BMJ Open

Protocol for the derivation and external validation of a 30-day mortality risk prediction model for older patients having emergency general surgery (PAUSE score—Probability of mortality Associated with Urgent/emergent general Surgery in oldEr patients score)

Simon Feng [1], Carl Van Walraven,[2] Manoj Lalu [1,2] Husein Moloo,[2,3] Reilly Musselman,[3] Daniel I McIsaac[1,2]

For numbered affiliations see end of article.

Correspondence to
Dr Simon Feng; sfeng@toh.ca

ABSTRACT

Introduction People 65 years and older represent the fastest growing segment of the surgical population. Older age is associated with doubling of risk when undergoing emergency general surgery (EGS) procedures and often coexists with medical complexity and considerations of end-of-life care, creating prognostic and decisional uncertainty. Combined with the time-sensitive nature of EGS, it is challenging to gauge perioperative risk and ensure that clinical decisions are aligned with the patient values. Current preoperative risk prediction models for older EGS patients have major limitations regarding derivation and validation, and do not address the specific risk profile of older patients. Accurate and externally validated models specific to older patients are needed to inform care and decision making.

Methods and analysis We will derive, internally and externally validate a multivariable model to predict 30-day mortality in EGS patients >65 years old. Our derivation sample will be individuals enrolled in the National Surgical Quality Improvement Program (NSQIP) database between 2012 and 2016 having 1 of 7 core EGS procedures. Postulated predictor variables have been identified based on previous research, clinical and epidemiological knowledge. Our model will be derived using logistic regression penalised with elastic net regularisation and ensembled using bootstrap aggregation. The resulting model will be internally validated using k-fold cross-validation and bootstrap validation techniques and externally validated using population-based health administrative data. Discrimination and calibration will be reported at each step.

Ethics and dissemination Ethics for NSQIP data use was obtained from the Ottawa Hospital Research Ethics Board; external validation will use routinely collected anonymised data legally exempt from research ethics review. The final risk score will be published in a peer-reviewed journal. We plan to further disseminate the model as an online calculator or application for clinical use. Future research

Strengths and limitations of this study

► The predictor variables will be selected based on pre-existing research, clinical and epidemiological knowledge, and will include novel and strong predictors such as frailty.

► By combining elastic net regularisation and ensemble model building techniques, our analytical approach should minimise overfitting.

► The model will undergo internal validation (to quantify model optimism) as well as external validation (to assess whether it will maintain accuracy in different settings and future use).

► The derivation data will not include physiological variables and hospital indicators; future research will be needed to determine if such variables improve predictive accuracy.

► The model will be derived and validated using a large number of unselected, representative patients from a broad assortment of regions and hospitals in North America.

will be required to test the clinical application of the final model.

INTRODUCTION

Approximately 15% of people in Western countries are age 65 or older, yet this age group represents over 30% of the 550 000 people who require emergency general surgery (EGS) each year in North America.[1–3] Thirty-day morbidity (40%) and mortality (10%) rates in this population are twofold to fivefold higher compared with people younger than 65.[4–7]

Older people considering EGS often need to quickly make complex clinical decisions in the face of potentially poor expected outcomes. Accurate risk stratification for older EGS patients is a key component of informed consent and patient decision making. Therefore, supporting patients and families to make health decisions that are congruent with their values and goals of care is a key consideration.[8 9] Prognostic uncertainty has been highlighted as a key barrier to communication and providing goal-concordant care in older EGS patients,[10 11] and the best practice guidelines emerging from the National Emergency Laparotomy Audit (NELA) recommend preoperative documentation of risk for all EGS patients, as well as specific processes of care based on risk assessment.[12]

Unfortunately, current risk prediction models are inadequate for risk prediction in older EGS patients. Several multivariable models have been derived and tested to predict risk across all EGS patients, and at least two systematic reviews of such tools have been published.[13 14] However, both content and methodological issues substantially limit the applicability of existing risk models when applied to older EGS patients. Existing prediction models that are not specific to EGS are driven primarily by procedure type, procedure urgency and age.[13 14] When applied to an older EGS population, these models are less likely to discriminate well.[15 16] In addition, models specifically derived for EGS patients suffer from at least three significant limitations. First, none have been derived or validated specifically in older EGS patients. This has resulted in age acting as a significant driver of risk, which could lead to decreased discrimination when prediction is limited to older people. Second, geriatric-specific risk factors, and in particular frailty, have not been considered as predictors. This is a significant limitation because frailty is a strong and consistent predictor of outcomes in EGS patients that has been shown to out-perform the National Surgical Quality Improvement Program (NSQIP) Universal Risk Calculator among older EGS patients.[17] Finally, existing EGS-specific models have been developed with significant methodological limitations. One recommended model, the Emergency Surgery Acuity Score (ESAS),[18] was derived using univariate screening and stepwise variable selection, two techniques that can introduce significant bias and data overfitting.[19] The ESAS tool was also derived without clear consideration of calibration, a key aspect of predictive accuracy when a model is intended to support clinical decision making.[20] Another recently developed tool, the NELA risk model, was derived and internally validated using recommended methodology, but combined preintraoperative and intraoperative variables, making it unsuitable for preoperative decision making.[12] Furthermore, it includes NELA-specific covariates (such as NELA audit year), which would not be available for non-NELA patients. Finally, neither of the ESAS or NELA tools have been externally validated.

Therefore, the creation of an accurate risk prediction tool that can be applied in a timely manner, which contains risk factors specific to older patients, is urgently required to address the prognostic uncertainty that challenges decision making for older EGS patients. The proposed project will address current limitations in risk models for older EGS patients using the best practice methodologies. It will use frameworks that have been proposed, which focus on discrimination (how well can a model differentiate a high-risk from a low-risk individual), calibration (do predicted outcome probabilities closely align with observed outcome rates) and overall accuracy (the difference between predicted and observed risk).[21] This prediction model will also be derived and undergo validation testing to evaluate whether future predictions in different settings will maintain the same accuracy.[19] The model will be derived using methods to ensure inclusion of all variables that add predictive value while avoiding overfitting. These include: (1) variable selection based on previous research and pre-existing knowledge,[19] (2) penalised regression techniques that will shrink regression coefficients to avoid poor generalisation,[22 23] (3) ensemble modelling to simulate variation across data patient populations,[24] (4) robust internal validation that considers trade-offs between bias and variance[19 24] and (5) external validation to support generalisability.[19] We believe that these steps will support our objective of producing a robust, generalisable and clinically applicable model to predict mortality risk in older people considering EGS.

METHODS AND ANALYSIS
Design and data sources
This will be a retrospective observational study using data from a multicentred prospective surgical registry to derive and validate a mortality risk prediction model. Derivation and internal validation data will come from the NSQIP Participant Use File (PUF) between 2012 and 2016. The PUF is a prospectively collected dataset from all participating NSQIP hospitals in North America. Data are collected in each hospital by trained nurse abstractors, using standardised methods and definitions. The NSQIP PUF contains rich preoperative variables pertaining to patient comorbidities, surgical urgency and outcomes (including vital status), collected up to 30 days after surgery. External validation data will be from ICES (formerly known as the Institue of Clinical Evaluative Sciences; from the years 2017–2019 to avoid overlap with the NSQIP population). ICES is an independent health research institute that houses the health administrative data for the province of Ontario, Canada, which has a universal health insurance programme that provides all residents with coverage for physician and hospital care. ICES data are collected from all hospitals (>80 acute care centres) using standardised methods and formats, and are de-identified, cleaned and labelled using standardised central practices. Linkage across data sets is accomplished using deterministic techniques based on an anonymised unique identifier. Results will be reported according to

Table 1 Inclusion and exclusion criteria for NSQIP PUF data set

Inclusion	Exclusion
Enrolment in NSQIP PUF	No receipt of a surgical procedure
Age 65 or greater	
Appendectomy, cholecystectomy, laparotomy, lysis of adhesions, peptic ulcer repair, bowel resection performed on a non-elective basis	

NSQIP PUF, National Surgical Quality Improvement Program Participant Use File.

the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis statement.[25]

Patient and public involvement

Patients and the public will not be involved in the design, conduct, reporting or dissemination of this research.

Study population

The study will include patients 65 years or older undergoing EGS procedures, based on a core set of surgeries as specified by Scott *et al*.[1] This core set of EGS procedures comprises seven specific surgeries (appendectomy, cholecystectomy, laparotomy, lysis of adhesions, large bowel resection, small bowel resection and peptic ulcer repairs) that account for approximately 80% of all procedures, mortality, complications and costs related to EGS in the USA (table 1).[1] These procedures will be identified using Current Procedural Terminology (CPT) codes in the NSQIP. These CPT codes have already been translated into Canadian Classification of Intervention codes (a Canada-specific modification to International Classification of Diseases, 10th Edition procedure codes that are used in Canadian administrative data),[3] which will support external validation at ICES.[3]

Sample size

Our sample size was based on inclusion of all older people having target EGS procedures between 2012 and 2016 in the NSQIP. We anticipate that this sample will be more than adequately powered to derive our prediction model. Based on previous research, we anticipate identification of approximately 15 000 cases per year, which should provide approximately 60 000 cases overall.[18] With an anticipated 30-day mortality rate of 10%, this would provide approximately 6000 outcomes.[18] Finally, with 52 variables prespecified for inclusion, we should have greater than 100 events per variable. In the external validation data source, we anticipate identifying 6000 cases per year, and over 3 years should therefore identify approximately 18 000 cases with approximately 1800 outcomes.[3]

Outcomes

The outcome of interest is all cause 30-day mortality following the index surgical procedure. This will be identified from the NSQIP using the 'Days from Operation to Death' variable and from vital statistics data at ICES, which captures the death date of all residents of Ontario.

Predictor variables

After reviewing existing generic and EGS-specific risk models, as well as considering data fields available in the NSQIP PUF and at ICES, we prespecified a list of preoperative predictors (online supplementary appendix A). These include demographic variables (age, sex), comorbidities (disseminated cancer, diabetes, hypertension, heart failure, chronic obstructive pulmonary disease, dyspnoea, smoking, dialysis, ascites, steroid use for a chronic condition, history of weight loss), functional health status, preoperative place of residence (nursing home vs community), frailty (based on the Risk Analysis Index for the NSQIP data and the preoperative Frailty Index at ICES, both previously validated frailty instruments[26–28]), American Society of Anesthesiologists classification, indicators of acute illness (acute renal failure, sepsis, preoperative ventilation[29]), minimally invasive vs open surgery and the specific surgical procedure. Dichotomisation of continuous variables will be avoided to minimise loss of information.[21] All continuous variables will be centred at the mean and standardised using the SD, followed by evaluation of fractional polynomial transformations (within an unpenalised logistic regression framework) to identify the optimal continuous format.[30] We will also include prespecified interaction terms between frailty and procedure,[3 12] procedure and minimally invasive versus open approach, and sex and procedure.[31]

Descriptive statistics

Descriptive statistics for the NSQIP and ICES datasets will be computed based on outcomes status (ie, for participants who die within 30 days of surgery compared with those who do not). Continuous variables will be described using means and SDs (prior to centring and standardisation), binary or categorical variables will be described using proportions. Estimation of differences between outcome groups will be performed using absolute standardised differences.[32] Descriptive comparisons will not be used to guide variable selection.

Derivation of the prediction model

We will derive an ensemble multivariable prediction model via bootstrap aggregation (also known as 'bagging'[24]) using logistic regression that is penalised with elastic net regularisation.[24] Simulation studies demonstrate that penalised (or shrunken) regression coefficients reduce data overfitting and resultant model optimism.[22] Elastic net regularisation is a variable selection technique that combines ridge regression and least absolute shrinkage and selection operator (lasso) regression.[33 34] Elastic net regularisation is appropriate for our modelling scenario

because we have numerous predictor variables, many of which that are likely to be collinear. In this scenario, ridge regression in isolation tends to shrink collinear coefficients to a common value, whereas lasso in isolation can indiscriminately eliminate certain strong predictors that demonstrate collinearity.[24] Therefore, elastic net regularisation should yield a model that maintains important, but correlated, predictors while supporting parsimony through the elimination of weakly predictive variables. Bootstrap aggregation is a machine learning technique in the ensemble model family that averages regression coefficients across all bootstrap samples to account for variability in the underlying data.[34] This should further reduce data overfitting in our prediction model, translating into better performance in internal and external validation.

To derive the model, the NSQIP PUF data will first be used to tune the parameters of the elastic net model (ie, λ (the amount of shrinkage applied) and α (the mixture of lasso and ridge techniques applied, where a value of 0 would be entirely ridge and a value of 1 entirely lasso)) via 10-fold cross-validation to identify values that minimise the explained deviance. We will then generate 5000 bootstrap samples, using 1:1 sampling with replacement. Logistic regression models, penalised with elastic net regularisation, will then be run in each training bootstrap sample and will consider each of our prespecified predictor variables. The ensemble model will be constructed using the median value from the bootstrap distribution of the regression coefficient for each predictor variable that was selected in at least 95% of bootstrap samples. The percentile method will be used to estimate 95% CIs around each regression coefficient.

Once derived, predictive performance in the derivation set will then be estimated in keeping with recommended best practices.[20] Model discrimination (ie, the extent to which people who die will be assigned a higher predicted risk of death than people who do not die) will be measured using the area under the curve (AUC) statistic, where values closer to 1 (maximum value) denote higher discrimination and values close to 0.5 (minimum value) suggest negligible discrimination. While no irrefutable cutoffs exist, models with an AUC <0.7 may not be suitable to support decision making, while values >0.8 provide strong discrimination.[35] Calibration (ie, the extent that predicted outcome probabilities match observed outcomes rates) will be assessed using visual inspection of Loess-smoothed plots of observed outcome rates versus expected probabilities.[36] Austin and Steyerberg have shown that this technique is suitable in large samples, whereas the Hosmer-Lemeshow goodness-of-fit test p values are biased downward due to the test's reliance on the sample size-dependent χ^2 distribution.[36 37] This technique also allows assessment of calibration across the range of predicted probabilities. Overall accuracy of the model will be evaluated with the max-rescaled Brier score, which measures the squared differences between observed and predicted outcomes

(for the max-rescaled score, values approaching 1 signify a perfect model and smaller values denote worse performance).[20 38]

Internal validation of the prediction model

To estimate possible optimism and overfitting, we will perform two internal validation steps. First, we will use 10-fold cross-validation by splitting the data into 10 equal parts. The model will be retrained on 9/10th of the data and accuracy statistics will be subsequently validated on the remaining 1/10th. This will be repeated 10 times using each subset of the data. This technique provides an average estimate of overfitting, and k=10 has been shown to balance the concerns of variance and bias in internal validation.[24 39] We will also perform a bootstrap internal validation across 5000 samples drawn 1:1 with replacement, which tends to provide an estimate of optimism with lower variance compared with cross-validation.[19] The range of AUCs will be reported for k-fold cross validation and a 95% CI for the bootstrap internal validation.

External validation of the prediction model

The model will be externally validated in a separate set of electronic health data at ICES collected from 2017–2019. The ICES external validation set will be created using linked clinical and administrative datasets for Ontario patients using encrypted patient-specific identifiers. The specific datasets used and the mapping of NSQIP to ICES data elements is provided in online supplementary appendix A. To conduct the external validation, the regression coefficients from the internally derived model will be imported to ICES and will be used to score the external validation set. This will allow prediction of expected outcome probabilities, which in combination with observed outcomes, will allow for measurement of discrimination, calibration, and max-rescaled Brier score in the external data set.

Missing data

Outcome data will be complete for all participants. Missing predictor data will be handled using complete case analysis where <1% of the values for a given predictor are missing (a threshold below some rule of thumb recommendations of 5%).[40] Predictors missing >1% data will not be considered for the main analysis. While multiple imputation is the preferred method to incorporate cases with missing data in a prediction model, to our knowledge there are no available techniques to apply and pool elastic net penalisation to multiply imputed datasets. However, to estimate the possible impact and causes of missing data, we will assess and report measured characteristics between people with missing and non-missing data. Furthermore, we will perform unpenalised logistic regression with complete-case analysis and using multiple imputed missing values to explore whether missing values may substantially impact the predictive performance of our model.

Exploratory analyses

Because laboratory data are not available for all records at ICES, our primary approach will be to construct our model without laboratory data. However, we will perform an exploratory analysis in the derivation data with internal validation where lab values (online supplementary appendix B) are considered in model building. We will then assess the relative change in the AUC, calibration and max-rescaled Brier score with inclusion of laboratory data compared with the primary model without laboratory data.

Software

The regression analyses, model derivation, accuracy analysis, internal validation and subgroup analyses will be completed using R programming language and statistical software (V.3.5.2). External validation will be performed using SAS V.9.4.

Ethics and dissemination

As this study will use routinely collected patient data, individual patient is waived. Due to the anonymised and de-identified nature of the data required from ICES, the external validation phase is legally exempt from research ethics board review. The results and discussion of this work will be disseminated through peer-reviewed conferences and journals. Future plans for dissemination and knowledge translation include development of an online or app-based electronic calculator that will allow direct entry of clinical data to derive personalised predictions for patient care. Future research will be required to assess the clinical impact of this risk model.

LIMITATIONS

Our derivation and internal validation data will be from the NSQIP database. Therefore, two forms of selection bias could be present. First, only hospitals participating in NSQIP will contribute cases, and these hospitals may have outcomes that differ from non-NSQIP hospitals. Second, all NSQIP participants will have undergone surgery, limiting the results of our findings to older individuals who will be having surgery (as opposed to those undergoing conservative management). External validation at ICES will address the former, but not the latter, limitation. Additionally, both datasets are from North America, which may limit generalisability to other jurisdictions. Laboratory values will not be included in the primary model as they are not available for all cases at ICES. This will be addressed with an exploratory analysis to assess the possible impact of including laboratory values, but only internal validation of this exploratory model will be available. Physiological variables are not available in either the NSQIP or ICES data but would be expected to be strong predictive variables. While all chosen variables are available in NSQIP and ICES data at a conceptual level, specific definitions may vary. While we will make all possible efforts to directly align definitions (including

diagnoses and timing), some specific definitions (such as frailty) will differ between datasets. In the NSQIP data, frailty will be evaluated using the Risk Analysis Index, while at ICES we will use the validated preoperative Frailty Index. While this could lead to a decrease in external validation accuracy, this will address a real-world challenge with model transportability, as well as issues of instrument heterogeneity in frailty assessment. Fortunately, head-to-head comparisons of different frailty instruments in perioperative settings have demonstrated similar effect sizes between instruments across different outcomes.[41 42]

All prediction models are subject to data overfitting and optimism. While we will use variable prespecification and elastic net regularisation and ensemble modelling to reduce these risks, they cannot be entirely eliminated. We have specified mortality as our primary outcome, which is an outcome of primary importance for older people[43] and is accurately measured in electronic data. Other outcomes, such as time spent in hospital, functional recovery and independence, are also high-priority outcomes that will require attention in future predictive modelling studies. Furthermore, the NSQIP provides only 30-day mortality data; longer-term mortality (eg, 90 days) may better reflect meaningful survival after surgery.[44]

CONCLUSIONS

Older patients are disproportionately represented in EGS populations and experience high rates of postoperative morbidity and mortality. These patients are often at the end of life, with many qualifying for palliative care. This results in routine occurrence of complex clinical decision making regarding whether to proceed with emergency surgery. Accurate risk stratification is a key component of informed decision making in these time-sensitive scenarios. However, there is currently a dearth of high-quality risk assessment tools specific to older people considering EGS.

The planned study aims to fulfil the need for an accurate and generalisable risk prediction tool for older people considering EGS. This project will use established frameworks to create and evaluate a tool that we hope will predict mortality risk. Through prespecified internal and external validation, these efforts should result in the creation of a robust, generalisable and clinically applicable risk prediction model that can be used to inform clinical and patient decision making for older people considering EGS.

Author affiliations

[1]Anesthesiology and Pain Medicine, The Ottawa Hospital, Ottawa, Ontario, Canada
[2]Clinical Epidemiology Program, The Ottawa Hospital Research Institute, Ottawa, Ontario, Canada
[3]Surgery, The Ottawa Hospital, Ottawa, Ontario, Canada

Twitter Manoj Lalu @manojlalu

ORCID iDs
Simon Feng http://orcid.org/0000-0003-4674-2424
Manoj Lalu http://orcid.org/0000-0002-0322-382X

REFERENCES

1 Scott JW, Olufajo OA, Brat GA, et al. Use of national burden to define operative emergency general surgery. *JAMA Surg* 2016;151:e160480.

2 Wohlgemut JM, Ramsay G, Jansen JO. The changing face of emergency general surgery: a 20-year analysis of secular trends in demographics, diagnoses, operations, and outcomes. *Ann Surg* 2018;1.

3 McIsaac D, Moloo H, Bryson G, et al. The association of frailty with outcomes and resource use after emergency general surgery: a population-based cohort study. *Anesth Analg* 2017;124:1653–61.

4 Kuy S, Sosa JA, Roman SA, et al. Age matters: a study of clinical and economic outcomes following cholecystectomy in elderly Americans. *Am J Surg* 2011;201:789–96.

5 Ingraham AM, Cohen ME, Raval MV, et al. Variation in quality of care after emergency general surgery procedures in the elderly. *J Am Coll Surg* 2011;212:1039–48.

6 McIsaac DI, Abdulla K, Yang H, et al. Association of delay of urgent or emergency surgery with mortality and use of health care resources: a propensity score–matched observational cohort study. *Can Med Assoc J* 2017;189:E905–12.

7 Turrentine FE, Wang H, Simpson VB, et al. Morbidity, and mortality in elderly patients. *J Am Coll Surg* 2006;203:865–77.

8 Hargraves I, LeBlanc A, Shah ND, et al. Shared decision making: the need for Patient-Clinician conversation, not just information. *Health Aff* 2016;35:627–9.

9 Cooper Z, Lilley EJ, Bollens-Lund E, et al. High burden of palliative care needs of older adults during emergency major abdominal surgery. *J Am Geriatr Soc* 2018;66:2072–8.

10 Cooper Z, Koritsanszky LA, Cauley CE, et al. Recommendations for best communication practices to facilitate Goal-concordant care for seriously ill older patients with emergency surgical conditions. *Ann Surg* 2016;263:1–6.

11 Cooper Z, Courtwright A, Karlage A, et al. Pitfalls in communication that lead to Nonbeneficial emergency surgery in elderly patients with serious illness: description of the problem and elements of a solution. *Ann Surg* 2014;260:949–57.

12 Eugene N, Oliver CM, Bassett MG, et al. Development and internal validation of a novel risk adjustment model for adult patients undergoing emergency laparotomy surgery: the National emergency laparotomy audit risk model. *Br J Anaesth* 2018;121:739–48.

13 Oliver CM, Walker E, Giannaris S, et al. Risk assessment tools validated for patients undergoing emergency laparotomy: a systematic review. *Br J Anaesth* 2015;115:849–60.

14 Havens JM, Columbus AB, Seshadri AJ, et al. Risk stratification tools in emergency general surgery. *Trauma Surg Acute Care Open* 2018;3:e000160.

15 Nieboer D, van der Ploeg T, Steyerberg EW. Assessing discriminative performance at external validation of clinical prediction models. *PLoS One* 2016;11:e0148820.

16 Austin PC, Steyerberg EW. Interpreting the concordance statistic of a logistic regression model: relation to the variance and odds ratio of a continuous explanatory variable. *BMC Med Res Methodol* 2012;12:82.

17 Sadoum M, Zangbar B, Rhee PM, et al. NSQIP surgical risk calculator and frailty in emergency general surgery: a measure of surgical resilience patterns of postoperative care after surgery in freestanding ambulatory surgery centers in South Carolina 2015.

18 Sangji NF, Bohnen JD, Ramly EP, et al. Derivation and validation of a novel emergency surgery acuity score (ESAs). *J Trauma Acute Care Surg* 2016;81:213–20.

19 Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996;15:361–87.

20 Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models : A framework for some traditional and novel measures. *Epidemiology* 2010;21:128–38.

21 Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 2010;21:128–38.

22 Harrell FE, Habbema JDF, Steyerberg EW, et al. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. *Stat Med* 2000;19:1059–79.

23 Steyerberg EW, Eijkemans MJC, Habbema JDF. Application of shrinkage techniques in logistic regression analysis: a case study. *Stat Neerl* 2001;55:76–88.

24 Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning*. 2nd edn. Springer, 2009.

25 Collins GS, Reitsma JB, Altman DG, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med* 2015;162:55.

26 Shah R, Attwood K, Arya S, et al. Association of frailty with failure to rescue after low-risk and high-risk inpatient surgery. *JAMA Surg* 2018;153:e180214.

27 Hall DE, Arya S, Schmid KK, et al. Development and initial validation of the risk analysis index for measuring frailty in surgical populations. *JAMA Surg* 2017;152:175.

28 McIsaac DI, Harris EP, Hladkowicz E, et al. Prospective comparison of preoperative predictive performance between 3 leading frailty instruments. *Anesth Analg* 2019;1.

29 Bilimoria KY, Liu Y, Paruch JL, et al. Development and evaluation of the universal ACS NSQIP surgical risk calculator: a decision aid and informed consent tool for patients and surgeons. *J Am Coll Surg* 2013;217:833–42.

30 Sauerbrei W, Meier-Hirmer C, Benner A, et al. Multivariable regression model building by using fractional polynomials: description of SAS, STATA and R programs. *Comput Stat Data Anal* 2006;50:3464–85.

31 Grewal K, Wijeysundera DN, Carroll J, et al. Gender differences in mortality following non-cardiovascular surgery: an observational study. *Can J Anesth/J Can Anesth* 2012;59:255–62.

32 Austin PC. Using the standardized difference to compare the prevalence of a binary variable between two groups in observational research. *Commun Stat Simul Comput* 2009;38:1228–34.

33 Pavlou M, Ambler G, Seaman S, et al. Review and evaluation of penalised regression methods for risk prediction in low-dimensional data with few events. *Stat Med* 2016;35:1159–77.

34 Pavlou M, Ambler G, Seaman SR, et al. How to develop a more accurate risk prediction model when there are few events. *Br Med J* 2015;351.

35 Hosmer DW, Lemeshow S, Regression AL, et al. *Applied logistic regression*. New York, NY: John Wiley & Sons, 1989.

36 Austin PC, Steyerberg EW. Graphical assessment of internal and external calibration of logistic regression models by using Loess smoothers. *Stat Med* 2014;33:517–35.

37 Hosmer DW, Hosmer T, Le Cessie S, et al. A comparison of goodness-of-fit tests for the logistic regression model. *Stat Med* 1995;1997:965–80.

38 Rufibach K. Letters to the editor use of Brier score to assess binary predictions. *J Clin Epidemiol* 2010;63:938–42.

39 Jung Y, Hu J. A K-fold averaging cross-validation procedure. *J Nonparametr Stat* 2015;27:167–79.

40 Jakobsen JC, Gluud C, Wetterslev J, et al. When and how should multiple imputation be used for handling missing data in randomised clinical trials – a practical guide with flowcharts. *BMC Med Res Methodol* 2017;17:162.

41 McIsaac DI, Taljaard M, Bryson GL, et al. Frailty as a predictor of death or new disability after surgery. *Ann Surg* 2018;1.

42 Cooper Z, Rogers SO, Ngo L, *et al*. Comparison of frailty measures as predictors of outcomes after orthopedic surgery. *J Am Geriatr Soc* 2016;64:2464–71.

43 Akpan A, Roberts C, Bandeen-Roche K, *et al*. Standard set of health outcome measures for older persons. *BMC Geriatr* 2018;18:1–10.

44 Jammer I, Wickboldt N, Sander M, *et al*. Standards for definitions and use of outcome measures for clinical effectiveness research in perioperative medicine: European perioperative clinical outcome (EPCO) definitions: a statement from the ESA-ESICM joint Taskforce on perioperative outcome measures. *Eur J Anaesthesiol* 2015;32:88–105.