Title page

Title: Predicting Postoperative Morbidity in Adult Elective Surgical Patients using the Surgical Outcome Risk Tool (SORT)

Authors:

Dr D. J. N. Wong [1,2,3,*]

Dr C. M. Oliver [1,2]

Dr S. R. Moonesinghe [1,2,3]

1. UCL/UCLH Surgical Outcome Research Centre (SOuRCe), 3rd Floor, Maple Link Corridor, University College Hospital, 235 Euston Road, London NW1 2BU, UK
2. National Institute of Academic Anaesthesia Health Services Research Centre, Royal College of Anaesthetists, Churchill House, 35 Red Lion Square, London WC1R 4SG, UK
3. Department of Applied Health Research, University College London, 1–19 Torrington Place, London WC1E 7HB, UK

[*] Dr D. J. N. Wong is the Corresponding Author. Address for correspondence: Department of Applied Health Research, University College London, 1–19 Torrington Place, London WC1E 7HB, UK; email: danny.wong@ucl.ac.uk

Abstract

Background

The Surgical Outcome Risk Tool (SORT) is a risk stratification tool that predicts perioperative mortality. We construct a new recalibrated model based on SORT to predict the risk of developing postoperative morbidity.

Methods

We analysed prospectively collected data from a single-centre cohort of adult patients undergoing major elective surgery. The data set was split randomly into derivation and validation samples. We used logistic regression to construct a model in the derivation sample to predict postoperative morbidity as defined using the validated Postoperative Morbidity Survey (POMS) assessed at one week after surgery. Performance of this "SORT-morbidity" model was then tested in the validation sample, and compared against the Physiological and Operative Severity Score for the enUmeration of Mortality and morbidity (POSSUM).

Results

The SORT-morbidity model was constructed using a derivation sample of 1056 patients, and validated in 527 patients. SORT-morbidity was well-calibrated in the validation sample, as assessed using calibration plots and the Hosmer-Lemeshow Test (χ^2 = 4.87, p = 0.77). It showed acceptable discrimination by Receiver Operator Characteristic (ROC) curve analysis (Area Under the ROC curve, AUROC = 0.72, 95% CI 0.67–0.77). This compared favourably with POSSUM (AUROC = 0.66, 95% CI 0.60–0.71), while remaining simpler to use. Linear shrinkage factors were estimated, which allow the SORT-morbidity model to predict a range of alternative morbidity outcomes with greater accuracy, including low- and high-grade morbidity, and POMS at later time-points.

Conclusions

SORT-morbidity can be used preoperatively, with clinical judgement, to predict postoperative morbidity risk in major elective surgery.

Keywords

Morbidity, Risk Assessment, Postoperative Complications

Introduction

Accurate perioperative risk assessment at an individual patient level supports clinical decision-making and clear communication of risks when consenting for surgery. Additionally, at a hospital or provider level, adjustment for patient case-mix enables surgical outcomes to be meaningfully compared between institutions for health service evaluation or within institutions for clinical audit. A number of risk stratification tools currently exist in clinical practice for both these purposes.[1]

While there are many tools to predict or assess risk based on mortality, for example the Portsmouth modification of the Physiological and Operative Severity Score for the enUmeration of Mortality and morbidity (P-POSSUM)[2] and Surgical Risk Scale (SRS)[3], there are fewer tools available to predict morbidity outcomes. This is despite morbidity being an important outcome of concern to both patients and clinicians. Morbidity following major surgery is firstly a more common outcome than death making it a potentially more sensitive measure by which to compare individual, team or provider performance.[4] Secondly, morbidity can impact significantly on quality of life and long-term survival, making it an important target for quality improvement.[5] Indeed the P-POSSUM was a variation of an earlier tool, the Physiological and Operative Severity Score for the enUmeration of Mortality and morbidity (POSSUM), which was originally developed in the 1990s to address the lack of tools for predicting morbidity outcomes at the time.[1 6] POSSUM remains one of the most frequently used tools for predicting morbidity risk.

The Surgical Outcome Risk Tool (SORT) was developed following the 2011 National Confidential Enquiry into Patient Outcome and Death (NCEPOD) report *Knowing the Risk*, to enable better identification of patients at high risk of postoperative morbidity and mortality, clearer documentation of risks and better discussions about risks with patients before surgery.[7 8] SORT is a parsimonious model using six routinely collected data items, designed to preoperatively predict an individual's probability of 30-day mortality. It compared favourably with other previously validated risk stratification tools, namely the American Society of Anesthesiologists Physical Status (ASA-PS) grade and SRS and has been externally validated recently in a cohort of patients undergoing hip fracture surgery.[9]

Therefore, our aims in this paper are to develop and validate a new model to predict the likelihood of postoperative morbidity using predictor variables found in SORT, and then compare its performance against POSSUM.

Methods

Patient population

Data were prospectively collected from 1934 consecutive adult patients who had undergone a variety of elective major inpatient operations at University College London Hospital between June 2009 and May 2012. Ethics approval was not required as the routine collection of this data was approved as service evaluation by the local research and development office.

Demographic and perioperative data collection was conducted by trained research staff, independent of the clinical care teams treating the patients. Risk variable, co-morbidity, hospital length of stay (LOS) and mortality data were obtained from case note/electronic record review. POMS outcomes were measured by research staff visiting the patients at the bedside, and reviewing case notes. Baseline co-morbidities were recorded for each patient, and, as with the original development and validation of SORT, free-text surgical procedure descriptions were categorised by surgical specialty, and by surgical severity, based on the reference manual for AXA PPP Healthcare Specialist Procedure Codes (examples of AXA PPP procedure coding for surgical severity can be found in Supplementary Data, Table 1).[7 10]

Postoperative Morbidity Survey as the morbidity outcome

The Postoperative Morbidity Survey (POMS) was prospectively administered to patients in the cohort who remained in hospital at the following six time-points: postoperative Day 3, Day 5, Day 7 or 8, Day 14 or 15, Day 21 and Day 28. POMS at the 1-week and 2-week time-points were pragmatically allowed to be recorded at either Day 7or 8, and Day 14 or 15, respectively, in order to reduce the administrative burden placed on research staff.

POMS is an 18-item survey of short-term postoperative morbidity encompassing nine organ-system domains (Supplementary Data, Table 2) and has been multiply validated.[5 11–13] It was developed to identify morbidity of the nature that would prolong LOS, in other words, patients were unlikely to be able to return home if they exhibit POMS-defined morbidity.

Previous studies suggest that POMS-defined morbidity is a measure of "true" morbidity when measured beyond postoperative Day 5.[12 14] Early exploration of the data showed that the median LOS in our cohort was 6 days (Table 1). We assumed that this LOS represented an uncomplicated postoperative course—i.e. even if patients experienced POMS-defined morbidity on Days 3 or 5, this might be expected as part of the usual postoperative course for their surgery. Hence, POMS at either Day 7 or 8 had face validity in this cohort for representing a postoperative course that met with complication(s). We therefore chose the presence of any POMS-defined morbidity, recorded at the third time-point (postoperative Day 7 or 8), as our morbidity outcome measure.

Treatment of cases for analysis

Cases with erroneously duplicated or missing data were excluded, and only cases with complete predictor variable and POMS outcome data were included for analysis. No imputation of missing data was performed. Patients who remained in hospital longer than 7 days but did not have POMS outcomes recorded on Day 7 were then excluded from the data set used to derive a new SORT model for predicting morbidity (SORT-morbidity).

Patient characteristics of the excluded cases were compared to the original data set to look for differences between the groups.

Model derivation and internal validation for a new SORT-morbidity model

The data set was split randomly into two samples: a model derivation sample of approximately two-thirds of the cohort, and a validation sample consisting of the remaining one-third, similar to methods described previously.[7] Based on the six data items from the original SORT model for mortality, multivariable logistic regression models were fitted on the derivation sample with a binary composite outcome variable of POMS-defined morbidity (i.e. if any one of the nine organ-system domains was positive, the patient was considered positive for POMS-defined morbidity) on postoperative Day 7 or 8. Patients who were discharged before postoperative Day 7 were assumed to have no POMS-defined morbidity.

The following logit formula was used for model-fitting:

$$ln(R\ /\ (1-R)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n,$$

where R is the probability of a patient having POMS-defined morbidity 1 week following surgery, $\beta_0, \beta_1, \ldots, \beta_n$ are the model coefficients and $x_1, \ldots, x_n$ are the predictor variables.

For ease of reference, Protopapa's original SORT-mortality model takes the following form:

$$\begin{gathered} ln(R/(1-R)) = \\ -7.366 \\ +1.411\times(\mathrm{ASA-PS\ III}) \\ +2.388\times(\mathrm{ASA-PS\ IV}) \end{gathered}$$

$$
\begin{gathered}
+4.081\times(\text{ASA} - \text{PS V}) \\
+1.236\times(\text{Surgical Urgency } = \text{ Expedited}) \\
+1.657\times(\text{Surgical Urgency } = \text{ Urgent}) \\
+2.452\times(\text{Surgical Urgency } = \text{ Immediate}) \\
+0.712\times(\text{High Risk Surgical Specialty}) \\
+0.381\times(\text{Surgical Severity } = \text{ Xmajor/complex}) \\
+0.667\times(\text{Malignancy}) \\
+0.777\times(\text{Age } 65 - 79 \text{ years}) \\
+1.591\times(\text{Age } 80 \text{ years or more}),
\end{gathered}
$$

where High Risk Surgical Specialties include Gastrointestinal, Thoracic or Vascular Surgery, and R in this model is the risk of 30-day mortality.

The original SORT model used the following predictor variables: ASA-PS Grade (III, IV or V); Surgical Urgency (Expedited, Urgent or Immediate); High-risk Specialties (Gastrointestinal, Thoracic or Vascular surgery); Surgical Severity (Xmajor/Complex); Malignancy; and Age (65–79 or ≥80 years). More detailed definitions of the variables included in the SORT-mortality model are included in Supplementary Data (Supplementary Table 1). We did not include Surgical Urgency in our model-fitting as the patients in this cohort underwent elective procedures only.

Based on the original SORT model, sequential adjustments were manually made to the predictor variables, and new model coefficients were obtained by fitting against the model derivation sample of the cohort to obtain a number of candidate models (Supplementary Data, Table 3). Goodness-of-fit was then assessed between models by comparing their Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC).[15–18]

To perform internal validation, each candidate SORT-morbidity model was then tested in the validation data set by assessing model calibration and discrimination, using calibration plots and the Hosmer-Lemeshow Test for the former, and Receiver Operator Characteristic (ROC) analysis for the latter.[1 19–22] Areas Under the ROC curve (AUROC) were calculated, and DeLong's method was applied to calculate confidence intervals. The Hosmer–Lemeshow test compares observed and predicted risk across the range of predicted risk, with non-significant Chi-squared test results ($p > 0.05$) indicating a well-calibrated model. The AUROC takes values between 0.5 and 1.0, where less than 0.7 identifies a model with poor performance, 0.7–0.9 indicates acceptable or moderate performance, and over 0.9 indicates high performance.

A final parsimonious model (the simplest model that makes an accurate prediction with as few predictor variables as possible) was then chosen from this process. We selected our final model based on a balanced assessment of the goodness-of-fit measures (low AIC and BIC values), calibration (non-significant p-value for Hosmer-Lemeshow test and visual inspection of calibration plots in the validation sample), and discrimination (high values of AUROC in the validation sample). We then compared the performance of our final model to POSSUM.

Sensitivity analyses and model recalibration for predicting alternative morbidity outcomes

We conducted a sensitivity analysis to test the performance of SORT-morbidity for predicting different outcomes, due to concerns that POMS-defined morbidity outcomes may be too sensitive. These concerns arose because of some less severe and relatively transient morbidity contained within the POMS outcome definitions, such as the presence of a urinary catheter and the use of anti-emetics.

Sensitivity analyses for regression models are typically performed by assessing the accuracy of the outcome prediction based on changes in the variability of the predictor variables. However, in this study we decided to test the accuracy of our model's outcome prediction if the definition used for outcome changed, in order to give the reader an understanding of the limitations of our model. This was done by assessing the calibration and discrimination of SORT-morbidity for predicting: 1) a modified POMS outcome in our patient cohort with lower- and higher-grade morbidity; and 2) POMS outcomes at time-points beyond 1 week after surgery.

For the first step, we mapped the individual POMS organ-system sub-domain items against Clavien-Dindo definitions of postoperative complications[23 24], and assigned Clavien-Dindo grades to each POMS sub-domain. This process yielded a list of transient, low-grade POMS morbidity (equivalent to Clavien-Dindo Grade 1), which comprise: the presence of a urinary catheter (renal), presence of a fever (infectious), inability to tolerate oral diet (gastrointestinal), vomiting and abdominal distension (gastrointestinal), and pain requiring opioid analgesia (pain). All other POMS-defined morbidity sub-domains were considered high-grade (Clavien-Dindo Grade 2 or more). There was a consensus among the authors as to what were considered low-grade and high-grade POMS outcomes. We then assessed SORT-morbidity performance in predicting the modified low-grade and high-grade POMS outcomes. Details of how the POMS organ-system sub-domains were mapped against Clavien-Dindo grades can be found in Supplementary Data (Supplementary Table 4).

For the second step, we assessed SORT-morbidity performance in predicting more delayed POMS-defined morbidity on Day 14 or 15, Day 21, and Day 28.

Lastly, using methods described by Harrell[25], we obtained estimates of uniform linear shrinkage factors for mis-calibrations identified in steps 1) and 2) by fitting the following logit formula:

$$\mathrm{logit}(\mathrm{P}(\mathrm{Y})) = \alpha + \beta \times \mathrm{logit}(\mathrm{p}),$$

where P(Y) is the probability of the modified outcomes either of low- or high-grade POMS, or POMS outcomes at later time-points, logit(p) is the SORT-morbidity formula as derived earlier, α is the recalibrated intercept and β is the recalibrated slope. By applying the shrinkage factors, calibration would then be improved, and the accuracy of SORT-morbidity for predicting these other outcomes would also be improved.

The linear shrinkage factors estimated by this method can then applied directly onto the SORT-morbidity equation to uniformly modify all the coefficients, $\beta_0, \beta_1, \ldots, \beta_n$. This technique therefore accomplishes two objectives in our study: firstly, the shrinkage factors quantify the degree to which SORT-morbidity is mis-calibrated when predicting other outcomes of different severity; and secondly, the shrinkage factors can then be used for adjustments to the SORT-morbidity equation to arrive at a better accuracy for predicting these other outcomes.

Statistical analysis

Statistical analyses were performed using R version 3.3.1 (R Foundation for Statistical Computing, Vienna, Austria), with the following external packages enabled: *caret*, *PredictABEL*, *pROC*, *rms*. For normally distributed data, means and standard deviations are reported. For non-normally distributed data, medians and interquartile ranges (IQR) are reported. A p-value of ≤ 0.05 was considered statistically significant. Logistic regression models were fitted using the *glm* function and shrinkage factors were estimated using the *val.prob* function from the *rms* package in R.

Results

Baseline patient characteristics and outcomes

Following exclusions 1583 patients were used to perform model derivation and internal validation for morbidity prediction. Figure 1 shows the numbers of patients excluded from the study at each stage of analysis. Baseline patient characteristics of the study cohort and patients excluded from model derivation and validation are summarised in Table 1. The median age of the study patients was 62.6, with a range of 17 to 95.4 years. The majority of patients (58.0%) were female. The majority of the patients in the study sample underwent Orthopaedic procedures (n = 873, 45.1%), with a substantial proportion undergoing Abdominal (either Colorectal or Upper GI) procedures (n = 885, 39.3%). Mortality in this patient sample was low, with only 6 deaths within 30 days of surgery recorded (0.31%). Median length of stay was 6 days (interquartile range: 4–10). Compared with the total patient cohort, patients with missing data had a higher age, longer average length of stay, and a different distribution of ASA-PS grades. Table 2 shows the proportion of patients with POMS-defined morbidity at each time point that POMS was administered, as a percentage of the total patient cohort.

Model derivation and internal validation for a new SORT-morbidity model

After randomly splitting the patient cohort, multivariable logistic regression models were fitted with predictor variables based on the original SORT model data categories in a derivation sample of 1056 patients, with the outcome variable set as the presence of POMS-defined morbidity on postoperative Day 7 or 8.

The coefficients for the final SORT-morbidity model are summarised in Table 3, where the original SORT model formula is displayed in the table caption for comparison. Intermediate models fitted during the step-wise manual adjustments of model variables are summarised in Supplementary Data (Supplementary Table 3), along with their corresponding performance statistics. Details of the Hosmer-Lemeshow tests, and the tables of predicted and observed morbidity outcomes by risk quantile for SORT-morbidity and POSSUM are also included in Supplementary Data.

The novel SORT-morbidity model demonstrated good calibration in the validation sample (n = 527) as shown by the calibration plot (Figure 2(A)) and the Hosmer-Lemeshow test ($\chi^2 = 4.87$, p = 0.77), and acceptable discrimination (AUROC = 0.72, 95% CI 0.67–0.77, Figure 2(B)). SORT-morbidity compared favourably with POSSUM, which showed poorer calibration (Hosmer-Lemeshow $\chi^2 = 34.45$, p < 0.001, Figure 2) and poor discrimination (AUROC = 0.66, 95% CI 0.60–0.71). However, DeLong's test for two correlated ROC curves did not show a statistically significant difference in discrimination between SORT-morbidity and POSSUM (z = 1.85, p = 0.06). These findings suggest that SORT-morbidity is as accurate as POSSUM for predicting POMS-defined morbidity at 1 week following surgery.

Sensitivity analyses and model recalibration for predicting alternative morbidity outcomes

In assessing the performance of SORT-morbidity for modified POMS outcomes, our model showed poor calibration (Hosmer-Lemeshow test results for both were statistically significant, both p <0.01) and acceptable discrimination for predicting both low-grade and high-grade POMS morbidity (AUROC = 0.75, 95% CI 0.72–0.78; and AUROC = 0.72, 95% CI 0.69–0.76, respectively).

Similar to our findings for predicting high-grade POMS morbidity, SORT-morbidity was good-to-acceptable at discrimination, but poorly calibrated for Day 14 or 15, Day 21 and Day 28 POMS outcomes (Figure 3). The AUROC for these outcomes ranged from 0.73 for Day 14 or 15 morbidity to 0.81 for Day 28 morbidity.

We estimated linear shrinkage factors to be applied to the SORT-morbidity model in order to improve its accuracy when predicting low- and high-grade POMS morbidity, and morbidity at alternative time-points. Table 4 shows the recalibration intercepts, α, and slopes, β, that can be applied to the SORT-morbidity formula, as well as the corresponding AUROCs for the recalibrated formulae.

Discussion

Principal findings

We have developed and internally validated the SORT-morbidity model, a new refitted multivariable logistic regression model using the original SORT data items as predictors for POMS-defined morbidity one week after surgery. Our new model compares favourably with POSSUM and can be used in conjunction with the original SORT mortality model to inform clinical decisions and the consent process in major surgery.

Additionally, we have estimated linear shrinkage factors which can be applied to SORT-morbidity to improve its prediction for a range of other morbidity outcomes, including higher-grades of morbidity and POMS-defined morbidity at later time-points. With the application of these shrinkage factors, SORT-morbidity is able to therefore predict a range of different morbidity outcomes with acceptable discrimination.

Strengths and limitations of the study

In this study, we demonstrate that the performance of the new SORT-morbidity model is at least comparable with POSSUM, which is currently one of the most widely used risk prediction models for morbidity.[1] SORT-morbidity requires substantially fewer predictor variables than POSSUM, making it easier to use in routine clinical practice. Furthermore, we rigorously test the prediction performance of SORT-morbidity against a number of different morbidity outcome definitions, and estimate shrinkage factors which can be applied to the model to improve its performance for predicting these alternative outcomes.

The biggest limitation of our study is that it was conducted in a single-centre cohort of patients undergoing elective surgery which had only 6 (0.31%) deaths within 30 days of surgery. This is a low-mortality cohort compared with the 0.36–0.67% 30-day mortality for elective surgical cohorts reported elsewhere in the literature.[8 26 27] We were therefore unable to perform an external validation of SORT for the prediction of 30-day mortality.

Our cohort exhibited some missing data, in both predictor and outcome variables. We opted for complete-case analysis instead of imputation of missing data. We considered the number cases with missing predictor variables to be small: no single predictor variable had more than 5% of data missing. Examination of the missing data suggested that missing data did not fulfil the Missing Completely At Random (MCAR) or Missing At Random (MAR) assumptions that are required for imputation. The variable with the largest amount of missing data was the POMS outcome variable (Figure 1). We suspected that the mechanism underlying the degree of missingness in our data was related to how unwell the patients were, thus making the data likely to be Missing Not At Random (MNAR). The evidence for this was that patients with missing data appeared to be older, have higher ASA-PS grades and had longer LOS (Table 1). Imputation under such circumstances potentially leads misleading results and may increase the bias in statistical estimates.

Defining morbidity according to POMS

We chose to define morbidity using POMS at one week following surgery in our analysis. The strengths and weaknesses of POMS therefore merit discussion.

There is currently no consensus on the best way to measure postoperative morbidity and surgical complications.[28 29] In their seminal paper describing the development of POSSUM, Copeland *et al* defined morbidity as any recorded occurrence of a list of complications appearing within 6-weeks follow-up after surgery.[6] When reporting their development of P-POSSUM, Prytherch *et al* cited difficulties over defining morbidity and uncertainty over accuracy in recording complications as the reasons for not including morbidity prediction in their model.[2]

POMS is one of the only validated measures of short-term morbidity, and allows for a binary outcome to be modelled (presence or absence of morbidity).[11] It is a highly sensitive measure, and detects associated morbidity of a magnitude which requires the patient to remain in hospital. POMS criteria are prescriptive, and have previously been shown to have very high inter-rater agreement.[12] Although POMS domains include some relatively minor morbidity (such as the presence of a urinary catheter) —which, particularly in the first few days after surgery, may be indicative of normal processes and pathways after major surgery rather than reflecting the occurrence of complications. In our study population, the median length of stay was 6 days, suggesting that any morbidity observed after this time point may be as a consequence of morbidity leading to a deviation from the usual patient pathway. Previous research has shown the presence of prolonged POMS-defined morbidity to be correlated with poorer long-term outcomes.[14]

We propose that POMS offers a more reliable and repeatable means of measuring morbidity than the original Copeland criteria, and one which is suitable for the modern-day evaluation of postoperative outcomes for the purposes of quality improvement and comparative audit. However, we acknowledge that POMS at Day 7 or 8 may be perceived as being too sensitive an outcome measure for some clinicians. The findings from our sensitivity analysis mitigate against this shortcoming by providing estimates of how miscalibrated SORT-morbidity is for predicting alternative morbidity outcomes, through the calculation of shrinkage factors (intercept and slope corrections).

Our results in relation to existing literature

In the original SORT development study by Protopapa *et al*, SORT performance was compared against ASA-PS and a modified version of SRS, and was found to be more accurate than either of the comparators.[7]

Marufu *et al* recently reported an AUROC of 0.70 when validating SORT against an emergency hip fracture population.[9] They compared SORT in their study against the Nottingham Hip Fracture Score (NHFS), which showed equivalent discrimination but better calibration. This suggests that SORT is better used as a risk stratification tool for heterogeneous surgical patient groups, and that other tools may be more appropriate for more homogeneous or surgery-specific case-mixes.[30]

Clinical Implications

P-POSSUM and POSSUM are the most frequently and widely-validated risk stratification model for heterogeneous populations in the literature.[1] It therefore has widespread familiarity amongst anaesthetists and surgeons internationally. However, their value in preoperative prognostication is reduced by three factors: firstly, the large number (18) of variables required by their models; secondly, a number of their required variables are only available after surgery; and thirdly, some of the predictor variables are blood test results, which not all patients may have available preoperatively.[1 7]

SORT and SORT-morbidity are both more parsimonious models compared to P-POSSUM and POSSUM, requiring fewer data items to make a morbidity prediction than POSSUM. Furthermore, SORT and SORT-morbidity require only preoperative data variables in order to make risk predictions for mortality and morbidity respectively, and in particular, variables which are all known at the time of preoperative assessment in the outpatient clinic, when laboratory variables may not yet be available. These factors therefore make them simpler and more practical to use for preoperative morbidity risk prediction than POSSUM. Risk stratification tools may be more likely to be adopted if they are simple, and when the performance of two or more models are similar, the simplest (most parsimonious) model would be preferable for clinical use.

SORT has been made available for bedside use as an online risk prediction calculator accessible via an internet browser and more recently via smartphone applications developed for both iOS and Android devices.[31 32] Although SORT-morbidity differs from SORT in how some of the data variables are used in the model —for example, SORT-morbidity uses ASA-PS as a categorical variable with 4 categories (ASA I, ASA II, ASA III, or ASA ≥IV) as opposed to 3 categories as in the original SORT model (ASA I–II, ASA III, or ASA ≥IV)— it does not require any new data to be collected. It would therefore be straightforward to incorporate the new SORT-morbidity prediction model into these electronic tools, and they can easily be used at the bedside in conjunction with careful clinical assessment to support decision-making and promote informed patient consent preoperatively.

SORT-morbidity when used together with linear shrinkage factors, therefore provides a potentially powerful tool for perioperative shared decision-making by predicting the risk of developing a range of postoperative morbidity outcomes, over a number of time-points.

There is much added benefit in being able to predict morbidity rather than just mortality, both from an institutional-level and patient-level perspective. At the institutional-level, hospital resource utilisation increases with increased patient morbidity following surgery. The ability to accurately predict those patients who might experience prolonged lengths of stay and more interventions as a result of complications allows for better planning and resource allocation. At the patient-level, morbidity risk prediction allows for more informed discussions between patients and clinicians when deciding to undergo surgery. During these discussions, the risk of suffering morbidity and mortality can thus be better weighed up against the potential benefit surgery in improving quality and duration of life.

Unanswered questions and future research

External validation of the new SORT-morbidity model is needed to assess its generalisability in other patient populations. There would also be scope in refitting the SORT-morbidity model to include surgical urgency as a predictor variable. SORT-morbidity could also be recalibrated predict morbidity outcomes according to other definition systems, for example the Clavien-Dindo classification of surgical complications, which produces an ordinal outcome variable, ranging from Grade 1 (least severe complications) to Grade 5 (most severe).[23 24]

In the future, consistent measurement of postoperative morbidity and mortality outcomes and comparing performance between institutions is likely to be increasingly important as healthcare costs increase with an ageing population. SORT and SORT-morbidity can therefore be used together for local departmental audits of practice to assess how well patients recover from surgery against predicted mortality and morbidity. They would also have value in case-mix risk-adjustment and benchmarking in national registries for tracking institutional quality and perioperative outcomes. Examples of such registries include the American College of Surgeons National Surgical Quality Improvement Program (ACS-NSQIP) in the U.S.A., and the Perioperative Quality Improvement Programme (PQIP) currently under development in the U.K.[33 34] Both tools would need periodic recalibration when used within such contexts to ensure they remain valid for surgical populations over time.

Conclusion

We show that SORT-morbidity can be used preoperatively to predict postoperative morbidity risk. Risk stratification tools offer clinicians a means of providing more accurate information to patients, help to guide perioperative care decisions, and allow for case-mix adjustments between institutions for audit and research purposes. As mobile digital devices become more widely available, accurate, simple and parsimonious risk stratification tools such as SORT and SORT-morbidity will become increasingly accessible to clinicians for use at the bedside.

Details of authors contributions

The study was conceived by C.M.O. and S.R.M. Study data were prospectively collected by UCLH clinical staff and research nurses as part of routine service evaluation. Data linkage and cleaning was performed by D.J.N.W. and C.M.O. Analysis was performed by D.J.N.W. and C.M.O., with input to analysis from S.R.M. The manuscript was drafted by D.J.N.W. and subsequently revised after critical review by all authors.

Acknowledgements

The authors thank the UCL/UCLH SOuRCe research nurses, Denise Wyndham and Ernesto Bettini, for all data collected.

Declaration of interests

S.R.M. was senior author on the original SORT development and validation manuscript. She is Director of the NIAA Health Services Research Centre and the UCLH Surgical Outcomes Research Centre and Associate National Clinical Director for elective care at NHS England.

Funding

The UCLH Surgical Outcomes Research Centre is funded by the University College Hospitals Foundation Trust/University College London National Institute for Health Research (NIHR) Comprehensive Biomedical Research Centre, which received a portion of its funding from the UK Department of Health Biomedical Research Centre's funding scheme; S.R.M. receives funding from the BRC and is a member of the faculty. S.R.M. is also an Improvement Science

Fellow funded by the Health Foundation and supported for her role as Director of the NIAA Health Services Research Centre by funding from the Royal College of Anaesthetists.

D.J.N.W. receives a clinical salary from The London Clinic hospital.

References

1. 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

2. Prytherch DR, Whiteley MS, Higgins B, Weaver PC, Prout WG, Powell SJ. POSSUM and portsmouth POSSUM for predicting mortality. *Br J Surg* 1998;**85**:1217–20

3. Sutton R, Bann S, Brooks M, Sarin S. The surgical risk scale as an improved tool for risk-adjusted analysis in comparative surgical audit. *Br J Surg* 2002;**89**:763–8

4. Walker K, Neuburger J, Groene O, Cromwell DA, Meulen J van der. Public reporting of surgeon outcomes: Low numbers of procedures lead to false complacency. *The Lancet* 2013;**382**:1674–7

5. Davies SJ, Francis J, Dilley J, Wilson RJT, Howell SJ, Allgar V. Measuring outcomes after major abdominal surgery during hospitalization: Reliability and validity of the postoperative morbidity survey. *Perioperative Medicine* 2013;**2**:1

6. Copeland GP, Jones D, Walters M. POSSUM: A scoring system for surgical audit. *Br J Surg* 1991;**78**:355–60

7. Protopapa KL, Simpson JC, Smith NCE, Moonesinghe SR. Development and validation of the surgical outcome risk tool (SORT). *Br J Surg* 2014;**101**:1774–83

8. Findlay GP, Goodwin APL, Protopapa K, Smith NCE, Mason M. Knowing the risk: A review of the peri-operative care of surgical patients. National Confidential Enquiry into Patient Outcome; Death (NCEPOD); 2011.

9. Marufu TC, White SM, Griffiths R, Moonesinghe SR, Moppett IK. Prediction of 30-day mortality after hip fracture surgery by the nottingham hip fracture score and the surgical outcome risk tool. *Anaesthesia* 2016;n/a–a

10. AXA PPP healthcare: Specialist procedure codes. 2016. Available from: https://online.axappphealthcare.co.uk/SpecialistForms/SpecialistCode.mvc/Print?source=contracted

11. Bennett-Guerrero E, Welsby I, Dunn TJ, et al. The use of a postoperative morbidity survey to evaluate patients with prolonged hospitalization after routine, moderate-risk, elective surgery: *Anesthesia & Analgesia* 1999;**89**:514–9

12. Grocott MPW, Browne JP, Van der Meulen J, et al. The postoperative morbidity survey was validated and used to describe morbidity after major surgery. *Journal of Clinical Epidemiology* 2007;**60**:919–28

13. Goodman BA, Batterham AM, Kothmann E, et al. Validity of the postoperative morbidity survey after abdominal aortic aneurysm repair—a prospective observational study. *Perioper Med (Lond)* 2015;**4**

14. Moonesinghe SR, Harris S, Mythen MG, et al. Survival after postoperative morbidity: A longitudinal observational cohort study. *Br J Anaesth* 2014;**113**:977–84

15. Posada D, Buckley TR. Model selection and model averaging in phylogenetics: Advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. *Syst Biol* 2004;**53**:793–808

16. Akaike H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 1974;**19**:716–23

17. Akaike H. Information theory and an extension of the maximum likelihood principle. In: Parzen E, Tanabe K, Kitagawa G, editors. *Selected papers of hirotugu akaike* Springer New York; 1998. p. 199–213

18. Schwarz G. Estimating the dimension of a model. *Ann Statist* 1978;**6**:461–4

19. Hosmer DW, Hosmer T, Le Cessie S, Lemeshow S. A comparison of goodness-of-fit tests for the logistic regression model. *Stat Med* 1997;**16**:965–80

20. Metz CE, Herman BA, Roe CA. Statistical comparison of two ROC-curve estimates obtained from partially-paired datasets. *Med Decis Making* 1998;**18**:110–21

21. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics* 1988;**44**:837–45

22. Swets JA. Measuring the accuracy of diagnostic systems. *Science* 1988;**240**:1285–93

23. Dindo D, Demartines N, Clavien P-A. Classification of surgical complications. *Ann Surg* 2004;**240**:205–13

24. Clavien PA, Barkun J, Oliveira ML de, et al. The clavien-dindo classification of surgical complications: Five-year experience. *Ann Surg* 2009;**250**:187–96

25. Harrell F. Regression modeling strategies - with applications to linear models, logistic regression, and survival analysis. Springer; 2001.

26. Pearse RM, Harrison DA, James P, et al. Identification and characterisation of the high-risk surgical population in the united kingdom. *Crit Care* 2006;**10**:R81

27. Aylin P, Alexandrescu R, Jen MH, Mayer EK, Bottle A. Day of week of procedure and 30 day mortality for elective surgery: Retrospective analysis of hospital episode statistics. *BMJ* 2013;**346**:f2424

28. Myles PS, Grocott MPW, Boney O, et al. Standardizing end points in perioperative trials: Towards a core and extended outcome set. *Br J Anaesth* 2016;**116**:586–9

29. Boney O, Moonesinghe SR, Myles PS, Grocott MPW. Standardizing endpoints in perioperative research. *Can J Anesth/J Can Anesth* 2016;**63**:159–68

30. Kehlet H, Jørgensen CC. Predicting postoperative morbidity: In what procedures and what patients? *Anesthesiology* 2014;**120**:1297

31. NCEPOD, SOuRCe. Surgical outcome risk tool (SORT). 2015. Available from: http://www.sortsurgery.com/SORT_home

32. NCEPOD surgical outcome risk tool. Cranworth medical. 2016. Available from: http://www.cranworthmedical.co.uk/medical-apps/ncepod-surgical-outcome-risk-tool/

33. ACS national surgical quality improvement program. American college of surgeons. 2016. Available from: https://www.facs.org/quality-programs/acs-nsqip

34. Moonesinghe SR, Grocott MPW. Towards a national perioperative quality improvement programme (PQIP). *Bulletin of the Royal College of Anaesthetists* 2015;12–3

Table 1

Table 1 Caption: Descriptive data for the total study population, derivation and validation samples, and patients excluded from analysis due to missing data, including 30-day mortality and length of hospital stay. Numbers in parentheses represent sample percentages. Numbers in square brackets represent either median ages or median lengths of hospital stay with interquartile ranges (IQR) for each sample.

Variable	Total Patient Cohort, No. of patients (%)	Derivation Sample for new Morbidity model, No. of patients (%)	Validation Sample, No. of patients (%)	Patients excluded due to missing data, No. of patients (%)
Total	1934(100)	1056(100)	527(100)	351(100)
Age				
[Median, IQR]	[62.6, IQR:48.7–71.7]	[61.5, IQR:47–71.1]	[61.6, IQR:49.2–71.4]	[66.0, IQR:56.4–73.1]
<65	1099(56.8)	635(60.1)	308(58.4)	156(44.4)
≥65 and <80	646(33.4)	322(30.5)	179(34.0)	145(41.3)
≥80	174(9.0)	99(9.4)	40(7.6)	35(10.0)
Sex				
Female	1121(58.0)	614(58.1)	303(57.5)	204(58.1)
ASA-PS Grade				
1	313(16.2)	185(17.5)	84(15.9)	44(12.5)
2	1161(60.0)	623(59.0)	322(61.1)	216(61.5)
3	432(22.3)	238(22.5)	108(20.5)	86(24.5)
4	26(1.3)	10(0.9)	12(2.3)	4(1.1)
5	1(0.1)	0(0.0)	1(0.2)	0(0.0)
Surgical Specialty				
Orthopaedic	873(45.1)	470(44.5)	243(46.1)	160(45.6)
Colorectal	652(33.7)	332(31.4)	161(30.6)	159(45.3)
Upper GI	108(5.6)	68(6.4)	34(6.5)	6(1.7)
Vascular	122(6.3)	70(6.6)	33(6.3)	19(5.4)
Bariatric	125(6.5)	82(7.8)	40(7.6)	3(0.9)
Other	53(2.7)	34(3.2)	16(3.0)	3(0.9)
Severity				
Minor	108(5.6)	66(6.2)	31(5.9)	11(3.1)
Intermediate	108(5.6)	53(5.0)	29(5.5)	26(7.4)

Major	211(10.9)	107(10.1)	60(11.4)	44(12.5)
Xmajor/complex	1507(77.9)	830(78.6)	407(77.2)	270(76.9)
Co-morbidities				
None Documented	1009(52.2)	561(53.1)	282(53.5)	166(47.3)
Arrhythmia	104(5.4)	52(4.9)	27(5.1)	25(7.1)
Prev. Myocardial Infarct	81(4.2)	41(3.9)	25(4.7)	15(4.3)
Congestive Cardiac Failure	23(1.2)	11(1.0)	5(0.9)	7(2.0)
Peripheral Vascular Disease	97(5.0)	54(5.1)	24(4.6)	19(5.4)
Cerebrovascular Disease	92(4.8)	51(4.8)	18(3.4)	23(6.6)
Chronic Obstructive Pulmonary Disease	220(11.4)	113(10.7)	61(11.6)	46(13.1)
Connective Tissue Disease	48(2.5)	16(1.5)	13(2.5)	19(5.4)
Peptic Ulcer Disease	39(2.0)	19(1.8)	13(2.5)	7(2.0)
Diabetes Mellitus (uncomplicated)	185(9.6)	107(10.1)	52(9.9)	26(7.4)
Diabetes Mellitus (with end-organ involvement)	1(0.1)	0(0.0)	0(0.0)	1(0.3)
Chronic Kidney Disease (moderate to severe)	16(0.8)	8(0.8)	4(0.8)	4(1.1)
Hemiplegia	6(0.3)	3(0.3)	2(0.4)	1(0.3)
Malignancy	103(5.3)	58(5.5)	23(4.4)	22(6.3)
Liver Disease	18(0.9)	14(1.3)	2(0.4)	2(0.6)
Smoking (active)	334(17.3)	187(17.7)	80(15.2)	67(19.1)
Smoking (previously)	538(27.8)	309(29.3)	136(25.8)	93(26.5)
30-day Mortality	6(0.3)	2(0.2)	1(0.2)	3(0.9)
Length of Hospital Stay				
[Median days, IQR]	[6, IQR:4–10]	[6, IQR:4–11]	[5, IQR:4–10]	[7, IQR:7–9]
0–3 days	376(19.4)	246(23.3)	122(23.1)	8(2.3)

4–6 days	608(31.4)	387(36.6)	201(38.1)	20(5.7)
7–13 days	629(32.5)	236(22.3)	120(22.8)	273(77.8)
14–20 days	142(7.3)	86(8.1)	38(7.2)	18(5.1)
21–27 days	52(2.7)	33(3.1)	9(1.7)	10(2.8)
≥28 days	127(6.6)	68(6.4)	37(7.0)	22(6.3)

Table 2

Table 2 Caption: Breakdown of POMS outcomes at different time-points and by organ-system domains. Percentages in parentheses represent the number of patients remaining in hospital at a particular time-point suffering from POMS-defined morbidity as a proportion of the total number of patients in the whole cohort (total n = 1934).

	Day 3	Day 5	Day 7 or 8	Day 14 or 15	Day 21	Day 28
Total number of patients still in hospital, Day 0 = 1934 (100%)	1763 (91.2%)	1373 (71%)	950 (49.1%)	321 (16.6%)	179 (9.3%)	127 (6.6%)
Patients with POMS-defined morbidity (%)	1124 (58.1%)	648 (33.5%)	403 (20.8%)	155 (8%)	68 (3.5%)	47 (2.4%)
Pulmonary (%)	436 (22.5%)	160 (8.3%)	74 (3.8%)	29 (1.5%)	17 (0.9%)	11 (0.6%)
Infectious (%)	216 (11.2%)	159 (8.2%)	128 (6.6%)	58 (3%)	31 (1.6%)	22 (1.1%)
Renal (%)	711 (36.8%)	377 (19.5%)	196 (10.1%)	66 (3.4%)	33 (1.7%)	19 (1%)
Gastrointestinal (%)	520 (26.9%)	322 (16.6%)	222 (11.5%)	84 (4.3%)	40 (2.1%)	29 (1.5%)
Cardiovascular (%)	68 (3.5%)	38 (2%)	36 (1.9%)	15 (0.8%)	7 (0.4%)	6 (0.3%)
Neurological (%)	42 (2.2%)	34 (1.8%)	21 (1.1%)	7 (0.4%)	6 (0.3%)	6 (0.3%)
Haematological (%)	57 (2.9%)	23 (1.2%)	11 (0.6%)	4 (0.2%)	4 (0.2%)	2 (0.1%)
Wound (%)	207 (10.7%)	120 (6.2%)	81 (4.2%)	55 (2.8%)	23 (1.2%)	19 (1%)
Pain (%)	674 (34.9%)	172 (8.9%)	84 (4.3%)	33 (1.7%)	17 (0.9%)	9 (0.5%)
Missing POMS outcomes (%)	266 (14.6%)	282 (15.4%)	243 (13.3%)	73 (4%)	71 (3.9%)	53 (2.9%)

Table 3

Table 3 Caption: Coefficients for the final SORT-morbidity Prediction Model.

	Estimate	Std. Error	z value	p value
ASA = I	Reference			
ASA = II	0.332	0.229	1.453	0.146
ASA = III	1.140	0.257	4.426	0.000
ASA ≥ IV	1.223	0.743	1.646	0.100
Surgical Specialty (Orthopaedic)	Reference			
Surgical Specialty (Colorectal)	1.658	0.221	7.491	0.000
Surgical Specialty (Upper GI)	-0.929	0.439	-2.116	0.034
Surgical Specialty (Vascular)	0.296	0.316	0.937	0.349
Surgical Specialty (Bariatric)	-1.065	0.454	-2.345	0.019
Surgical Specialty (Other)	0.181	0.530	0.341	0.733
Surgical Severity (Minor/Intermediate/Major)	Reference			
Surgical Severity (Xmajor/Complex)	1.238	0.233	5.308	0.000
No Malignancy	Reference			
Malignancy	0.897	0.328	2.735	0.006
Age 16 to 64 years	Reference			
Age 65 to 79 years	0.118	0.183	0.647	0.518
Age 80 years or more	0.550	0.255	2.155	0.031
Constant	-3.228	0.325	-9.932	0.000

Table 4

Table 4 Caption: Shrinkage factors to be applied to SORT-morbidity formula in order to improve the accuracy for predicting alternative morbidity outcomes. AUROC for the recalibrated formulae are also displayed. An example of how to use the shrinkage factors follows: if applying the SORT-morbidity formula yields a probability (p) of having POMS-defined morbidity on Day 7 or 8 of 50%; then logit(p) = 0, and logit(P(Y)) = -1.478 + 0.894(0), where P(Y) is the probability of having POMS-defined morbidity on Day 14 or 15. P(Y) in this example would therefore be 18.6%.

Outcome	Intercept, α	Slope, β	AUROC (95% CI)
Low-grade POMS morbidity	-0.316	1.008	0.75 (0.72–0.78)
High-grade POMS morbidity	-0.874	0.827	0.72 (0.69–0.76)
Day 14/15 POMS	-1.478	0.894	0.73 (0.69–0.78)
Day 21 POMS	-2.327	1.081	0.79 (0.74–0.85)
Day 28 POMS	-2.770	1.048	0.81 (0.76–0.87)

Figure 1

Figure 1 Caption: Flow diagram summarising cases included and excluded from analysis.

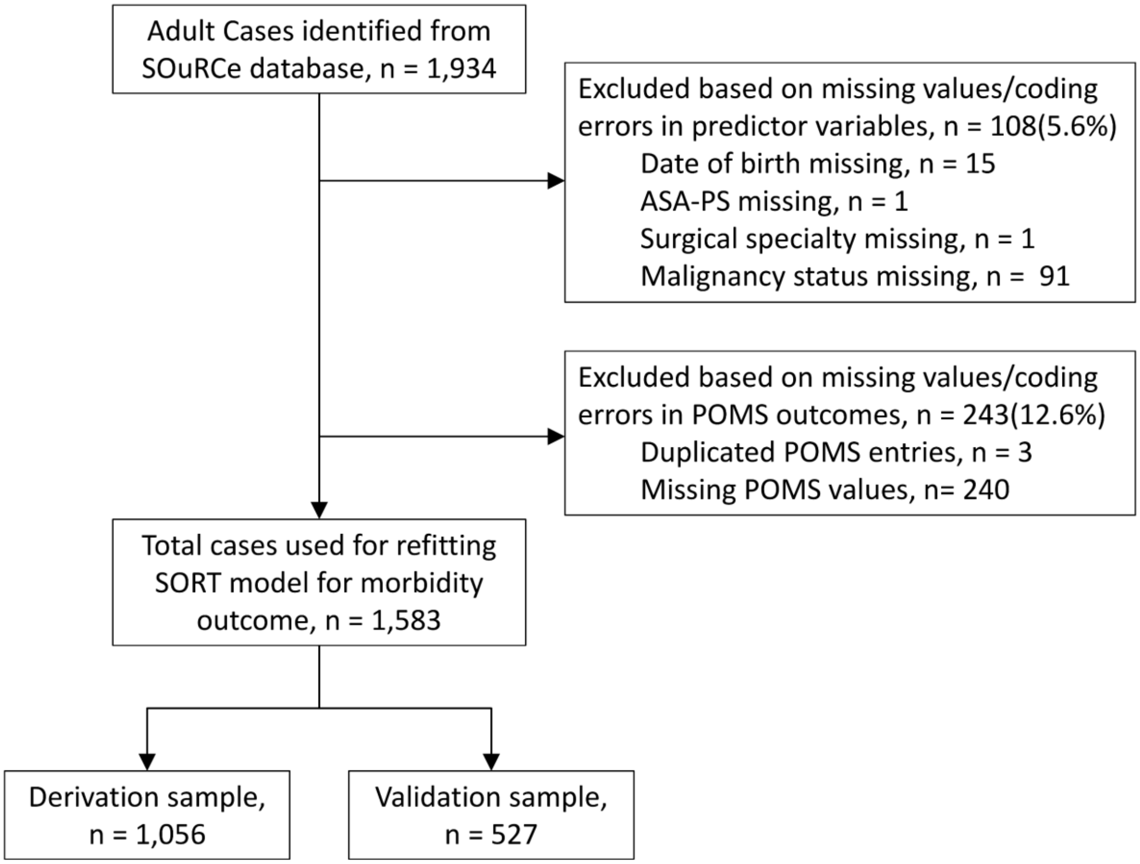

Adult Cases identified from SOuRCe database, n = 1,934
Excluded based on missing values/coding errors in predictor variables, n = 108(5.6%)
Date of birth missing, n = 15
ASA-PS missing, n = 1
Surgical specialty missing, n = 1
Malignancy status missing, n = 91
Excluded based on missing values/coding errors in POMS outcomes, n = 243(12.6%)
Duplicated POMS entries, n = 3
Missing POMS values, n= 240
Total cases used for refitting SORT model for morbidity outcome, n = 1,583
Derivation sample, n = 1,056
Validation sample, n = 527

Figure 2

Figure 2 Caption: (A) Calibration plot of SORT-morbidity compared to POSSUM: Observed versus predicted occurrence of Day 7 or 8 morbidity at varying levels of risk in the validation cohort of 527 patients; (B) ROC curve plot of SORT-morbidity compared against POSSUM (AUROC = 0.72, and 0.66 respectively) tested in the validation cohort. ROC: Receiver Operator Characteristic; AUROC: Area Under the ROC curve.

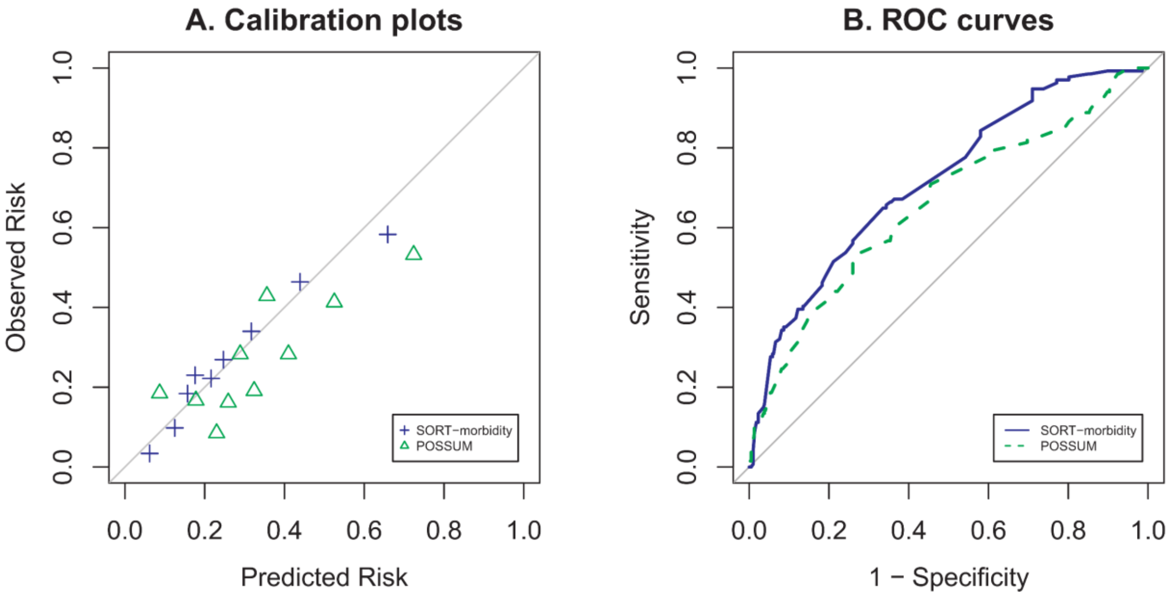
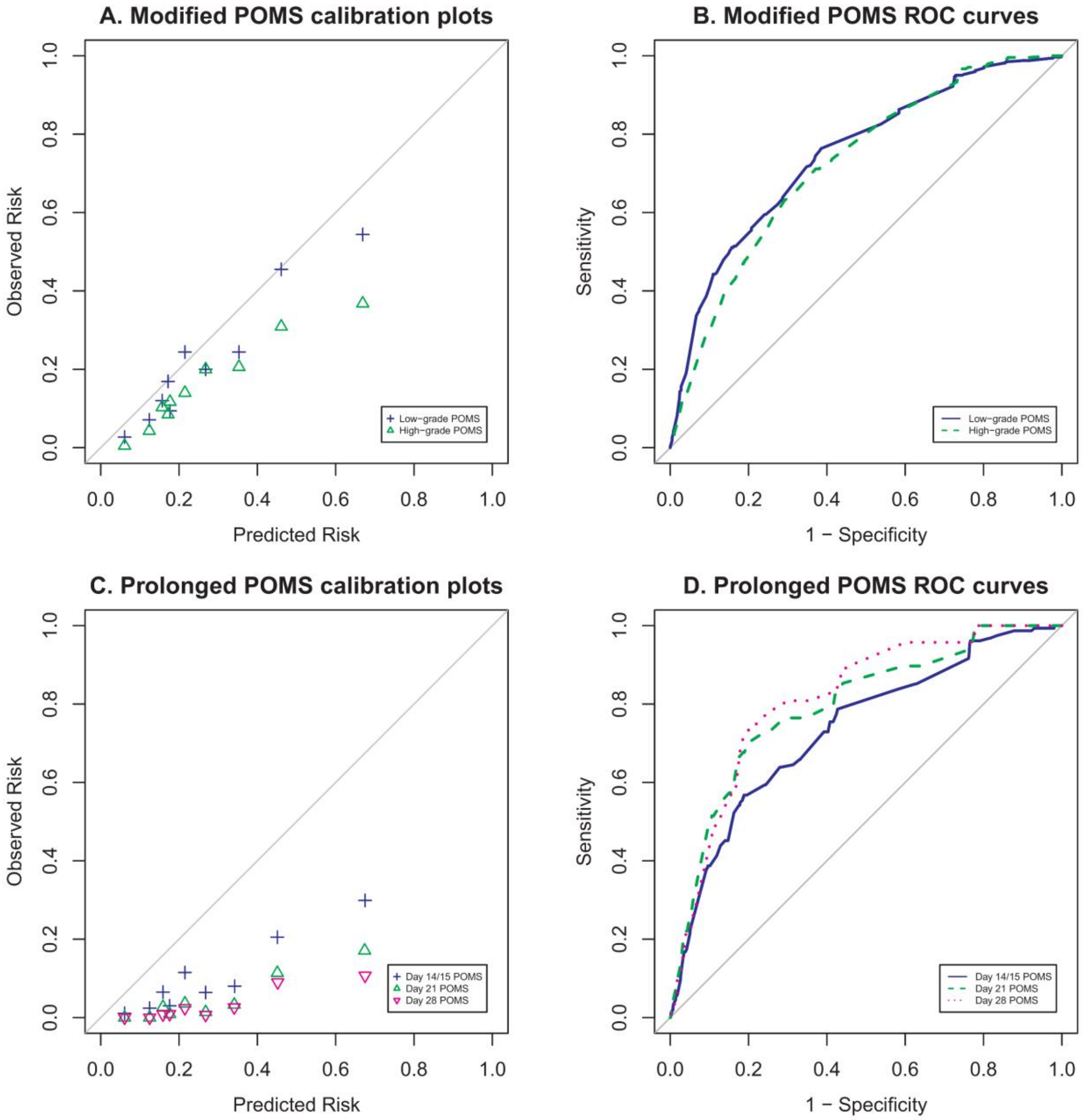


Figure 3

Figure 3 Caption: Sensitivity Analyses. (A) Calibration plot of SORT-morbidity for predicting modified POMS, comparing low-grade and high-grade POMS morbidity outcomes; (B) ROC curve plot showing SORTmorbidity discrimination performance for predicting low-grade and high-grade POMS morbidity outcomes (AUROC = 0.75, and 0.72, respectively);(C) Calibration plot of SORT-morbidity for predicting Day 14/15, Day 21, and Day 28 POMS outcomes; (D) ROC curve plot showing SORT-morbidity discrimination performance for predicting Day 14/15, Day 21, and Day 28 POMS outcomes (AUROC = 0.73, 0.79 and 0.81, respectively). ROC: Receiver Operator Characteristic; AUROC: Area Under the ROC curve.



Supplementary Data

Supplementary Table 1

Supplementary Table 1 Caption: Surgical Outcome Risk Tool variable definitions, adapted from Protopapa et al.[7]

SORT Variable	Definition	Further Detail
ASA-PS Grade	**Grade I**: A normal healthy patient. **Grade II**: A patient with mild systemic disease. **Grade III**: A patient with severe systemic disease. **Grade IV**: A patient with severe systemic disease that is a constant threat to life. **Grade V**: A moribund patient who is not expected to survive without the operation.	From the American Society of Anesthesiologists Physical Status Classification System
Surgical Urgency	**Immediate**: Immediate life, limb or organ-saving intervention – resuscitation simultaneous with intervention. Normally within minutes of decision to operate. **Urgent**: Intervention for acute onset or clinical deterioration of potentially life-threatening conditions, for those conditions that may threaten the survival of limb or organ, for fixation of many fractures and for relief of pain or other distressing symptoms. Normally within hours of decision to operate. **Expedited**: Patient requiring early treatment where the condition is not an immediate threat to life, limb or organ survival. Normally within days of decision to operate. **Elective**: Intervention planned or booked	From the National Confidential Enquiry into Patient Outcome and Death Classification of Intervention, http://www.ncepod.org.uk/classification.html

	in advance of routine admission to hospital. Timing to suit patient, hospital and staff.	
High Risk Surgical Specialty	Any of the following specialties: Gastrointestinal, Thoracic or Vascular Surgery	
Surgical Severity	Taken from a list of over 2000 procedure severity codes with 5 categories. **Minor**: e.g. Dilation of stricture of rectum, Change of cast under general anaesthetic. **Intermediate**: e.g. Laparoscopic repair of incisional or ventral hernia requiring mesh, Closed reduction of fracture of short bone (including cast or percutaneous K-wires). **Major**: e.g. Laparoscopic appendicectomy, Closed reduction of fracture of long bone with external fixation (excluding fixation by cast or percutaneous K-wires). **Xmajor**: e.g. Right hemicolectomy, Locked intramedullary nailing of fractured long bone. **Complex**: e.g. Total excision of colon and	Full list of procedures can be found on AXA PPP Healthcare website, https://online.axappphealthcare.co.uk/SpecialistForms/SpecialistCode.mvc/Print?source=contracted

	ileorectal anastomosis, Revision of uncemented or cemented total hip replacement without adjunctive procedures.	
Malignancy	Presence or absence of active malignancy within past 5 years	
Age	Three age categories: **<65 years**, **65–79 years**, **≥80**.	

Supplementary Table 2

Supplementary Table 2 Caption: Postoperative Morbidity Survey (POMS) Domains, adapted from Grocott et al.[12]

Morbidity Domain	**Criteria**	**Source of Data**
Pulmonary	New requirement for oxygen or respiratory support.	Patient observation or Treatment chart
Infectious	Currently on antibiotics or temperature >38°C in the last 24hr.	Treatment chart or Observation chart
Renal	Presence of oliguria <500 mL/24hr, increased serum creatinine (>30% from preoperative level) or urinary catheter in situ.	Fluid balance chart or Biochemistry result
Gastrointestinal	Unable to tolerate an enteral diet for any reason, including nausea, vomiting, and abdominal distension, or use of antiemetic.	Patient questioning, Fluid balance chart or Treatment chart
Cardiovascular	Diagnostic tests or therapy within the last 24 hr for any of the following: new myocardial infarction or ischaemia, hypotension (requiring pharmacological or fluid therapy >200 mL/hr), atrial or ventricular arrhythmias, cardiogenic pulmonary oedema, or thrombotic event requiring anticoagulation.	Treatment chart or Note review
Neurological	New focal neurological deficit, confusion, delirium, or coma.	Note review or Patient questioning
Haematological	Requirement for any of the following within the last 24 hr: packed erythrocytes, platelets, fresh-frozen plasma, or cryoprecipitate.	Treatment chart or Fluid balance chart
Wound	Wound dehiscence requiring surgical exploration or drainage of pus from the operation wound with or without isolation of organisms.	Note review or Pathology result
Pain	New postoperative pain significant enough to require parenteral opioids or regional analgesia.	Treatment chart or Patient questioning

Supplementary Table 3

Supplementary Table 3 Caption: Summary of the Multivariable Logistic Regression models derived in the study. Final model chosen is in ***bold****.*

In constructing our 6 candidate models, we started with a model closest to the original SORT-mortality model (Model 1), deriving empirically-weighted coefficients for this. We then made serial adjustments to the representations of the different predictor variables, one variable at a time, recording the performance measures for each step-wise change to the model. We selected our final model based on a balance of the following performance measures: goodness-of-fit measures (AIC and BIC), calibration in the validation sample (as judged by Hosmer-Lemeshow and calibration plots) and discrimination in the validation sample (as assessed by AUROC).

The permutations chosen for the different representations is exhaustive within the constraints of scientific plausibility, as well as within the arbitrary constraints imposed upon model-development with the aim of working with the data variables found in the SORT-mortality model. ASA 1 and 2 could plausibly be combined into one category because the mortality rates associated with the two grades does not differ greatly (0.0% vs 0.5% in a UK population, data from the paper by Protopapa et al), whereas ASA 3 and >4 conferred a 3.2% mortality and >16.5% mortality respectively.[7] We chose to combine ASA 4 and 5 because our study cohort only had 1 patient with ASA 5, which would be in keeping with an elective surgical cohort. With regards to different representations of age: The SORT-mortality calculator (http://www.sortsurgery.com/index.php?) asks for patients ages in the specific range categories which we tested in our models, and we did not want to significantly depart from this range in order to maintain compatibility with the data items collected for SORT-mortality predictions. We tried one first order interaction and found that it only resulted in a minor improvement in AIC at the cost of a significant worsening in BIC, with no improvement in discrimination, and therefore stopped at that point in our model construction, satisfied that further interaction terms would similarly penalise the model. This interaction was tested as it was posited that patients may undergo different types of operation at different age groups, for example bariatric surgery would be rarer at higher age, while colorectal and orthopaedic surgery rates would be increased at the middle age range.

Model Number	**Model Variables**	**Akaike Information Criterion (AIC)**	**Bayesian Information Criterion (BIC)**	**Hosmer-Lemeshow χ^2 statistic (p-value)**	**AUROC (95% CI)**
1	ASA = 3; ASA ≥ 4; "High-risk" Surgical Specialty; Xmajor/Complex Surgical Severity; Malignancy; Age 65–79 yrs; Age ≥ 80 yrs	1142.32	1182.01	7.09 (0.53)	0.63 (0.57–0.68)
2	ASA = 3; ASA ≥ 4; Colorectal or Upper GI Surgery; Xmajor/Complex Surgical Severity;	1115.37	1155.07	15.37 (0.05)	0.69 (0.64–0.74)

	Malignancy; Age 65–79 yrs; Age ≥ 80 yrs				
3	ASA = 3; ASA ≥ 4; Surgical Specialty (as a categorical variable with 6 categories); Xmajor/Complex Surgical Severity; Malignancy; Age 65–79 yrs; Age ≥ 80 yrs	1068.49	1128.04	10.1 (0.26)	0.72 (0.67–0.77)
4	**ASA = 2; ASA = 3; ASA ≥ 4; Surgical Specialty (as a categorical variable with 6 categories); Xmajor/Complex Surgical Severity; Malignancy; Age 65–79 yrs; Age ≥ 80 yrs**	**1068.31**	**1132.82**	**4.87 (0.77)**	**0.72 (0.67–0.77)**
5	ASA (as a discrete numerical variable); Surgical Specialty (as a categorical variable with 6 categories); Xmajor/Complex Surgical Severity; Malignancy; Age 65–79 yrs; Age ≥ 80 yrs	1066.82	1121.4	4.37 (0.82)	0.71 (0.66–0.76)
6	ASA (as a discrete numerical variable); Surgical Specialty (as a categorical variable with 6 categories); Xmajor/Complex Surgical Severity; Malignancy; Age 65–79 yrs; Age ≥ 80 yrs; Age x Surgical Specialty Interaction	1065.11	1164.36	3.42 (0.91)	0.72 (0.67–0.77)

Supplementary Table 4

Supplementary Table 4 Caption: Details of how the POMS organ-system sub-domains were mapped against Clavien-Dindo grades.

POMS Organ-system	POMS Sub-domain	Assigned Clavien-Dindo Grade
Pulmonary	New requirement for oxygen	2
Pulmonary	New requirement for respiratory support	2
Infectious	Currently on antibiotics	2
Infectious	Temperature >38°C in the last 24hr	1
Renal	Urinary catheter in situ	1
Renal	Increased serum creatinine (>30% from preoperative level)	2
Renal	Presence of oliguria <500 mL/24hr	2
Gastrointestinal	Unable to tolerate an enteral diet for any reason	1
Gastrointestinal	Vomiting or abdominal distension, or use of anti-emetics	1
Cardiovascular	Thrombotic event requiring anticoagulation (new)	2
Cardiovascular	Atrial or ventricular arrhythmias (new)	2
Cardiovascular	Hypotension (requiring pharmacological or fluid therapy >200 mL/hr)	2
Cardiovascular	New myocardial infarction or ischaemia	2
Cardiovascular	Cardiogenic pulmonary oedema	2
Neurological	New coma	3
Neurological	New confusion or delirium	2
Neurological	New focal neurological deficit	2
Haematological	Platelet, fresh-frozen plasma, or cryoprecipitate transfusion in last 24hrs	2
Haematological	Packed erythrocyte transfusion in the last 24hrs	2
Wound	Wound dehiscence requiring surgical exploration or drainage of pus from the operation wound with or without isolation of organisms	2
Pain	New pain significant enough to require parenteral opioids	1
Pain	New pain significant enough to require regional analgesia	2

Supplementary Table 5

Supplementary Table 5 Caption: Hosmer-Lemeshow Test Results for SORT-morbidity. Observed versus predicted POMS-defined morbidity are compared in the validation cohort per quantile of predicted risk.

Risk quantile	Total no. of patients in the quantile	Predicted risk of morbidity in the quantile	Observed risk of morbidity in the quantile	Number of patients with predicted morbidity in the quantile	Number of patients with observed morbidity in the quantile
[0.0213,0.0772)	59	0.062	0.034	3.66	2
[0.0772,0.1380)	51	0.125	0.098	6.35	5
[0.1380,0.1719)	76	0.157	0.184	11.96	14
[0.1719,0.1858)	100	0.176	0.23	17.55	23
[0.1858,0.2422)	63	0.216	0.222	13.59	14
[0.2422,0.2993)	26	0.247	0.269	6.43	7
[0.2993,0.3648)	47	0.317	0.34	14.89	16
[0.3648,0.5198)	69	0.439	0.464	30.3	32
[0.5198,0.8608]	36	0.659	0.583	23.71	21

Supplementary Table 6

Supplementary Table 6 Caption: Hosmer-Lemeshow Test Results for POSSUM. Observed versus predicted POMS-defined morbidity are compared in the validation cohort per quantile of predicted risk.

Risk quantile	Total no. of patients in the quantile	Predicted risk of morbidity in the quantile	Observed risk of morbidity in the quantile	Number of patients with predicted morbidity in the quantile	Number of patients with observed morbidity in the quantile
[0.0547,0.123)	54	0.087	0.185	4.72	10
[0.1235,0.209)	84	0.178	0.167	14.98	14
[0.2092,0.237)	47	0.23	0.085	10.79	4
[0.2369,0.267)	68	0.259	0.162	17.64	11
[0.2670,0.299)	53	0.289	0.283	15.31	15
[0.2994,0.334)	47	0.324	0.191	15.22	9
[0.3340,0.371)	28	0.356	0.429	9.96	12
[0.3705,0.448)	53	0.41	0.283	21.74	15
[0.4477,0.606)	46	0.525	0.413	24.17	19
[0.6059,0.965]	47	0.724	0.532	34.02	25