BMC Medical Informatics and Decision Making

RESEARCH

Open Access

Comparing conventional and Bayesian workflows for clinical outcome prediction modelling with an exemplar cohort study of severe COVID-19 infection incorporating clinical biomarker test results

Brian Sullivan[1*], Edward Barker[1], Louis MacGregor[1], Leo Gorman[8], Philip Williams[3], Ranjeet Bhamber[1], Matt Thomas[7], Stefan Gurney[4], Catherine Hyams[1], Alastair Whiteway[6], Jennifer A. Cooper[1], Chris McWilliams[2], Katy Turner[1], Andrew W. Dowsey[1†] and Mahableshwar Albur[1,5*†]

Abstract

Purpose Assessing risk factors and creating prediction models from real-world medical data is challenging, requiring numerous modelling decisions with clinical guidance. Logistic regression is a common model for such studies, for which we advocate the use of Bayesian methods that can jointly deliver probabilistic risk factor inference and prediction. As an exemplar, we compare Bayesian logistic regression with horseshoe priors and Projective Prediction variable selection with the established frequentist LASSO approach, to predict severe COVID-19 outcomes (death or ICU admittance) from demographic and laboratory biomarker data. Our study serves as guidance on data curation, variable selection, and performance assessment with cross-validation.

Methods Our source data is based on a retrospective observational cohort design with records from three National Health Service (NHS) Trusts in southwest England, UK. Models were fit to predict severe outcomes within 28 days after admission to hospital (or a positive PCR result if already admitted) using demographic data and the first result from 30 biomarker tests collected within 3 days after admission (or testing positive if already admitted).

Results Patients included hospitalized adults positive for COVID-19 from March to October 2020, 756 total patients: Mean age 71, 45% female, 31% (n=234) had a severe outcome, of whom 88% (n=206) died. Patients were split into training (n=534) and external validation groups (n=222). Using our Bayesian pipeline, we show a reduced variable model using Age, Urea, Prothrombin time (PT) C-reactive protein (CRP), and Neutrophil-Lymphocyte ratio (NLR) has better predictive performance (median external AUC: 0.71, 95% Quantile [0.7, 0.72]) relative to a GLM using all variables (external AUC: 0.67 [0.63, 0.71]).

†Andrew W. Dowsey and Mahableshwar Albur contributed equally to this work.

*Correspondence:
Brian Sullivan
brian.sullivan@bristol.ac.uk
Mahableshwar Albur
mahableshwar.albur@nbt.nhs.uk
Full list of author information is available at the end of the article

Conclusion Urea, PT, CRP, and NLR have been highlighted by other studies, and respectively suggest that hypovolemia, derangement of circulation via clotting, and inflammation are strong predictive risk factors of severity. This study provides guidance on conventional and Bayesian regression and prediction modelling with complex clinical data.

Keywords Projective prediction, Bayesian, Logistic regression, Risk factors, COVID-19

Introduction

Estimating predictive risk factors for disease outcomes with explainable statistical models is desirable for clinical use and decision making. Clinical resources are typically limited, and variable selection techniques that can reduce complex multivariate models to ones with a smaller subset are useful as they can offer similar performance without the resource cost of collecting additional test results. We provide a guide for modern Bayesian approaches for joint risk factor analysis and variable selection demonstrated in a patient dataset obtained from UK hospitals during the first wave of the COVID-19 pandemic.

We analyze a range of laboratory blood marker values across metabolic pathways affected by COVID-19 infection and evaluate predictive models of severe outcomes. We: (a) Examine statistical associations of routinely measured blood biomarkers, and age and gender, to predict severe COVID-19 outcomes; (b) Develop cross-validated logistic regression prediction models using the candidate biomarkers, highlighting biomarkers worthy of future research. (c) Employ variable selection techniques, comparing the least absolute shrinkage and selection operator (LASSO) frequentist method [1] to the recent Projective Prediction approach [2] on Bayesian logistic regression models with horseshoe priors to illustrate the process of creating a reduced model that maintains similar performance while being more feasible to implement clinically; (d) We demonstrate a balance between best analytic practices and pragmatic solutions for clinical data curation and statistical modelling decisions emphasizing the benefits of the proposed Bayesian workflow.

While our paper is methodological, we detail several aspects of COVID-19 and surrounding research to motivate circumstances around the dataset we obtained and the types of clinical decisions made. Further, clinical considerations motivate why variable selection can play a crucial factor in modeling. Globally, COVID-19 has resulted in hundreds of millions of cases and millions of deaths (WHO Coronavirus (COVID-19) Dashboard https://covid19.who.int/). COVID-19 has a wide spectrum of clinical features ranging from asymptomatic to severe systemic illness with a significant attributable mortality, while clinical manifestations are variable especially in the most vulnerable groups and immunocompromised people [3]. COVID-19 is a multi-system disease resulting in the derangements of homeostasis affecting pulmonary, cardiovascular, coagulation, haematological, oxygenation, hepatic, renal and fluid balance [4–6]. During the first wave of the pandemic, the majority of people with COVID-19 had mild or no symptoms, but an estimate of one in five to one in 10 needed hospitalisation [7]. Early identification of hospitalised COVID-19 patients who are likely to deteriorate, i.e. transfer to ICU or who may die, is vital for clinical decision making.

Several prediction models have evaluated case-level factors that might predict poor outcomes (critical illness or death). A recent living systematic review [8] identified 265 prognostic models for mortality and 84 for progression to severe or critical state. The majority of the studies looked at vital signs, age, comorbidities, and radiological features. According to the review, models were unlikely to include a broad range of variables concerning co-infection, biochemical factors (outside of C-reactive protein), and other haematological factors on an individual patient level. Further, most prognostic models did not describe the target population or care setting adequately, did not fully describe the regression equation, showed high or unclear risk of bias and/or were inadequately evaluated for performance. These drawbacks highlight a need to demonstrate sound practices for severity prediction modeling. Collins et al. and Riley et al. have written a compelling series of such recommendations on many shortcomings of clinical prediction models and steps to remedy [9–11].

We compiled a COVID-19 dataset that is novel in the broad number of blood biomarkers included from clinical laboratory testing, supported by routine patient demographic information. Our dataset was captured with the intent to create a clinical severity score to complement those using physiological data for use during the pandemic, but it became apparent that the dataset was not adequately powered to definitively answer this question. We deviate from suggestions from Riley et al. and Collins et al. concerning sample sizes for data as our dataset is limited in the number of severe outcome examples. We emphasize that the work here is primarily a guide and not intended to make a definitive statement for COVID-19 prediction models. The experiences gained over the course of this research led us to refocus our attention on demonstrating our statistical workflows for this complex

data. We highlight two methodologically sound contemporary models from Knight et al. and Carr et al. [12, 13] with better powered studies, but neither record the same biomarkers as each other or our dataset (making direct comparison difficult), nor do they use Bayesian approaches for modeling or variable selection, the present work's strength.

Methods

Overview

Using complex clinical data, we use logistic regression to predict the likelihood of a severe outcome (death or transfer to ICU) for a patient based on demographic information and any available patient biomarker data collected during a 3-day time window starting from being admitted to hospital with COVID-19 or testing positive if already admitted. If a patient transferred to ICU during the 3-day window, we only consider data collected prior to transfer. We highlight the benefits of a Bayesian approach and then focus on variable selection. Clinicians must balance time, money, and equipment access all while trying to deliver high quality patient care. Models that deliver good prediction performance with a small amount of biomarker data are valued for their efficiency.

Study cohort and demographics

Following a retrospective observational cohort design, anonymized data were obtained from Laboratory Information Management Systems (LIMS) linking patient data for laboratory markers to key clinical outcomes. Three hospitals in the Southwest region of England, UK, participated in the study, two of which were tertiary teaching hospitals and the third was a district general hospital (DGH).

The study underwent a rigorous ethical and regulatory approval process, following an Integrated Research Application System application [IRAS project ID: 283439], a favourable written authorization was gained from NHS Research Ethics Service, Wales Research Ethics Committee 7, c/o Public Health Wales, Building 1, Jobswell Road, St David's Park, SA31 3HB on 11/09/2020. Our research complies with the declaration of Helsinki with anonymized data and ethical review, as explained below informed consent for data sharing was waived due to overriding public interest. See: https://www.hra.nhs.uk/about-us/committees-and-services/res-and-recs/

The requirement for informed consent was waived by NHS Research Ethics Service, Wales Research Ethics Committee 7 (see above), given overriding public interest in the research. Furthermore, during project development prior to ethics review, a public and patient involvement meeting conducted at North Bristol NHS Trust by author MH received similar support. North Bristol NHS Trust and University Hospitals Bristol and Weston NHS Foundation Trust signed data sharing agreements confirming this waiver. See: https://www.england.nhs.uk/publication/information-sharing-policy All data were fully anonymized before they were transferred to the research team for analysis.

A system-wide data search was conducted on the LIMS for all patients who tested positive for SARS-CoV-2 by polymerase chain reaction (PCR) at these three hospitals during the first wave of COVID-19 pandemic. Data were collected from records between March 1, 2020 to October 31, 2020, with research data access authorized from January 1, 2021 to present day. Serial laboratory data collected as a part of standard of care of patients admitted with/for COVID-19 were included: bacteriology, virology, mycology, haematology, and biochemistry. All patients testing negative for SARS CoV-2 by PCR were excluded. All laboratory markers including clinical outcomes from LIMS were extracted and the final dataset was anonymized with no patient identifying data to link back.

Inclusion and exclusion criteria

To be included in the study we had several mandatory criteria. We included all adult patients admitted to the study's hospitals between March to October 2020 and tested positive for SARS CoV-2 by PCR. Pediatric patients (<18 years old) were excluded. Hospital staff/healthcare workers and their house-hold contacts were excluded prior to data transfer (as it was marked on COVID-19 test requests). Figure 1 depicts the decision flow for inclusion and exclusion of patient data. Furthermore, all patients required age, gender, complete admission/discharge records, and records of their outcome with COVID-19. If a patient had multiple admissions, only the most recent admission since a positive COVID-19 test was considered. Despite our data request constraints, the data transferred contained records outside our criteria. For example, not all patients had records indicating a positive COVID-19 test; we speculate there was a data processing or human error. When combined with restrictions on biomarker data, this considerably narrowed our data set from 1159 patients to 736 who met all criteria, as detailed in the flow chart.

Predictors (data covariates)

Our dataset includes a variety of clinical severity indices, microbiological, immunological, haematological and biochemistry parameters used as predictive variables in the regression models. A full list of recorded data items is shown in Table 1

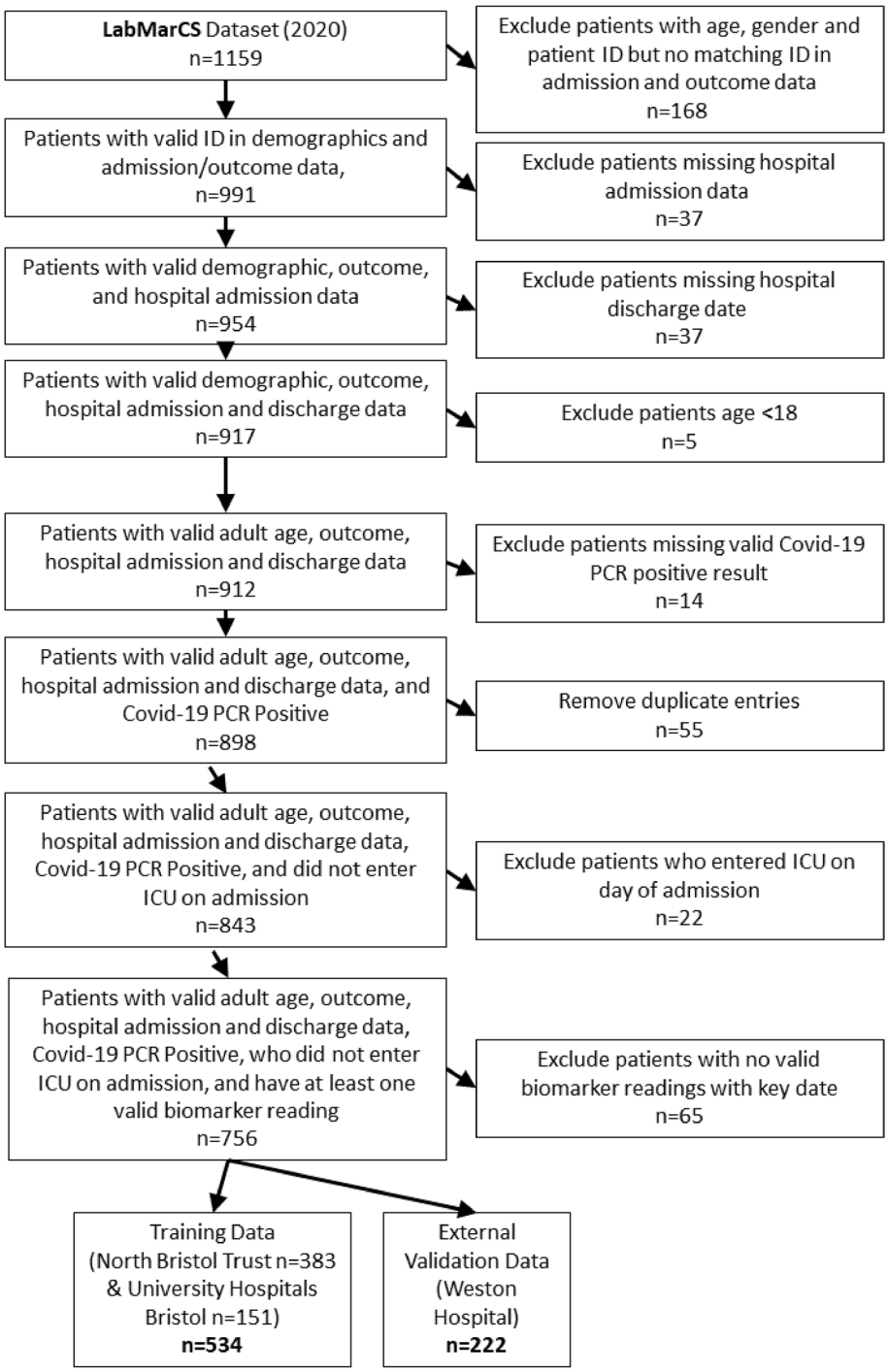

Fig. 1 Flowchart of patient exclusion and inclusion criteria. The initial set of 1159 candidate patients was narrowed to a training set (n=534) and an external validation set (n=222)

Outcomes

For all sites, the primary prediction outcome was death or transfer to the ICU within 28 days after the key date. This key date was either the point of admission to hospital, or the date of the first positive COVID-19 PCR test result if the patient was already admitted. 28 days was chosen due to clinical convention and advice from our clinical colleagues, using a different time window may be justified in other circumstances. The distribution of severe outcomes after the key date is right skewed, with

Table 1 Variables recorded in the study dataset, including plain text description, abbreviation, place of record, frequency in the dataset, and criteria used for converting continuous readings into categorical values

Biomarker	Abbreviation	Comment	# Readings	% of Patients With Reading	Reference range/criteria	Categories	Comments
Blood Clotting Tests							
Activated partial thromboplastin time	APTT	Admisson	385	51%	Normal between 21-33 seconds	Normal, Abnormal	Normal: <33; Mild: 33-49.5; Moderate: 49.5-82.5; Severe: >82.5
Prothrombin Time	PT	Admisson	399	53%	Normal between 9.5-13 seconds	Normal, Abnormal	Abnormal: >=13
Blood Gas Tests							
Carbon Dioxide	CO2	Arterial/ Point of care	127	17%	Normal: 4.6-6.4 seconds	Normal, Abnormal	Abnormal if outside range
Lactate	poctLAC	Arterial/ Point of care	127	17%	0.5-2.2 mmol/L	Normal, Abnormal	Abnormal if <0.5 or >2.2
Oxygen	O2	Arterial/ Point of care	127	17%	11.0-14.4 seconds	Normal, Abnormal	Abnormal if <11 or >14.4
Bicarbonate Excess	BE	Arterial or Venous / Point of care	394	52%	22-29	Normal, Abnormal	Abnormal if outside range
pH acid/base scale	pH	Arterial or Venous / Point of care	393	52%	7.35-7.45	Normal, Abnormal	Abnormal if outside these bounds
Coinfection Battery							
Blood Culture	bc_coinfection	Admisson	756	100%	34 bacterial strains tested	Positive, Negative	Postive if one or more postive
Respiratory	resp_coinfection	Admisson	756	100%	34 bacterial strains tested	Positive, Negative	Postive if one or more postive
Urine	urine_coinfection	Admisson	756	100%	34 bacterial strains tested	Positive, Negative	Postive if one or more postive
Viral	viral_coinfection	Admisson	756	100%	10 viral infections tested	Positive, Negative	Postive if one or more postive
Diabetes							
Glucose	Glucose	Point of Care / Record Often Not Digitized	197	26%	Non-fasting: 3.0-7.8 mmol/L	Normal, Abnormal	Abnormal if outside range
Full Blood Count Tests							
Hemoglobin	HB	Admisson	735	97%	Male 130-170 g/L, Female 120-150 g/L	Normal, Mild, Moderate, Severe	
Platelet Count	PLT	Admisson	733	97%	150-450 10^9/L	Normal, Mild, Moderate, Severe	Normal: >150; Mild: 100-150; Moderate: 50-100; Severe: <50
Lymphocytes	Lymphocyte	Admisson	735	97%	1.5-4.5 10^9/L	Normal, Mild, Moderate, Severe	Normal 1.5-4.5; Mild 1-1.5; Moderate 0.5-1; Severe: <0.5 or >4.5
Neutrophils	Neutrophil	Admisson	750	99%	2.0-7.5 10^9/L	Normal, Mild, Moderate, Severe	Normal 2-7.5; Mild 1-2; Moderate: 0.5-1; Severe: <0.5 or > 7.5
Neutrophil - Lymphocyte Ratio	NLR	Admisson	735	97%	0.78 and 3.53	Normal, Mild, Moderate, Severe	Normal: <3; Mild: 3-8; Moderate: 8-18; Severe: >18
White Cell Count	WCC	Admisson	735	97%	4.0-11.0 10^9/L	Normal, Mild, Moderate, Severe	Normal: 4-11; Mild: 1-4; Moderate: 0.5-1; Severe: <0.5 and >11
Urea & Electrolytes Tests							
C-Reactive Protein	CRP	Admisson	722	96%	6 mg/L	Normal, Abnormal	Abnormal if greater than criteria
Estimated Glomerular Filtration Rate	eGFR	Admisson	672	89%	91	Normal, Abnormal	Abnormal if less than criteria
Urea	urea	Admisson		0%	2.5-7 10^9/L	Normal, Abnormal	Abnormal if outside these bounds
Investigatory Tests							
Brain / B-type natriuretic peptide	BNP	Cardiac Function	39	5%	Men under 70: <100pg/ml, Women under 70: <150 pg/ml, All 70yr and over: <300 pg/ml	Normal, Abnormal	Abnormal if greater than age/gender specific criteria
D-Dimer	DDM		89	12%	Age (Years) D-dimer (ng/ml) <60 <500 61-70 <600 71-80 <700 81-90 <800 >90 <900	Normal, Abnormal	Abnormal if greater than age-specific criteria
Ferritin	FER		94	12%	Male: 33-490, Female(0-44): 15-445, Female(45+yrs): 30-470	Normal, Mild, Moderate, Severe	Normal: <age/gender appropriate criteria; Mild: >criteria-735; Moderate: 735-2450; Severe: >2450
Fibrinogen	fib		86	11%	1.8-4.0 g/L	Normal, Mild, Severe.	Normal: >1.8; Mild: 1-1.8; Severe: <1
Glycated haemoglobin	HBA1c	Diabetes	17	2%	47 mmol/mol	Normal, Abnormal	Abnormal if greater than criteria
Lactate dehydrogenase	LDH	Investigatory	50	7%	240-480 IU/L	Normal, Mild, Moderate, Severe	Normal: <=480; Mild: >480-720; Moderate: >720-1440; Severe: >1440
Procalcitonin	PCT	ITU / Bacterial Infection	9	1%	Normal range: <0.2ng/mL	Normal, Abnormal	Abnormal: >=0.2
Triglycerides	trig	Investigatory	5	1%	0.5-1.7 mmol/L	Normal, Abnormal	Abnormal if outside these bounds
Troponin-T	trop	Cardiac Function	159	21%	13ng/L	Normal, Abnormal	Abnormal if greater than criteria
Covid-19 Test							
Covid CT	Covid CT		756	100%	Threshold unique to type of test. Lab reports categorical 'positive' variable alongside CT value	Positive, Negative	Only positives included in current study
Other Data							
Age	Age		756	100%		Continuous	All ages >=18
Gender	Gender		756	100%		Male, Female	
Covid Positive on Admission	OnAdmission		756	100%		True, False	Tested only in univariate evaluation
Outcome	Outcome		756	100%		Discharge, ICU, Death	

75% occurring within 10 days, with a mean of 7.6 days and standard deviation of 5.5.

Patient timelines

The collected laboratory biomarkers are continuous measures and provide a time-series representation of the course of a patient's admission. Figure 2 shows an example of a single patient's readings over the course of 18 days between testing positive for COVID-19 and being released from hospital care. This provides a representative example of the heterogeneity seen in our dataset, i.e. not all tests are taken and others are taken regularly or intermittently (further examples in Supplementary Materials A2-A6).

Transformation of biomarker data

Prediction modelling of irregularly sampled time-series data is a challenging open research question [14]. In this study we focused on established and available tools for conventional and Bayesian prediction. To balance inclusion of biomarker data not available on the day of admission and the need for clinical decisions to be guided soon after admission, we chose to consider the first value recorded for each biomarker within three days after their 'key date'. We additionally considered the worst or best readings within 1, 5 or 7 days after the key date, and found the first reading within 3 days after the key date to offer a reasonable compromise between prediction performance and speed to inform decision making. Systematic exploration of these parameters would be worthwhile to optimize performance in coordination with clinical needs, but is the beyond the scope of this paper.

In addition, we transformed continuous biomarkers into categorical variables via reference ranges for clinical use in the typical healthy population ranges, see Table 1. These categories are actively in use at laboratories at the participating trusts and were arrived at through a combination of clinician advice, handbooks [15], and guidance from lab test manufacturers. Such transforms are not a trivial decision and there are merits to both hand-crafted transforms informed by domain experts (as we have chosen) versus data driven approaches. On one hand, clinical experience has delineated useful categories of biomarker readings, but it is not evident a priori that such categories are removing nuance present in a continuous measure, especially in the case of a novel disease. Furthermore, a transform could be learned across studies or tuned for a particular dataset. However, this requires sufficient representative data and may require further choices of transformation or non-linear modelling approaches. As an example, Fig. 3 shows the histogram of readings for all values recorded for Neutrophils, including clinical

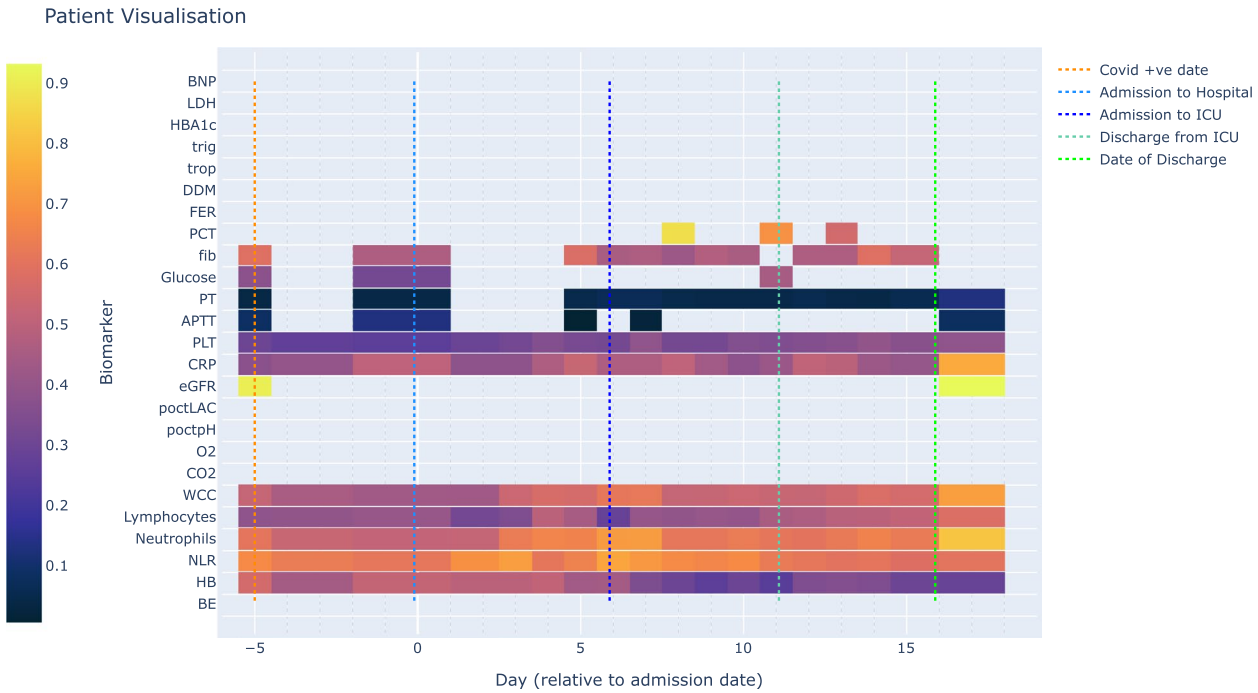

Fig. 2 Example of a single patient's time series laboratory biomarker data. Covid +ve indicates the time of a COVID-19 positive test. See Table 1 for biomarker abbreviations. Biomarkers vary widely in units of measurement. As a simple indication of upward and downward movement of readings, variation in biomarker measures are visualised from low (purple/black) to high (orange/yellow) created by subtracting the minimum value and normalizing the readings to span 0 to 1

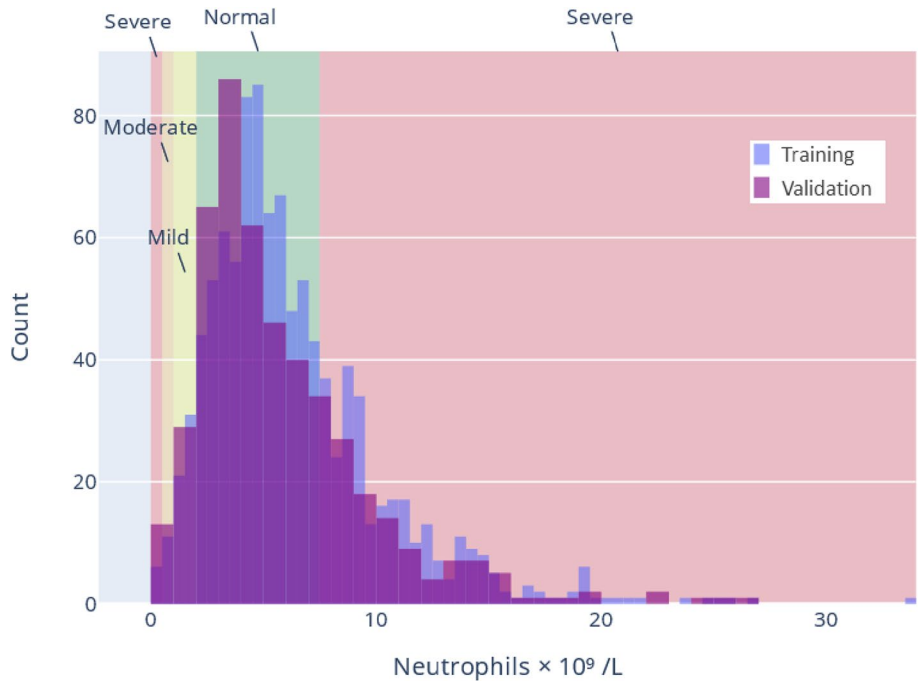

Fig. 3 Example distribution of biomarker readings for Neutrophil training and external validation data. Vertical lines indicate clinical thresholds for bounds on Normal, Mild, Moderate, and Severe categorization

thresholds to transform into categorical data. No missing data imputation was performed, instead missingness was coded as as an additional category 'Test not taken'. The distribution of how many patients fell into each category per biomarker can be found in Table 2

For further elaboration of these modelling choices and the challenges, please see the Discussion.

Statistical analysis

Analytics were carried out using the R statistical language (v4.4.1) and R Studio (2024.09.0). We used the following packages: Standard logistic regression analyses used the R Stats GLM package (v4.4.1); LASSO analyses, GLMnet (v4.1.8); and for Bayesian analyses, BRMS (v2.22.0) and ProjPred (v2.8.0). Source code for this analysis pipeline can be found at https://github.com/biospi/LABMARCS.

Analysis of individual biomarkers

Before running full regression models, we examined the independent contribution of individual biomarkers in the training dataset predicting ICU entry or death via standard logistic regressions and Bayesian logistic regressions with either a flat (aka uniform) or horseshoe prior. This allowed calculation of p-values and odds ratios for each biomarker. A 5-fold cross-validation (repeated 20 times was run for each biomarker to estimate median AUC and 95% interquartile intervals. Stratified cross-validation data shuffling was pre-computed per biomarker so models used the same starting data. When performing basic cross-validation, it is possible that some folds end up with few or zero data-points for an outcome so that convergence becomes poor or impossible (which is exacerbated by some biomarkers having few readings). Conversely, stratified cross-validation guarantees that the outcome occurs in the same proportion in each fold and in the same proportion as in the overall training data distribution for that biomarker. To achieve this, per biomarker, patients with and without the outcome were separated and then these groups were shuffled and split into 5 equal subgroups. These groups can then paired at random, ensure training and test datasets have the same proportion of patients with a severe outcome as in full the sample for that biomarker. This substantively improves the chance of convergence for biomarkers with high data missingness.

Here, only complete cases of training data available for each biomarker were considered, i.e. we did not include data for variables marked 'Test not taken', to focus on the predictive power of observed test results. The prediction power of the underlying test can be confounded by the clinical decision to order the test only in certain patient circumstances. However, in some cases 'Test Not Taken' is much more common (and as described in the results motivating exclusion of the biomarker) and would not represent the predictive power of the biomarker result. In proceeding sections, we allow for a 'Test not

Table 2 Individual biomarker evaluation including descriptive statistics, unadjusted *p*-values, and logistic regression model outcomes (Standard, Bayesian with flat prior, and Bayes with horseshoe prior), including age and gender (except univariate age and gender models)

					Standard Logistic GLM		Bayesian Logistic (Flat Prior)	Bayesian Logistic (Horse Shoe Prior)
Biomarker	**Binary Categorical Variable**	**% of Patients with Biomarker Recording**	**# TRUE (% of TRUE Observations with Severe Outcome)**	**# FALSE (% of FALSE Observations with Severe Outcome)**	**P Value**	**Odds Ratio [2.5%, 97.5%]**	**Odds Ratio [2.5%, 97.5%]**	**Odds Ratio [2.5%, 97.5%]**
Demographics / Other								
Age	-	100%	-	-	8.39E-07	1.02 [1.01, 1.04]	1.02 [1.01, 1.04]	1.02 [1.01, 1.04]
Gender	Female	100%	257 (26%)	333 (32%)	0.08	0.72 [0.50, 1.03]	0.72 [0.50, 1.03]	0.79 [0.54, 1.07]
Age & Gender	-	100%	-	-	2.83E-05	1.02 [1.01, 1.04]	1.02 [1.01, 1.04]	1.02 [1.01, 1.04]
	Female		257 (26%)	333 (32%)	0.06	0.70 [0.49, 1.02]	0.70 [0.48, 1.02]	0.83 [0.55, 1.07]
Nosocomial Transmission	TRUE	100%	240 (30%)	350 (29%)	0.65	0.92 [0.63, 1.33]	0.91 [0.63, 1.32]	0.94 [0.70, 1.21]
Blood Clotting Tests								
Activated partial thromboplastin time	Mild	54%	30 (63%)	291 (320%)	2.44E-03	3.44 [1.57, 7.88]	3.6 [1.52, 8.30]	2.79 [1.12, 6.19]
	Moderate	54%	4 (100%)	317 (34%)	0.98	9.91E+06 [0.00, NA]	5.3E+106 [7.9E+05, Inf]	6.84 [0.92, 218.16]
Prothrombin Time	Abnormal	56%	45 (58%)	288 (31%)	2.96E-03	MarkerReferenceRanges_Oc	2.18 [1.01, 4.67]	2.26 [1.01, 4.63]
Blood Gas Tests								
Carbon Dioxide	Abnormal	21%	68 (59%)	57 (51%)	0.33	1.44 [0.70, 2.99]	1.46 [0.69, 3.01]	1.08 [0.82, 1.92]
Lactate	Abnormal	21%	13 (54%)	112 (55%)	0.96	1.03 [0.32, 3.44]	1.03 [0.32, 3.50]	1.00 [0.58, 1.64]
Oxygen	Abnormal	21%	105 (55%)	20 (55%)	0.98	1.01 [0.38, 2.66]	1.01 [0.37, 2.69]	1.00 [0.65, 1.54]
Bicarbonate Excess	Abnormal	64%	123 (38%)	252 (31%)	0.26	1.30 [0.82, 2.05]	1.30 [0.82, 2.00]	1.11 [0.86, 1.70]
pH acid/base scale	Abnormal	63%	136 (46%)	238 (26%)	1.05E-04	2.45 [1.56, 3.87]	2.48 [1.59, 3.96]	2.23 [1.33, 3.57]
Coinfection								
Blood Culture	TRUE	100%	5 (0%)	585 (30%)	0.98	3.20E-07 [NA, 2.94E+22]	2.17E-140 [0, 2E-05]	0.45 [0.01, 1.26]
Respiratory	TRUE	100%	6 (50%)	584 (29%)	0.20	2.95 [0.52, 16.62]	2.94 [0.48, 18.21]	1.30 [0.69, 5.62]
Urine	TRUE	100%	12 (25%)	579 (30%)	0.63	0.72 [0.15, 2.53]	0.64 [0.14, 2.46]	0.93 [0.36, 1.56]
Viral	TRUE	100%	7 (71%)	583 (29%)	0.06	4.95 [1.04, 35.13]	5.86 [1.06, 44.82]	1.92 [0.82, 12.53]
Diabetes								
Glucose	Abnormal	30%	49 (45%)	126 (32%)	0.11	1.77 [0.88, 3.54]	1.80 [0.90, 3.63]	1.31 [0.85, 2.60]
Full Blood Count Tests								
Hemoglobin	Mild	92%	176 (36%)	368 (27%)	0.13	1.38 [0.91, 2.08]	1.37 [0.91, 2.08]	1.14 [0.90, 1.74]
	Moderate		48 (33%)	495 (30%)	0.62	1.19 [0.59, 2.29]	1.17 [0.59, 2.31]	1.02 [0.68, 1.57]
	Severe		11 (55%)	532 (30%)	0.03	4.08 [1.16, 15.06]	4.22 [1.20, 15.15]	1.63 [0.83, 7.16]
Platelet Count	Mild	92%	67 (39%)	474 (29%)	0.07	1.65 [0.95, 2.83]	1.63 [0.93, 2.86]	1.33 [0.90, 2.31]
	Moderate		17 (65%)	524 (29%)	0.01	4.21 [1.54, 12.65]	4.40 [1.53, 13.35]	2.58 [0.97, 8.33]
	Severe		4 (75%)	537 (30%)	0.12	6.16 [0.76, 126.83]	9.34 [0.84, 305.58]	1.86 [0.69, 17.37]
Lymphocytes	Mild	92%	151 (27%)	392 (31%)	0.12	1.69 [0.89, 3.34]	1.71 [0.89, 3.26]	1.06 [0.73, 1.86]
	Moderate		217 (31%)	326 (30%)	0.03	1.96 [1.07, 3.75]	1.99 [1.08, 3.70]	1.21 [0.88, 2.18]
	Severe		84 (48%)	459 (27%)	4.99E-04	3.48 [1.75, 7.17]	3.53 [1.74, 7.11]	2.00 [1.00, 4.35]
Neutrophils	Mild	92%	23 (13%)	520 (31%)	0.23	0.47 [0.11, 1.43]	0.41 [0.09, 1.34]	0.71 [0.22, 1.29]
	Moderate		3 (33%)	540 (30%)	0.67	1.71 [0.08, 19.15]	1.29 [0.04, 21.86]	1.07 [0.30, 3.80]
	Severe		143 (41%)	400 (26%)	1.88E-03	1.92 [1.27, 2.91]	1.93 [1.28, 2.92]	1.75 [1.07, 2.70]
Neutrophil - Lymphocyte Ratio	Mild	92%	237 (28%)	306 (32%)	3.69E-03	2.50 [1.38, 4.79]	2.60 [1.40, 5.05]	1.83 [0.99, 3.58]
	Moderate		137 (39%)	406 (27%)	3.18E-05	3.97 [2.12, 7.81]	4.13 [2.19, 8.23]	2.90 [1.42, 5.80]
	Severe		54 (54%)	489 (28%)	2.61E-06	6.38 [2.99, 14.14]	6.72 [3.06, 15.22]	4.48 [1.91, 10.15]
White Cell Count	Mild	92%	57 (23%)	486 (31%)	0.34	0.72 [0.36, 1.38]	0.70 [0.35, 1.32]	0.84 [0.47, 1.23]
	Moderate		2 (50%)	541 (30%)	0.45	3.03 [0.11, 83.24]	3.11 [0.08, 117.32]	1.15 [0.43, 5.00]
	Severe		85 (42%)	458 (28%)	0.02	1.84 [1.12, 3.00]	1.83 [1.10, 3.00]	1.48 [0.97, 2.68]
Urea & Electrolytes Tests								
C-Reactive Protein	Abnormal	91%	489 (33%)	47 (4%)	1.49E-03	10.23 [3.08, 63.44]	13.12 [3.48, 77.20]	7.81 [2.44, 31.55]
Estimated Glomerular Filtration Rate	Abnormal	82%	350 (38%)	131 (18%)	0.06	1.76 [0.98, 3.23]	1.79 [0.98, 3.28]	1.42 [0.93, 2.83]
Urea	Abnormal	89%	262 (47%)	264 (15%)	4.23E-11	4.27 [2.79, 6.63]	4.33 [2.85, 6.79]	4.13 [2.69, 6.33]
Investigatory Tests								
Brain / B-type natriuretic peptide	Abnormal	7%	30 (53%)	14 (29%)	0.13	3.91 [0.73, 27.00]	4.78 [0.76, 34.31]	1.52 [0.74, 9.03]
D-Dimer	Abnormal	12%	52 (42%)	18 (33%)	0.67	1.29 [0.40, 4.43]	1.34 [0.40, 4.88]	1.11 [0.61, 2.62]
Ferritin	Mild	14%	11 (64%)	72 (39%)	0.09	3.61 [0.84, 17.70]	4.00 [0.85, 19.54]	1.30 [0.78, 5.08]
	Moderate	14%	28 (46%)	55 (40%)	0.27	1.79 [0.64, 5.15]	1.85 [0.66, 5.30]	1.09 [0.74, 2.36]
	Severe	14%	6 (33%)	77 (43%)	0.94	0.93 [0.11, 5.90]	0.82 [0.09, 5.71]	0.95 [0.35, 1.89]
Fibrinogen*	Mild	5%	4 (75%)	26 (46%)	0.10	11.27 [0.85, 360.85]	24.77 [1.06, 1.29E+03]	1.40 [0.67, 11.84]
	Severe	5%	3 (67%)	27 (48%)	0.40	3.41 [0.23, 105.85]	3.41 [0.23, 105.85]	1.15 [0.46, 5.30]
Glycated haemoglobin*	Abnormal	3%	11 (9%)	4 (0%)	1.00	2.98E+08 [0, NA]	8.30E+11 [0.22, 4E+46]	1.28 [0.32, 15.50]
Lactate dehydrogenase*	Mild	6%	12 (67%)	25 (56%)	0.49	2.61 [0.19, 71.00]	3.67 [0.18, 166.51]	1.14 [0.58, 3.98]
	Moderate	6%	16 (63%)	21 (57%)	0.78	1.81 [0.09, 77.89]	1.81 [0.09, 77.89]	1.01 [0.39, 2.45]
	Severe	6%	5 (60%)	32 (59%)	0.34	4.63 [0.22, 178.20]	7.51 [0.19, 465.95]	1.06 [0.43, 3.67]
Procalcitonin*	Abnormal	4%	21 (86%)	4 (100%)	1.00	1.15E-07 [NA, 1.6E+184]	2.3E-07 [1.05E-31, 5.16]	0.80 [0.08, 2.72]
Triglycerides*	Abnormal	3%	10 (90%)	5 (100%)	1.00	1.68E-09 [NA, Inf]	4.3E-07 [1.98E-30, 1.31]	0.74 [0.04, 3.05]
Troponin-T	Abnormal	24%	91 (44%)	51 (22%)	0.03	2.96 [1.17, 7.96]	3.07 [1.25, 7.91]	1.71 [0.91, 5.40]

* Biomarkers not included in subsequent models due to small sample size, and recorded only in ICU (PCT)

The True and False columns describe the number/percentage of severe outcomes for cases where the particular biomarker or demographic reading is true or false. For example, there were 257 patients who were women who had a severe outcome, and conversely there were 333 patients who were not women (i.e. men) who had a severe outcome. Regressions were fit using all associated dummy variables for a given biomarker (e.g. normal, mild, moderate, severe) and using only complete cases of training data, i.e. not using a variable for 'Test not taken'. Categorical variables use a reading of 'Normal' as a reference in the fitted model, except 'Male' used as the reference category for gender

taken' category, effectively removing the complete cases requirement. Each individual biomarker model includes age and gender (except univariate age and gender models) and were compared against a standard model including only age and gender. Regressions were fit using all associated dummy variables for a given biomarker (e.g. 'Mild', 'Moderate', 'Severe') using 'Normal' as the reference.

Analysis using all valid biomarker data

After individual biomarker evaluation, logistic regression models considering all valid biomarkers (Prediction using individual variables section) and demographic variables were fit to the data. Their predictions were tested via internal and external validation using the stratified cross-validation procedures detailed above, except models were fit using all available training data using 'Test Not Taken' for absent data. The models include a standard logistic regression, a logistic regression regularised with LASSO, and two Bayesian models using a flat and a horseshoe prior [16].

Analysis using reduced variable models

While a model using all biomarker data may have strong predictive power, it is clinically desirable to have a strong prediction with the least amount of biomarkers possible to save on time, money and other resources devoted to biomarker collection and analysis [17, 18]. We used two methodologies to choose reduced variable models to predict COVID-19 severe outcomes, LASSO and Bayesian Projective Prediction.

LASSO is an optimization constraint that shrinks parameters according to their unexplained variance with respect to the outcome variable, reduces over-fitting, and enables variable selection [1]. The optimal degree of regularisation is determined by tuning parameter λ within each cross-validation fold through a nested cross-validation step. LASSO has a drawback of having biased coefficient and log-odds estimates, as such after evaluating LASSO models there is a need to run a standard logistic regression model on the reduced biomarker panel selected with the LASSO in order to reduce bias in reporting risk factor effect sizes.

To evaluate LASSO coefficient estimates, we performed repeated nested stratified cross-validation (5-folds for the inner LASSO loop; 5-folds for the outer loop, and 20 repeats). For a particular dataset fit, LASSO optimises for a sparse representation with many coefficients set to zero. Across cross-validated trials these variables will vary. LASSO fits are statistically biased and are better suited as a guide for variable selection, with a reduced variable standard logistic regression used to infer odds ratios. As recommended in Heinze et al. [19], we consider the frequency of how often a particular biomarker has non-zero log-odds coefficients and count across cross-validation trials. There is no set rule for how to translate these frequencies into a set of reduced variables. We suggest to only consider variables that have non-zero coefficients at least 50% of the time, but this is merely a heuristic.

For determining unbiased effect sizes for the reduced variable set with a standard GLM, it was decided that if at least one categorical level for a particular biomarker (e.g. 'Severe') was selected by the LASSO, all levels for that biomarker were included in the model. This resulted in a final set of 'LASSO inspired' variables that were then fit with standard logistic GLM. Note this approach, and more generally fitting multiple models to the same dataset, is subject to the problem of selective inference (aka multiple comparison error), see [20, 21] and the related R package [22]. This is a limitation that is improved by the Bayesian approach described below.

The second variable selection method explored was Bayesian Projective Prediction [2], a technique for assessing reduced variable models against a complete 'reference' model, which in our case is a Bayesian logistic regression with a horseshoe prior [16]. Priors such as the horseshoe can be applied to provide adaptive shrinkage to covariates in Bayesian models directly so that full posterior distributions of odds estimates can be generated in an unbiased way. Unlike the LASSO, this does not shrink coefficients to zero exactly as the inherent uncertainty is not ignored. To perform hard variable selection, the recent approach of Projective Prediction can be used to compare the fit of sub-models of the reference model through projections and approximate leave-one-out (LOO) cross-validation. Under the hood, Projective Prediction uses forward search to select submodels for comparison, but retains the Bayesian inference for coefficient ranking and odds-ratio estimates. The projective prediction package allows systematic evaluation of the trade-off between AUC performance and the ranked contribution from the variables included in the model. The experimenter must decide at what performance level to cut-off the ranked variables. We chose to examine when increases in AUC asymptoted and used any biomarkers that did not have 'Test Not Taken' as the highest ranked predictor. This reduced set of variables was used for the submodel projection. Projective prediction allows the flexibility to train one model on all valid available data, perform variable selection, and then use any projected sub-model with reduced variables to predict outcomes for novel data. Projective prediction models were evaluated using cross-validation procedures described in prior sections. Note, the analysis of the projective prediction model using all training data uses LOO for variable selection, which is computationally intensive. To speed variable selection computation during our cross-validation analysis, we used 'naive' variable selection, which only considers the training data from current fold as is, and does not perform any further internal cross-validation (the projective prediction package allows naive, k-fold, and LOO).

Results

Cohort description

756 of 1159 patients (73%) patients testing positive for SARS-CoV-2 were eligible given our inclusion criteria, see Fig. 1. Of these patients, 57% were hospitalised for COVID-19 (n=433), the remainder (n=323) had nosocomial infection. For our statistical models, the training cohort (n=534) was defined as all adults admitted to hospital and testing positive for SARS-Cov-2 by PCR, or testing positive while already admitted between March 1 to October 31, 2020. For external validation, we held the DGH cohort (n=222) out of training. This cohort was selected as the hospital is in another county compared to the trusts used in the training data. To avoid over-fitting to local idiosyncrasies, ideally, the external validation data set would differ on a national or international level. Given our limited data, this was the best external validation possible. Patients in the training set had a mean age of 70, were 44% female, and 28% had severe outcomes. The external validation set had a mean age of 74, were 47% female, and 37% had a severe outcome. There were statistically significant differences (tested via Wilcoxon Mann-Whitney U test evaluated at a significance level of 0.05), with the external validation set having a larger incidence of severe outcomes ($W = 64296$, p-value $= 0.02$), and an older population ($W = 68074$, p-value < 0.001). Gender was statistically similar ($W = 57480$, p-value $= 0.44$)

Prediction using individual variables

Table 2 shows descriptive statistics on individual biomarker readings and their odds ratio contributions in a 5-fold 20-repeat stratified cross-validated logistic regression including the particular biomarker and age and gender. Our approach uses complete cases to estimate the predictive capacity of biomarker test results (avoiding 'Test Not Taken') but this may introduce bias as discussed. Table 3 details performance using the area under the receiver operating characteristic curve (AUC) metric, comparing biomarker models (a particular biomarker plus age and gender) to a model using only age and gender. A simple age and gender model acts as a foil to illustrate the worth of a biomarker over easily collected but often predictive variables. Due to the categorical representation of the biomarkers, individual levels may be significant while another is not (e.g. 'Severe' is a predictor, but 'Mild' is not). Statistically significant predictors (i.e. odds ratios deviating from one with p-value at 0.05 or lower) associated with increasing risk of a severe outcome (as shown in Table 2) include age, and the biomarkers: Activated Partial Thromboplastin Time (Mild), Prothrombin time (Abnormal), blood pH (Abnormal), Haemoglobin (Severe), Platelet count (Moderate), Lymphocytes (Moderate, Severe), Neutrophils (Severe), Neutrophil-Lymphocyte Ratio (Mild, Moderate, Severe), C-Reactive Protein (Abnormal), Urea (Abnormal), and Troponin-T (Abnormal). Nosocomial transmission was included due to the high number of cases in our cohort but was not a significant predictor and excluded from further analyses. Due to small numbers preventing cross-validation, Triglycerides, Glycated Haemoglobin, Procalcitonin (also invalid due to being recorded only in ICU), Fibrinogen, and Lactose Dehydrogenase were excluded from further analysis and require future research.

Regression models using all valid biomarker data

Each model was evaluated via 5-fold stratified cross-validation with 20 repeats (100 models total). As such, each model is trained with a randomised sample of 80% of the training data set (n=427). Internal validation evaluates model predictions on the 20% (n=107) held out. External validation uses the same model, but is instead tested on the never trained on external validation data set on recorded a separate hospital (n=222). Missing data for each biomarker is coded as 'Test Not Taken' and is included as a predictor variable. Table 4 shows the performance of these models (AUC, Sensitivity, Specificity).

To estimate variability in model performance and allow comparison between models, we compute inter-quantile AUC difference ranges using 5-fold 20-repeat cross-validation of models. While Delong's method [23] is also be used to compare between models, it tests only for a significant difference between the AUCs of two trained models. Conversely, cross-validation (or bootstrap) considers also variability in model training due to sample variance by providing a comparison across models for each of many data splits. For each data split, we compute the AUC for a given model and then compute the delta to the reference model (Bayesian horseshoe), thus allowing the comparison of 95% intervals. Cross-validation results provide 95% inter-quantile ranges that clearly illustrate that in general, all models perform similarly, with a median AUC ranging from 0.76–0.82 in internal validation, and ranging from 0.67–0.71 in external validation. While the LASSO inspired GLM model has the best median internal AUC difference (0.02 better than the Bayesian horseshoe reference), all models overlap in their 95% AUC difference intervals. When considering external validation, the median AUC difference tends to be smallest or even slightly positive for the Bayesian methods, but all models overlap within the 95% bounds of the reference model, except the LASSO model. Note the LASSO model also has higher variation in AUC difference indicating the model's performance is not very consistent across cross-validation folds. The calibration

Table 3 Predictive performance of the individual biomarker models in Table 2 as described by the median area under the curve (AUC) in receiver operating curve (ROC) analysis and median difference between an Age and Gender reference model and the same model (negative values indicate the reference has worse performance) with the particular biomarker included (except univariate age and gender models)

	Standard Logistic GLM		Bayesian Logistic (Flat Prior)		Bayesian Logistic (Horse Shoe Prior)	
	Cross-Validated 80/20 Split		Cross-Validated 80/20 Split		Cross-Validated 80/20 Split	
Demographic / Biomarker	**Median AUC [2.5%,97.5%]**	**Median AUC Difference to Age & Gender Standard [2.5%,97.5%]**	**Median AUC [2.5%,97.5%]**	**Median AUC Difference to Age & Gender Standard [2.5%,97.5%]**	**Median AUC [2.5%,97.5%]**	**Median AUC Difference to Age & Gender Standard [2.5%,97.5%]**
Demographics / Other						
Age	0.62 [0.52, 0.74]	0.00 [-0.10, 0.03]	0.61 [0.45, 0.74]	0.01 [-0.10, 0.06]	0.61 [0.51, 0.73]	0.00 [-0.13, 0.03]
Gender	0.54 [0.45, 0.62]	0.07 [-0.06, 0.17]	0.54 [0.46, 0.64]	0.08 [-0.08, 0.20]	0.54 [0.47, 0.62]	0.07 [-0.11, 0.17]
Age & Gender	0.62 [0.51, 0.71]	0.00, [0.00, 0.00]	0.62 [0.51, 0.73]	0.00, [-0.02, 0.01]	0.61 [0.52, 0.72]	0.00, [-0.02, 0.02]
Nosocomial Transmission	0.61 [0.47, 0.69]	0.00 [-0.05, 0.04]	0.61 [0.48, 0.69]	0.00 [-0.02, 0.04]	0.61 [0.43, 0.70]	0.00 [-0.11, 0.02]
Blood Clotting Tests						
Activated partial thromboplastin time	0.65 [0.46, 0.78]	-0.04 [-0.23, 0.04]	0.65 [0.47, 0.78]	-0.04 [-0.24, 0.06]	0.64 [0.45, 0.77]	-0.04 [-0.15, 0.06]
Prothrombin Time	0.64 [0.46, 0.77]	-0.03 [-0.22, 0.06]	0.64 [0.47, 0.76]	-0.03 [-0.16, 0.05]	0.64 [0.50, 0.75]	-0.03 [-0.13, 0.04]
Blood Gas Tests						
Carbon Dioxide	0.54 [0.42, 0.69]	0.01 [-0.16, 0.24]	0.55 [0.42, 0.68]	0.00 [-0.15, 0.21]	0.56 [0.43, 0.69]	0.00 [-0.17, 0.13]
Lactate	0.57 [0.42, 0.72]	0.00 [-0.20, 0.10]	0.56 [0.40, 0.72]	-0.01 [-0.16, 0.19]	0.57 [0.42, 0.76]	0.00 [-0.23, 0.18]
Oxygen	0.57 [0.43, 0.74]	0.00 [-0.21, 0.14]	0.56 [0.42, 0.71]	0.00 [-0.16, 0.12]	0.54 [0.42, 0.70]	0.03 [-0.15, 0.18]
Bicarbonate Excess	0.57 [0.47, 0.70]	0.00 [-0.09, 0.11]	0.57 [0.45, 0.70]	0.00 [-0.08, 0.13]	0.58 [0.43, 0.72]	0.00 [-0.05, 0.13]
pH acid/base scale	0.65 [0.48, 0.75]	-0.07 [-0.18, 0.08]	0.65 [0.48, 0.74]	-0.07 [-0.26, 0.07]	0.64 [0.47, 0.75]	-0.06 [-0.21, 0.09]
Coinfection						
Blood Culture	0.62 [0.52, 0.71]	-0.01 [-0.02, 0.00]	0.62 [0.52, 0.72]	-0.01 [-0.03, 0.01]	0.62 [0.51, 0.72]	-0.00 [-0.13, 0.02]
Respiratory	0.61 [0.51, 0.73]	-0.00 [-0.02, 0.01]	0.62 [0.50, 0.73]	-0.00 [-0.05, 0.02]	0.62 [0.50, 0.73]	-0.00 [-0.02, 0.02]
Urine	0.61 [0.47, 0.70]	0.00 [-0.00, 0.04]	0.61 [0.48, 0.70]	0.00 [-0.03, 0.04]	0.62 [0.50, 0.71]	0.00 [-0.05, 0.02]
Viral	0.65 [0.56, 0.76]	0 [-0.07, 0.02]	0.64 [0.56, 0.75]	0 [-0.06, 0.02]	0.64 [0.56, 0.75]	0 [-0.04, 0.02]
Diabetes						
Glucose	0.69 [0.43, 0.84]	-0.01 [-0.09, 0.07]	0.7 [0.43, 0.84]	-0.01 [-0.07, 0.07]	0.68 [0.44, 0.85]	-0.01 [-0.05, 0.09]
Full Blood Count Tests						
Hemoglobin	0.68 [0.57, 0.78]	0 [-0.03, 0.05]	0.68 [0.57, 0.78]	0 [-0.03, 0.06]	0.69 [0.59, 0.76]	0 [-0.02, 0.03]
Platelet Count	0.7 [0.58, 0.78]	-0.02 [-0.06, 0.06]	0.69 [0.58, 0.78]	-0.02 [-0.05, 0.06]	0.69 [0.6, 0.78]	-0.01 [-0.04, 0.03]
Lymphocytes	0.7 [0.56, 0.79]	-0.02 [-0.08, 0.05]	0.7 [0.55, 0.79]	-0.02 [-0.08, 0.06]	0.7 [0.59, 0.79]	-0.01 [-0.04, 0.03]
Neutrophils	0.67 [0.55, 0.77]	-0.01 [-0.06, 0.08]	0.67 [0.55, 0.76]	-0.01 [-0.07, 0.08]	0.67 [0.54, 0.76]	-0.01 [-0.05, 0.05]
Neutrophil - Lymphocyte Ratio	0.72 [0.6, 0.8]	-0.03 [-0.09, 0.02]	0.72 [0.61, 0.8]	-0.04 [-0.09, 0.02]	0.71 [0.62, 0.79]	-0.03 [-0.06, 0.02]
White Cell Count	0.68 [0.58, 0.77]	0.01 [-0.03, 0.05]	0.68 [0.57, 0.77]	0 [-0.03, 0.05]	0.68 [0.6, 0.77]	-0.01 [-0.03, 0.04]
Urea & Electrolytes Tests						
C-Reactive Protein	0.7 [0.55, 0.78]	-0.03 [-0.08, 0.01]	0.7 [0.55, 0.78]	-0.03 [-0.07, 0.02]	0.7 [0.57, 0.77]	-0.03 [-0.08, 0.01]
Estimated Glomerular Filtration Rate	0.69 [0.47, 0.8]	-0.01 [-0.04, 0.04]	0.68 [0.47, 0.8]	-0.01 [-0.04, 0.04]	0.69 [0.47, 0.8]	0 [-0.02, 0.04]
Urea	0.74 [0.62, 0.82]	-0.07 [-0.18, 0.03]	0.74 [0.63, 0.82]	-0.07 [-0.19, 0.05]	0.74 [0.62, 0.83]	-0.07 [-0.17, 0.03]
Investigatory Tests						
Brain / B-type natriuretic peptide	0.8 [0.5, 1]	0 [-0.2, 0.2]	0.8 [0.47, 1]	0 [-0.2, 0.12]	0.77 [0.47, 1]	0 [-0.2, 0.13]
D-Dimer	0.7 [0.37, 0.94]	0 [-0.21, 0.19]	0.71 [0.38, 0.94]	0 [-0.21, 0.26]	0.7 [0.42, 1]	0 [-0.12, 0.08]
Ferritin	0.63 [0.42, 0.82]	0 [-0.18, 0.25]	0.64 [0.42, 0.81]	0 [-0.17, 0.32]	0.61 [0.4, 0.82]	0 [-0.17, 0.23]
Fibrinogen*	0.8 [0.33, 1]	0 [-0.33, 0.33]	0.8 [0.33, 1]	0 [-0.33, 0.33]	0.8 [0.33, 1]	0 [-0.33, 0.33]
Glycated haemoglobin*	NA	NA	NA	NA	NA	NA
Lactate dehydrogenase*	0.86 [0.5, 1]	0 [-0.56, 0.33]	0.88 [0.5, 1]	0 [-0.44, 0.33]	0.88 [0.5, 1]	0 [-0.33, 0.33]
Procalcitonin*	NA	NA	NA	NA	NA	NA
Triglycerides*	NA	NA	NA	NA	NA	NA
Troponin-T	0.67 [0.42, 0.8]	-0.02 [-0.14, 0.26]	0.67 [0.42, 0.81]	-0.02 [-0.16, 0.18]	0.68 [0.43, 0.83]	-0.03 [-0.19, 0.11]

* Biomarkers not included in subsequent models due to small sample size, and recorded only in ICU (PCT)

Regressions were fit using all associated dummy variables for a given biomarker (e.g. mild, moderate, severe) and using only complete cases of training data (n=590), i.e. not using a variable for 'Test not taken'. 95% inter-quantile ranges calculated via 5-fold cross-validation with 20 repeats (100 models total). Categorical variables use a reading of 'Normal' as a reference in the fitted model, except 'Male' used as the reference category for gender

of the models is varied on the internal training data, with the GLM with LASSO regularization and Bayesian and projective prediction models having the best performance. However, the flat and horseshoe Bayesian models appear to overestimate the presence of severe events as indicated by the calibration-in-the-large values. External validation calibration is worse across models with most underestimating the presence of severe events. While the 4-biomarker projective prediction model has good AUC performance the external calibration slope is quite

Table 4 Internal and external cross-validated performance of models trained using valid biomarker data

Model	AUC [2.5%, 97.5%]	AUC Difference [2.5%, 97.5%]	Specificity at 90% Sensitivity [2.5%, 97.5%]	Specificity at 95% Sensitivity [2.5%, 97.5%]	Calibration Slope [2.5%, 97.5%]	Calibration Intercept [2.5%, 97.5%]	Calibration in the Large [2.5%, 97.5%]
Internal Validation							
Standard Logistic GLM	0.76 [0.63, 0.84]	-0.04 [-0.15, 0.01]	0.37 [0.02, 0.64]	0.15 [0, 0.52]	0.59 [0.21, 0.93]	0.11 [-0.04, 0.29]	0 [-0.06, 0.06]
GLM with LASSO regularisation	0.8 [0.5, 0.86]	-0.00 [-0.28, 0.02]	0.51 [0, 0.71]	0.37 [0, 0.61]	0.89 [0.26, 1.35]	-0.01 [-0.18, 0.28]	0 [-0.03, 0.22]
Bayesian GLM (Flat Prior)	0.76 [0.64, 0.84]	-0.04 [-0.14, 0.02]	0.39 [0, 0.6]	0.16 [0, 0.52]	0.51 [0.02, 0.9]	0.11 [-0.03, 0.34]	0.06 [0.02, 0.1]
Bayesian GLM (Horse Shoe Prior)	0.8 [0.74, 0.86]	*Reference*	0.55 [0.31, 0.69]	0.42 [0.2, 0.62]	0.88 [0.06, 1.35]	0 [-0.17, 0.31]	0.13 [0.1, 0.14]
LASSO inspired GLM (16 biomarkers)	0.82 [0.75, 0.89]	0.02 [-0.03, 0.07]	0.57 [0.3, 0.73]	0.39 [0.02, 0.68]	0.76 [0.41, 1.13]	0.05 [-0.12, 0.22]	0 [-0.04, 0.05]
Projective Prediction (26 Biomarkers)	0.8 [0.74, 0.88]	0.00 [-0.01, 0.08]	0.56 [0.32, 0.71]	0.42 [0.19, 0.64]	0.88 [0.2, 1.33]	-0.01 [-0.15, 0.27]	0 [-0.03, 0.03]
Projective Prediction (4 Biomarkers)	0.8 [0.79, 0.8]	-0.00 [-0.06, 0.06]	0.5 [0.44, 0.55]	0.37 [0.3, 0.42]	0.98 [0.51, 1.13]	-0.03 [-0.1, 0.14]	0 [-0.01, 0.01]
External Validation							
Standard Logistic GLM	0.67 [0.63, 0.71]	-0.03 [-0.08, 0.01]	0.23 [0.08, 0.35]	0.1 [0.02, 0.21]	0.42 [0.27, 0.55]	0.21 [0.16, 0.28]	-0.03 [-0.07, -0.01]
GLM with LASSO regularisation	0.7 [0.5, 0.72]	-0.01 [-0.21, 0.01]	0.3 [0, 0.4]	0.19 [0, 0.24]	0.88 [0.65, 1.07]	0.07 [0.01, 0.37]	-0.08 [-0.1, 0.13]
Bayesian GLM (Flat Prior)	0.66 [0.62, 0.71]	-0.04 [-0.08, 0.00]	0.24 [0.09, 0.34]	0.1 [0, 0.2]	0.34 [0.16, 0.51]	0.24 [0.16, 0.32]	0.02 [-0.01, 0.04]
Bayesian GLM (Horse Shoe Prior)	0.71 [0.69, 0.72]	*Reference*	0.34 [0.27, 0.41]	0.21 [0.16, 0.28]	0.84 [0.34, 1.04]	0.07 [0.01, 0.25]	0.05 [0.05, 0.06]
LASSO inspired GLM (16 biomarkers)	0.68 [0.67, 0.69]	-0.02 [-0.04, -0.01]	0.27 [0.21, 0.33]	0.11 [0.08, 0.14]	0.6 [0.45, 0.73]	0.17 [0.12, 0.22]	-0.09 [-0.11, -0.07]
Projective Prediction (26 Biomarkers)	0.7 [0.68, 0.72]	-0.00 [-0.02, 0.01]	0.33 [0.25, 0.41]	0.21 [0.13, 0.27]	0.84 [0.27, 1.03]	0.07 [0.01, 0.27]	-0.07 [-0.1, -0.05]
Projective Prediction (4 Biomarkers)	0.71 [0.7, 0.72]	0.01 [-0.01, 0.02]	0.31 [0.21, 0.39]	0.16 [0.1, 0.24]	0.19 [-0.07, 0.49]	0.32 [0.21, 0.41]	-0.08 [-0.09, -0.07]

95% inter-quantile ranges are presented for each estimate. Specificity is obtained by evaluating at a set sensitivity of either 90% or 95%. All reduced variable models include age, and a stated number of biomarkers. The reduced variable standard GLM uses age and 16 biomarkers that had non-zero coefficients on >=50% LASSO Cross-validation trials. If at least one categorical level for a particular biomarker (e.g. severe) met this requirement, all levels for that biomarker were included in the model. The 4 biomarker projective prediction model uses all categorical levels for Urea, PT, CRP, and NLR. Pairwise AUC difference is presented in comparison to the Bayesian (Horse shoe prior) model

low (0.19) which appears to be due to poor estimates for patients with high probabilities of a severe outcome, see Supplementary Materials A9.

Reduced variable models

The models detailed above are moderately good predictors of severe COVID-19 outcomes, but for clinicians with limited time and resources, reduced models can balance predictive performance with ease of clinical use by using only the most informative biomarkers. To address this, we use two variable selection approaches, LASSO and projective prediction, that allow the creation of reduced models with fewer biomarkers but similar performance to the larger models.

LASSO models

After performing 5-fold 20 repeat cross-validation we examined the frequency of how often a particular biomarker has a coefficient greater than zero and count across cross-validation trials. Supplementary Figure A10 shows the frequency of variables having a coefficient great than zero in the cross-validated LASSO analysis. If we select variables that appear at least 50% of the time, our reduced model would include: Age, BE (abnormal), CRP (abnormal), eGFT (abnormal), HB (severe), PLT (mild, moderate), Lymphocytes (Severe), Neutrophils (Mild, Severe), NLR (Severe), APTT (mild, moderate), 0xygen (abnormal), PT (abnormal), blood pH (abnormal), Urea (abnormal), and positive viral, respiratory, and blood culture co-infections.

For the LASSO inspired reduced variable standard GLM, this resulted in a model using the 16 biomarkers above and age for all categorical levels, and was evaluated via both cross-validation and as fit to all available training data. This model had performance similar to the models using all valid biomarker data, with a median external validation AUC of 0.68 [0.67, 0.69], see Table 4.

Note, 'Test Not Taken' was a significant predictor for some biomarkers on over 50% of cross-validation trials (see Supplementary Figure A10). The potential significance of missing data is complex and is addressed in the Discussion section. Due to this confounding, biomarkers whose top predictive contribution was from 'Test Not Taken' were excluded from both LASSO reduced variable models and projective prediction models described below.

Projective prediction models

When all biomarkers are considered, projective prediction ranks all variables in descending order of contribution to AUC performance. We considered the top 20 including: Urea (abnormal), Age, PT (abnormal), CRP (abnormal), NLR (Severe), APPT (moderate), PLT (mild, moderate), Neutrophils (mild, severe), Lymphocytes (severe), blood co-infection, hemoglobin (severe), blood pH (abnormal). Thus age and 11 biomarkers were candidates for a reduced model. Several predictors of 'Test Not Taken' were in the AUC ranking. However, as mentioned above, these biomarkers are set aside due to this confound. Supplementary Figure A11 shows the projective prediction ranking the AUC contribution. A model using a projection incorporating all biomarker and demographic data is equivalent to the standard Bayesian GLM we evaluated in the prior section, see Table 4.

Reduced variable projections were evaluated by manual inspection of AUC performance among groups of

models using the top biomarkers. Guided by the projective prediction ranking, we ran a model using only the top biomarker, using only the top two, the top three, and so on. As described above, we omit biomarkers with significant contributions from 'Test Not Taken' and include all categorical levels for a given biomarker as long as one level is highly ranked. Ultimately, we found a four biomarker projective prediction model using age and including urea, prothrombin time, neutrophil-lymphocyte ratios, and C-reactive protein had similar performance to larger models with a median internal validation AUC of 0.8 [0.79, 0.8], and external validation AUC of 0.71 [0.7, 0.72], as shown in Table 4. Odds ratios for the full Bayesian model and the reduced 4-biomarker model can be found in Supplementary Materials A12.

The 11 coefficients and intercept present can be substituted into a standard logistic equation. The calibration of the model is reasonably good on the training data but has poor calibration on the external validation set, see Supplementary Figure A13.

Discussion

Summary

Building prediction models using real world clinical data offers many challenges. There are numerous decision points required to curate data and many choices that require domain expertise. We use a COVID-19 dataset with novel biomarker data to illustrate many of these challenges. Furthermore, if models are to be used clinically they must be feasible given the many resource constraints clinicians face. In principle, a model like ours (with a larger training set, testing, and translation into a clinical score) could have been used to guide clinicians on how to triage patients and direct prophylactic measures (though these were minimal at the time) and help anticipate which patients would be more at risk. While our model is not of use at this stage with COVID-19, our methodology would generalise to other infectious diseases.

We demonstrate methods for Bayesian variable selection in logistic regression using projective prediction and compare to a LASSO approach. While Bayesian models and projective prediction are more computationally intensive than standard approaches, they offer small but consistent AUC gains. A Bayesian approach also provides unbiased coefficients compared to LASSO, and projective prediction provides a systematic method to evaluate the contribution of model variables by AUC contribution and guide variable selection. Below we detail many of the methodological challenges faced.

Challenges of complex medical data

Data curation is challenging as clinical data are heterogeneous in multiple ways. Biomarkers are recorded for different reasons, e.g. routine upon admission, investigatory tests, or tests primarily or exclusively taken in ICU. Further, some biomarkers are typically recorded together (but not always) as part of a test suite, including: Urea and electrolytes, full blood count, COVID-19 and co-infection swab test, blood clotting, and blood gas tests (arterial or venous). The schedule when these markers are recorded varies by patient and clinical decision, leading to records being present in highly varying amounts, e.g. only 3% up to 100% of patients depending on the particular biomarker, see Supplementary Materials A1.

Modelling choices

When constructing and evaluating models, there are many choice points that should be explicitly highlighted with justification, be it based on convenience, computational complexity, clinical advice, or a heuristic. We regularly consulted our clinical partners for choices on the transformation of variables, the time window to consider, why data was missing and patient inclusion criteria. Complex data sets can be modelled with a variety of approaches, as described below we considered a number of time windows, ways to aggregate multi-day data, and data imputation procedures before forming a consensus with the presented models. Our approach emphasizes explainable risk factors, predictive performance and highlights the benefits of the Bayesian variable selection technique projective prediction for practical clinical use. However, non-linear approaches such as decision forests, boosting, or neural networks are all valid options if these features are not prioritized.

Missing data

Missingness, in the context of this study and in healthcare data more generally, can sometimes be informative and missing not at random, with the presence or absence of a test correlated with the its measurement or the study outcome. Imputation of missing data relies on key statistical assumptions that imputed variables are missing at random (MAR) or missing completely at random. Conversations with our clinical co-authors established some routinely collected biomarkers might be inferred to be MAR. However, the routines identified were specific to a small a subset of our cohort and not likely to extrapolate. Clinicians advised that tests not being taken are almost always a clinical decision and therefor not random, as such we ultimately erred to be conservative and avoid all imputation, and instead include the presence/absence of missing values as a covariate itself [24, 25]. As such, in

the current study we chose to use placeholders for 'Test not taken' if there was no recorded value for a particular biomarker within the evaluated 3-day window after the key date.

This approach allows the possibility that a 'Test Not Taken' may be a significant predictor. This has many potential meanings, as it may convey that when a patient is doing well and unlikely to experience a severe outcome, clinicians are unlikely to request some biomarker tests. Alternatively, if a patient is in palliative care and has a poor prognosis, a clinician may consider further testing unnecessary. As such, the likelihood of a test being administered may follow an inverted-U function as patients to healthy or too ill may not have tests administered. Furthermore, as our data was collected early in the pandemic, there may be other underlying clinical decisions or resource limitations that drove why some tests were taken but not others. Lastly, because we only consider results from within the first 3 days after a patients date, it may be that some tests were simply taken later in a patient's stay due to operational constraints, and hence may be more predictive as they were taken closer to the outcome. When these instances occurred, we were conservative and excluded biomarkers with 'Test Not Taken' as the most informative category from our reduced variable models.

Data transforms - time windows

In the early days of the COVID-19 pandemic clinicians desired a way to triage patients near admission to help manage resources. If a good prediction on patient outcome could be made on or near the time of admission, this could greatly help divert resources to the correct patients. However, not all tests are administered on admission. To balance inclusion of test data not available on the day of admission and the need for clinical decisions to be guided soon after admission, we chose to consider the first value recorded for each biomarkers within three days after their 'key date', i.e. date of admission if already COVID-19 positive, or if already in hospital, the date of testing COVID-19 positive. However, given the richness of the time series data collected, further research into models that leverage this extra information is needed.

Focusing on early detection reflects the intent for the model to improve early stage clinical decision making when potential treatments or changes in care may be introduced. This focus on the first reading in a 3-day interval loses information, but greatly simplifies the modelling approach. Note, this choice is not without risk of reducing statistical power, increasing the risk of false positives, and underestimation of the extent of variation in biomarker readings and outcomes between groups [26]. It is likely that representing biomarker data as time-series (assuming regular measures across patients) would add considerable information for continuous monitoring.

Data transforms - continuous vs. categorical

A key modelling decision must be made on whether to use continuous data or transformed categorical data. Clinicians often use biomarker thresholds to provide semantic categories (e.g. normal, mild, moderate, severe) which sometimes use non-linear or discontinuous mappings that require special care if using continuous data. While clinical thresholds are likely established with evidence, it may be the case that thresholds for one use may not apply to a novel use. This led [12, 27] to use machine learning approaches to build categorisation models on continuous biomarker data dependent on the training data at hand. However, using machine learning to establish categorisation thresholds on our biomarker data is difficult with a small training data set and the heterogeneity of biomarker recordings. If missing data imputation is performed, it raises another decision point on whether to impute the continuous or the transformed categorical data.

Another important factor to recognise is that some biomarkers lack a linear relationship between a reading and a semantic category. Biomarkers can have a lower and upper bound for what is considered normal, and both below and above this range reflects clinically meaningful yet sometimes separate abnormalities. The modelling needs to factor in non-linearity when persevering continuous data or trying to map to a categorical space. In our position, categorical transformation had an advantage, as they allowed us to collaborate with ICU consultants while using pre-established clinically acceptable ranges to define our categorisation, see Table 1. Categorization is worth critical consideration in model planning and potentially worth revisiting. For example, with eGFR we simply consider kidney function as normal or abnormal, but test results can be put into more fine-grained categories to label the severity of kidney failure.

Training and external validation data selection

There are multiple ways that our data set could be split between training and external validation sets, e.g. randomly sampling 1/3 of the data to hold out as an external validation set. Random selection of training data should in principle generate data more representative of the external validation set left out. However, hospitals may have differing practices and non-stratified randomization may inflate performance at the cost of real world generalisation. We chose to separate our training and external validation datasets by hospital to provide a stronger test of generalisation that should mimic generalisation to

novel hospitals completely outside the original training data .

Model performance evaluation and dissemination

There are a variety of ways statistical model performance can be evaluated. Here we have chose here to emphasize cross-validated estimates of AUC, sensitivity, and specificity. Inter-quartile intervals over these measures reveal that the variety of models perform in similar ways. With a larger data set, trade-offs may become more apparent. Model calibration on the external validation set is a clear weak point. While the models have a reasonable calibration for training data, generalization performance is weak and suggestive of the lack of sufficient data.

Comparison to contemporary models

We found several biomarkers previously highlighted by other groups to have significant predictive power, including: Urea, Neutrophil-Lymphocyte Ratio (NLR), Lymphoctyes, PT, eGFR, and CRP. Our highly reduced 3-biomarker model (plus age) uses Urea (highlighted by all prior models), NLR (highlighted by [13, 27, 28]), and PT (highlighted by [27, 29]). These biomarkers highlight aspects of hypovolaemia (UREA), inflammation (NLR and CRP), and blood clotting factors (PT) that are consistently altered in patients with severe outcomes. A direct comparison with other models is not possible due to differing variables, but our external validation performance (Full model AUC: 0.7, 3-biomarker model AUC: 0.67) suffers compared to Knight et al (AUC: 0.77) and is similar to Carr et al (AUC: 0.69 to 0.79 dependent on the training dataset). While our current model is not state of the art, with a larger more diverse dataset, our methods should achieve such results and allow possible inclusion of some biomarkers not included in the present model, as well physiological bedside measures captured by Knight et al. and Carr et al. but not present in our own.

Advantages of Bayesian modelling

While the predictive performance across models presented here is generally similar within 95% bounds, the Bayesian horseshoe model has slightly better median AUC difference cross-validated predictive performance. Reasons for researchers to favor Bayesian approaches should include that coefficients estimated via Bayes should on average deliver better predictive performance than standard GLM [30]. Additionally, if a sparse model is needed, a horseshoe prior can provide advantages similar to LASSO without biased coefficient estimates enabling joint probabilistic modelling of prediction and risk factor inference. Computationally, Bayesian techniques can be slow due the Hamiltonian Monte Carlo used to sample the coefficient space. If one is interested in variable selection, projective prediction offers the ability to take a single Bayesian model fit, run a variable selection algorithm to rank variable contributions, and then arbitrarily create sub-model projections with any number of original variables. While the initial model fit and variable selection are computationally intensive, sub-model projections are fast to create and performance test. Bayesian logistic regression with variable selection has the flexibility of providing both conventional risk factor analysis and prediction, but approaches like deep learning [31] or ensemble methods [32] can offer superior prediction performance due to their non-linear nature. However, there are trade-offs, deep learning works best with large datasets (unlike ours) and does not have intuitive regression coefficients for explainability. Ensemble methods (e.g., gradient-boosted decision trees) can achieve high performance with smaller datasets but also with some sacrifice in explainability. Further, neither of these approaches have a statistically rigorous variable selection method similar to projective prediction, though models can be augmented with regularization terms to encourage sparsity. However, tools like SHAP [33] are becoming more mature and can offer a model agnostic way to view contribution of both variables and samples to model performance, and is well worth exploring. Ultimately, we favored the clinical explainability offered by logistic regression and ease of use with the ProjPred package for variable selection, even if it does sacrifice some performance compared to non-linear techniques. We encourage researchers to try a variety of models depending on requirements for balancing data set size, performance, and explainability.

Conclusion

Limitations: This is a retrospective cohort study in Southwest England where case numbers have varied widely, and were below national incidence rates during the first wave. This results in less precise parameter estimates for prediction models (less power/smaller sample size) and likely reduced generalizability of the model to other settings. The timing of biomarker collection was highly varied both within and between patients, with many types of readings missing.

Strengths: The primary strength of our study is the granularity of serial laboratory data available linked to clinical outcomes. This study was performed during the first wave where there was the original Wuhan strain circulating amongst the unvaccinated naïve population without any specific immunomodulating therapies such as steroids or antiviral agents, reflecting the "true" homeostasis derangements at a population level.

In particular, this study describes the variety of challenges present in complex medical data sets and how

researchers need to balance the aim of statistically sound practices with the pragmatics and limitations of observational datasets like these. We highlight the benefits of recent Bayesian methodology for variable selection. Our study reiterates the predictive value of previously identified biomarkers for COVID-19 severity assessment (e.g. age, urea, prothrombin time, c-Reactive protein, and neutrophil-lymphocyte ratio). Both the full and reduced variable models have moderately good training performance, but improved external validation is needed for all models to be clinically viable. The methods presented here should generalize well to a larger dataset and serve as a guide.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12911-025-02955-3.

Supplementary Material 1.

Acknowledgements
This research was supported by the National Institute for Health and Care Research (NIHR) Applied Research Collaboration West (NIHR ARC West). The views expressed in this article are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care.

Authors' contributions
B.S. co-wrote and revised the manuscript, conducted data curation and all statistical analyses, and prepared all figures. E.B. reviewed the manuscript, conducted data curation, and co-prepared Figs. 2, 3, and A1-A7. L.M. co-wrote the manuscript, conducted initial data curation, analysis, and visualization. P.W. co-designed the study and was involved in data collection, revised the manuscript, provided clinical guidance on curation and statistical analyses, and revised the manuscript. R.B. co-designed the study and was involved in data curation. M.T. co-designed the study and was involved in data collection. S.G. co-designed the study and was involved in data collection. C.H. co-designed the study and was involved in data collection. A.W. co-designed the study and was involved in data collection. J.A.C. co-designed the study and was involved in data collection. C.W. co-designed the study and was involved in data curation. K.T. co-designed the study. A.W.D. reviewed the manuscript, consulted on data curation and all statistical analyses, and figure preparation P.W. co-designed the study and was involved in data collection, reviewed the manuscript, provided clinical guidance on curation and statistical analyses, and revised the manuscript. L.G. reviewed the manuscript and helped in co-preparation of tables 1-4, A8, A12.

Funding
This work was supported by Health Data Research UK via the Better Care Partnership Southwest (HDR CF0129) awarded to Drs Turner and Dowsey; Medical Research Council Research Grant MR/T005408/1 awarded to Dr Williams; and the Elizabeth Blackwell Institute for Health Research, University of Bristol and the Wellcome Trust Institutional Strategic Support Fund (204813/Z/16/Z) awarded to Drs Turner and Dowsey. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data availability
Due to strict data governance of medical records, the data in this study cannot be directly shared by the authors. Interested parties should contact the corresponding authors and bnssg.research@nhs.net to arrange a data sharing discussion with the Bristol, North Somerset, and South Gloucestershire Integrated Care Board who steward the data, for more information please consult the website: BNSSG Research and Evidence.

Declarations

Ethics approval and consent to participate
The study underwent a rigorous ethical and regulatory approval process, following an Integrated Research Application System application [IRAS project ID: 283439], a favourable written authorization was gained from NHS Research Ethics Service, Wales Research Ethics Committee 7, c/o Public Health Wales, Building 1, Jobswell Road, St David's Park, SA31 3HB on 11/09/2020. Our research complies with the declaration of Helsinki with anonymized data and ethical review, as explained below informed consent for data sharing was waived due to overriding public interest. See: https://www.hra.nhs.uk/about-us/committees-and-services/res-and-recs/
The requirement for informed consent was waived by NHS Research Ethics Service, Wales Research Ethics Committee 7 (see above), given overriding public interest in the research. Furthermore, during project development prior to ethics review, a public and patient involvement meeting conducted at North Bristol NHS Trust by author MH received similar support. North Bristol NHS Trust and University Hospitals Bristol and Weston NHS Foundation Trust signed data sharing agreements confirming this waiver. See: https://www.england.nhs.uk/publication/information-sharing-policy All data were fully anonymized before they were transferred to the research team for analysis.

Consent for publication
NHS Research Ethics Service, Wales Research Ethics Committee 7 and North Bristol NHS Trust and University Hospitals Bristol and Weston NHS Foundation Trust (see above) waived the requirement for informed patient consent for publication.

Competing interests
Dr Catherine Hyams is the Principal Investigator of the AvonCAP study, which is a University of Bristol sponsored study funded by Pfizer. This does not alter our adherence to BMC policies on sharing data and materials.

Author details
[1]Department of Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK. [2]Department of Engineering Mathematics, Faculty of Engineering, University of Bristol, Bristol, UK. [3]Department of Microbiology, University Hospitals Bristol and Weston NHS Foundation Trust, Bristol, UK. [4]Intensive Care Unit, University Hospitals Bristol and Weston NHS Foundation Trust, Bristol, UK. [5]Severn Infection Sciences, Southmead Hospital, North Bristol NHS Trust, Bristol, UK. [6]Department of Clinical Heamatology, Southmead Hospital, North Bristol NHS Trust, Bristol, UK. [7]Intensive Care Unit, Southmead Hospital, North Bristol NHS Trust, Bristol, UK. [8]Jean Golding Institute, University of Bristol, Bristol, UK.

Received: 18 December 2023 Accepted: 26 February 2025
Published online: 10 March 2025

References

1. Tibshirani R. Regression Shrinkage and Selection Via the Lasso. J R Stat Soc Ser B (Methodol). 1996;58(1):267–88. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.
2. Piironen J, Paasiniemi M, Vehtari A. Projective Inference in High-Dimensional Problems: Prediction and Feature Selection. Electron J Stat. 2020;14(1):2155–97. https://doi.org/10.1214/20-EJS1711.
3. Docherty AB, Harrison EM, Green CA, Hardwick HE, Pius R, Norman L, ... Semple MG. Features of 20 133 UK patients in hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: prospective observational cohort study. BMJ. 2020;369:1–12.
4. Esakandari H, Nabi-Afjadi M, Fakkari-Afjadi J, Farahmandian N, Miresmaeili SM, Bahreini E. A comprehensive review of COVID-19 characteristics. Biol Proced Online. 2020;22(1):1–10.
5. Yuki K, Fujiogi M, Koutsogiannaki S. COVID-19 pathophysiology: A review. Clin Immunol. 2020;215:108427.
6. Zaim S, Chong JH, Sankaranarayanan V, Harky A. COVID-19 and multiorgan response. Curr Probl Cardiol. 2020;45(8):100618.

7. Verity R, Okell LC, Dorigatti I, Winskill P, Whittaker C, Imai N, et al. Estimates of the severity of coronavirus disease 2019: a model-based analysis. Lancet Infect Dis. 2020;20(6):669–77.
8. Wynants L, Calster BV, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction Models for Diagnosis and Prognosis of Covid-19: Systematic Review and Critical Appraisal. BMJ. 2020;369:m1328. https://doi.org/10.1136/bmj.m1328.
9. Collins GS, Dhiman P, Ma J, Schlussel MM, Archer L, Van Calster B, ... Riley RD. Evaluation of clinical prediction models (part 1): from development to external validation. BMJ. 2024;384:1–11.
10. Riley RD, Archer L, Snell KI, Ensor J, Dhiman P, Martin GP, et al. Evaluation of clinical prediction models (part 2): how to undertake an external validation study. BMJ. 2024;384:1–12.
11. Riley RD, Snell KI, Archer L, Ensor J, Debray TP, Van Calster B, et al. Evaluation of clinical prediction models (part 3): calculating the sample size required for an external validation study. BMJ. 2024;384:1–9.
12. Knight SR, Ho A, Pius R, Buchan I, Carson G, Drake TM, et al. Risk stratification of patients admitted to hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: development and validation of the 4C Mortality Score. BMJ. 2020;370:1–13.
13. Carr E, Bendayan R, Bean D, Stammers M, Wang W, Zhang H, et al. Evaluation and Improvement of the National Early Warning Score (NEWS2) for COVID-19: A Multi-Hospital Study. BMC Med. 2021;19(1):23. https://doi.org/10.1186/s12916-020-01893-3.
14. van der Schaar Lab M. Time series in healthcare: challenges and solutions. 2022. https://www.vanderschaar-lab.com/time-series-in-healthcare/. Accessed 12 Dec 2023.
15. Provan D, Krentz AJ, Longmore JM, editors. Oxford Handbook of Clinical and Laboratory Investigation. Oxford Medical Publications. Oxford and New York: Oxford University Press; 2002.
16. Carvalho CM, Polson NG, Scott JG. Handling sparsity via the horseshoe. Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS), Clearwater Beach, Florida, USA. Volume 5 of Journal of Machine Learning Research: W&CP 5; 2009. pp. 73–80.
17. Shipe ME, Deppen SA, Farjah F, Grogan EL. Developing prediction models for clinical use using logistic regression: an overview. J Thorac Dis. 2019;11(Suppl 4):S574.
18. Chowdhury MZI, Turin TC. Variable selection strategies and its importance in clinical prediction modelling. Fam Med Community Health. 2020;8(1):e000262. 1–7.
19. Heinze G, Wallisch C, Dunkler D. Variable selection-a review and recommendations for the practicing statistician. Biom J. 2018;60(3):431–49.
20. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Ser B (Methodol). 1995;57(1):289–300.
21. Taylor J, Tibshirani RJ. Statistical Learning and Selective Inference. Proc Natl Acad Sci. 2015;112(25):7629–34. https://doi.org/10.1073/pnas.1507583112.
22. Tibshirani R, Tibshirani R, Taylor J, Loftus J, Reid S. selectiveInference: Tools for Post-Selection Inference. Github. 2019. https://github.com/selective-inference/R-software. Accessed 12 Dec 2023.
23. 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.
24. Groenwold RH. Informative missingness in electronic health record systems: the curse of knowing. Diagn Prognostic Res. 2020;4(1):1–6.
25. Van Buuren S. Flexible imputation of missing data. Boca Raton: CRC Press; 2018.
26. Altman DG, Royston P. The Cost of Dichotomising Continuous Variables. BMJ Brit Med J. 2006;332(7549):1080.
27. Zhou J, Lee S, Wang X, Li Y, Wu WKK, Liu T, et al. Development of a multivariable prediction model for severe COVID-19 disease: a population-based study from Hong Kong. NPJ Digit Med. 2021;4(1):1–9.
28. Liu J, Liu Y, Xiang P, Pu L, Xiong H, Li C, et al. Neutrophil-to-Lymphocyte Ratio Predicts Severe Illness Patients with 2019 Novel Coronavirus in the Early Stage. medRxiv. 2020;2020.02.10.20021584. https://doi.org/10.1101/2020.02.10.20021584.
29. Tekle E, Gelaw Y, Dagnew M, Gelaw A, Negash M, Kassa E, et al. Risk stratification and prognostic value of prothrombin time and activated partial thromboplastin time among COVID-19 patients. PLoS ONE. 2022;17(8):e0272216.
30. Bolstad WM, Curran JM. Introduction to Bayesian statistics. Hoboken: Wiley; 2016.
31. Bailly A, Blanc C, Francis É, Guillotin T, Jamal F, Wakim B, et al. Effects of dataset size and interactions on the prediction performance of logistic regression and deep learning models. Comput Methods Prog Biomed. 2022;213:106504.
32. Zhang Z, Chen L, Xu P, Hong Y. Predictive analytics with ensemble modeling in laparoscopic surgery: a technical note. Laparosc Endoscopic Robot Surg. 2022;5(1):25–34.
33. Lundberg SM, Lee SI. A Unified Approach to Interpreting Model Predictions. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al., editors. Advances in Neural Information Processing Systems 30. Curran Associates, Inc.; 2017. pp. 4765–74. http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf. Accessed 3 Mar 2025.

Publisher's Note