Machine learning approach to dynamic risk modeling of mortality in COVID-19: a UK Biobank cohort study

Mohammad A. Dabbah [1 (0000-0003-1950-9053)], Angus B. Reed [1 (0000-0003-2751-2535)], Adam T.C. Booth [1 (0000-0003-3319-3585)], Arrash Yassaee [1 (0000-0002-8423-9902)], Alex Despotovic [1,2 (0000-0001-8137-911X)], Benjamin Klasmer [1 (0000-0002-2510-4696)], Emily Binning [1 (0000-0001-9920-5943)], Mert Aral [1 (0000-0003-1950-9053)], David Plans* [1,3 (0000-0002-0476-3342)], Alain B. Labrique [4 (0000-0003-2502-7819)], Diwakar Mohan [4 (0000-0002-7532-366X)]

1. Huma Therapeutics Limited, London, United Kingdom
2. Faculty of Medicine, University of Belgrade, Serbia
3. University of Exeter, SITE, Exeter, United Kingdom
4. Johns Hopkins Bloomberg School Public Health, Baltimore, Maryland, United States

*** Corresponding Author:** David Plans (david.plans@huma.com)

Disclosure statement

M.A.D., A.B.R., A.T.C.B., A.Y., A.D., B.K., E.B., M.A. and D.P. are employees of Huma Therapeutics Ltd. D.M. & A.L. declare that they have no conflicts of interest to report.

Funding

This research was funded by Huma Therapeutics Ltd.

Keywords

COVID-19; SARS-CoV-2; risk factors; prediction; mortality; machine learning; clinical characteristics; TRIPOD

Acknowledgements

The authors would like to thank Davide Morelli for contributions in model development. This research has been conducted using data from UK Biobank, a major biomedical database (www.ukbiobank.ac.uk).

Author contributions

All authors have approved the final version of the manuscript submitted. All authors agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. M.A.D. conceived and designed the study, interpreted the results, developed the computation models, analysed the data, and wrote and reviewed the manuscript. A.B.R. and A.T.C.B. conceived and designed the study, interpreted the results, and wrote and reviewed the manuscript. B.K. interpreted the results, developed the computation models, analysed the data, and wrote and reviewed the manuscript. A.Y. and A.D. interpreted the results, and wrote and reviewed the manuscript. E.B. and M.A. conceived and designed the study and reviewed the manuscript. D.M., A.L., and D.P. interpreted the results and reviewed the manuscript.

Abstract

The COVID-19 pandemic has resulted in over two million deaths globally. There is an urgent need for robust, scalable monitoring tools supporting resource allocation and stratification of high-risk patients. This research aims to develop and validate prediction models, using the UK Biobank to estimate COVID-19 mortality risk in confirmed cases. We developed a random forest classification model using baseline characteristics, pre-existing conditions, symptoms, and vital signs, such that the score could dynamically assess risk of mortality with disease deterioration (AUC: 0.92). The design and feature selection of the framework lends itself to deployment in remote settings. Possible applications include supporting individual-level risk profiling and monitoring disease progression across high volumes of patients with COVID-19, especially in hospital-at-home settings.

The COVID-19 pandemic has precipitated over 100 million confirmed cases and 2.3 million deaths globally[1]. The impact of the pandemic has not been limited to healthcare systems: a ripple effect has resulted in wide-ranging economic and social disruption[2]. Interventions to reduce transmission, such as lockdowns, travel restrictions, and re-allocation of health resources, are critical to limiting the impact[3]. Although large-scale vaccination programmes have begun, many countries globally will not have widespread access to vaccines until 2023, meaning that non-pharmaceutical interventions are likely to remain indispensable national strategies for some time[4].

COVID-19 shows highly varied clinical presentation. A significant proportion (17-45%) of cases are asymptomatic and require no specific care[5,6]. Conversely, reviews of severe complications have found that up to 32% of hospitalized COVID-19 patients are admitted to ICU[7]. Between these two extremes, typical symptoms include fever, continuous cough, anosmia, and dyspnoea, which may range from requiring only self-management at home to inpatient care. Understanding which individuals are most vulnerable to severe disease, and thereby in most need of resources, is critical to limit the impact of the virus.

Decision-making at all levels requires an understanding of individuals' risk of severe disease. Various patient characteristics, comorbidities, and lifestyle factors have been linked to greater risk of death and/or severe illness following infection[8–10]. Furthermore, socioeconomic factors have also been linked as risk factors for COVID-19 mortality [11,12]. Once patients are infected with SARS-CoV-2, additional physiological parameters, such as symptoms and vital signs, can inform real-time prognostication[13]. Laboratory testing and imaging can also inform risk stratification for early, aggressive intervention, though this data is only accessible to hospital inpatients, who are likely to be already severely affected[14,15].

Robust, predictive models for acquisition and prognosis of COVID-19[16–18] and resource management[19,20] have been developed to support risk stratification and population management at-scale, offering important insights for organizational decision-making. However, the individual is currently overlooked, and granular, patient-specific risk-scoring could potentially unify decision-making at all levels. Existing individualized risk scores, however, often conflate risk of COVID-19 acquisition with risk of mortality following infection[16,17], which can limit their utility in patient management.

For prediction models to achieve impact at scale, assessment of risk factors should be inexpensive and accessible to the general population, ideally without the need for specialized testing or hospital visits. Such risk prediction tools, enabling improved patient triage, could be used to further increase the efficiency of, and confidence in, hospital-at-home solutions, which have shown promise in reducing hospital burden throughout the pandemic[21]. Risk scores in these circumstances need to be dynamic and contemporaneous, ideally incorporating symptoms and vital sign data to maximise utility to clinical and research teams. Therefore, the primary aim of this study is to develop and validate a population-based prediction model, using a large, rich dataset and a selective, clinically informed approach, which dynamically estimates the COVID-19 mortality risk in confirmed diagnoses.

Results

Summary population

There were 7,536 adults (aged 51–85 yrs, mean: 67.4, SD: 8.8) from the UK Biobank (UKB) included in the analysis, of whom 496 (6.6%) had died as a result of COVID-19. The mean age of survivors was 66.8 years (SD: 8.7), compared to 75.9 years (SD: 8.7) for those that died. The most common pre-existing conditions in patients were hypertension (37.6%), osteoarthritis (24.3%), and chronic ischemic heart disease (13.2%) (Table 1).

Characteristic		n (%) [count] All Participants	Survived	Died
	Total	7,536	7,040 (93.4)	496 (6.6)
Demographic				
Male sex		3,664	3,327 (90.8)	337 (9.2)
Age (yrs), mean (SD)		67.4 (8.8) [7,536]	66.8 (8.7) [7,040]	75.9 (5.6) [496]
Lifestyle and anthropometrics				
Body mass index, mean (SD)		28.5 (5.1) [7,471]	28.4 (5.0) [6,986]	30.0 (5.6) [485]
Waist circumference (cm), mean (SD)		92.9 (14.1) [7,485]	92.4 (13.9) [6,999]	100.3 (14.8) [489]
Hip circumference (cm), mean (SD)		104.8 (9.7) [7,482]	104.6 (9.6) [6,996]	106.6 (11.2) [489]
Body weight (kg), mean (SD)		81.2 (16.9) [7,484]	80.9 (16.7) [6,996]	86.1 (18.2) [488]
Obesity (BMI > 30)		908	802 (88.3)	106 (11.7)
Standing height (cm), mean (SD)		168.6 (9.2) [7,478]	168.6 (9.2) [6,991]	169.1 (9.3) [487]
Blood group				
Unknown		22	21 (95.5)	1 (4.5)
AA		613	571 (93.1)	42 (6.9)
AB		294	279 (94.9)	15 (5.1)
AO		2,711	2,548 (94)	163 (6)
BB		51	46 (90.2)	5 (9.8)
BO		701	664 (94.7)	37 (5.3)
OO		2,913	2,704 (92.8)	209 (7.2)
Sleep duration (hrs), mean (SD)		7.0 (1.4) [7,522]	7.0 (1.4) [7,027]	7.2 (1.7) [495]
Alcohol intake				
Unknown		22	21 (95.5)	1 (4.5)
Daily or almost daily		1,120	1,041 (92.9)	79 (7.1)
Three or four times a week		1,494	1,414 (94.6)	80 (5.4)
Once or twice a week		2,030	1,911 (94.1)	119 (5.9)
One to three times a month		853	813 (95.3)	40 (4.7)
Special occasions only		881	811 (92.1)	70 (7.9)
Never		1,135	1,029 (90.7)	106 (9.3)
Smoking status				
Unknown		22	21 (95.5)	1 (4.5)
Never		1,135	1,029 (90.7)	106 (9.3)
Previous		2,695	2,467 (91.5)	228 (8.5)
Current		4,091	3,878 (94.8)	213 (5.2)
Gait and mobility issues		729	552 (75.7)	177 (24.3)
Medication and treatment				
Allergy to antibiotics		781	713 (91.3)	68 (8.7)
Long-term use of anticoagulants		739	611 (82.7)	128 (17.3)
Radiation therapy		191	163 (85.3)	28 (14.7)
Maintenance chemotherapy		347	302 (87)	45 (13)
Chemotherapy		187	157 (84)	30 (16)
Pre-existing medical conditions				
General diseases of the circulatory system		882	740 (83.9)	142 (16.1)
Chronic ischemic heart disease		995	846 (85)	149 (15)
Atrial fibrillation		745	610 (81.9)	135 (18.1)
Hypertension		2,834	2,479 (87.5)	355 (12.5)
Hypotension		257	197 (76.7)	60 (23.3)
Stroke		574	467 (81.4)	107 (18.6)
General diseases of the respiratory system		118	100 (84.7)	18 (15.3)
Asthma		1,034	948 (91.7)	86 (8.3)
Chronic obstructive pulmonary disease		502	400 (79.7)	102 (20.3)
Interstitial lung disease		79	51 (64.6)	28 (35.4)
Respiratory failure				
less than 1 month		252	149 (59.1)	103 (40.9)
between 1 and 12 months		136	88 (64.7)	48 (35.3)
more than 12 months		116	82 (70.7)	34 (29.3)
Non-bacterial pneumonia				
less than 1 month		686	453 (66)	233 (34)
between 1 and 12 months		402	283 (70.4)	119 (29.6)
more than 12 months		470	376 (80)	94 (20)
Bacterial pneumonia				
less than 1 month		624	409 (65.5)	215 (34.5)
between 1 and 12 months		275	185 (67.3)	90 (32.7)
more than 12 months		38	32 (84.2)	6 (15.8)
General diseases of the nervous system		455	387 (85.1)	68 (14.9)

Parkinson's disease	119	87 (73.1)	32 (26.9)
MND, MS, or HD	15	13 (86.7)	2 (13.3)
Dementia	391	294 (75.2)	97 (24.8)
Haematological Cancer			
less than 12 months	69	40 (58)	29 (42)
between 12 and 60 months	66	46 (69.7)	20 (30.3)
more than 60 months	80	59 (73.8)	21 (26.3)
Non-haematological Cancer			
less than 12 months	137	113 (82.5)	24 (17.5)
between 12 and 60 months	412	381 (92.5)	31 (7.5)
more than 60 months	634	579 (91.3)	55 (8.7)
Diabetes (Type 1)	102	75 (73.5)	27 (26.5)
Diabetes (Type 2)	977	812 (83.1)	165 (16.9)
Osteoarthritis	1,831	1,655 (90.4)	176 (9.6)
Depression and anxiety disorder	1,001	891 (89)	110 (11)
Rheumatoid arthritis	228	191 (83.8)	37 (16.2)
Anemia	893	740 (82.9)	153 (17.1)
Urinary tract infection			
less than 1 month	81	62 (76.5)	19 (23.5)
between 1 and 12 months	129	100 (77.5)	29 (22.5)
more than 12 months	629	512 (81.4)	117 (18.6)
Acute kidney failure			
less than 1 month	224	135 (60.3)	89 (39.7)
between 1 and 12 months	208	139 (66.8)	69 (33.2)
more than 12 months	323	233 (72.1)	90 (27.9)
Any bacterial infection			
less than 1 month	141	92 (65.2)	49 (34.8)
between 1 and 12 months	164	117 (71.3)	47 (28.7)
more than 12 months	355	286 (80.6)	69 (19.4)
Diverticulum	1,142	1,029 (90.1)	113 (9.9)
Haemorrhoids	721	682 (94.6)	39 (5.4)
Irritable bowel syndrome	295	270 (91.5)	25 (8.5)
Gastroenteritis			
less than 1 month	133	114 (85.7)	19 (14.3)
between 1 and 12 months	121	99 (81.8)	22 (18.2)
more than 12 months	1,178	1,058 (89.8)	120 (10.2)
Symptoms			
Joint pain	823	725 (88.1)	98 (11.9)
Delirium	206	145 (70.4)	61 (29.6)
Hematemesis	379	339 (89.4)	40 (10.6)
Syncope and collapse	15	13 (86.7)	2 (13.3)
Dyspnea	193	164 (85)	29 (15)
Cough	62	52 (83.9)	10 (16.1)
Myalgia	172	155 (90.1)	17 (9.9)
Nausea and vomiting	33	24 (72.7)	9 (27.3)
Chest pain	587	530 (90.3)	57 (9.7)
Hematuria	35	30 (85.7)	5 (14.3)
Malaise and fatigue	40	33 (82.5)	7 (17.5)
Vital signs			
Diastolic blood pressure, mean (SD)	77.9 (12.2) [123]	77.2 (10.9) [104]	81.9 (17.4) [19]
Systolic blood pressure, mean (SD)	129.3 (19.2) [124]	128.2 (17.6) [104]	135.1 (25.7) [20]
Heart rate, mean (SD)	84.7 (17.5) [80]	84.0 (16.9) [71]	90.9 (22.0) [9]
Body temperature, mean (SD) *	37.5 (1.2) [41]	37.7 (1.1) [37]	36.1 (0.9) [4]
Oxygen saturation, mean (SD) *	94.7 (3.3) [20]	94.4 (3.6) [16]	95.8 (1.5) [4]
Respiratory rate, mean (SD) *	24.1 (7.4) [18]	24.8 (8.5) [11]	22.9 (5.8) [7]

Table 1. Descriptive characteristics of the UK Biobank cohort with positive COVID-19 test results. Pre-existing medical conditions included only when reported more than one week prior to COVID-19 positive test result. Symptoms and vitals included only from primary care (GP) records when reported within +/- two weeks of COVID-19 positive test result. MND = motor neurone disease; MS = multiple sclerosis; HD = Huntington's disease. * Oxygen saturation, respiratory rate, and body temperature were included in the initial analysis, however, they were removed from the model due to low data availability.

The evaluation of the model is organised into two parts: the machine learning data-driven approach; and the machine learning and clinically selective approach. From an optimisation perspective, the objective of the model is to reduce the full feature array to a minimal subgroup (**Figure 1**) while maintaining a high prediction accuracy for COVID-19 mortality.

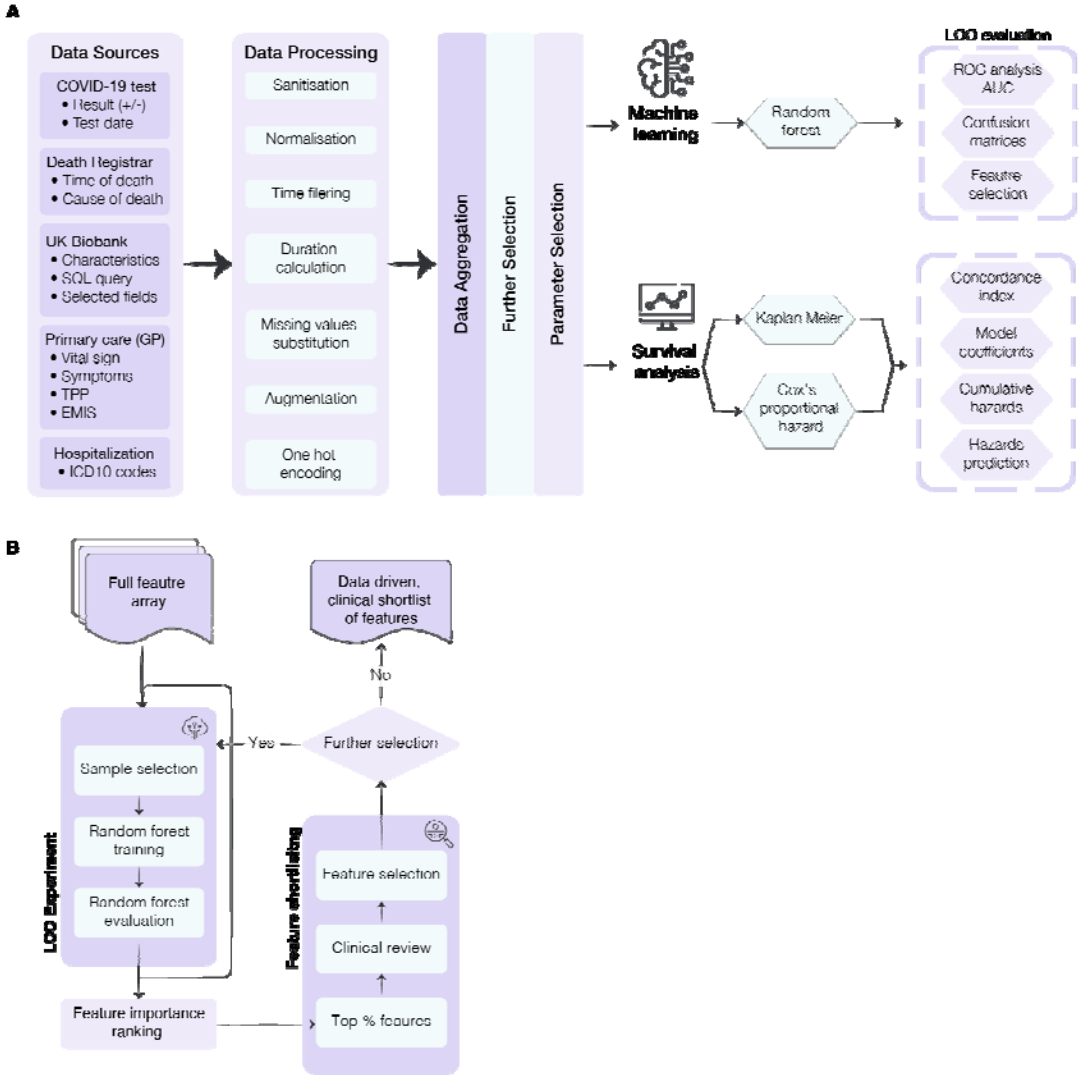

Figure 1. Workflow for model development and feature selection. A) Conceptual diagram of the data ingestion pipeline and analysis methods. To combine databases, several data pre-processing steps were carried out, including: sanitisation (eliminating redacted records and nuanced entries); normalization (scaling values to ensure fitting with a reasonable range for further processing); time filtering; duration calculation (computing the time interval between testing positive and mortality); missing value substitution (replacing missing values or records with the mean value of the UK Biobank database); augmentation (bringing all data for each subject into a single unified record); and one-hot-encoding (codifying the presence of a pre-existing condition or symptom into a binary sequence for each subject). This data ingestion process standardized the input features and attributes for all subjects in this study regardless of their unique and variable conditions, symptoms, vital signs, and records. B) Illustration of the data-driven and clinically reviewed feature refinement process. AUC = area under the curve; GP = general practice; LOO = Leave-One-Out; ROC = receiver operating characteristic.

By reducing the feature array to the 100 top-ranked features out of ~12,000 features, the performance of the Random Forest (RF) model improved. The receiver operating characteristic (ROC) curves in **Figure 2A** demonstrate the model performance after shortlisting the features to 64 characteristics. The feature selection process (**Figure 1B**) ensured the combination of data-driven insights with clinical experience. The shortlisted features included: 3 vital signs; 11 symptoms; 32 pre-existing clinical conditions; 5 medications and treatments; and 13 patient characteristics (**Table 1**).

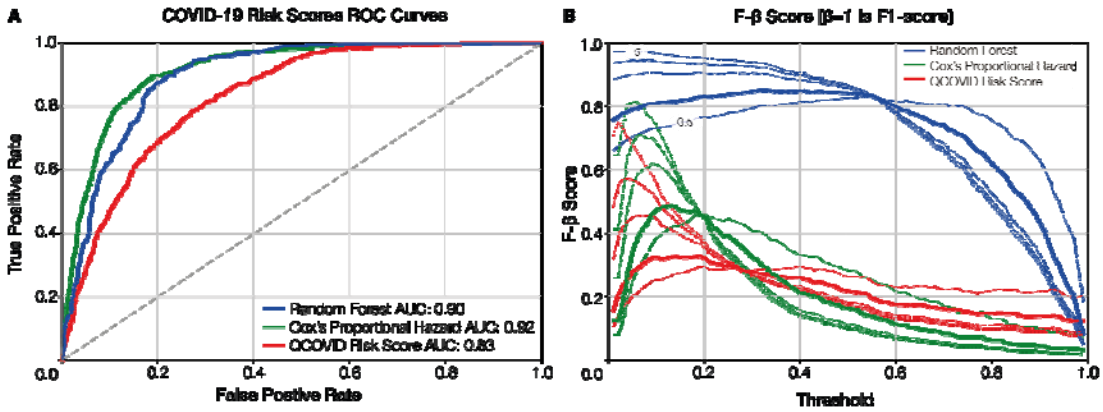

Figure 2. Model performance evaluation showing: (A) the receiver operating characteristic (ROC) curve comparison shown for our Random Forest (RF) and Cox models against QCOVID; (B) the F-β score generated at β=1 (F1-score in bold), β=[0.5, 2, 3, 5], shown in decreasing size dashed line. AUC = area under the curve.

Clinical refinement led to improved accuracy (AUC: 0.90). Furthermore, by reducing the number of features, the model overcomes the curse of dimensionality, where beforehand the full feature array was far greater in size than the available samples. While a Cox Proportional Hazard (CPH) model was trained using the final set of RF-defined variables to maximize explainability of the RF, it reached a higher AUC of 0.92. The top risk factors were age, acute kidney failure (<1 month), and waist circumference (**Figure 3**). Detailed results can be found in **Supplementary Table 3**. To test for overfitting due to prominent features and limitations in the dataset, the model was re-processed without age, which had minimal effect on model performance (AUC: 0.89, **Supplementary Figure 2**). **Figure 2A** also shows the ROC curves for both the RF and CPH [22] models against the QCOVID model [16].

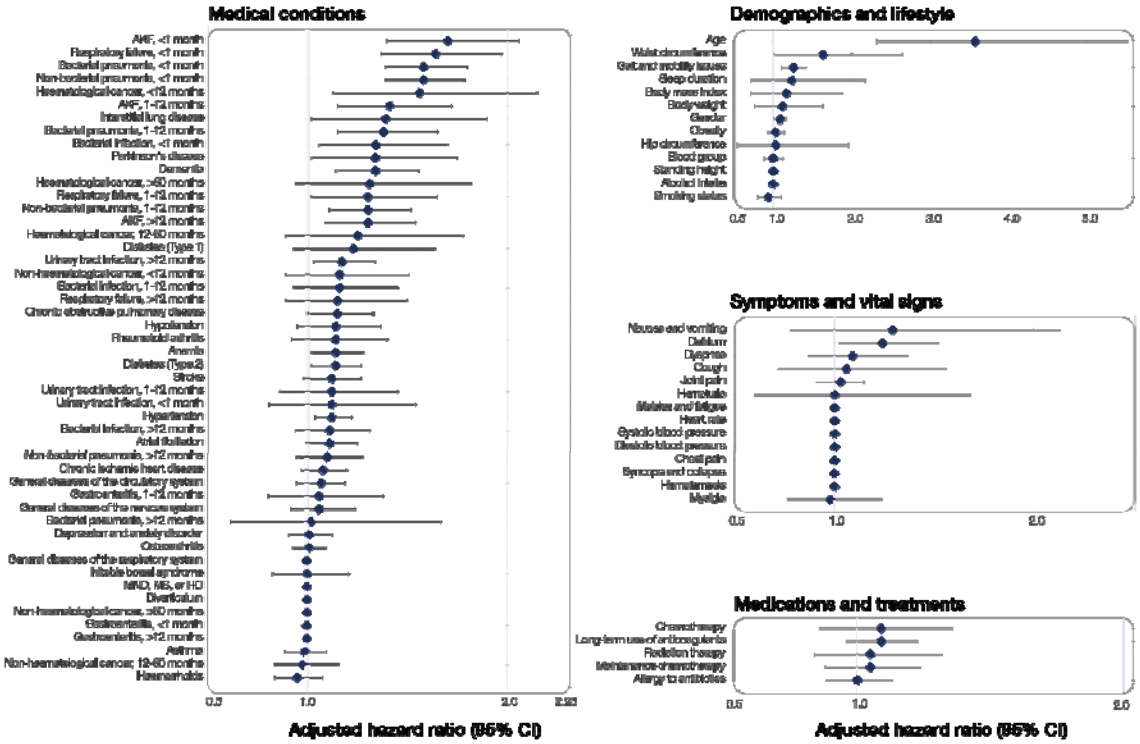

Figure 3. Plot of Cox model coefficients of COVID-19 mortality in UK Biobank cohort. Values show HR ± 95%CI. AKF = acute kidney failure, MND = motor neurone disease, MS = multiple sclerosis, HD = Huntington's disease, HR = hazard ratio, CI = confidence interval.

As shown, the ROC curves for the RF and CPH are very comparable with a slight advantage for the CPH. From **Figure 2A**, it can be seen that when QCOVID is applied to the UKB dataset it performs well and achieves an AUC of 0.83, showcasing robustness to unseen data. To explore the performance further, it is essential to look at the robustness of the generated models. **Figure 2B** illustrates the use of F-β statistical analysis to examine the performance of the various models. As expected, despite the CPH having a slightly greater AUC score, it is clear that the RF has much more stable performance. Moreover, it can be seen that both the CPH and QCOVID models achieve optimal F-β scores when β is small. However, for the RF model, the F-β scores are considerably larger than its comparators and are more consistent across the range of thresholds, thereby demonstrating greater stability and increased capabilities regarding recall (i.e. minimizing false-negatives).

Discussion

This study developed and validated machine learning models to predict mortality in patients with COVID-19 using comprehensive data from 7,536 COVID-19 patients in the UKB. The results show that using easily

accessible patient characteristics, brief medical history, symptoms, and vital signs can predict mortality in patients with COVID-19 with excellent levels of accuracy (AUC: 0.92).

The features selected in the presented model mirror much of the current clinical understanding regarding factors associated with severe COVID-19 outcomes. While research on COVID-19 is rapidly evolving, it is well established that age[23] and obesity[24] are significant variables. In the data-driven model, we found that age was the most critical factor in determining outcome, in line with meta-analyses on outcomes in COVID-19 patients[10,23]. Obesity is also well reflected in our model, with body mass index (BMI) and related measures featuring dominantly in the data-driven ranking. Potentially mediated by proinflammatory states induced by adiposity, the pathophysiological link between adiposity and severe COVID-19 outcomes is not fully understood, body composition may provide more granular risk profiling than BMI alone[25,26]. The highest ranking pre-existing conditions were found to be type 2 diabetes and hypertension. While both conditions are linked to obesity, diabetes has been reported to confer significant independent mortality risk[27]. Evidence for the role of hypertension as an independent risk factor, however, is not fully established, with meta-analyses finding hypertension non-significant after adjusting for other risk factors[10,28,29].

Alongside the specific COVID-19 data releases, the unique value of the UKB can be attributed to its well-established, longitudinal background dataset. Encompassing non-traditional health data, including anthropometric measurements and lifestyle insights, allows for the assessment of commonly overlooked, yet easily collectable, variables to enrich the already-documented clinical factors. The ability to capture a deeper phenotype of the individual prior to infection has proved integral to our model's performance, in line with the findings of other disease-specific prediction models developed on the UKB[30–32]. Notably, we identified baseline hip circumference, waist circumference, weight, and height to be valuable independent of BMI and obesity, accounting for four of the top-ten RF ranked features (Supplementary Figure 3). Moreover, while baseline sleep duration has been demonstrated to be highly predictive of all-cause mortality[33], cardiovascular diseases[34], and type 2 diabetes[35], this marks the first instance of its significant predictive influence within COVID-19 prognosis.

As a result of the data-driven approach, our most interesting findings concern the impact of prior urinary tract infection (UTI), respiratory failure, acute kidney failure, bacterial and non-bacterial pneumonias, and other bacterial infections. In the case of all but UTI, dividing each feature into time groupings by their proximity to the COVID-19 diagnosis highlights attenuated risk the more distant the event, returning to approximately baseline when >12 months for both the respiratory conditions and other infections. Interestingly, the outlying significance of acute kidney failure at >12 months suggests the impact of damage to this organ may be more integral to COVID-19 prognosis than that of the respiratory system. This is supported by findings related to UTIs, where they appear as a less damaging, but persistent, risk factor regardless of the time since diagnosis. Respiratory and kidney complications are a hallmark of severe and fatal COVID-19, thus the observation that a history of severe conditions affecting these organs effectively forecasts COVID-19 prognosis is logical[36]. To date, however, the relationship between less severe urogenital factors and COVID-19 has not been effectively assessed. A recent systematic review on urological manifestations found urinary symptoms were absent from all included studies[37]. Where data has been collected, sample sizes have been too low to draw strong conclusions. Though the occurrence of *de novo* urinary symptoms has been documented without noticeable impact on

prognosis[38,39], it has been previously suggested, and more recently evidenced, that the presence of pre-existing urinary conditions may be associated with a poorer disease prognosis proportional to their severity[40,41]. Our investigation provides the first reliable evidence that a history of UTI is predictive of greater COVID-19 mortality risk. We hypothesise that the underlying nature of this association reflects poorer baseline health status which may not previously have been of clinical significance in the absence of a highly infective, fatal pathogen such as SARS-CoV-2.

The approach taken in the development of this model is a symbiosis of machine learning and traditional statistical modeling, boosting the performance and acceptability of the resultant algorithm. The results show that both the RF and CPH models are comparable in terms of accuracy. However, the RF was integral to the CPH's construction by searching through the large feature space and selecting the most important of the original ~12,000. Moreover, the RF model is more resilient to overfitting the data, and this could explain the improved F1-scores. Given the different performance characteristics of the RF and CPH models, an ensemble of the two models is recommended to ensure greater stability and performance. By using an ensemble approach, predictions from each model are less susceptible to error, with averaging across the predictions reducing individual error.

Our model's critical component is the distinction of variables with respect to their time of onset. Classifying variables in a time-dependent fashion enables discrimination between events that occurred prior to COVID-19 infection and those which occurred during the course of the disease, either shortly before or following diagnosis (i.e. segregating pre-existing conditions, symptoms, and complications). This was especially important as several of our novel features are also established complications of COVID-19. Studies have emphasised the need for distinguishing pre-existing conditions from complications of COVID-19 infection and their respective impact on prognosis[42,43] but, to our knowledge, no predictive models for this disease have stratified variables in such a way. Applied in the context of patient management and enriched by the explainability of variable time-filtering, our results could help clarify crucial aspects of patients' past medical history and their relation to predicted prognosis. Models which forecast infection risk as a component of their mortality prediction have been criticised for generalizing human behaviour, which results in underestimation of risk factors and leaves their calibration extremely vulnerable to changes in local population dynamics[44]. One strength of our model is that the risk of mortality is predicated on the assumption of a positive COVID-19 test, avoiding the associated ambiguity of multi-event prediction and enabling its use in clinical practice.

The COVID-19 pandemic has resulted in extraordinary acceptance of digital technology in healthcare[45]. As clinician:patient ratios are increasingly stretched[46], risk assessment tools can support the streamlining of clinical time and resource prioritization, whether on a national, organizational, or patient level. Models such as those presented, can support the latter by monitoring patients at-scale and identifying those at-risk of severe illness, in real-time, and without requiring specialist equipment or clinical input. Notably, our focus on variables which are not presently assessed by clinicians to stratify patients in remote monitoring settings means such a model enriches the standard-of-care, rather than attempting to replace it through the amalgamation of currently utilised data.

Algorithm performance may be further improved by inclusion of passive, continuous variables via smartphones or wearables. Establishing our model in a prospective healthcare setting may enable this when coupled with high quality, continuous vital sign information and replete data on the course of symptomatology. Similar digital phenotyping has also shown potential in predicting COVID-19 infection at early symptoms onset[47,48]. We believe a combination of these two types of digital tools, in union with dedicated hospital-at-home services, may become considered standard practice in infectious disease management, particularly during historically resource-intense periods, such as annual influenza outbreaks.

While the use of the UKB is a key strength in the development of the model, there are associated limitations which may impact the generalizability of the model. The UKB cohort trends towards being healthier and wealthier than the general population, which poses a notable limitation when modeling noncommunicable diseases[49]. As COVID-19 acquisition, however, is determined by exposure, this limitation is minimised in our investigation. Separately, the UKB COVID-19 data subset is less likely to capture asymptomatic or non-severe cases, in part as such individuals may not have received a test or sought medical treatment, but predominantly owing to UKB's enrichment for older age resulting in lesser rates of such presentation. The restricted age distribution (51-85 years) may further limit generalization of our findings to outside of this age range, however, ONS figures show those aged 50+ have accounted for 98.06% of all COVID-19-related deaths in England and Wales[50].

Although age is clearly an important feature, our sensitivity analysis (Supplementary Figure 2) demonstrated negligible performance drop, likely because much of the risk associated with older age is captured within other included features. One reason for using uniform leave-one-out (LOO) training is to overcome such issues of feature reliance and generalize the model as much as possible. The F-score in **Figure 2B** illustrates this robustness, however, this must be tested on a separate representative dataset for a conclusive answer. Several included risk factors (lifestyle and anthropometric data) were assessed over 10-years prior to the time of COVID-19 infection and may have since changed. Deterioration in these factors may be expected over time and, therefore, the findings may be biased towards the null. Despite these limitations, our robust development approach, paired with deep individual phenotyping, strengthens the evidence towards effective COVID-19 risk profiling. In addition to the limitations of the dataset, it is likely that there are regional variances in COVID-19 outcomes. As such, the model would strongly benefit from external validation, especially with the continued emergence of disruptive SARS-CoV-2 variants[51].

While the model presented outperforms QCOVID (AUC: 0.92 vs. 0.83), and best efforts were made in the comparison, it cannot be considered a direct comparison. In replication of the QCOVID algorithms, variables were mapped to related fields in the UKB, however, we were unable to confirm these were fully paired. Moreover, as the UKB is not linked to GP databases in the same manner, there were some missing variables (Supplementary Table 2). Importantly, contrasting with our purpose of supporting patient management, QCOVID is designed for population risk stratification to aid public health decision-making, and was used to exemplify the necessity of specific model design for specific purposes.

Conclusion

In conclusion, we present a comprehensive, robust model based on readily accessible factors. In our novel analysis, we combine data-driven model development and clinical refinement to produce a model that uniquely incorporates time-to-event, symptoms, and vital signs. The design and feature selection of the framework lends itself for deployment in a digital setting. Possible applications of this include supporting individual-level risk profiling and monitoring deterioration in high volumes of COVID-19 patients, particularly in hospital-at-home settings.

Online Methods

Study population

The development and validation of the risk model was carried out using the UKB. The UKB is a large cohort study with rich phenotype mapping of participants, including over 500,000 individuals aged between 40- and 69-years-old at recruitment, between 2006 and 2010, from across England, Scotland, and Wales[52]. The open dataset contains detailed health data and outcomes obtained prospectively from electronic health records and self-reported health measures from on-site testing over the past 15-years. The current analysis was approved under the UKB application number 55668. Ethical approval was granted by the national research ethics committee (REC 16/NW/0274) for the overall UK Biobank cohort.

COVID-19 Status and Sample Selection

For this study, only participants with a positive RT-PCR COVID-19 test, from English assessment centres who were alive on 19th December 2020 were included (Supplementary Figure 1). Public Health England provided data on SARS-CoV-2 tests, including the specimen date, location, and result[53]. COVID-19 test result data were available for the period 16th March 2020 to 12th December 2020, and were linked with hospital admission, primary care, and death records. In total, 83,148 COVID-19 tests were conducted on 46,450 participants in the available cohort. Of these, 37,951 were excluded due to negative test results. Overall, 7,536 participants tested positive of which 7,040 were survivors and 496 non-survivors. Deaths were defined as COVID-19-related if ICD-10 codes U07.1 or U07.2 were present on the death certificate. No COVID-19 test data were available for UKB assessment centers in Scotland and Wales, thus data from these centers were not included.

Time Filtering

Considering the chronology of medical events is critical to distinguish between, for example, pre-existing conditions and complications resulting from COVID-19. Specific attributes, therefore, can be included or excluded in the prediction model for various use cases. This study focuses on developing a model to predict mortality for COVID-19 patients before hospital admission. Accordingly, inclusion of respiratory failure (ICD-10: J96.9), for example, as a symptom or complication to predict mortality has limited use, as such events would

demand hospital admission. Conversely, it is valuable to include personal history of respiratory failure as a prognostic indicator. Thus, we implemented a time filter for all features which were not demographics, symptoms, or vital signs, excluding any data recorded later than one-week prior to patients' positive COVID-19 test. This accounted for the circumstance whereby a patient may have been admitted for severe symptoms of COVID-19 prior to receiving a test. Further time filtering of <1 month, 1-12 months, and >12 months was applied to specific acute features to provide more granular insight. Similarly, it is important to consider only relevant symptoms and vital signs corresponding to the period of COVID-19 infection. Thus, a two-week window pre- and post- the first COVID-19 positive test was implemented.

COVID-19 Mortality Model

Feature Selection

The data ingestion pipeline, **Figure 1A**, generates an array of ~12,000 dimensions (including patient characteristics, pre-existing conditions, symptoms, and vital signs). Owing to the small number of samples in the dataset and the importance of obtaining an unbiased model, a LOO cross-validation experiment, which is closely related to the jack-knife estimation method[54], was used to search the full feature array for the most relevant features. LOO iterates through every sample in the dataset, whereby at each step the current sample was used to evaluate the model trained on the remaining dataset. At each iteration the samples of all classes were balanced to ensure unbiased training and, following evaluation, the model was discarded and a new model trained. A RF model was chosen due to its inherent ability to extract features, handle high dimensionality data, and generalize well to unseen data[55]. During each step of the LOO cross-validation, a ranked list of features was extracted and averaged across the entire experiment to obtain a final shortlist of features that produced the highest accuracy, further cross-checked by clinical expertise. **Figure 1B** illustrates the production of shortlisted features driven by data, and their validation and review based on clinical judgement.

Clinical feature selection was informed by a review of ranked feature importance in RF model. The highest ranked 1,000 features were screened by at least two reviewers. Any disagreements were settled by consensus with input of additional reviewers. Features were excluded where: *i*) they could not be readily obtained through self-reporting or measured outside of the clinical setting; *ii*) there was high confounding with higher ranked features; *iii*) clinical consensus concluded that the feature's rank was more likely to be explained by database bias. Subsequently, features which were closely related (e.g. cancer diagnoses) were grouped together. Supplementary ICD-10 codes were included and, where possible, generalized (**Supplementary Table 1**).

Experimental Setup

The LOO evaluation was selected to maximize the value of the available datasets. By evaluating one sample at each iteration, the rest of the samples could be used for training the model. By iterating on the entire dataset, each sample took turns to be the evaluation sample. At each iteration, the previously trained model was completely discarded and a new one trained. Another advantage of using the LOO was to ensure fair machine learning training by having enough samples to represent the different prediction classes uniformly.

In this study, the prediction classes were two: COVID-19 survivors (n=7,040) and non-survivors (n=496). At each LOO iteration, two groups of equal sample size were randomly selected without replacement for training. The evaluation sample outcome and RF likelihood value were aggregated from all iterations. After aggregating all the evaluation results from the LOO experiment, the ROC curve analysis was carried out, and the AUC computed as a measure of accuracy[55]. Furthermore, the F-β statistic was used to evaluate the robustness of the model. When β is 1, this becomes the F1-score, which gives equal weights to recall and precision. A smaller β value gives more weight to precision, minimising false-positive errors, while a larger β value gives more weight to recall, minimising false-negative errors. The F-score range is [0, 1], where a score of 1 is a perfect performance.

The machine learning algorithm used in this study is the RF, which is an ensemble meta-estimator constructed from several decision trees[55]. These trees were fitted to the data using the bootstrap aggregation method (or *bagging*), which is robust and resilient to over-fitting[56]. The Gini impurity was used to compute the model likelihood of prediction. To quantify the prediction uncertainty of the RF model, a Monte Carlo approach was used to compute the confidence interval of each prediction. A CPH model[22] trained on the same subset of features was constructed and assessed to maximise explainability of the RF model.

QCOVID Comparison

We compared our model against QCOVID, a leading risk prediction model for infection and subsequent death due to COVID-19, which was developed by fitting a sub-distribution hazard model on the QResearch database[16]. Predictor variables reported in QCOVID were mapped to comparable features in the UKB dataset. The UKB dataset did not include all of the relevant variables used in the QCOVID algorithm, hence chemotherapy grades and medication variables were excluded in our analysis (**Supplementary Table 2**). QCOVID risk equations for mortality were then implemented for both male and female cohorts. To ensure a fair comparison between models, QCOVID risk equations were evaluated on the UKB dataset using the same methods described above.

This article was written following the TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis) guidelines[57], which are further elaborated in **Supplementary Table 5.**

References

1. World Health Organization. WHO Coronavirus Disease (COVID-19) Dashboard. https://covid19.who.int (2020).

2. Nicola, M. *et al.* The socio-economic implications of the coronavirus pandemic (COVID-19): A review. *Int. J. Surg. Lond. Engl.* **78**, 185–193 (2020).

3. Walker, P. G. T. *et al.* The impact of COVID-19 and strategies for mitigation and suppression in low- and middle-income countries. *Science* (2020) doi:10.1126/science.abc0035.

4. Reuters, Inc. China, India's COVID-19 vaccinations to stretch to late 2022: study | The Journal Pioneer. http://www.journalpioneer.com/news/world/china-indias-covid-19-vaccinations-to-stretch-to-late-2022-study-545388/.

5. Oran, D. P. & Topol, E. J. Prevalence of Asymptomatic SARS-CoV-2 Infection. *Ann. Intern. Med.* **173**, 362–367 (2020).

6. Byambasuren, O. *et al*. Estimating the extent of asymptomatic COVID-19 and its potential for community transmission: Systematic review and meta-analysis. *Off. J. Assoc. Med. Microbiol. Infect. Dis. Can.* **5**, 223–234 (2020).

7. Abate, S. M., Ahmed Ali, S., Mantfardo, B. & Basu, B. Rate of Intensive Care Unit admission and outcomes among patients with coronavirus: A systematic review and Meta-analysis. *PLOS ONE* **15**, e0235653 (2020).

8. Atkins, J. L. *et al*. Preexisting Comorbidities Predicting COVID-19 and Mortality in the UK Biobank Community Cohort. *J. Gerontol. Ser. A* **75**, 2224–2230 (2020).

9. Li, B. The Association Between Symptom Onset and Length of Hospital Stay in 2019 Novel Coronavirus Pneumonia Cases Without Epidemiological Trace. *J. Natl. Med. Assoc.* (2020) doi:10.1016/j.jnma.2020.05.015.

10. Booth, A. *et al*. Population risk factors for severe disease and mortality in COVID-19: A global systematic review and meta-analysis. *medRxiv* 2020.12.21.20248610 (2020) doi:10.1101/2020.12.21.20248610.

11. Holman, N. *et al.* Risk factors for COVID-19-related mortality in people with type 1 and type 2 diabetes in England: a population-based cohort study. *Lancet Diabetes Endocrinol*. **8**, 823–833 (2020).

12. Public Health England. Disparities in the risk and outcomes of COVID-19. 92 (2020).

13. Rechtman, E., Curtin, P., Navarro, E., Nirenberg, S. & Horton, M. K. Vital signs assessed in initial clinical encounters predict COVID-19 mortality in an NYC hospital system. *Sci. Rep.* **10**, 21545 (2020).

14. Zhou F *et al*. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *Lancet* **395**, 1054–1062 (2020).

15. Foy, B. H. *et al*. Association of Red Blood Cell Distribution Width With Mortality Risk in Hospitalized Adults With SARS-CoV-2 Infection. *JAMA Netw. Open* **3**, e2022058 (2020).

16. Clift, A. K. *et al*. Living risk prediction algorithm (QCOVID) for risk of hospital admission and mortality from coronavirus 19 in adults: national derivation and validation cohort study. *BMJ* **371**, (2020).

17. Jin, J. *et al*. Individual and community-level risk for COVID-19 mortality in the United States. *Nat. Med.* 1–6 (2020) doi:10.1038/s41591-020-01191-8.

18. Barda, N. *et al*. Developing a COVID-19 mortality risk prediction model when individual-level data are not available. *Nat. Commun.* **11**, 4439 (2020).

19. Knight, S. R. *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* **370**, (2020).

20. Qian, Z., Alaa, A. M. & van der Schaar, M. CPAS: the UK's national machine learning-based hospital capacity planning system for COVID-19. *Mach. Learn.* **110**, 15–35 (2021).

21. Shah, S. S., Gvozdanovic, A., Knight, M. & Gagnon, J. Mobile App–Based Remote Patient Monitoring in Acute Medical Conditions: Prospective Feasibility Study Exploring Digital Health Solutions on Clinical Workload During the COVID Crisis. *JMIR Form. Res.* **5**, e23190 (2021).

22. Cox, D. R. Regression Models and Life-Tables. *J. R. Stat. Soc. Ser. B Methodol.* **34**, 187–220 (1972).

23. Bonanad, C. *et al*. The Effect of Age on Mortality in Patients With COVID-19: A Meta-Analysis With 611,583 Subjects. *J. Am. Med. Dir. Assoc.* **21**, 915–918 (2020).

24. Stefan, N., Birkenfeld, A. L. & Schulze, M. B. Global pandemics interconnected — obesity, impaired metabolic health and COVID-19. *Nat. Rev. Endocrinol.* 1–15 (2021) doi:10.1038/s41574-020-00462-1.

25. Petersen, A. *et al*. The role of visceral adiposity in the severity of COVID-19: Highlights from a unicenter cross-sectional pilot study in Germany. *Metabolism* **110**, 154317 (2020).

26. Watanabe, M. *et al*. Obesity and SARS-CoV-2: A population to safeguard. *Diabetes Metab. Res. Rev.* e3325 (2020) doi:10.1002/dmrr.3325.

27. Targher, G. *et al*. Patients with diabetes are at higher risk for severe illness from COVID-19. *Diabetes Metab*. **46**, 335–337 (2020).

28. Cummings, M. J. *et al.* Epidemiology, clinical course, and outcomes of critically ill adults with COVID-19 in New York City: a prospective cohort study. *MedRxiv Prepr. Serv. Health Sci.* (2020) doi:10.1101/2020.04.15.20067157.

29. Williamson, E. J. *et al.* Factors associated with COVID-19-related death using OpenSAFELY. *Nature* (2020) doi:10.1038/s41586-020-2521-4.

30. Alaa, A. M., Bolton, T., Angelantonio, E. D., Rudd, J. H. F. & Schaar, M. van der. Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants. *PLOS ONE* **14**, e0213653 (2019).

31. Rezaee, M., Putrenko, I., Takeh, A., Ganna, A. & Ingelsson, E. Development and validation of risk prediction models for multiple cardiovascular diseases and Type 2 diabetes. *PLOS ONE* **15**, e0235758 (2020).

32. Sanikini, H. *et al.* Anthropometry, body fat composition and reproductive factors and risk of oesophageal and gastric cancer by subtype and subsite in the UK Biobank cohort. *PLOS ONE* **15**, e0240413 (2020).

33. Cappuccio, F. P., D'Elia, L., Strazzullo, P. & Miller, M. A. Sleep Duration and All-Cause Mortality: A Systematic Review and Meta-Analysis of Prospective Studies. *Sleep* **33**, 585–592 (2010).

34. Cappuccio, F. P., Cooper, D., D'Elia, L., Strazzullo, P. & Miller, M. A. Sleep duration predicts cardiovascular outcomes: a systematic review and meta-analysis of prospective studies. *Eur. Heart J.* **32**, 1484–1492 (2011).

35. Gangwisch, J. E. *et al*. Sleep Duration as a Risk Factor for Diabetes Incidence in a Large US Sample. *Sleep* **30**, 1667 (2007).

36. Singhal, T. A Review of Coronavirus Disease-2019 (COVID-19). *Indian J. Pediatr.* **87**, 281–286 (2020).

37. Chan, V. W.-S. *et al.* A systematic review on COVID-19: urological manifestations, viral RNA detection and special considerations in urological conditions. *World J. Urol.* 1–12 (2020) doi:10.1007/s00345-020-03246-4.

38. Dhar, N. *et al.* De Novo Urinary Symptoms Associated With COVID-19: COVID-19-Associated Cystitis. *J. Clin. Med. Res.* **12**, 681–682 (2020).

39. Mumm, J.-N. *et al.* Urinary Frequency as a Possibly Overlooked Symptom in COVID-19 Patients: Does SARS-CoV-2 Cause Viral Cystitis? *Eur. Urol.* **78**, 624–628 (2020).

40. Wu Zhang-song, Zhang Zhi-qiang, & Wu Song. Focus on the Crosstalk between COVID-19 and Urogenital Systems. *J. Urol.* **204**, 7–8 (2020).

41. Karabulut, I. *et al.* The Effect of the Presence of Lower Urinary System Symptoms on the Prognosis of COVID-19: Preliminary Results of a Prospective Study. *Urol. Int.* **104**, 853–858 (2020).

42. Guan, W.-J. *et al.* Comorbidity and its impact on 1590 patients with COVID-19 in China: a nationwide analysis. *Eur. Respir. J.* **55**, (2020).

43. Wang, Z. *et al.* Clinical symptoms, comorbidities and complications in severe and non-severe patients with COVID-19. *Medicine (Baltimore)* **99**, (2020).

44. Sperrin, M. & McMillan, B. Prediction models for covid-19 outcomes. *BMJ* **371**, m3777 (2020).

45. Whitelaw, S., Mamas, M. A., Topol, E. & Spall, H. G. C. V. Applications of digital technology in COVID-19 pandemic planning and response. *Lancet Digit. Health* **2**, e435–e440 (2020).

46. Lasater, K. B. *et al.* Chronic hospital nurse understaffing meets COVID-19: an observational study. *BMJ Qual. Saf.* (2020) doi:10.1136/bmjqs-2020-011512.

47. Mishra, T. *et al.* Pre-symptomatic detection of COVID-19 from smartwatch data. *Nat. Biomed. Eng.* **4**, 1208–1220 (2020).

48. Miller, D. J. *et al.* Analyzing changes in respiratory rate to predict the risk of COVID-19 infection. *PloS One* **15**, e0243693 (2020).

49. Fry, A. *et al.* Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *Am. J. Epidemiol.* **186**, 1026–1034 (2017).

50. Office for National Statistics. Deaths registered weekly in England and Wales, provisional. https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/datasets/weeklyprovisionalfiguresondeathsregisteredinenglandandwales.

51. Williams, T. C. & Burgers, W. A. SARS-CoV-2 evolution and vaccines: cause for concern? *Lancet Respir. Med.* **0**, (2021).

52. Sudlow, C. *et al.* UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Med.* **12**, e1001779 (2015).

53. Armstrong, J. *et al.* Dynamic linkage of COVID-19 test results between Public Health England's Second Generation Surveillance System and UK Biobank. *Microb. Genomics* **6**, e000397 (2020).

54. Efron, B. *The jackknife, the bootstrap, and other resampling plans.* (Society for Industrial and Applied Mathematics, 1982).

55. Breiman, L. Random Forests. *Mach. Learn.* **45**, 5–32 (2001).

56. Breiman, L. Bagging Predictors. *Mach. Learn.* **24**, 123–140 (1996).

57. Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. *BMC Med.* **13**, 1 (2015).