8 (0.5–1.2)	0.7 (0.3–1.1)	0.8 (0.4–1.2)	0.7 (0.3–1.1)	0.8 (0.4–1.2)
Cardiac failure						
None	Reference					
Drug therapy	0.3 (0.2, 0.4)	0.4 (0.3–0.5)	0.2 (0.1–0.3)	0.3 (0.2–0.4)	0.2 (0.1–0.3)	0.3 (0.2–0.4)
Signs of heart failure	0.7 (0.2–1.1)	0.8 (0.4–1.2)	0.7 (0.2–1.1)	0.8 (0.4–1.2)	0.7 (0.2–1.1)	0.8 (0.4–1.2)
EQ-5D-5L pain/discomfort						
None/mild	Reference					
Moderate	0.2 (0.1–0.3)	0.2 (0.1–0.3)	0.2 (0.1–0.3)	0.2 (0.1–0.3)	0.2 (0.1–0.3)	0.2 (0.1–0.3)
Severe/extreme	0.4 (0.3–0.6)	0.4 (0.3–0.6)	0.5 (0.3–0.6)	0.4 (0.3–0.6)	0.5 (0.3–0.6)	0.4 (0.3–0.6)
ASA physical status						
≤ 2	Reference					
3	0.4 (0.3–0.5)	0.5 (0.4–0.6)	0.4 (0.3–0.5)	0.5 (0.4–0.6)	0.4 (0.3–0.5)	0.5 (0.4–0.6)
4	0.5 (0.1–1.0)	0.8 (0.3–1.2)	0.7 (0.2–1.2)	0.9 (0.5–1.4)	0.7 (0.2–1.2)	0.9 (0.5–1.4)
Dyspnoea						
None	Reference					
On exertion	0.2 (0.1–0.3)		0.2 (0.1–0.3)		0.2 (0.1–0.3)	
At rest/on limited exertion	0.4 (0.2–0.6)		0.5 (0.3–0.7)		0.5 (0.3–0.7)	
Pre-operative haemoglobin; $g.l^{-1}$						
8.2–9.9	Reference					
10.0–13.8	−0.2 (−0.4 to −0.1)		−0.3 (−0.5 to −0.1)		−0.3 (−0.5 to −0.1)	
≥ 13.9	−0.4 (−0.5 to −0.2)		−0.4 (−0.5 to −0.2)		−0.4 (−0.5 to −0.2)	
Preceding operations past 30 days						
0	Reference					
≥ 1	0.3 (0.2–0.5)		0.3 (0.2–0.5)		0.3 (0.2–0.5)	
Age; y						
≤ 63	Reference		0		0	
64–79	0.1 (0.1–0.2)		0.2 (0.1–0.3)		0.2 (0.1–0.3)	
≥ 80	0.3 (0.2–0.4)		0.4 (0.2–0.5)		0.4 (0.2–0.5)	
Smoking history						
Never	Reference		0		0	
Unknown	−0.2 (−0.4–0.0)		0.1 (−0.1–0.3)		0.1 (−0.1–0.3)	
Quit > 6 months	0.2 (0.1–0.2)		0.1 (0.1–0.2)		0.1 (0.1–0.2)	
Current or quit < 6 months	0.3 (0.2–0.4)		0.2 (0.1–0.3)		0.2 (0.1–0.3)	

(continued)

Table 1 (*continued*)

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Interaction terms						
ASA-PS = 3 and Cardiac failure = drug therapy	−0.3 (−0.5 to −0.1)	−0.3 (−0.4 to −0.1)	−0.3 (−0.5 to −0.2)	−0.3 (−0.5 to −0.2)	−0.3 (−0.5 to −0.2)	−0.3 (−0.5 to −0.2)
Speciality = 2 and duration of surgery = 2–3			−0.9 (-1.7 to 0.0)	−0.9 (-1.7 to 0.1)	−0.9 (-1.7 to 0.0)	−0.9 (-1.7 to −0.1)
Specialty = 2 and duration of surgery ≥ 3		−0.9 (-1.8 to −0.1)				

n/a, not applicable; EQ5D5L, EuroQoL 5D5L.

Table 2 Metrics for the association of six models, which we derived from 4810 complications of Clavien–Dindo grade ≥ 2 after 18,735 operations (training cohort), with 1579 complications after 6248 operations (testing cohort). Lower values indicate better performance for Akaike's information criterion (AIC) and Brier score. For goodness of fit (Pearson's χ^2 test), a non-significant p value (> 0.04) indicates a well-calibrated model. Higher values indicate better performance for the area under the receiving operator characteristic (AUROC) curve.

Model	Variables	Hospital-specific		AIC	Brier score	Goodness of fit, Pearson's χ^2	AUROC (95%CI)
		Adjustment	Intercept				
1	11	No	n/a	6236	0.19	2301, p < 0.001	0.65 (0.64–0.66)
2	6	No	n/a	6260	0.19	523, p < 0.001	0.64 (0.63–0.65)
3	11	Yes	No	6269	0.18	2312, p < 0.001	0.64 (0.63–0.65)
4	6	Yes	No	6311	0.19	577, p < 0.001	0.62 (0.61–0.63)
5	11	Yes	Yes	5942	0.17	4714, p = 0.662	0.73 (0.72–0.74)
6	6	Yes	Yes	5987	0.18	2418, p = 0.484	0.72 (0.71–0.73)

n/a, not applicable.

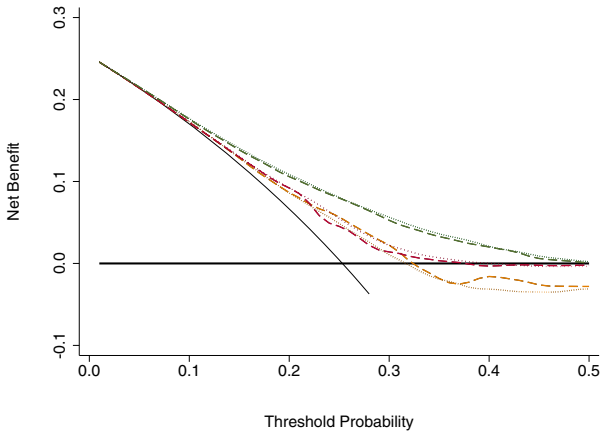

Figure 1 Decision curve analysis (DCA) plots for six novel models for prediction of postoperative complications. Model 1, orange dot; Model 2, orange dash; Model 3, red dot; Model 4, red dash; Model 5, green dot; Model 6, green dash; thick black line, no individuals receive treatment; thin black line, all individuals receive treatment. The `treat-none/treat-all´ lines intersect at the incidence. The benefit/utility of using a model to inform practice balances the value and consequences of an intervention based on the prediction. This benefit/utility is quantified by the area above the `treat-none/treat-all´ intersection.

proactively improve care and outcomes are missed. Perversely, this may be most relevant for the many, most easily `rescued´ patients, since current approaches focus resources on older, multiply comorbid patients, in whom these rescue efforts are most likely to fail to prevent death. Our comparison of decision curves shows substantially greater benefit from using our prediction model over SORT, a widely accepted mortality prediction tool, reinforcing the need for a better approach to pre-operative risk prediction.

Multiple tools for predicting postoperative complications and morbidity have been developed, but none, with the exception perhaps of ACS-NSQIP, have gained widespread uptake. With its proprietary algorithm, unfortunately we were unable to assess the ACS-NSQIP. Our analysis does, however, support observations that existing tools are either too inaccurate or too narrow in their focus to support peri-operative decision-making. In marked contrast with most existing prediction tools (for complications, morbidity or mortality), we identified and addressed hospital-level variation in complication rates as an important source of inaccuracy. Modelling of this variation resulted in

Table 3 Accuracy of existing and the best-performing novel risk tools in correctly predicting multisystem and organ-specific complications and morbidity. Existing tools were assessed in all eligible patients, whereas, for fair comparison, Model 5 was assessed only in the test cohort (n = 6248). For goodness of fit, a non-significant p value using Pearson's χ^2 (> 0.04) indicates a well-calibrated model. Lower values indicate better performance for Akaike's information criterion (AIC) and Brier score. Higher values indicate better performance for the area under the receiving operator characteristic (AUROC) curve.

		Model 5	SORT (morbidity)	SORT (mortality)	ARISCAT	RCRI
Generic complications						
Clavien–Dindo grade ≥ 2	Goodness of fit	4714, p = 0.662	1510, p < 0.001	110, p < 0.001		
	Brier score	0.174	0.228	0.255		
	AIC	5942	28,380	28,365		
	AUROC (95%CI)	0.73 (0.72–0.74)	0.49 (0.48–0.50)	0.58 (0.57–0.59)		
Day 7 POMS	Goodness of fit	4890, p = 0.09	1962, p < 0.001	1024, p < 0.001		
	Brier	0.183	0.215	0.240		
	AIC	6056	28,133	28,059		
	AUROC (95%CI)	0.67 (0.65–0.68)	0.50 (0.49–0.51)	0.58 (0.57–0.59)		
Organ system-specific complications						
PPC	Goodness of fit	5362, p = 0.08	1439, p < 0.001	992, p < 0.001	162, p = 0.44	
	Brier score	0.132	0.032	0.031	0.064	
	AIC	1477	6951	6911	6732	
	AUROC (95%CI)	0.67 (0.63–0.71)	0.49 (0.47–0.51)	0.67 (0.65–0.68)	0.65 (0.63–0.67)	
MACE	Goodness of fit	5294, p = 0.11	812, p < 0.001	554, p < 0.001		-
	Brier score	0.131	0.074	0.033		-
	AIC	1596	6584	6525		6476
	AUROC (95%CI)	0.70 (0.67–0.73)	0.53 (0.51–0.55)	0.67 (0.65–0.69)		0.61 (0.59–0.63)

SORT, surgical outcome risk tool; ARISCAT, assess respiratory risk in surgical patients in Catalonia risk score for postoperative pulmonary complications; RCRI, revised cardiac risk index; POMS, postoperative morbidity survey; PPC, postoperative pulmonary complication(s); MACE, major adverse cardiovascular event.

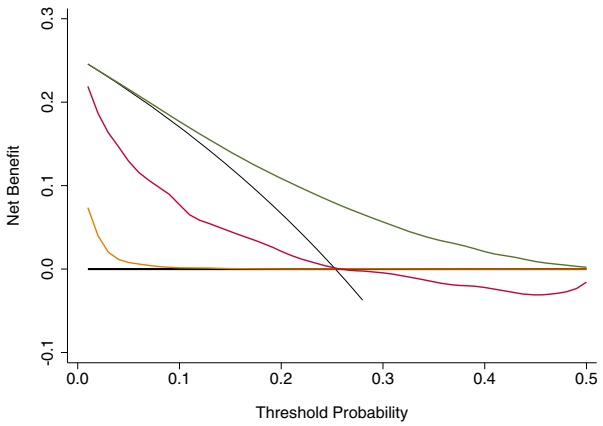

Figure 2 Decision curve analysis (DCA) plots for novel model 5 (green) and the surgical outcome risk tool (SORT) for morbidity (red) and mortality (orange). Thick black: no individuals receive treatment, Thin black: all individuals receive treatment. The `treat-none/treat-all´ lines intersect at the incidence. The benefit/utility of using a model to inform practice balances the value and consequences of an intervention based on the prediction. This benefit/utility is quantified by the area above the `treat-none/treat-all´ intersection.

a highly accurate model that showed substantially superior clinical benefit across a relevant range of risk (10–50%) in decision curve analysis.

The SORT morbidity tool was developed to predict postoperative morbidity but performed poorly in its first major assessment of external validity in a heterogeneous cohort. It has been reported to perform poorly in a colorectal cohort from the PQIP database [14]. Reasons for poor performance might include its single-centre development, internal validation and relative over-representation of orthopaedic surgery (which typically confers less risk than other surgical specialties).

Our analysis also highlights important, but frequently overlooked, modifiable risk factors for major postoperative complications in this high-granularity dataset. Pain, poor mobility, respiratory limitation, reduced haemoglobin reserves and < 6-month smoking abstinence were prioritised by our approach to variable selection using LASSO. These findings support the use of the targeted preventative interventions that we advocate over `recognise, relay and react´ approaches for reducing postoperative complications [29].

While we were successful in developing an accurate model that could offer substantial clinical benefit, our work and that of others to develop predictive models shows that current strategies are no longer feasible for several reasons. First, accurate models require levels of complexity that may compromise their clinical utility. Second, it is more informative to predict multiple specific outcomes than a single generic measure, but to do so accurately would require multiple calculators. Third, because models are specific to the location, time and population in which they were derived, they require ongoing updating and adapting to retain accuracy. This may be most relevant to the hospital-specific intercept at the core of our novel model, given that, at the time of publication, this cohort is more than three years in the past and from a world yet to experience COVID-19. Finally, the generation of point estimates, both for coefficients and probability, in current modelling approaches belies the precision of these values.

Individual solutions are available for some of these issues, but comprehensive approaches are now required [30]. With the availability of electronic patient records in increasing numbers of institutions, mechanisms to automate data flow and advanced modelling capabilities are already achievable [10]. The next generation of clinical prediction systems will automate the pull of electronic data, simultaneously predict multiple outcomes, perhaps using ensemble predictions to give certainty to estimates, adapt predictions to local populations and dynamically model fluctuations in individual, hospital and population risk. Perhaps most importantly, they must usefully integrate timely risk estimates into clinical support systems. Systems to automate the delivery of timely clinician recommendations embedded in the clinical workflow have already been shown to support improvements in clinical practice [11, 31].

The strengths of this study include: the use of a large, high-granularity and contemporary peri-operative dataset to develop complex multi-level prediction models; the systematic approach to developing novel predictive models; and rigorous testing of new and existing tools for a range of important complication types. We anticipate that our findings will be generalisable across healthcare systems due to the use of simplified pre-operative predictor variables, international definitions of complications and representativeness of patients, operative procedures and incidences of outcomes. The utility of the Model 5 tool should be assessed externally and temporally. There are, however, limitations to our approach. Some important risk calculators could not be assessed due to unavailable coefficients and intercepts or missing predictor variables. Some data items were missing or not collected, leading to discrepancies with existing outcome definitions. Some variables were not modelled, including the sequence and multiplicity of outcomes and socio-economic deprivation. Patients were restricted to those undergoing non-emergent surgery. Risk factors, most notably pain, function and smoking status, were self-reported. Subsequent assessment of the accuracy of our model should note that the Clavien–Dindo definitions used here differ somewhat from those proposed by Dindo et al. [16], including the omission (and potential inclusion as grade 2 complications) of some grade 1 interventions (administration of diuretics or electrolytes; physiotherapy; and wound infections opened at the bedside).

Major complications are common following major surgery and are associated with increased short- and long-term use of healthcare resources, disability and mortality. The ability to accurately predict complications, rather than death, will enable more effective preparation for surgery, responsive prevention and informed shared decision-making. We show that complications can be accurately predicted if important risk factors are included and advanced modelling and computation are used. Future efforts might externally validate this model and should focus on the development of automated prediction models embedded within electronic clinical systems.

Acknowledgements

The authors wish to acknowledge all members of the PQIP delivery team, including: A. Sahni; D. McGuckin; D. Gilhooly; C. Santos; J. Wilson; P. Martin; G. Singleton; K. Edwards; C. Vindrola-Padros; S. Warnakulasuriya; J. Dorey; I. Leemans; D. Martinez; J. Lourtie; I. Leeman; R. Baumber; A. Swift; A. Jackson; N. Fulop; A. Brent; K. Williams; M. Grocott; M. Mythen; D. Olive; C. Taylor; S. Drake; M. Swart; A. Bougeard; M. Bedford; A. Vallance; P. Singh; R. Vohra; A. Ignacka; O. Tucker; G. Aresu; M. Cripps; H. Ellicott; K. Samuel; and M. Chazapis. A. Putzu provided detailed feedback on the manuscript. Other collaborators who have supported the study through data acquisition or in an advisory capacity: https://www.rcoa.ac.uk/research/research-projects/perioperative-quality-improvement-programme-pqip. The authors would like to thank Dr J. Carlisle for creating the funnel plot in online Supporting Information Figure S1. The dataset used for these analyses is available on request to the PQIP committee, subject to approval criteria. Statistical code is available on request. No competing interests were declared.

References

1. Pearse RM, Moreno RP, Bauer P, et al. Mortality after surgery in Europe: a 7 day cohort study. *Lancet* 2012; **380**: 1059–65.

2. Wijeysundera DN, Pearse RM, Shulman MA, et al. Assessment of functional capacity before major non-cardiac surgery: an international, prospective cohort study. *Lancet* 2018; **391**: 2631–40.
3. Tillmann BW, Hallet J, Guttman MP, et al. A population-based analysis of long-term outcomes among older adults requiring unexpected intensive care unit admission after cancer surgery. *Annals of Surgical Oncology* 2021; **28**: 7014–24.
4. Vonlanthen R, Slankamenac K, Breitenstein S, et al. The impact of complications on costs of major surgical procedures a cost analysis of 1200 patients. *Annals of Surgery* 2011; **254**: 907–13.
5. Pinto A, Faiz O, Davis R, Almoudaris A, Vincent C. Surgical complications and their impact on patients' psychosocial well-being: a systematic review and meta-analysis. *BMJ Open* 2016; **6**: e007224.
6. Bilimoria KY, Liu YM, Paruch JL, Zhou L, Kmiecik TE, Ko CY, Cohen ME. 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**: 833–842.e1-3.
7. Teoh D, Halloway RN, Heim J, Vogel RI, Rivard C. Evaluation of the American College of Surgeons national surgical quality improvement program surgical risk calculator in gynecologic oncology patients undergoing minimally invasive surgery. *Journal of Minimally Invasive Gynecology* 2017; **24**: 48–54.
8. Cohen ME, Liu YM, Ko CY, Hall BL. An examination of American College of Surgeons NSQIP surgical risk calculator accuracy. *Journal of the American College of Surgeons* 2017; **224**: 787–795.e1.
9. Mazzarello S, McIsaac DI, Beattie WS, Fergusson DA, Lalu MM. Risk factors for failure to rescue in myocardial infarction after noncardiac surgery a cohort study. *Anesthesiology* 2020; **133**: 96–108.
10. 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.
11. Wijeysundera DN. Predicting outcomes: is there utility in risk scores? *Canadian Journal of Anesthesia* 2016; **63**: 148–58.
12. Collins GS, Reitsma JB, Altman DG, Moons KGM, Grp T. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *European Urology* 2015; **67**: 1142–51.
13. Moonesinghe SR, McGuckin D, Martin P, et al. The Perioperative Quality Improvement Programme (PQIP patient study): protocol for a UK multicentre, prospective cohort study to measure quality of care and outcomes after major surgery. *Perioperative Medicine* 2022; **11**: 37.
14. 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.
15. Oliver CM, Warnakulasuriya S, McGuckin D, et al. Delivery of drinking, eating and mobilising (DrEaMing) and its association with length of hospital stay after major noncardiac surgery: observational cohort study. *British Journal of Anaesthesia* 2022; **129**: 114–26.
16. Dindo D, Demartines N, Clavien PA. Classification of surgical complications - a new proposal with evaluation in a cohort of 6336 patients and results of a survey. *Annals of Surgery* 2004; **240**: 205–13.
17. Pelosi P, Ball L, Schultz MJ. How to optimize critical care resources in surgical patients: intensive care without physical borders. *Current Opinion in Critical Care* 2018; **24**: 581–7.
18. Gawande AA, Kwaan MR, Regenbogen SE, Lipsitz SA, Zinner MJ. An Apgar score for surgery. *Journal of the American College of Surgeons* 2007; **204**: 201–8.
19. 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.
20. Wong DJN, Oliver CM, Moonesinghe SR. Predicting postoperative morbidity in adult elective surgical patients using the Surgical Outcome Risk Tool (SORT). *British Journal of Anaesthesia* 2017; **119**: 95–105.
21. Nijbroek SG, Schultz MJ, Hemmes SNT. Prediction of postoperative pulmonary complications. *Current Opinion in Anesthesiology* 2019; **32**: 443–51.
22. Moonesinghe SR, Mythen MG, Das P, Rowan KM, Grocott MPW. Risk stratification tools for predicting morbidity and mortality in adult patients undergoing major surgery qualitative systematic review. *Anesthesiology* 2013; **119**: 959–81.
23. Cohn SL, Ros NF. Comparison of 4 cardiac risk calculators in predicting postoperative cardiac complications after noncardiac operations. *American Journal of Cardiology* 2018; **121**: 125–30.
24. Haga Y, Ikei S, Ogawa M. Estimation of Physiologic Ability and Surgical Stress (E-PASS) as a new prediction scoring system for postoperative morbidity and mortality following elective gastrointestinal surgery. *Surgery Today - The Japanese Journal of Surgery* 1999; **29**: 219–25.
25. Copeland GP, Jones D, Walters M. POSSUM: a scoring system for surgical audit. *British Journal of Surgery* 1991; **78**: 355–60.
26. Protopapa KL, Simpson JC, Smith NCE, Moonesinghe SR. Development and validation of the Surgical Outcome Risk Tool (SORT). *British Journal of Surgery* 2014; **101**: 1774–83.
27. Canet J, Gallart L, Gomar C, et al. Prediction of postoperative pulmonary complications in a population-based surgical cohort. *Anesthesiology* 2010; **113**: 1338–50.
28. Lee TH, Marcantonio ER, Mangione CM, et al. Derivation and prospective validation of a simple index for prediction of cardiac risk of major noncardiac surgery. *Circulation* 1999; **100**: 1043–9.
29. Centre for Perioperative Care. Preoperative Assessment and Optimisation for Adult Surgery. 2021. https://www.cpoc.org.uk/preoperative-assessment-and-optimisation-adult-surgery (accessed 17/01/2024).
30. Ladha KS, Wijeysundera DN. Perioperative outcomes: easier to predict but harder to change. *Canadian Journal of Anesthesia* 2019; **66**: 1014–7.
31. Kawamoto K, Houlihan CA, Balas EA, Lobach DF. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. *British Medical Journal* 2005; **330**: 765.

Supporting Information

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

Table S1. Candidate predictor variables for novel predictive models.

Table S2. Characteristics of patients for 24,983 scheduled operations used to train and test the model for postoperative complications.

Table S3. Incidence, severity and type of complications (as classified in the PQIP Case report form) recorded in patients identified as having postoperative Clavien–Dindo ≥ 2 complication(s).

Table S4. Surgical specialty classes and associated multivariate coefficients.

Table S5. Component variables for existing risk tools.

Table S6. Composite pulmonary and cardiovascular complication definitions.

Appendix S1. Method for calculating an individual's predicted likelihood of Clavien–Dindo ≥ 2 complications using the M5 novel model.

Figure S1. Funnel plot of hospital-level incidence of Clavien–Dindo ≥ 2 complications against volume of eligible procedures in patients undergoing one of four colorectal operations.

Figure S2. Calibration belts describing the accuracy of six novel generic models in estimating the percentage likelihood of Clavien–Dindo ≥ 2 complications. A diagonal line indicates perfect calibration.