A prognostic model for use before elective surgery to estimate the risk of postoperative pulmonary complications (GSU-Pulmonary Score): a development and validation study in three international cohorts

NIHR Global Health Research Unit on Global Surgery*, STARSurg Collaborative*

Summary

Background Pulmonary complications are the most common cause of death after surgery. This study aimed to derive and externally validate a novel prognostic model that can be used before elective surgery to estimate the risk of postoperative pulmonary complications and to support resource allocation and prioritisation during pandemic recovery.

Methods Data from an international, prospective cohort study were used to develop a novel prognostic risk model for pulmonary complications after elective surgery in adult patients (aged ≥18 years) across all operation and disease types. The primary outcome measure was postoperative pulmonary complications at 30 days after surgery, which was a composite of pneumonia, acute respiratory distress syndrome, and unexpected mechanical ventilation. Model development with candidate predictor variables was done in the GlobalSurg-CovidSurg Week dataset (global; October, 2020). Two structured machine learning techniques were explored (XGBoost and the least absolute shrinkage and selection operator [LASSO]), and the model with the best performance (GSU-Pulmonary Score) underwent internal validation using bootstrap resampling. The discrimination and calibration of the score were externally validated in two further prospective cohorts: CovidSurg-Cancer (worldwide; February to August, 2020, during the COVID-19 pandemic) and RECON (UK and Australasia; January to October, 2019, before the COVID-19 pandemic). The model was deployed as an online web application. The GlobalSurg-CovidSurg Week and CovidSurg-Cancer studies were registered with ClinicalTrials.gov, NCT04509986 and NCT04384926.

Findings Prognostic models were developed from 13 candidate predictor variables in data from 86,231 patients (1158 hospitals in 114 countries). External validation included 30,492 patients from CovidSurg-Cancer (726 hospitals in 75 countries) and 6789 from RECON (150 hospitals in three countries). The overall rates of pulmonary complications were 2.0% in derivation data, and 3.9% (CovidSurg-Cancer) and 4.7% (RECON) in the validation datasets. Penalised regression using LASSO had similar discrimination to XGBoost (area under the receiver operating curve [AUROC] 0.786, 95% CI 0.774–0.798 vs 0.785, 0.772–0.797), was more explainable, and required fewer covariables. The final GSU-Pulmonary Score included ten predictor variables and showed good discrimination and calibration upon internal validation (AUROC 0.773, 95% CI 0.751–0.795; Brier score 0.020, calibration in the large [CITL] 0.034, slope 0.954). The model performance was acceptable on external validation in CovidSurg-Cancer (AUROC 0.746, 95% CI 0.733–0.760; Brier score 0.036, CITL 0.109, slope 1.056), but with some miscalibration in RECON data (AUROC 0.716, 95% CI 0.689–0.744; Brier score 0.045, CITL 1.040, slope 1.009).

Interpretation This novel prognostic risk score uses simple predictor variables available at the time of a decision for elective surgery that can accurately stratify patients’ risk of postoperative pulmonary complications, including during SARS-CoV-2 outbreaks. It could inform surgical consent, resource allocation, and hospital-level prioritisation as elective surgery is upscaled to address global backlogs.

Funding National Institute for Health Research.

Copyright © 2024 The Author(s). Published by Elsevier Ltd. This is an Open Access article under the CC BY 4.0 license.

Introduction Pulmonary complications are a common and lethal sequela of elective surgery, affecting 2–62% of patients depending on the operative approach and casemix. As elective surgery is upscaled to address a growing global backlog from the COVID-19 pandemic, pulmonary events pose a major threat to patient safety. Such events were implicated in 37% of postoperative deaths before the pandemic, and as many as 66% of deaths during the pandemic due to additional risks of perioperative SARS-CoV-2 infection. Pulmonary complications of surgery have a major effect on hospital resource use and demand.
Research in context

Evidence before this study
The global backlog of elective surgery is now a major public health priority, but patients remain at risk of postoperative pulmonary complications, which are the most common cause of death after surgery. As elective capacity is upcaled to address waiting lists, individual risk scoring is needed to inform patient consent, case prioritisation, and resource allocation during the COVID-19 recovery period. We previously completed a systematic review of articles reporting risk prediction models for postoperative pulmonary complications. We searched MEDLINE and Embase from database inception to March 1, 2020 using the search terms “surgery”, “pulmonary complication”, and “postoperative pulmonary complication”; of the 29 models identified, we validated six in an external cohort. The findings were published in 2022 in The Lancet Digital Health. We then updated this search to include studies up to Nov 8, 2023. We identified 33 unique prognostic models. None of the six models that underwent external validation showed good discrimination (defined as the area under the receiver operating curve [AUROC] lower confidence interval estimate of ≥0·70) in identifying postoperative pulmonary complications. All six were at high or unclear risk of bias on PROBAST assessment. Existing risk scores did not evaluate structured machine learning methods, and often included intraoperative variables that were either not available at the time of the decision to operate, or were inaccessible in resource-constrained health systems.

Added value of this study
The GSU-Pulmonary Score is a novel prognostic risk score that aims to estimate the risk of postoperative pulmonary complications in adults undergoing elective surgery, based on ten preoperative variables routinely accessible in preoperative settings. To our knowledge, this is the first score to consider background community SARS-CoV-2 infection rate, allowing dynamic risk prediction in areas with ongoing SARS-CoV-2 transmission. The model was derived from a prospective international dataset of 86231 patients from a diverse range of hospitals and countries, and discrimination and calibration were explored in two external prospective datasets of 37281 patients in total. The discrimination of the model was superior to that of the best-performing score (ARISCAT) in our 2022 systematic review (AUROC 0·716 vs 0·700), with the caveat that ARISCAT included abdominal, abdominothoracic, and pelvic operations only.

Implications of all the available evidence
The GSU-Pulmonary Score includes simple, preoperative variables that can be deployed by a patient’s bedside or in an outpatient clinic in any country and setting, and considers baseline community SARS-CoV-2 infection rates. It can be implemented in individual risk assessment for patient prioritisation and resource management during the ongoing COVID-19 recovery period. It will support front-line clinicians and policy makers to safely scale capacity for elective surgery to address growing waiting lists around the world. In case of a future emerging threat from a high-risk respiratory virus, we anticipate that the model could be recalibrated to address changing risks of pulmonary complications, acting as a sleeping sentinel that underpins future risk profiling of SARS-CoV-2.

Methods
Study design
This prognostic model derivation and validation study used data from three international, prospective, multicentre cohort studies to develop a tool to estimate...
the risk of postoperative pulmonary complications for adult patients undergoing planned (elective) surgery. The derivation dataset was from the GlobalSurg-CovidSurg Week study, which aimed to determine the optimal timing for surgery following SARS-CoV-2 infection. A full description of methods and findings of this study has been published.9 This study dataset was selected for model derivation as it was the largest and most data-rich for model fitting, with the broadest inclusion criteria (ie, most generalisable). The protocol of GlobalSurg-CovidSurg Week was prospectively registered at ClinicalTrials.gov (NCT04509986). The model developed in this cohort was then externally validated in two independent datasets. The first was from CovidSurg-Cancer, a study of adult patients with a surgically curable cancer who underwent surgery during the COVID-19 pandemic.10 The study was done in accordance with a preregistered protocol (NCT04384926). The second validation dataset was from RECON, a study of pulmonary complications after major abdominal surgery.11

All three studies collected routine, anonymised data with no change to clinical care pathways, and secured approvals for participating hospitals in line with local and national regulations before entering data into the study. Informed patient consent was obtained if this was necessary to comply with local or national regulations. Country-specific ethics or relevant clinical governance approvals are available upon request. Collaborators were required to confirm evidence of relevant approvals before entering patient data in all three included studies. This approvals process was delegated to the site principal investigators at each participating centre, and one or more national leads in each participating country (appendix pp 33–213).

The study authors collaborated across the leadership teams for all three cohort studies and had direct access to the original datasets. This study is reported with reference to the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) guidelines.12 Two independent statisticians assessed risk of bias in the study using the prediction model risk of bias assessment tool (PROBAST).13

Participants

GlobalSurg-CovidSurg Week9 included all consecutive patients undergoing surgery between Oct 5 and Nov 1, 2020. Any centre worldwide performing inpatient surgery was eligible for inclusion. Participating centres could include children and adults from one or more specialty groups, so long as consecutive sampling was done. This included both elective and emergency surgery, and patients undergoing all operation types for any condition.

CovidSurg-Cancer10 included consecutive adult patients with a surgically curable cancer who underwent surgery from the emergence of COVID-19 up to Aug 31, 2020. Any hospital worldwide that performed elective cancer surgery in an area affected by the COVID-19 pandemic was eligible. Data from the 15 most common operated, solid tumour types were collected with colleagues collaborating across ten surgical oncology disciplines. Patients who had clinical symptoms consistent with COVID-19 or who were confirmed to have SARS-CoV-2 infection (by quantitative RT-PCR or thoracic CT imaging, or both) at the time of surgery were excluded.

RECON11 included consecutive adult patients undergoing a broad range of elective major abdominal surgeries, including abdominal visceral resection, reversal of stoma, open vascular surgery, anterior abdominal wall hernia repair, or transplant surgery, using any operative approach. Planned day-case procedures and abdominal surgeries without visceral resection were excluded. Eligible patients were identified at each participating hospital during two data collection periods: Jan 1 to April 30, 2019 (in the UK and Ireland), and Sept 1 to Oct 31, 2019 (in Australia).

This derivation and validation study included only adult patients (aged ≥18 years) undergoing planned (elective) surgery. Elective surgery was defined as any operation for which the decision for surgery was made before the hospital admission during which the operation took place. No restrictions on operation type, approach, or indication for surgery were imposed. Patients with suspected or confirmed SARS-CoV-2 infection at the time of surgery (ie, a positive nasopharyngeal swab with RT-PCR test confirming the presence of SARS-CoV-2, or strongly indicative signs or symptoms when testing facilities were not available) were excluded. Patients undergoing local excision or endoluminal surgery were excluded.

Data governance and data quality

Data were collected online and stored on a secure server running the Research Electronic Data Capture (REDCap, Vanderbilt University, Nashville, TN, USA) web application,14 based in the University of Birmingham, UK. Data quality rules were built into the dataset, and study definitions were provided inline with data entry to improve reproducibility. Patients were pseudonymised at the point of entry using a unique REDCap identifier, with linked patient identifiers stored securely by the participating site. Data quality was checked continuously, with data clarification requests sent to sites at twice monthly intervals during the study period, and during data cleaning, in which inconsistencies or missing data were identified. Previous international outcomes studies from our groups have shown that this collaborative methodology has greater than 95% case ascertainment and greater than 98% data accuracy.15 In these three studies, if a specialty within a hospital was unable to confirm consecutive enrolment, their specialty’s data were excluded from analysis.
Outcomes
The primary outcome was postoperative pulmonary complications (PPC) within 30 days of surgery (with day 0 as the day of surgery). PPC was defined as a composite of pneumonia, acute respiratory distress syndrome, and unplanned use of invasive or non-invasive mechanical ventilation. This was adopted for two reasons: to mirror this outcome definition in several major randomised trials and cohort studies, and to include outcomes plausibly linked to SARS-CoV-2 infection with relevance to practice during pandemic recovery. The 30-day PPC rate was the predefined primary outcome in the GlobalSurg-CovidSurg Week and CovidSurg-Cancer cohort study protocols and the key clinical outcome in the RECON study. Outcome assessment was done on or as close as possible after the 30th postoperative day by a trained member of the local clinical team. As front-line clinical investigators were aware of patients’ relevant medical history, masking of outcome assessment to some predictor variables was not feasible. Diagnosis was, however, supported using radiological assessments, reported locally by a radiologist masked to several of the predictor variables. Full definitions are included in the appendix (pp 16–21). The secondary outcome measure was the 30-day postoperative mortality rate (all-cause).

Predictor variables
Candidate predictor variables (n=53) were identified in a systematic review of existing risk scores and then refined with multidisciplinary clinical input, gaining consensus from surgery, anaesthesia, and critical care experts. To improve the adoption of the risk stratification tool, we planned to include only predictor variables that were universally available without the need for additional tests, and that were relevant to low-resource countries and hospitals. We selected variables that are routinely available to clinicians at the point of a decision for elective surgery (eg, preoperatively in an outpatient clinic or inpatient ward) to inform preoperative decision making and consent; we did not include intraoperative variables as they cannot influence clinical pathways determined preoperatively, or physiological parameters as they are challenging to collect reliably in diverse, multicountry settings. Therefore, 13 candidate patient-level predictors were selected a priori to be included in model development (appendix pp 16–21): (1) patient factors: age in years (collected in primary data as a categorical variable: 18–29, 30–39, 40–49, 50–59, 60–69, 70–79, 80–89, and ≥90), sex (female or male), American Society of Anesthesiologists (ASA) grade (1, 2, 3, or ≥4), Revised Cardiac Risk Index score (0, 1, 2, or ≥3), pre-existing respiratory disease (yes or no), and smoking status (non-smoker, stopped smoking >6 weeks before surgery, or current smoker); (2) disease factors: compartment (thoracic, abdominopelvic, head and neck, limb, or other), indication (benign or cancer), and preoperative SARS-CoV-2 test (not-performed or negative); (3) operation factors: anaesthetic type (general, regional, or local) and operation grade (minor or major, defined according to the BUPA operation severity classification system listed in the schedule of procedures); (4) location factors: hospital type (COVID-19-free surgical pathway or no defined pathway), country income (high-income country or HIC, upper-middle-income country or UMIC, or low-middle-income country or LMIC [including both lower-middle and low-income countries], according to the World Bank classification), and community SARS-CoV-2 risk (low or high, based on the 14-day cumulative case notification rate in the area of the hospital around the date of surgery).

Country income has a known association with risk of perioperative complications, such as surgical site infection, anastomotic leak, and mortality; the reasons for this are multifactorial but are considered to reflect a hospital’s or health system’s capacity to rescue. This describes the ability of hospitals to identify, intervene, and rescue patients from complications, requiring infrastructure, consumables, facilities, training, staffing, and contextually relevant protocols. Patients with recent or active preoperative SARS-CoV-2 infection were excluded. As we could not predict potential risk of postoperative SARS-CoV-2 infection at an individual patient level using only information available preoperatively, we used community SARS-CoV-2 rates as a proxy for a patient’s risk of nosocomial (ie, in-hospital from patients being treated for COVID-19) or community (ie, after discharge while recovering at home) transmission. Due to differences in detection of SARS-CoV-2 with the availability of facilities for testing and notification across income settings, we preplanned to handle the interaction between these two location factors by creating a stratified factor of community SARS-CoV-2 risk by income.

Predictor variables were recorded at the time of patient entry (ie, day of surgery or day 0). Where one of the candidate variables was not collected in a dataset, and its value was implicit from the study protocol, it was imputed (ie, preoperative SARS-CoV-2 test imputed as not-performed, and community SARS-CoV-2 risk imputed as low in the pre-pandemic RECON dataset).

A complete-case analysis was preplanned if missing data were both missing at random and in a low number of samples (<5%). If both assumptions were not met, we planned to impute missing predictor variable data using multiple imputation by chained equations. The number of patients with missing data was reported for each predictor variable across each of the three studies.

Model derivation
First, we assessed the adequacy of our sample size for use in new model development, using an estimate of the predicted AUROC (0.700) and prevalence (3.0%) from a contemporary systematic review. We estimated a
minimum sample size requirement of 21 299 patients, and 639 events with RStudio using finalfit, tidyeval, cvAUC, tidymodels, workflowsets, finetune, PROC, rms, glmnet, dcurves, and shiny packages.

Two models were trained using the derivation data during model comparison. The first model consisted of a penalised logistic regression model, the least absolute shrinkage and selection operator (LASSO), in which logistic regression with an extra penalty term shrinks covariate coefficients towards zero, allowing the generation of sparse models and concurrently performing feature selection.\textsuperscript{23} The second model to be applied was gradient boosting decision trees (ie, XGBoost), which is commonly recognised as a best-in-class machine learning algorithm for datasets of moderate size, particularly where significant class

---

**Figure 1:** Study flowchart

LASSO=least absolute shrinkage and selection operator.
### Patient factors

<table>
<thead>
<tr>
<th>Age, years</th>
<th>Derivation: GlobalSurg-CovidSurg Week (n=86231)</th>
<th>Validation: CovidSurg-Cancer (n=30492)</th>
<th>Validation: RECON (n=6789)</th>
</tr>
</thead>
<tbody>
<tr>
<td>18-29</td>
<td>9106 (10.6%)</td>
<td>636 (2.1%)</td>
<td>276 (4.1%)</td>
</tr>
<tr>
<td>30-39</td>
<td>12176 (14.1%)</td>
<td>1708 (5.6%)</td>
<td>555 (8.2%)</td>
</tr>
<tr>
<td>40-49</td>
<td>13272 (15.4%)</td>
<td>3745 (12.3%)</td>
<td>983 (14.5%)</td>
</tr>
<tr>
<td>50-59</td>
<td>15703 (18.2%)</td>
<td>6697 (22.0%)</td>
<td>1466 (21.6%)</td>
</tr>
<tr>
<td>60-69</td>
<td>16665 (19.3%)</td>
<td>8347 (27.4%)</td>
<td>1623 (23.9%)</td>
</tr>
<tr>
<td>70-79</td>
<td>17472 (19.5%)</td>
<td>6954 (22.8%)</td>
<td>1390 (20.5%)</td>
</tr>
<tr>
<td>80-89</td>
<td>5110 (5.9%)</td>
<td>2265 (7.4%)</td>
<td>484 (7.1%)</td>
</tr>
<tr>
<td>&gt;90</td>
<td>457 (0.5%)</td>
<td>140 (0.5%)</td>
<td>12 (0.2%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sex</th>
<th>Derivation: GlobalSurg-CovidSurg Week (n=86231)</th>
<th>Validation: CovidSurg-Cancer (n=30492)</th>
<th>Validation: RECON (n=6789)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>47219 (54.8%)</td>
<td>18329 (60.1%)</td>
<td>3905 (57.5%)</td>
</tr>
<tr>
<td>Male</td>
<td>39012 (45.2%)</td>
<td>12163 (39.9%)</td>
<td>2884 (42.5%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ASA grade</th>
<th>Derivation: GlobalSurg-CovidSurg Week (n=86231)</th>
<th>Validation: CovidSurg-Cancer (n=30492)</th>
<th>Validation: RECON (n=6789)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grade 1</td>
<td>21667 (25.2%)</td>
<td>6421 (21.1%)</td>
<td>1048 (15.4%)</td>
</tr>
<tr>
<td>Grade 2</td>
<td>43099 (50.0%)</td>
<td>15752 (51.7%)</td>
<td>3903 (57.5%)</td>
</tr>
<tr>
<td>Grade 3</td>
<td>19052 (22.1%)</td>
<td>7912 (26.0%)</td>
<td>1733 (25.5%)</td>
</tr>
<tr>
<td>Grade 4 and 5</td>
<td>2408 (2.8%)</td>
<td>397 (1.3%)</td>
<td>105 (1.5%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Revised Cardiac Risk Index</th>
<th>Derivation: GlobalSurg-CovidSurg Week (n=86231)</th>
<th>Validation: CovidSurg-Cancer (n=30492)</th>
<th>Validation: RECON (n=6789)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>38541 (44.7%)</td>
<td>9245 (30.3%)</td>
<td>4112 (60.6%)</td>
</tr>
<tr>
<td>1</td>
<td>37688 (43.7%)</td>
<td>15896 (52.1%)</td>
<td>2210 (32.6%)</td>
</tr>
<tr>
<td>2</td>
<td>7595 (8.8%)</td>
<td>4284 (14.1%)</td>
<td>448 (6.6%)</td>
</tr>
<tr>
<td>≥3</td>
<td>2407 (2.8%)</td>
<td>1067 (3.5%)</td>
<td>19 (0.3%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pre-existing respiratory disease</th>
<th>Derivation: GlobalSurg-CovidSurg Week (n=86231)</th>
<th>Validation: CovidSurg-Cancer (n=30492)</th>
<th>Validation: RECON (n=6789)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>77747 (90.2%)</td>
<td>27282 (89.5%)</td>
<td>5611 (82.6%)</td>
</tr>
<tr>
<td>Yes</td>
<td>8484 (9.8%)</td>
<td>3210 (10.5%)</td>
<td>1178 (17.4%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Current smoker</th>
<th>Derivation: GlobalSurg-CovidSurg Week (n=86231)</th>
<th>Validation: CovidSurg-Cancer (n=30492)</th>
<th>Validation: RECON (n=6789)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>72830 (84.5%)</td>
<td>27311 (89.6%)</td>
<td>5852 (86.2%)</td>
</tr>
<tr>
<td>Yes</td>
<td>13401 (15.5%)</td>
<td>3181 (10.4%)</td>
<td>937 (13.8%)</td>
</tr>
</tbody>
</table>

### Disease factors

<table>
<thead>
<tr>
<th>Compartment</th>
<th>Derivation: GlobalSurg-CovidSurg Week (n=86231)</th>
<th>Validation: CovidSurg-Cancer (n=30492)</th>
<th>Validation: RECON (n=6789)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abdominopelvic</td>
<td>43761 (50.7%)</td>
<td>16812 (55.2%)</td>
<td>6789 (100.0%)</td>
</tr>
<tr>
<td>Head and neck</td>
<td>8228 (9.5%)</td>
<td>3823 (12.5%)</td>
<td>NA</td>
</tr>
<tr>
<td>Limb</td>
<td>12514 (14.5%)</td>
<td>196 (0.6%)</td>
<td>NA</td>
</tr>
<tr>
<td>Other</td>
<td>17216 (20.3%)</td>
<td>7226 (23.7%)</td>
<td>NA</td>
</tr>
<tr>
<td>Thoracic</td>
<td>4212 (4.9%)</td>
<td>2430 (8.0%)</td>
<td>NA</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Indication</th>
<th>Derivation: GlobalSurg-CovidSurg Week (n=86231)</th>
<th>Validation: CovidSurg-Cancer (n=30492)</th>
<th>Validation: RECON (n=6789)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Benign</td>
<td>64822 (75.2%)</td>
<td>NA</td>
<td>3884 (57.2%)</td>
</tr>
<tr>
<td>Cancer</td>
<td>21409 (24.8%)</td>
<td>30492 (100.0%)*</td>
<td>2905 (42.8%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Preoperative SARS-CoV-2 test</th>
<th>Derivation: GlobalSurg-CovidSurg Week (n=86231)</th>
<th>Validation: CovidSurg-Cancer (n=30492)</th>
<th>Validation: RECON (n=6789)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative</td>
<td>61339 (71.1%)</td>
<td>14128 (46.3%)</td>
<td>6789 (100.0%)***</td>
</tr>
<tr>
<td>Not performed</td>
<td>24892 (28.9%)</td>
<td>16364 (53.7%)</td>
<td>NA</td>
</tr>
</tbody>
</table>

### Operation factors

<table>
<thead>
<tr>
<th>Anaesthesia type</th>
<th>Derivation: GlobalSurg-CovidSurg Week (n=86231)</th>
<th>Validation: CovidSurg-Cancer (n=30492)</th>
<th>Validation: RECON (n=6789)</th>
</tr>
</thead>
<tbody>
<tr>
<td>General</td>
<td>61235 (73.3%)</td>
<td>28771 (94.4%)</td>
<td>6789 (100.0%)</td>
</tr>
<tr>
<td>Regional or local</td>
<td>22996 (26.7%)</td>
<td>1721 (5.6%)</td>
<td>NA</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Operation grade</th>
<th>Derivation: GlobalSurg-CovidSurg Week (n=86231)</th>
<th>Validation: CovidSurg-Cancer (n=30492)</th>
<th>Validation: RECON (n=6789)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Major</td>
<td>53061 (61.5%)</td>
<td>24439 (80.1%)</td>
<td>6789 (100.0%)</td>
</tr>
<tr>
<td>Minor</td>
<td>33170 (38.5%)</td>
<td>6053 (19.9%)</td>
<td>NA</td>
</tr>
</tbody>
</table>

(Table 1 continues on next page)

---

*imbalance exists.* XGBoost can acknowledge more complex interactions and so is hypothesised to obtain better predictive capabilities. However, because of its complexity, the ability to explain and understand predictive reasoning is hindered (black box machine learning). The derivation set was split into a random 75% (training) and 25% (testing), and both models were fit in the training set through 10-times cross-validation and hyperparameter tuning using tune race analysis of variance with grid search. The hyperparameters tuned were lambda for LASSO and learning rate, maximum tree depth, loss reduction, and minimum number, with trees set to 1000 for XGBoost.

During the model selection phase, model performance and feature importance were compared between the two models. Performance was estimated using the model discrimination, which describes models’ ability to rank patients based on risk. Discrimination was evaluated using the AUROC. Confidence intervals were assessed from the cross-validated models using the cvAUC package of R version 1.1.4. Variable importance was assessed for both modelling approaches to better understand the model’s decision making. Final model selection was balanced between (in order of priority): (1) performance; (2) interpretation and explainability; (3) the number of predictor variables (ie, fewer predictor variables assumed to be easier to adopt in clinical practice). If the LASSO model was selected, we preplanned to refit this model using resampling (2000 bootstraps) and average the coefficients across all resamples. To generate a pragmatic prognostic index that could be used at a patient’s bedside, we planned to scale and round the averaged penalised regression coefficients (eg, mean coefficients multiplied by 3 and rounded). If the XGBoost model was selected, we planned to conduct further bootstrapping and feature importance analysis (eg, DALEX). The performance of the final model (hereafter the GSU-Pulmonary Score) was assessed using the testing split previously derived (internal validation) and both discrimination and calibration were measured. Discrimination was assessed with 95% CIs calculated by bootstrap resampling (2000 samples). Calibration, measuring the agreement between predicted and observed risk, was examined through the Brier score, calibration in the large (CITL), and calibration slope, and visualised through reliability diagrams across risk deciles.

**Model validation**

After internal validation, the GSU-Pulmonary Score was externally validated using two large, international, prospective datasets described above (CovidSurg-Cancer and RECON). We assessed the GSU-Pulmonary Score’s performance (discrimination and calibration) in the external datasets, overall and in three preplanned subgroups deemed to be of high clinical importance: (1) major versus minor operation grade; (2) low (ASA grade 1–2) versus high (ASA grade ≥3) risk; and (3)
Patient and public involvement
A patient advisory group was formed with Patients and Research Together at Bowel Research UK (registration number 1120460) to provide active input into both the GlobalSurg-CovidSurg Week and CovidSurg-Cancer studies. This group helped to prioritise this research question, provide important input on the protocol, and coproduce patient-facing materials to disseminate research findings. Our methodology for rapid, responsive patient involvement during COVID-19 has been published,32 and our updated patient-facing resources are open access and available online. All patient partners are included as equal coauthors in our collaborative author group.

Role of the funding source
The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Results
Data from 126,410 patients undergoing surgery in 119 countries were eligible (figure 1). This included 87,863 patients in 1158 hospitals from 114 countries in the derivation dataset, 31,205 patients in 726 hospitals from 75 countries in the CovidSurg-Cancer validation dataset, and 7342 patients in 150 hospitals from three countries in the RECON validation dataset. The overall rate of missing data was very low (<0.1% overall and <2% for all candidate predictor variables; appendix p 2) and appeared to be missing at random (appendix p II), so the preplanned complete-case analysis was done. The complete-case analysis included 86,231 patients in model development, and 30,492 from CovidSurg-Cancer and 6789 from RECON in external validation.

Table 1 presents differences in casemix and outcomes across the three cohorts. In the derivation data, 2.1% (95% CI 2.0–2.2; 1781/86,231) of patients had a pulmonary complication. The 30-day PPC rate was higher in both the CovidSurg-Cancer (3.9%, 95% CI 3.7–4.1; 1185/30,492) and RECON (4.7%, 95% CI 4.2–5.2; 318/6789) validation studies. The postoperative mortality rate was 0.7% (0.6–0.8; 601/86,231, three missing) in the derivation data, 1.1% (95% CI 0.9–1.2; 320/30,492, 24 missing) in CovidSurg-Cancer, and 0.4% (95% CI 0.3–0.6; 27/6789, 223 missing) in RECON.

86,231 entered the structured machine learning process, in which patients were split into test and training sets, and the performance of XGBoost and LASSO models was assessed (figure 1). Similar discrimination (appendix p 3) was observed between LASSO (AUROC 0.786, 95% CI 0.774–0.798) and XGBoost (AUROC 0.785, 95% CI 0.774–0.798) in both validation datasets.
### Postoperative Pulmonary Complications

<table>
<thead>
<tr>
<th>Age, years</th>
<th>No postoperative complications</th>
<th>Postoperative pulmonary complications</th>
<th>Odds ratio (univariable)</th>
<th>Mean coefficients (LASSO)</th>
<th>Odds ratio (LASSO)</th>
<th>Point score</th>
</tr>
</thead>
<tbody>
<tr>
<td>18–29</td>
<td>6751 (10·7)</td>
<td>66 (5·0)</td>
<td>1·01 (0·73 to 1·40)</td>
<td>0·019 (0·012 to 0·025)</td>
<td>1·019</td>
<td>0</td>
</tr>
<tr>
<td>30–39</td>
<td>9030 (14·3)</td>
<td>87 (6·6)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>40–49</td>
<td>9822 (15·5)</td>
<td>154 (11·6)</td>
<td>1·63 (1·25 to 2·13)</td>
<td>0·176 (0·17 to 0·182)</td>
<td>1·192</td>
<td>1</td>
</tr>
<tr>
<td>50–59</td>
<td>11571 (18·3)</td>
<td>227 (17·1)</td>
<td>2·04 (1·59 to 2·62)</td>
<td>0·158 (0·153 to 0·161)</td>
<td>1·171</td>
<td>0</td>
</tr>
<tr>
<td>60–69</td>
<td>12122 (19·1)</td>
<td>325 (25·3)</td>
<td>2·82 (2·27 to 3·66)</td>
<td>0·290 (0·291 to 0·304)</td>
<td>1·348</td>
<td>1</td>
</tr>
<tr>
<td>70–79</td>
<td>10035 (15·8)</td>
<td>309 (23·3)</td>
<td>3·20 (2·53 to 4·08)</td>
<td>0·382 (0·376 to 0·388)</td>
<td>1·465</td>
<td>1</td>
</tr>
<tr>
<td>80–89</td>
<td>3686 (5·8)</td>
<td>123 (10·0)</td>
<td>3·75 (2·86 to 4·93)</td>
<td>0·594 (0·587 to 0·601)</td>
<td>1·811</td>
<td>2</td>
</tr>
<tr>
<td>≥90</td>
<td>231 (0·5)</td>
<td>14 (1·1)</td>
<td>4·39 (3·27 to 7·55)</td>
<td>0·882 (0·867 to 0·897)</td>
<td>2·416</td>
<td>3</td>
</tr>
</tbody>
</table>

### Sex

<table>
<thead>
<tr>
<th></th>
<th>No</th>
<th>Yes</th>
<th>Odds ratio (LASSO)</th>
<th>Point score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>34908 (55·1)</td>
<td>535 (40·4)</td>
<td>0·996</td>
<td>0</td>
</tr>
<tr>
<td>Male</td>
<td>28440 (44·9)</td>
<td>790 (59·6)</td>
<td>1·15</td>
<td>0</td>
</tr>
</tbody>
</table>

### ASA grade

<table>
<thead>
<tr>
<th></th>
<th>No</th>
<th>Yes</th>
<th>Odds ratio (LASSO)</th>
<th>Point score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>16031 (25·3)</td>
<td>143 (10·8)</td>
<td>0·931</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>31963 (50·5)</td>
<td>443 (33·4)</td>
<td>1·55</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>13712 (21·6)</td>
<td>570 (40·3)</td>
<td>1·28</td>
<td>0</td>
</tr>
<tr>
<td>4 or 5</td>
<td>1642 (2·6)</td>
<td>169 (12·8)</td>
<td>1·15</td>
<td>0</td>
</tr>
</tbody>
</table>

### Revised Cardiac Risk Index

<table>
<thead>
<tr>
<th></th>
<th>No</th>
<th>Yes</th>
<th>Odds ratio (LASSO)</th>
<th>Point score</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>28632 (45·2)</td>
<td>263 (19·8)</td>
<td>0·907</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>27661 (43·7)</td>
<td>610 (46·0)</td>
<td>1·12</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>5415 (8·5)</td>
<td>303 (22·9)</td>
<td>0·60</td>
<td>0</td>
</tr>
<tr>
<td>≥3</td>
<td>1640 (2·6)</td>
<td>149 (11·2)</td>
<td>0·60</td>
<td>0</td>
</tr>
</tbody>
</table>

### Pre-existing respiratory disease

<table>
<thead>
<tr>
<th></th>
<th>No</th>
<th>Yes</th>
<th>Odds ratio (LASSO)</th>
<th>Point score</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>57217 (90·3)</td>
<td>1052 (79·4)</td>
<td>0·986</td>
<td>0</td>
</tr>
<tr>
<td>Yes</td>
<td>6131 (9·7)</td>
<td>273 (20·6)</td>
<td>1·03</td>
<td>0</td>
</tr>
</tbody>
</table>

### Current smoker*

<table>
<thead>
<tr>
<th></th>
<th>No</th>
<th>Yes</th>
<th>Odds ratio (LASSO)</th>
<th>Point score</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>53592 (84·6)</td>
<td>1071 (80·8)</td>
<td>0·94</td>
<td>0</td>
</tr>
<tr>
<td>Yes</td>
<td>9756 (15·4)</td>
<td>254 (19·2)</td>
<td>1·00</td>
<td>0</td>
</tr>
</tbody>
</table>

### Compartment

<table>
<thead>
<tr>
<th></th>
<th>No</th>
<th>Yes</th>
<th>Odds ratio (LASSO)</th>
<th>Point score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abdominopelvic</td>
<td>32281 (51·0)</td>
<td>649 (49·0)</td>
<td>0·98</td>
<td>0</td>
</tr>
<tr>
<td>Head and neck</td>
<td>6043 (9·5)</td>
<td>110 (8·3)</td>
<td>0·91</td>
<td>0</td>
</tr>
<tr>
<td>Limb</td>
<td>9208 (14·5)</td>
<td>111 (8·4)</td>
<td>0·90</td>
<td>0</td>
</tr>
<tr>
<td>Other</td>
<td>13010 (20·5)</td>
<td>141 (10·6)</td>
<td>0·90</td>
<td>0</td>
</tr>
<tr>
<td>Thoracic</td>
<td>2806 (4·4)</td>
<td>314 (23·7)</td>
<td>0·90</td>
<td>0</td>
</tr>
</tbody>
</table>

### Indication

<table>
<thead>
<tr>
<th></th>
<th>No</th>
<th>Yes</th>
<th>Odds ratio (LASSO)</th>
<th>Point score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Benign</td>
<td>47878 (75·6)</td>
<td>755 (57·0)</td>
<td>0·98</td>
<td>0</td>
</tr>
<tr>
<td>Cancer</td>
<td>15470 (24·4)</td>
<td>570 (43·0)</td>
<td>1·03</td>
<td>0</td>
</tr>
</tbody>
</table>

### Preoperative SARS-CoV-2 test*

<table>
<thead>
<tr>
<th></th>
<th>No</th>
<th>Yes</th>
<th>Odds ratio (LASSO)</th>
<th>Point score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative</td>
<td>45047 (71·1)</td>
<td>960 (72·5)</td>
<td>0·94</td>
<td>0</td>
</tr>
<tr>
<td>Not performed</td>
<td>18301 (28·9)</td>
<td>365 (27·5)</td>
<td>1·04</td>
<td>0</td>
</tr>
</tbody>
</table>

### Anaesthesia type

<table>
<thead>
<tr>
<th></th>
<th>No</th>
<th>Yes</th>
<th>Odds ratio (LASSO)</th>
<th>Point score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regional or local</td>
<td>17067 (26·9)</td>
<td>138 (10·4)</td>
<td>0·90</td>
<td>0</td>
</tr>
<tr>
<td>General</td>
<td>46281 (73·1)</td>
<td>1187 (89·6)</td>
<td>0·90</td>
<td>0</td>
</tr>
</tbody>
</table>

### Operation grade

<table>
<thead>
<tr>
<th></th>
<th>No</th>
<th>Yes</th>
<th>Odds ratio (LASSO)</th>
<th>Point score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minor</td>
<td>24750 (39·1)</td>
<td>197 (14·9)</td>
<td>0·80</td>
<td>0</td>
</tr>
<tr>
<td>Major</td>
<td>38598 (60·9)</td>
<td>1128 (85·1)</td>
<td>0·80</td>
<td>0</td>
</tr>
</tbody>
</table>

### Hospital type*

<table>
<thead>
<tr>
<th></th>
<th>No</th>
<th>Yes</th>
<th>Odds ratio (LASSO)</th>
<th>Point score</th>
</tr>
</thead>
<tbody>
<tr>
<td>COVID-19-free surgical pathway</td>
<td>52788 (83·3)</td>
<td>1077 (81·3)</td>
<td>0·90</td>
<td>0</td>
</tr>
<tr>
<td>Hospital with no defined pathway</td>
<td>10560 (16·7)</td>
<td>248 (18·7)</td>
<td>0·92</td>
<td>0</td>
</tr>
</tbody>
</table>

(Table 2 continues on next page)
Table 2: Final GSU-Pulmonary Score model description (derivation dataset)

<table>
<thead>
<tr>
<th>Country income × SARS-CoV-2 risk</th>
<th>No postoperative pulmonary complications</th>
<th>Postoperative pulmonary complications</th>
<th>Odds ratio (univariable)</th>
<th>Mean coefficients (LASSO)</th>
<th>Odds ratio (LASSO)</th>
<th>Point score</th>
</tr>
</thead>
<tbody>
<tr>
<td>High income, high SARS-CoV-2 risk</td>
<td>35553 (56.1)</td>
<td>705 (53.2)</td>
<td>1.30 (1.07 to 1.58)</td>
<td>0.289 (0.284 to 0.293)</td>
<td>1.334</td>
<td>1</td>
</tr>
<tr>
<td>High income, low SARS-CoV-2 risk</td>
<td>8055 (12.7)</td>
<td>123 (9.3)</td>
<td>1.32 (1.10 to 1.59)</td>
<td>0.285 (0.282 to 0.287)</td>
<td>1.334</td>
<td>1</td>
</tr>
<tr>
<td>Upper-middle income, high SARS-CoV-2 risk</td>
<td>6121 (9.7)</td>
<td>142 (10.7)</td>
<td>1.52 (1.19 to 1.94)</td>
<td>0.278 (0.273 to 0.283)</td>
<td>1.334</td>
<td>2</td>
</tr>
<tr>
<td>Upper-middle income, low SARS-CoV-2 risk</td>
<td>4555 (7.2)</td>
<td>154 (11.6)</td>
<td>2.21 (1.74 to 2.82)</td>
<td>1.039 (1.033 to 1.044)</td>
<td>2.825</td>
<td>3</td>
</tr>
<tr>
<td>Low-middle income, high SARS-CoV-2 risk</td>
<td>379 (0.6)</td>
<td>18 (1.4)</td>
<td>3.11 (1.82 to 5.02)</td>
<td>1.22 (1.208 to 1.232)</td>
<td>3.873</td>
<td>4</td>
</tr>
<tr>
<td>Low-middle income, low SARS-CoV-2 risk</td>
<td>8675 (13.7)</td>
<td>183 (13.8)</td>
<td>1.38 (1.10 to 1.74)</td>
<td>0.851 (0.846 to 0.857)</td>
<td>2.343</td>
<td>3</td>
</tr>
<tr>
<td>Intercept</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-6.836 (-6.844 to -6.827)</td>
<td>-21</td>
<td></td>
</tr>
</tbody>
</table>

Model description after 2000 bootstraps and averaged coefficients. To generate the final scaled calculator score, mean coefficients were multiplied by 3 and rounded. All categorical variables were dummyfied (ie, each different category acts as an independent variable). Where multiple factor levels (eg, age group) score 0 points, this reflects non-linear patterns of risk over different groups. During penalised regression using LASSO, the coefficients for all factor levels have shrunk towards 0 for some covariables. This indicates that they had provided little discriminative ability and therefore were not selected for inclusion in the final model (eg, current smoker). ASA=American Society of Anesthesiologists. LASSO=least absolute shrinkage and selection operator. *Covariables that were removed from the final model after penalised regression.

95% CI 0.772–0.797) models. Variable importance observed in both modelling approaches is summarised in the appendix (p 4). Based on our predefined model selection criteria, given the similar performance, more explainable nature, and increased simplicity of LASSO, we proceeded to develop this model as the GSU-Pulmonary Score. After resampling with 2000 bootstraps in the training set, beta coefficients were averaged and three variables with zero coefficients were dropped. The penalised regression coefficients were then scaled, generating the final pragmatic clinical scoring system, which included ten predictor variables. Results are summarised in table 2. For clinical application, individual point scores for each risk factor level are summed to give an overall point score out of a maximum of 23. The GSU-Pulmonary model was assessed in the derivation (testing) set with AUROC of 0.773 (95% CI 0.751–0.795; figure 2) and good calibration (Brier score 0.020, CITL 0.034, slope 0.954; figure 3).

We estimated a minimum sample size of 193 events and 6403 non-events for external validation based on the incidence of pulmonary complications in the derivation data. In assessment of the adequacy of sample size, both CovidSurg-Cancer (1185 events, 29303 non-events) and RECON (318 events, 6471 non-events) were sufficient for the spread of the linear predictor (log-odds predictor) was narrower in RECON (appendix pp 7, 9) than in CovidSurg-Cancer (appendix pp 8–9). In step two, we compared the calibration and discrimination of the model across the three datasets (figures 2, 3). In step three, we concluded that CovidSurg-Cancer appeared to evaluate (statistical) reproducibility of the GSU-Pulmonary Score whereas RECON appeared to favour (clinical) transportability.

Prognostic test accuracy metrics across a range of cutoff point scores to rule in or rule out PPC in the whole derivation set (appendix p 12) and cutoffs in validation sets (appendix pp 13–14) are provided. A cut-point score of 5 or less (30689 [35.6%-] of 23029 patients, PPC rate 0.4%) ruled out PPC with a sensitivity of 93.05%, specificity of 35.82%, positive predictive value of 2.96, and negative predictive value of 99.95. The distribution of point scores and risk estimates is shown in the appendix (p 9). Four risk strata were defined clinically as low risk (0–5 score, PPC rate in whole derivation set 0.45%, n=30689), intermediate risk (6–10 score, PPC rate 1.68%, n=43888), high risk (11–15 score, PPC rate 7.37%,
n=11027), and very high risk (≥16 score, PPC rate 14·67%, n=627). Risk groupings across the three datasets are shown in the appendix (p 6).

Finally, the GSU-Pulmonary Score calculator was deployed online. The independent PROBAST assessment indicated that the study was at low risk of bias and there were low concerns about applicability (appendix pp 22–30).

Decision curve analysis showed clinical use of the COVIDSurg Pulmonary Score over a range of threshold risks in CovidSurg-Cancer data (appendix p 10); however, our model had limited clinical use at low-risk thresholds in RECON data due to miscalibration and subsequent underestimation of risk. Therefore, we suggest caution when using the model to guide clinical management in patients with a low risk of pulmonary complications in non-COVID-19 data.

**Discussion**

This study has defined a novel risk stratification tool (GSU-Pulmonary Score) to estimate the risk of pulmonary complications in adults undergoing elective surgery. It requires simple predictor variables, which are readily available to clinicians across all resource settings and at the point of a decision for surgery. The study was designed and reported according to best-practice guidelines and was at a low risk of bias. It is the first scoring tool to consider background community SARS-CoV-2 infection rate, allowing dynamic risk prediction in areas with ongoing SARS-CoV-2 transmission and until universal vaccination coverage is achieved. The score was shown to have acceptable discrimination in two large, well powered, multicountry validation sets, with good calibration in CovidSurg-Cancer and miscalibration in RECON data. Pulmonary events are the major driver of death after elective surgery, and increased in frequency and severity during the COVID-19 pandemic.

We recommend that the GSU-Pulmonary Score be used to inform patient selection, consent, and resource allocation as elective surgery is upscaled during pandemic recovery efforts and beyond. Delays in elective health care represent a major threat to global public health and care for non-communicable disease, particularly for time-critical conditions such as cancer. As global health-care systems continue to upscale elective surgery to mitigate against further harm for these patients, accurate and accessible risk stratification tools will be essential to inform shared decision making about the timing of surgery. For patients classified at higher risk (including those who remain unvaccinated), vaccination (or a booster dose) can be offered, and...
surgery might be done under regional or local anaesthetic, or delayed while the patient is preoptimised (eg, preoperative chest physiotherapy, prophylactic mucolytics, or goal-directed haemodynamic therapy) or until community SARS-CoV-2 rates decline. Critical care providers could use this score to plan bed availability. For patients judged to be at lower risk, it is likely to be safe to continue surgery even when community SARS-CoV-2 rates are high. Decisions to postpone, continue, or cancel surgery are complex, multifactorial, and require deep shared decision making between the patient and the perioperative multidisciplinary team; the likelihood of clinical benefit of surgery would have significant influence on this process. It would therefore be inappropriate to specify exactly how the GSU-Pulmonary Score should be used in decision making. Instead, we present a range of cutoff point scores to rule in or rule out PPC that clinicians can explore during implementation. The incidence of PPC in all three studies was greater than 1·68% (intermediate-risk group cutoff), reflecting the fact that patients undergoing abdominal surgery are at intermediate-to-high risk at baseline. Of note, we did not include patients with recent preoperative SARS-CoV-2 infection in this study and the model is not directly applicable to this group; other data are available to inform clinical decision making here. Although some of the variability in PPC rates might have been due to inclusion of a small number of patients with undetected SARS-CoV-2, much of this variation would be captured by community SARS-CoV-2 transmission rates; preoperative SARS-CoV-2 testing had little effect on model discrimination and was dropped in penalised regression.

A systematic review and validation study in 2022 identified several other published prognostic models to estimate the risk of pulmonary complications after major abdominal surgery. In the same RECON cohort study data used in this study, no model showed good discrimination (an AUROC greater than 0·7). Here, in RECON study data, the GSU-Pulmonary Score displayed better discrimination than the best-performing score (Assess Respiratory Risk in Surgical Patients in Catalonia [ARISCAT]) identified in the systematic review (AUROC 0·716, 95% CI 0.689–0.744 vs 0·700, 0·683–0·717) and the Score for the Prediction of Postoperative Respiratory Complications ([SPORC]-1; AUROC 0·574, 95% CI 0·556–0·593), although with some miscalibration (CITL close to 1, and predicted risks systematically too low). RECON included major resectional abdominopelvic surgery in high-income countries only. Miscalibration might be due to unmeasured differences in casemix (eg, balance of major and complex-major surgery within the single major surgery predictor variable), differences in practice during the pandemic (eg, proactive risk modification), or system-level factors such as access to care and health behaviours in lower-resource settings (eg, barriers to travel to hospital after discharge, so PPC is less likely to be observed). The RECON dataset was included to improve relevance to countries and regions, which might in the future have very low or no ongoing SARS-CoV-2 transmission.

The GSU-Pulmonary Score has several notable advantages over previous prognostic tools for surgical patients. First, it considers variation in patients’ risk related to community SARS-CoV-2 case notification rates. Existing COVID-19 risk scores in general medical populations can only stratify risk for patients with active SARS-CoV-2 infection, so are applicable to only a small subset of elective surgery patients and are not designed to reflect the physiological insult of surgery. Second, it includes only resource-light predictor variables that do not require any additional tests or equipment (in contrast to preoperative oxygen saturation and serum haemoglobin in the ARISCAT score). It is applicable across resource settings and can be calculated remotely (eg, from electronic health records or during telemedicine consultation). Third, it includes only predictor variables available before the day of surgery, so can support preoperative decision making (in contrast to duration of surgery and emergency procedure in ARISCAT). Fourth, it has been developed and validated in large, international datasets with broad inclusion criteria, making it highly generalisable. Fifth, we did a model comparison study and selected a modelling technique that is readily interpretable by the front-line clinician, improving clinical utility. The finding that advanced structured machine learning approaches are not superior to simple regression, even in very large clinical datasets, has been replicated elsewhere. We also highlighted that different variables were most important across the two modelling approaches. This has important ethical implications for clinical risk prediction models in the future, whereby high-level decisions (eg, whether a patient should be offered surgery) might be based on different factors depending on the modelling approach. The interpretation framework for external validation studies suggests that the score is transferable across different perioperative patient groups, including beyond the COVID-19 pandemic.

There are also several limitations to this analysis. First, some clinically relevant predictor variables were not included in score development (eg, cardiac history, preoperative physiology, and cardiopulmonary exercise testing). This was a pragmatic decision to facilitate high data quality across large, prospective studies, and to ensure that the resulting model was applicable to low-resource settings. The output estimates should be implemented alongside holistic assessment of the patient’s status and with shared decision making. Data on ethnicity were also not collected due to lack of a global, culturally sensitive classification system. Improving reporting of ethnicity in model development is an important emerging area of research. Second, all clinical predictor variables were collected categorically.
This simplification might have reduced statistical power, and might explain a degree of miscalibration in RECON data. Third, data from the derivation and validation (CovidSurg-Cancer) sets came from the COVID-19 pandemic period. Although there continue to be areas of the world with high rates of SARS-CoV-2 infection, the individual risk of pulmonary complications in infected surgical patients is likely to be lower in the omicron-variant era and in vaccinated patients. Despite this, preliminary reports from the CovidSurg-3 study (NCT05161299; prospective data collection between December, 2021, and February, 2022) indicate that high-risk patients (eg, those aged >70 years, ASA grade 3–5) remain at very high risk of pulmonary complications and death with perioperative SARS-CoV-2 infection. To explore the implications of this on model performance, we also undertook model validation in a non-COVID dataset (RECON). Model re-calibration might be required in the future as the community gains a deeper understanding of the contemporary phenotype of COVID-19 in a surgical population. The directionality of effect for high versus low SARS-CoV-2 risk in upper-middle-income countries was reversed in comparison with other income groups; this might reflect within-group variation in access to testing and reporting rather than a true biological effect. Fourth, we dealt with missing data (<0.1% overall) using complete-case analysis. However, several clinical and simulation studies have shown little effect of multiple imputation methods where there are very low rates of missing data. Fifth, it was not possible to mask front-line clinical staff who recorded outcome assessment to several predictor variables, as these were readily apparent. However, the primary outcome measure required clinical imaging studies (chest x-ray or CT) for diagnosis that were reported by independent radiologists, minimising risk of bias. We adopted a composite outcome measure, and the model might not discriminate equally for each included endpoint. Sixth, the score cannot be applied to patients undergoing procedures under deep sedation, who are at risk of pulmonary complications. Seventh, death presented a competing risk for pulmonary complications and shared a common causal pathway for some patients. We showed robust performance of the score in a sensitivity analysis, but were unable to account for this during model derivation. Eighth, more models and preprocessing could have been implemented, including using methods for managing class imbalance (ie, oversampling or under-sampling, or Synthetic Minority Over-Sampling Technique). However, these are criticised in clinical datasets due to risk of miscalibration (strong overestimation of probability to belong to the minority class), and we opted not to adopt these here. Ninth, we adopted two techniques for model comparison but did not include other interpretable model types that can account for non-linear covariates, such as generalised additive modelling. This could be an important consideration in future work. Tenth, only one of three cohort studies (RECON) explicitly preplanned the development and validation of a prognostic model in its study protocol. Finally, we have not been able to compare performance of the GSU-Pulmonary Score with other models in a dataset, which includes all operation types and indications; our inferences about superior performance of the model relate to abdominal surgery only, and we have not compared it with SPORC-2, which was published after study completion. This highlights an area for future comparisons.
Acknowledgments
We thank the Royal College of Surgeons of England COVID-19 recovery research group for their support.

References