Uncovering clinical risk factors and prediction of severe COVID-19: A machine learning approach based on UK Biobank data

Kenneth Chi-Yin Wong, Yong Xiang, Liangying Yin, Hon-Cheong So

Submitted to: JMIR Public Health and Surveillance
on: April 12, 2021

Table of Contents

Uncovering clinical risk factors and prediction of severe COVID-19: A machine learning approach based on UK Biobank data

Kenneth Chi-Yin Wong[1] MSc; Yong Xiang[1] MBBS; Liangying Yin[1] MPhil; Hon-Cheong So[1, 2, 3, 4, 5, 6, 7] MBBS, PhD

[1]School of Biomedical Sciences, The Chinese University of Hong Kong Hong Kong CN
[2]Margaret K.L. Cheung Research Centre for Management of Parkinsonism, The Chinese University of Hong Kong Hong Kong CN
[3]Brain and Mind Institute, The Chinese University of Hong Kong Hong Kong CN
[4]KIZ-CUHK Joint Laboratory of Bioresources and Molecular Research of Common Diseases, Kunming Institute of Zoology and The Chinese University of Hong Kong Kunming CN
[5]CUHK Shenzhen Research Institute Shenzhen CN
[6]Department of Psychiatry, The Chinese University of Hong Kong Hong Kong CN
[7]Hong Kong Branch of the Chinese Academy of Sciences Center for Excellence in Animal Evolution and Genetics, The Chinese University of Hong Kong Hong Kong CN

Corresponding Author:
Hon-Cheong So MBBS, PhD
School of Biomedical Sciences, The Chinese University of Hong Kong
RM 520A, Lo Kwee Seong Biomedical Sciences Buildiing
Chinese University of Hong Kong
Hong Kong
CN

Abstract

Background: COVID-19 is a major public health concern. Given the extent of the pandemic, it is urgent to identify risk factors associated with disease severity. More accurate prediction of those at risk of developing severe infections is of high clinical importance.

Objective: Based on the UK-Biobank(UKBB), we aimed to build machine learning(ML) models to predict the risk of developing severe or fatal infections, and uncover major risk factors involved.

Methods: We first restricted the analysis to infected subjects(N=7846), then performed analysis at a population level, considering those with no known infection as controls(N controls=465,728). Hospitalization was used as a proxy for severity. Totally 97 clinical variables(collected prior to COVID-19 outbreak) covering demographic variables, comorbidities, blood measurements(e.g. hematological/liver/renal function/metabolic parameters), anthropometric measures and other risk factors(e.g. smoking/drinking) were included as predictors. We also constructed a simplified('lite') prediction model using 27 covariates that can be more easily obtained(demographic and comorbidity data). XGboost(gradient-bosted trees) was used for prediction and predictive performance assessed by cross-validation. Variable importance was quantified by Shapley values and accuracy gain. Shapley dependency and interaction plots were used to evaluate the pattern of relationship between risk factors and outcomes.

Results: Totally 2386 severe and 477 fatal cases were identified. For analysis among infected individuals (N=7846),our prediction model achieved AUCs of 0.723(95% CI:0.711-0.736) and 0.814(CI:0.791-0.838) for severe and fatal infections respectively. The top five contributing factors for severity were age, number of drugs taken(cnt_tx), cystatin C(reflecting renal function), waist-hip ratio(WHR) and Townsend Deprivation index(TDI). For mortality, the top features were age, testosterone, cnt_tx, waist circumference(WC) and red cell distribution width(RDW).

In analyses involving the whole UKBB population, corresponding AUCs for severity and fatality were 0.696(CI:0.684-0.708) and 0.802(CI:0.778-0.826) respectively. The same top five risk factors were identified for both outcomes, namely age, cnt_tx, WC, WHR and TDI. Apart from the above features, Type 2 diabetes(T2DM), HbA1c and apolipoprotein A were ranked among the top 10 in at least two (out of four) analyses. Age, cystatin C, TDI and cnt_tx were among the top 10 across all four analyses.

For the 'lite' models, the predictive performances are broadly similar, with estimated AUCs of 0.716, 0.818, 0.696 and 0.811 respectively. The top-ranked variables were similar to above, including e.g. age, cnt_tx, WC, male and T2DM.

Conclusions: We identified a number of baseline clinical risk factors for severe/fatal infection by ML. For example, age, central

obesity, impaired renal function, multi-comorbidities and cardiometabolic abnormalities may predispose to poorer outcomes. The presented prediction models may be useful at a population level to identify those susceptible to developing severe/fatal infections, facilitating targeted prevention strategies. A risk prediction tool is also available online. Further replications in independent cohorts are required to verify our findings. Clinical Trial: NA

(JMIR Preprints 12/04/2021:29544)
DOI: https://doi.org/10.2196/preprints.29544

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

Please make my preprint PDF available to anyone at any time (recommended).

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

✓ **No, I do not wish to publish my submitted manuscript as a preprint.**

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Uncovering clinical risk factors and prediction of severe COVID-19: A machine learning approach based on UK Biobank data

Kenneth Chi-Yin WONG[1], Yong XIANG[1], Liangying YIN[1], Hon-Cheong SO[1-7*]

[1]School of Biomedical Sciences, The Chinese University of Hong Kong, Shatin, Hong Kong
[2] KIZ-CUHK Joint Laboratory of Bioresources and Molecular Research of Common Diseases, Kunming Institute of Zoology and The Chinese University of Hong Kong, China
[3]CUHK Shenzhen Research Institute, Shenzhen, China
[4]Department of Psychiatry, The Chinese University of Hong Kong, Shatin, Hong Kong
[5]Margaret K.L. Cheung Research Centre for Management of Parkinsonism, The Chinese University of Hong Kong, Shatin, Hong Kong
[6]Brain and Mind Institute, The Chinese University of Hong Kong, Shatin, Hong Kong
[7]Hong Kong Branch of the Chinese Academy of Sciences Center for Excellence in Animal Evolution and Genetics, The Chinese University of Hong Kong, Hong Kong SAR, China

**Corresponding author*

Correspondence to: Hon-Cheong So, MBBS, PhD. Lo Kwee-Seong Integrated Biomedical Sciences Building, The Chinese University of Hong Kong, Shatin, Hong Kong. Tel: +852 3943 9255; E-mail: hcso@cuhk.edu.hk

Abstract

Background: COVID-19 is a major public health concern. Given the extent of the pandemic, it is urgent to identify risk factors associated with disease severity. More accurate prediction of those at risk of developing severe infections is of high clinical importance.

Objective: Based on the UK-Biobank(UKBB), we aimed to build machine learning(ML) models to predict the risk of developing severe or fatal infections, and uncover major risk factors involved.

Methods: We first restricted the analysis to infected subjects(N=7846), then performed analysis at a population level, considering those with no known infection as controls($N_{controls}$=465,728). Hospitalization was used as a proxy for severity. Totally 97 clinical variables(collected prior to COVID-19 outbreak) covering demographic variables, comorbidities, blood measurements(e.g. hematological/liver/renal function/metabolic parameters), anthropometric measures and other risk factors(e.g. smoking/drinking) were included as predictors. We also constructed a simplified('lite') prediction model using 27 covariates that can be more easily obtained(demographic and comorbidity data). XGboost(gradient-boosted trees) was used for prediction and predictive performance assessed by cross-validation. Variable importance was quantified by Shapley values (ShapVal), permutation importance(PermImp) and accuracy gain. Shapley dependency and interaction plots

were used to evaluate the pattern of relationships between risk factors and outcomes.

Results: Totally 2386 severe and 477 fatal cases were identified. For analyses within infected individuals (*N*=7846), our prediction model achieved AUC-ROC of 0.723(95% CI:0.711-0.736) and 0.814(CI:0.791-0.838) for severe and fatal infections respectively. The top five contributing factors(sorted by ShapVal) for severity were age, number of drugs taken(cnt_tx), cystatin C(reflecting renal function), waist-hip ratio(WHR) and Townsend Deprivation index(TDI). For mortality, the top features were age, testosterone, cnt_tx, waist circumference(WC) and red cell distribution width(RDW).

For analyses involving the whole UKBB population, AUCs for severity and fatality were 0.696(CI:0.684-0.708) and 0.825(CI: 0.802-0.848) respectively. The same top five risk factors were identified for both outcomes, namely age, cnt_tx, WC, WHR and TDI. Apart from the above, age, cystatin C, TDI and cnt_tx were among the top-10 across all four analyses. Other diseases top-ranked by ShapVal or PermImp included Type 2 diabetes (T2DM), coronary artery disease, atrial fibrillation and dementia, among others.

For the 'lite' models, predictive performances were broadly similar, with estimated AUCs of 0.716, 0.818, 0.696 and 0.830 respectively. The top-ranked variables were similar to above, including e.g. age, cnt_tx, WC, sex(male) and T2DM.

Conclusions: We identified numerous baseline clinical risk factors for severe/fatal infection by XGboost. For example, age, central obesity, impaired renal function, multi-comorbidities and cardiometabolic abnormalities may predispose to poorer outcomes. The prediction models may be useful at a population level to identify those susceptible to developing severe/fatal infections, facilitating targeted prevention strategies. A risk prediction tool is also available online. Further replications in independent cohorts are required to verify our findings.

Introduction

Coronavirus Disease 2019 (COVID-19) has resulted in a pandemic affecting more than a hundred countries worldwide [1-3]. More than 177 million confirmed cases and 3.8 million fatalities have been reported worldwide as at 19th June 2021 (https://coronavirus.jhu.edu/map.html), while a large number of mild or asymptomatic cases may remain undetected. Given the extent of the pandemic, it is urgent to identify risk factors that may be associated with severe disease, and to gain deeper understanding into its pathophysiology. Accurate prediction of those at risk of developing severe diseases is also clinically important.

Machine learning (ML) approaches are powerful tools to predict disease outcomes and have been increasingly applied in biomedical research. In this study we employed boosted trees (with XGboost) to predict disease outcomes and identify risk factors. This ML approach can capture complex, non-linear and interactions between variables, hence leading to better predictive power in many circumstances. In view of the COVID-19 pandemic, many ML models have been developed for diagnostic or prognostic purposes. To highlight a few studies, for instance, Bayat et al. developed a prediction model for COVID-19 infection based on 75,991 veteran patients who were tested for the virus. The prediction was based on boosted trees and predictors included vital signs, hematology measurements and blood biochemistries. Knight et al.[4] built a model to predict in-hospital mortality for hospitalized COVID-19 patients, based on demographics, comorbidities, vital signs and blood test results. A variety of methods including XGboost, generalized additive model and LASSO were employed. Chung et al.[5] employed deep neural networks (DNN) to predict severity of COVID-19 infection based on basic patient information, comorbidities, vital signs, clinical symptoms and complete blood count. Wynants et al. has performed a systematic review of COVID-19 related prediction models up to 1 July 2020, covering 169 studies describing 232 prediction models. Several recent reviews have also summarized applications of ML methods in the study of COVID-19 (e.g. ref[6-9]).

Here we made use of the UK Biobank (UKBB) data to build ML models to predict severity and fatality from COVID-19, and evaluated the contributing risk factors. We built prediction models not only for infected patients but also at a general population level. While predictive performance is the main concern in most previous studies, we argue that ML models can also provide important insight into individual contributing factors and the pattern of complex relationships between risk factors and the outcome. While many have studied risk factors on COVID-19 susceptibility or severity in the UKBB[10-12] or other cohorts (e.g. see

Refs[6, 13-16]), most relied on conventional linear models. As such, non-linear effects and interactions between variables may be missed.

We note that in the UKBB, clinical data were collected years before the outbreak of infection in 2020, which may be a limitation. Ideally, the predictors should be measured at the time when the model is intended to be applied (e.g. at admission). However, we believe building ML models with previously collected clinical data is useful for reasons detailed below. First of all, using previously collected clinical features may facilitate the identification of potential causal risk factors. As the predictors are collected prior to the outbreak, there is no concern of reverse causality. In practice, infection itself will lead to changes in many clinical parameters (e.g. glucose, inflammatory markers, liver/renal functions etc.); hence it is often difficult to tell the direction of effect in cross-sectional studies. We hypothesize that this study will identify general or 'baseline' risk factors or laboratory measurements that may be (causally) predictive of outcome. Secondly, the UKBB is a huge population-based sample (N ~ 500,000), and the rich clinical data collected previously enables ML models to be developed at the general population level. Importantly, there is a relative lack of such population-level ML prediction models to identify who may be at risk of developing severe COVID-19 infections. We hope this study will fill the gap, as this may have implications for prioritizing individuals for specific prevention strategies (e.g. vaccination) and diagnostic testing under limited resources.

In this study we performed four sets of analysis. In the first two sets, we built ML models to predict severity and mortality of COVID-19 within those who are tested positive for the virus. In this setting, predictive performance is of secondary concern (as predictors were not assessed at or during admission), but the predictive performance can shed light on to what extent *baseline* (pre-diagnostic) clinical characteristics contribute to severe infections. In the other two sets of analysis, we predicted severity and mortality of COVID-19 at the population level, considering subjects not known to be infected as 'controls'. Our objectives are two-fold. The first is to build prediction models for severity and mortality from COVID-19. In addition, we will uncover how different risk factors and their interactions impact on disease severity.

Methods

UK Biobank data

The UK Biobank is a large-scale prospective cohort comprising ~500,000 subjects aged 40–69 years when they were recruited in 2006–2010. Given that the first case of COVDI-19 in the UK was recorded on 31 Jan 2020, subjects with recorded mortality before 31 Jan 2020 (28,931 out of 502,524 subjects) were excluded. For a very small number of subjects (N=19) whose cause of mortality was COVID-19 (ICD code U07.1) but with negative test result(s) within one week, they were excluded from subsequent analyses. The current age of subjects included in our analyses ranged from 50 to 87 years, with 50.8% being older than 70. The present analysis was conducted under the project number 28732. For details of the UK Biobank data, please also refer to ref[17].

COVID-19 phenotypes

COVID-19 outcome data were downloaded from data portal provided by the UKBB. Details of data release are provided at http://biobank.ndph.ox.ac.uk/showcase/exinfo.cgi?src=COVID19. Briefly, the latest COVID test results were extracted on 30 Dec 2020 (last update on 14 Dec 2020). The dataset also included an indicator on whether the patient was an inpatient when the specimen was taken. We consider inpatient (hospitalization) status as a proxy for severity, as more sophisticated indicators of severity cannot be reliably derived yet. We noted that only ~10.2% (468,235 out of 4,581,006 infected cases, from https://coronavirus.data.gov.uk/ as at 16 Jun 2021) of patients were admitted in the UK; as such it is likely that only the more severe cases were hospitalized. Hospitalization has also been considered as an outcome measure in many studies, including studies of vaccination effectiveness[18-21], risk prediction[22, 23], and genetic/clinical risk factors[24, 25] underlying severe COVID-19.

In general, we required both test result and origin to be 1 (indicating positive test and inpatient origin) to qualify as an 'inpatient' case. For a small number of subjects with inpatient origin=0 and result=1, but changed to origin=1 with result=0 within 2 weeks' time (based on the fact that median duration of viral persistence is ~2 weeks [26]), we still considered those as inpatient cases (i.e. assume the hospitalization was

related to the infection). For all other patients with at least one positive SARS-CoV-2 test result, they were considered as 'outpatient'.

Data on mortality and cause of mortality were also extracted (with latest update on 14 Dec 2020). Subjects with recorded cause of mortality as "U07.1" was considered a fatal infection with laboratory-confirmed COVID-19. Please also refer to http://biobank.ndph.ox.ac.uk/showcase/exinfo.cgi?src=COVID19_tests on relevant details. We defined a case as 'severe COVID-19' if the subject is an inpatient and/or if the cause of mortality is U07.1.

Sets of analysis

Four sets of analysis were performed. The first two sets were restricted to test-positive cases (total *N*=7846). 'Severe COVID-19' (*N*=2386) and death (*N*=477) due to COVID-19 were treated as outcomes. Since only pre-diagnostic clinical data were available, the main objective of this analysis was to identify baseline risk factors for severe/fatal illness among the infected. We then performed another two sets of analysis with the same outcomes, but the 'unaffected' group was composed of the general population (*N*=465,728) who did not have a diagnosis of COVID-19 or were tested negative. The four sets of analysis were also referred to as cohorts A to D as shown in Table 1. We also constructed gender-specific prediction models.

Variables included in analysis

We extracted a total of 97 clinical variables of potential relevance based on the literature. For details, please refer to Table S1d and the references therein. The prediction model using all 97 variables will be referred to as the 'full' model, as opposed to a simplified model ('lite' model; see below) based on mainly demographic data and medical history that can be more readily obtained. Among the 97 variables, 21 were categorical and 76 were quantitative traits. The missing rates of variables were all below 20%. We included a wide range of clinical features here, with an objective to uncover potential novel risk factors for the disease. The ML model we employed (XGboost) tends to have a low bias and high variance, however with proper tuning of hyper-parameters and regularization, overfitting can be largely avoided even when a large number of predictors are included[27].

The full list of variables is shown in Table S1. Briefly, we included basic demographic variables (e.g. age, sex, ethnic group, socioeconomic status as indicated by the Townsend deprivation index), comorbidities (e.g. heart diseases, type 1 and 2 diabetes mellitus [T1DM/T2DM], hypertension [HT], asthma/chronic obstructive pulmonary disease [COPD], cancer, dementia and psychiatric disorders), indicators of general health (number of medications taken [cnt_tx], number of illnesses etc.), blood measurements (hematology, liver and renal function measures, metabolic parameters such as lipid levels, HbA1c etc.), anthropometric measures (e.g. waist circumference, waist-hip ratio, body mass index[BMI] etc.) and lifestyle risk factors (e.g. smoking, drinking habits etc.). For disease traits, they were defined based on ICD-10 diagnoses (UKBB data-field 41270), self-reported illnesses (UKBB data-field 20002) and data from follow-ups. Subjects with no records of the relevant disease from either self-reports or ICD-10 diagnoses were regarded as having no history of the disease.

Imputation

Missing values of remaining features were imputed with the R package missRanger. The program is based on missForest[28], which is an iterative imputation approach based on random forest (RF). It has been widely used and has been shown to produce low imputation errors and good performance in predictive models[29]. The main difference between missRanger and missForest is that the former uses the R package 'ranger' to build RFs, which can lead to a large improvement in speed. Predictive mean matching (pmm) was also employed to avoid imputation with values not present in the original data. We employed the default parameters (pmm.k = 3, num.trees = 100) and default settings of ranger. Out-of-bag errors (in terms of classification errors or normalized root-mean-squared error) were computed which provides a guide to imputation accuracy.

We have also attempted to use MICE (Multiple Imputation by Chained Equation) for imputation. For our dataset with ~500K subjects, MICE stopped after running for 6 hours due to memory overflow error (>64GB), whereas missRanger finished the imputation within 3 hours successfully. We considered the computational

burden of MICE as too high and therefore employed missRanger in our analyses.

Several studies have compared MissForest with MICE, and there are several advantages of missForest. For categorical variables, imputation accuracy of missForest is likely to be higher than MICE[30]. MissForest also runs considerably faster than MICE and is especially suitable for imputation settings where complex interactions and non-linear relationships are likely[28]. Stekhoven et al.[28] reported superior performance of missForest compared to MICE, with reduction in the proportion of falsely classified entries (PFC) of up to 60%. In another comparison study, missForest and MICE performed similarly but it was reported that highly correlated variables may lead to significant problems with MICE[31].

XGboost prediction model

XGboost with gradient-boosted trees was employed for building prediction models. Analysis was performed by the R package 'xgboost'. We employed a 5-fold nested cross-validation strategy to develop and test the model. To avoid overoptimistic results due to choosing the best set of hyper-parameters based on test performance, the test sets were *not* involved in hyper-parameter tuning.

In each iteration, we divided the data into 5 folds, among which 1/5 was reserved for testing only. For the remaining 4/5 of the data, we further sampled 4/5 for training and 1/5 for hyper-parameter tuning. The best prediction model was applied to the test set. The process was repeated five times. A grid-search procedure was used to search for the best combination of hyper-parameters (e.g. tree depth, learning rate, regularization parameters for L1/L2 penalty etc.). The full range of hyper-parameters chosen or for grid-search is given in Table S6.

Building a simplified 'lite' model

The 'full' model described above covers a wide range of predictors but some features (such as blood biochemistries) may not be readily accessible. For easier implementation in practice, we also built a simplified prediction model (also referred to as the 'lite' model) based on a reduced set of 27 predictors. The reduced set of variables were chosen based on the ease of being assessed or measured, which included comorbidities (see above), anthropometric measures (BMI, weight, waist circumference), demographic variables (e.g. age, sex, ethnic group) and general indicators of health (number of medications taken, number of illnesses).

Evaluating predictive performance and calibration

To evaluate the predictive performance of the prediction models, we computed the area under the receiving operating characteristic curve (AUC-ROC), which is very widely used in clinical prediction studies. We also calculated other measures including the area under the precision-recall curve (AUC-PRC), F1-score, accuracy and Matthews correlation coefficient (MCC). The cutoff of predicted probability for calculating the latter three measures was determined by optimizing the geometric mean of sensitivity and specificity.

In addition to good ability to discriminate cases from non-cases, it is also important that the predicted event probabilities match with the observed probabilities (also known as calibration of a model). We assessed calibration by several measures, including the Hosmer–Lemeshow test, Expected Calibration Error (ECE) and Maximum Calibration Error (MCE) [32-35] across 10 equally-sized bins by discretizing the predicted probabilities. We also attempted three approaches to further improve calibration, including Platt scaling, isotonic regression and beta calibration [36-39]. The objective is to rescale the predicted probabilities such that they are closer to the actual probabilities of the outcome[40].

Identifying and quantifying the effects of important predictors

In this work we primarily employed Shapley value (ShapVal)[41, 42] to assess variable importance, which is a measure based on game theory to assess the contribution of each feature. ShapVal has been shown to represent a consistent and locally accurate contribution of each feature[43]. ShapVal enables local explanation of the model as they could be computed for each observation, but can also provide global importance measures. On the other hand, gain and split count may produce inconsistent estimates of global importance as shown in Lundberg et al [43].

Intuitively, the ShapVal of the *i*-th feature (for subject *k*) is the contribution of this feature to prediction of outcome for the individual, averaging over all possible orderings of the features (as the contribution may differ when variables enter the prediction algorithm in different orders). We ranked the global importance of features based on mean absolute ShapVal as described in previous works [41, 42]. We also attempted an alternative approach similar to 'permutation importance' proposed in ref[44]. This method involves permuting the outcome vector to model the distribution of ShapVal under the null, and comparing the null ShapVals with the observed ShapVal. We derived a p-value from permutation as an alternative indicator of feature importance. A total of 500 permutations were performed for each model. To verify the validity of the permutation procedure especially under imbalanced case-control data, we also carried out a small-scale simulation study. A dataset with 50,000 subjects and 10 covariates (x_1, x_2…x_{10}) were generated, where the first covariate x_1 was linearly correlated with the outcome. The control:case ratio was set at 976:1, same as that for cohort D. Type I error and power were assessed by repeating the entire permutation procedure for 100 randomly generated datasets (please see Supplementary Text for details).

A related index is the Shapley *interaction* value[42], which computes the difference in Shapley value of feature *i* with and without another feature *j*. ShapVal were averaged across 5 folds. Besides, we also included the 'gain' measure for reference, which is the reduction of loss or impurity contributed by all splits by a specific variable.

An advantage of Shapley value is that it is calculated for each individual, so how each risk factor affects a specific person's risk of infection/severity can be estimated as well. To illustrate this concept, we also produced decision plots for subjects at the highest, median and lowest risk of each cohort.

Cluster analysis based on Shapley value

We also performed cluster analysis based on ShapVal to identify subgroup of patients who share similar clinical risk factors with respect to severity of infection. As introduced in[45], this approach may be considered a form of 'supervised' clustering, as the outcome (severe/fatal disease) is also taken into account in the clustering process. Unlike a traditional clustering approach based on risk factors, this approach has important advantages. Firstly, the clusters derived may be more clinically relevant as the *outcome* is also considered, reducing the chance that irrelevant features contribute to the subgrouping. (An irrelevant feature will have relatively small variations in ShapVal and will not contribute substantially to clustering). Secondly, this approach essentially considers all features on the same 'scale', as ShapVal are computed with respect to the outcome. Input features are often of different units and scales and but ShapVal considers feature contributions to the outcome as the unit of measure. Due to computational cost concerns, here we only performed clustering on cohorts A (non-severe vs severe infection) and B (fatal vs non-fatal infection).

K-means sparse clustering

Here we performed k-means sparse clustering to uncover underlying patient subgroups based on ShapVal of risk factors. As the number of features included is large but not all may contribute to the underlying subgroups, we employed *sparse* clustering which incorporates feature selection in the clustering process. The R package "sparcl" was employed. To perform sparse k-means clustering, we need to predetermine the number of clusters and tuning parameter (L1 penalty) for feature selection[46]. The optimal number of clusters was assumed to be the same as that in k-means clustering, which was determined by the silhouette index. The tuning parameter (L1 bound) was set to range between 2 to 6 with an interval of 0.4. Then the gap statistic[47] was used to determine the optimal tuning parameter.

Results

An overview of the sample sizes in each set of analysis is presented in Table 1. Please also refer to Table S1a and S1b for a detailed summary of case counts and covariates.

Simulation results for the permutation testing approach

Simulation results for the validity of permutation p-values are presented in Table S8. We observed no inflation

of type I error (false positive rate) despite the imbalanced case:control ratio. At a *p*-value threshold of 0.05, the proportion of results with $p<0.05$ for x_2 *to* x_{10} (variables with null effect) remained less than 0.05 for different effect sizes of the predictor (please also refer to Supplementary Text).

Prediction performance of the XGboost model for risk and severity of infection

AUC-ROC and other results

We performed 5-fold CV and the average AUC under the ROC curve is given in Table 1 and Table S2. Here we describe the results for the full models first. We observed better predictive performances in cohorts B (fatal cases vs outpatient cases) and D (fatal cases vs population with no known infection) where fatalities from COVID-19 were modeled. The corresponding mean AUC-ROC were 0.814 (95% CI: 0.791-0.838) and 0.825 (CI 0.802-0.848) respectively. The mean AUC-ROC for cohort A (hospitalized/fatal cases vs other cases) was 0.723 (CI: 0.711 – 0.736) and that for cohort C (hospitalized/fatal cases vs population with no known infection) was 0.696 (CI: 0.684 - 0.708).

As for the 'lite' models which included a reduced set of predictors, the predictive performances in terms of AUC are broadly similar, with estimated AUC-ROC of 0.716, 0.818, 0.696 and 0.830 respectively.

The results of other predictive indices are listed in Table S2b. Estimates of AUC-PRC were the highest for cohorts A and B (0.535 and 0.171 respectively) and much lower for cohorts C and D (0.007 and 0.006 respectively). This is expected due to the much higher prevalence of outcome in the first two cohorts. AUC-PRC may be approximated by the average precision (please refer to [48] for further details).

We also conducted sex-stratified analysis (Table S2). The resulting AUC-ROC were similar to the overall analysis in males (except for cohort D), but were generally lower in females. This may be partially explained by lower number of severe and fatal cases in females, which leads to greater difficulty in model training.

Proportion of cases explained by individuals at the top k% of predicted risk

We also computed the proportion of cases explained by individuals at the highest *k*% of predicted risks (Table 2). For example, considering the full model, for prediction of mortality among infected individuals (cohort B), subjects at the highest 5%, 10% and 20% of predicted risks explain 17.4%, 32.7% and 56.2% of total fatalities respectively. As for prediction in the population (cohort D), subjects at the highest 5%, 10% and 20% of predicted risks explain 32.5%, 45.7% and 63.5% of total fatalities respectively. For prediction of severe disease among the infected (cohort A), subjects at the highest 5%, 10% and 20% of predicted risks explain 11.2%, 21.6% and 38.2% of total cases respectively, whilst more than half (53.3%) of cases are explained by people at the top 30% of predicted risks. For prediction of severe cases in the population (cohort C), the corresponding figures were 19.7%, 29.3% and 42.7% respectively, and more than half (52.8%) of cases are explained by people at the top 30% of predicted risks. Similar figures were observed for full and lite models in general.

These results showed in general a strong enrichment of cases among those predicted to have high risks, indicating good model discriminatory ability.

Relative risk of actual outcome probabilities, comparing those at the highest and lowest k% of predicted risks

We also computed the relative risk (RR) of infection or severe disease by comparing individuals at the highest and lowest *k*% of predicted risks (Table 2). For example, considering the full model, if we compare the actual probability of outcome at the top decile (top 10%) against those at the bottom decile of predicted risks, the RR was 4.74, 158.2, 6.98, 218.02 respectively for cohorts A to D. If we compare the top 20% against the lowest 20% of predicted risks, the corresponding RRs were 4.00, 22.42, 4.67, 30.30 respectively. The RRs for the lite model were similar for cohorts A and C, but were smaller for cohorts B and D when the comparison was made at the more extreme ends of predicted risks.

We observed large RRs for cohorts B and D, suggesting that the prediction models were able to discriminate individuals at the highest and lowest risks of fatality very well. RRs for cohorts B and D were much larger than those for cohorts A and C, indicating that the model predicted fatality better than severe disease.

Calibration

As for calibration, please refer to Figures S6-S7. For full models, cohort A was well-calibrated (without using other methods for re-calibration) with ECE of 0.022 and MCE of 0.044 only. For other models, the ECE and MCE were generally larger, probably due to large difficulty in calibration with a much lower probability of the

outcome. The ECEs (after re-calibration by one of the three methods) were 0.11, 0.14 and 0.02 respectively for cohorts B to D. H-L test was non-significant in cohorts C and D. For the 'lite' models, the ECEs were 0.017, 0.043, 0.024 and 0.089 respectively for cohorts A to D, with non-significant H-L test results except for cohort B.

Results from cluster analysis based on ShapVal

Figure 4 and Figure S11 show the results based on sparse k-means clustering. We performed clustering separately in cases and controls to uncover patient subgroups with different clinical background. Here we focus on clustering results within cases. As the number of variables is large, we only showed the variables that were statistically significant (p<0.05 from t-test or ANOVA) across the clusters in the figures. For cohort A, we found two clusters as the optimal solution. The first cluster has higher ShapVal for most risk factors, especially age, but also cnt_tx, HbA1c, cystatin C, HDL-C and HT. ShapVal for WHR was positive for the 1st group but negative for 2nd group. The first cluster may represent a subgroup of severe cases with a larger number of clinical risk factors/comorbidities and advanced age, while the second cluster may be a distinct group with less conventional risk factors (especially obesity), yet are susceptible to severe infections perhaps due to other (unmeasured) factors, such as genetics.

Considering cohort B cases (fatal infections), the optimal solution comprised three clusters. Interestingly, the 1st and 3rd cluster seemed to be markedly different with respect to their risk factor profiles. Mean ShapVal for age were largely negative for the 1st cluster but highly positive for the other two clusters. On the other hand, mean ShapVal for waist circumference was markedly higher and positive for the 1st cluster. The 3rd cluster was characterized by the highest mean ShapVal for age, and higher (positive) ShapVal for mainly cnt_tx, HbA1c and T2DM. The results suggest that there may exist pathophysiologically distinct subgroups of patients with fatal infection. The 1st cluster represents a subgroup with younger age but higher proportion of obesity. The 3rd cluster represents another subgroup with advanced age, more comorbidities and higher proportion of glucose abnormalities or T2DM. The 2nd cluster is in between.

Important contributing variables identified

Here we primarily report the results of the full model as a more complete set of predictors is included. The Shapley dependence plots (ranked by mean absolute ShapVal) of the top 15 features (full model) are shown in Figure 1 and those of top 6 features for the lite model are presented in Figure 2. For more complete plots (up to 30 variables) with ranking by mean abs(ShapVal) or permutation p-values, please refer to Figures S1-4.

Full ShapVal analysis results on all variables are given in Tables S3a-c. Top 10 variables (ranked by either ShpaVal or permutation p-value) from the full model are presented in Tables 3-4 while top 5 from the lite model are presented in Tables 5-6. We also included variable importance by gain and plots are presented in Figure S5a and S5b.

As for interaction analyses, top results are presented in Table 7 and full results in Tables S4-5. Plots are presented in Figure 3 (top 2 interacting pairs from each model) and Figures S8-9 (top 6 interacting pairs).

Note that ShapVal are measured on the log-odds scale. Every unit increase of ShapVal corresponds to an odds ratio (OR) of exp(1)=2.72. Positive ShapVal indicates increase in the odds of outcome and vice versa.

Cohort A (hospitalized/fatal cases vs outpatient cases)

The top 5 contributing features by ShapVal included age, number of medications received (cnt_tx), cystatin C, Townsend Deprivation index (TDI), waist-hip ratio (WHR), followed by HbA1c. Higher levels of these risk factors generally lead to higher disease severity among the infected. Interestingly, Shapley dependence plots revealed potential *non-linear* and 'threshold' effects of risk factors on the outcome. For example, age of ~65 or above was associated with a markedly increased risk of severe/fatal infection. Markedly elevated risks were also observed for HbA1c > ~40 mmol/mol and number of drugs received >= ~5. Impaired renal function (raised cystatin C above ~1 mg/L) was also linked to worse outcomes. For WHR, levels of ~0.9 or higher appeared to be associated with a marked increase in risks. For other features please also refer to Figure 1. We note that at the extreme ends of variables, the observations are often sparse so the trend shown by the loess curve may not be reliable (this also applies to other cohorts). Variable importance based on gain revealed

similar patterns of important features (Figure S5).

If we consider the 'p-value' or permutation importance (PermImp) measure, variables with top 10 (absolute) ShapVal also showed significant p-values. T2DM was among the top 10 by PermImp but not ShapVal. Depression and coronary artery disease (CAD) also showed low p-values (p<0.02) but were not listed among the top 30 by ShapVal.

Regarding interactions between variables, most of the top interacting pairs involved age (Figure 3, Tables S4-5). For example, younger individuals were observed to have more extreme ShapVal at similar ranges of cnt_tx. The effect of WHR on severity was more marked among the elderly, and the same was true for HDL cholesterol (low HDL is a risk factor).

Model B (fatal cases vs outpatient cases)
The top 5 contributing variables by ShapVal included age, testosterone (which may reflect the effect of gender), cnt_tx, waist circumference (WC) and red cell distribution width (RDW), which were followed by cystatin C, TDI, pulse rate, systolic blood pressure (SBP) and percentage of lymphocytes. Again certain non-linear and 'threshold' effects appeared to be present for many top-ranked features. For age, the risk for mortality was more marked beyond ~65. Higher levels of all the above risk factors (RFs) (except percentage of lymphocytes which showed a U-shaped relationship) were associated with higher mortality, but the effects were non-linear. Regarding the top results based on PermImp, 8 out of 10 predictors ranked high by ShapVal also had the lowest p-values (lowest p-value = 0.002 since we performed 500 permutations). Other top-ranked features (p-value=0.002) included HbA1c, type 1 and T2DM, weight, mean platelet volume etc.

Variable importance based on gain yielded similar results (Figure S5). As for interactions between the variables, again interactions were most prominent with age (Figure 3). For example, the effects of waist circumference and BMI (when exceeding a threshold of around 110 cm and 35 kg/m^2 respectively) on mortality were more prominent among younger individuals. The effects of testosterone and HbA1c however were more marked in older subjects.

Model C (hospitalized/fatal cases vs *population with no known infection)*
Based on ShapVal, WHR was the top contributing variable and WC was ranked 5th, suggesting central obesity may be a stronger predictor for severe disease than BMI alone (BMI was ranked 13th by ShapVal). Similar to before, TDI and age were ranked among the top. For age, slightly unexpectedly, a U-shaped curve was observed, which suggests lowest risk at the age group of ~65-70. Note that model C may also capture RFs related to susceptibility to infection. It is possible, for instance, that younger subjects had higher risks of exposure due to work or social interactions. Among the top 10, two are related to general multi-comorbidities (cnt_tx/cnt_noncancer). Increased cystatin C and lower apolipoprotien A were also associated with higher susceptibility to severe infections, and HT and T2DM were also among the top 10. Considering PermImp as the ranking criteria, COPD, depression and dementia were observed to have the lowest permutation p-values (p=0.002) though not top-listed by ShapVal.

Interaction plot (Figure 3) shows WHR may interact with age, with elderly individuals showing more prominent effects from changes in WHR.

Model D (fatal cases vs *population with no known infection)*
Based on ShapVal, age was the top feature, followed by TDI, WHR, number of drugs taken and WC. Other top features included cystatin C, testosterone, hypertension, RBC distribution width and pulse rate. Higher levels of these features (or presence of comorbidity) generally lead to higher mortality risks. Based on PermImp, T2DM, dementia and COPD were the most highly ranked (ignoring features that are already listed in the top 10 by ShapVal).

Shapley interaction analysis suggested that the top interacting pairs involved age and some of top contributing features (Figure 3). The effects of testosterone (likely also reflects gender effects) and TDI were more prominent among the elderly, while the effect of BMI was larger in the younger age groups. Also, the protective effect of having no hypertension was more marked in the younger age group (~50-65) but not in the

elderly (Figure S8).

As for important variables from the sex-stratified analysis, the top variables were similar which included e.g. age, WC/WHR, cystatin C, number of medications received, socioeconomic status (as reflected by TDI), among others (Table S3c).

PermImp compared to ShapVal

Overall speaking, the PermImp measure tends to rank binary traits higher than ShapVal. Of note, several diseases were consistently top-listed by PermImp across the 4 cohorts (though some were not highlighted by ShapVal), including CAD, AF, T2DM, dementia, which were among the top 10 in at least 3 cohorts in Table 4. Other diseases that were listed at least twice included depression, COPD, stroke and heart failure.

Results from the 'Lite' model

Here we highlight top contributing features for the 'lite' models consisting of 27 predictors (Table S3b). Remarkably, the top 3 features (ranked by ShapVal) were consistent across all four cohorts. These features included age, cnt_tx and WC (WHR was not included in the lite model as WC is easier to measure). Of note, sex and T2DM were ranked among the top 6 across all cohorts.

If we consider PermImp as the ranking criteria (further ranked by ShapVal if PermImp is equal), age, cnt_tx and WC were still highly ranked and listed among the top 5 in at least 3 cohorts (Table S3b). T2DM was ranked among the top 5 in all cohorts. Other potential risk factors included dementia (top 10 across 3 cohorts) as well as atrial fibrillation (AF), COPD and CAD (top 10 across 2 cohorts).

Results from the logistic model

As discussed above, we primarily focused on the XGboost ML model as it can capture non-linear relationships and interactions between predictors. Here we also performed our analyses with logistic regression (LR) for comparison. For prediction performance (Table S7), the AUC-ROC of the full LR model were 0.728 (CI: 0.715–0.741), 0.810 (CI 0.786-0.834), 0.712 (CI 0.701 – 0.724) and 0.833 (CI 0.810-0.856) respectively for cohorts A to D. For the 'lite' model (using 27 predictors only), the AUC-ROC of the LR approach were 0.722 (CI 0.709-0.735), 0.824 (CI 0.801-0.848), 0.697 (CI 0.685-0.709) and 0.834 (CI 0.812-0.857) respectively (Table S7a). These figures were very close to those obtained by XGboost, although AUC-ROC using LR were slightly higher in general (median difference = 0.005). If we compute the relative risk of subjects at the highest and lowest k% of predicted risks, the results were generally similar (Table S7b). For cohort D and the full model of cohort B, XGboost performed better than LR at the extreme ends of predicted risks, with observed risk=0 (i.e. no cases were observed) for those predicted at the lowest 5% of risk (Table 2).

While prediction is one of our goals, uncovering important contributing factors and their relationship to COVID-19 severity is a major objective of this study. In fact, the latter is considered our primary objective when considering the analyses within infected patients (cohorts A and B). As LR assumes linearity on a log-odds scale, it could not capture non-linear relationships or 'threshold effects' of variables on disease severity.

Individual Shapley decision plots and online calculator

We also showed individual Shapley decision plot for three subjects with the highest, median and lowest predicted risks in each cohort (Fig S10). The y-axis is based on a log-odds scale.

To facilitate further research and studies on risk prediction models, we also constructed an online risk calculation tool (for 'lite' model) at https://labsocuhk.ddns.net:8890/covid19/ . The online tool can also construct a Shapley decision plot based on individual risk factors.

Discussions

In this study we have performed four sets of analysis, predicting severe or fatal COVID-19 infection among affected individuals or in the population. We observed good predictive power from the XGboost ML models, especially for the prediction of mortality. We also identified risk factors for increased severity or mortality, and uncovered possible non-linear effects of some features, which may be clinically relevant and shed light on disease mechanisms.

Prediction of severity/mortality

In general, our prediction models achieved reasonably good predictive power. The models predicted mortality (AUC ~81-83%) better than severity of disease. As discussed earlier, in the absence of better alternatives, hospitalization (test performed as inpatient) was used as a proxy for severity. However, reasons or criteria for hospitalization may vary across individuals or hospitals, and some tests may be performed in in-patients for surveillance or due to other confirmed/suspected cases in the ward. As a result, hospitalized patients could also include some with mild or moderate illnesses, which may also impair the prediction performance. On the other hand, mortality from infection is a more objective outcome. Other studies (e.g.[49-51]) have also defined 'severe' or 'critical' disease based on intensive care unit (ICU) admission and/or need for ventilatory support. However, we could not find sufficiently detailed clinical data to support such a classification at the time of this analysis.

Discriminatory power of the models and clinical implications

By assessing the proportion of cases explained by those at the top k% of predicted risks, we observed in general a strong enrichment of cases among those with high predicted risks, indicating good discriminative ability of the models and suggesting the possibility to focus on the highest-risk group for targeted preventions or treatment. Similar strong enrichment was also observed for the lite model with fewer predictors. We also observed large relative risks of the actual outcomes when comparing subjects at high vs low percentiles of predicted risks. For example, for the prediction of mortality among the infected, the RR was up to 158 times (~20% vs 0.1%) when comparing and top and bottom deciles using the full model, and 28.38 times when considering the simplified model (~21% vs 0.8%). These results suggest that the prediction models may be used for risk stratification and prioritizing those at higher risks of deterioration, for early medical attention or admission. As the 'lite' model only relies on demographic data and information on comorbidities, risk stratification may be conducted even at the start of the illness without other blood or imaging results.

Previous relevant works

A number of studies have focused on prediction of severity/mortality of COVID-19 (corresponding to our prediction in cohorts A and B) and were reviewed in ref [6]. For cohort A (prediction of severity among infected), the AUC is 72.3%, which is moderate but not as good as many previous ML models for severity prediction[6]. The AUC for prediction of mortality is much higher (AUC = 81.4%), although we noted some studies have reported higher predictive power from clinical symptoms, blood biochemistry on admission and imaging features[6]. We understand that without access to the above features, predictive performance may be inferior. On the other hand, due to heterogeneity of clinical samples, treatment approaches, model evaluation methods and other features across studies, direct comparisons of predictive performance across studies may be difficult. Here we are not aimed at deriving a highly accurate prediction model; the main purpose is to identify general or 'baseline' risk factors for severe disease, thereby gaining insight into disease pathophysiology. However, we also showed that such clinical features or blood measurements, even when collected much earlier in time, may still be highly predictive of outcomes and hence may be incorporated into existing prediction algorithms. The models here may also be useful when blood results or imaging are not available (e.g. before admission) and the goal is to quickly classify a patient's risk.

For cohorts C and D, the general population (with no known infection) was treated as 'controls'. Compared to cohorts A and B, the identified risk factors may also increase the overall susceptibility to infection. The AUC for cohort C (severe/fatal disease) is ~70% but is much higher when mortality is considered as the outcome (AUC~83%). To our knowledge, there are still very few predictive models built at a *general population level* to identify susceptible individuals; this work is among the first to employ an ML approach to risk prediction of COVID-19/severe infection at a population level. DeCaprio et al.[52] proposed an ML model to assess the vulnerability to COVID-19 in the population. However, due to limited data, no actual COVID-19 patients were included and 'proxy' outcomes were used instead. Models were built from mainly demographic and comorbidity data to predict hospitalization due to acute respiratory distress syndrome, pneumonia, influenza, acute bronchitis and other respiratory tract infections.

Another very recent study ('QCOVID' study) from the UK[53] utilized general practice records from 6.08 million adults (age 19 to 100) as derivation cohort and 2.17 million adults as the validation set. Mortality from COVID-19 was the primary outcome and a survival model (sub-distribution hazard model)[54] was used to

predict mortalities. The predictors included demographic (e.g. age, TDI, ethnicity), lifestyle (e.g. BMI, smoking) and a large range of comorbid conditions. The resulting Harrell's C (comparable to AUC) was 0.928. However, we note that the QCOVID study included subjects of a much younger age range (19 or above), which will improve predictive performance as age is by far the most important predictor of mortality, with markedly reduced risks in younger subjects. For example, if we refer to age-specific predictive performance (Supplementary Table C of the paper), Harrell's C for mortality were 0.678, 0.831, 0.812 and 0.814 in 50-59, 60-69, 70-79 and 80+ year olds respectively, for males in the first follow-up period (24 Jan to 30 Apr 2020). For females, the corresponding numbers were 0.618, 0.77, 0.866 and 0.821. These numbers reflect lower predictive power when restricted to a narrower age range. One main difference between the present work and the above study is that we employed an *XGboost machine learning* approach which is able to capture also non-linear and more complex interaction effects. As shown in our Shapley dependence plots, the models were able to reveal non-linear effects in a data-driven manner. We also included a number of blood measurements to shed light on potential new risk factors and mechanisms underlying the disease. The QCOVID study employed a survival model (sub-distribution hazard) that accounts for time-to-event and competing risks; however, the proportional hazards assumption is required which may not hold due to restrictions/interventions introduced during the period (i.e. time-dependent associations may be present).

A few other studies have investigated risk factors (especially comorbidities) for COVID-19 infection in the UKBB. For example, Atkins et al. [10] studied elderly subjects (age>65) in UKBB, and found that hypertension, history of falls, CAD, T2DM and asthma as the top comorbidities among hospitalized cases. The analysis was restricted to the elderly population however. In a more recent work, McQueenie et al. [11] studied multi-comorbidities and polypharmacy on infection risks. Having >=2 long-term conditions, cardiometabolic disorders and polypharmacy were associated with heightened risks of infection. Among individuals with multi-comorbidities, severe obesity and impaired renal function may lead to increased risks. Another study of primary care patients in the UK revealed that deprivation, male, older age, ethnicity (being black) and chronic renal disease were associated with higher risks of being tested positive. Another large-scale British primary care study of more than 17 million subjects revealed similar risk factors as above [55]. There is also a relatively large literature on the study of risk factors associated with severe or fatal disease [13-16, 56-59]. Some commonly reported risk factors included age, sex, obesity, diabetes, hypertension, renal, cardiometabolic and respiratory disorders. As discussed above, an important difference between the above epidemiological studies and the current work is that we employed XGboost, an ML approach that can uncover non-linear and interaction effects, while other studies mostly employed regression models that assumes linear and additive effects of covariates. We also performed a comprehensive analysis including four models covering different outcomes and both infected and population cohorts.

Comparison with logistic model

We have performed logistic regression to compare with XGboost on cohorts A and B. The differences in predictive performance appeared to be small. As the number of cases (especially fatalities) is relatively small in this dataset, this may limit the predictive performance of more complex models like XGboost, which may be expected to improve with larger case numbers. An important advantage of XGboost is that it can detect non-linear relationships when compared to LR. In addition, XGboost may handle multi-collinearity better than LR. Assuming two highly correlated features A and B, for each specific tree usually only one variable will be used and as the trees are sequential, the focus of the model will be usually on one but not both features [60]. Hence XGboost also handles multi-collinearity well, which is important here as many clinical variables are correlated. XGboost also directly models interaction between variables. It is much more difficult for LR to model interactions due to the rapid increase in feature space when interaction terms are included.

Highlights of potential risk factors

For the limit of space, we shall only highlight the top 5-10 risk factors ranked by ShapVal here. Across the four cohorts, age and cardiometabolic risk factors predominate the top risk factors. Age and WHR/WC was ranked among top 5 across all four cohorts. The number of medications taken was among top 5 across all cohorts, and cystatin C (reflecting renal function) was among the top 10 across all cohorts. HbA1c was a top-10 risk factor for cohort A, and T2DM was also highly ranked across multiple cohorts especially when PermImp was considered. Townsend deprivation index (reflecting socioeconomic status) was among the top

10 in most cohorts. As described above, results from the 'lite' models were generally in line with those from the full models, with age, WC and cnt_tx consistently ranked as the top 3.

Obesity has been observed to be a major risk factor for susceptibility or severity of infection in the UK Biobank [12, 61], and in many other studies[62, 63]. The observation that waist circumference/WHR were highly ranked suggests that *central* obesity is a major risk factor and may be a better predictor of severity than BMI alone.

Another major risk factor we identified is impaired renal function (IRF), as reflected by elevated risks with raised urea and cystatin C. Several studies also suggested IRF increases risk of mortality [59, 64, 65], although it is probably not as widely recognized as cardiometabolic disorders as a major risk factor. Since COVID-19 itself may lead to renal failure, our findings specifically suggest that underlying or baseline IRF is an important risk factor. The high ranking of cystatin C also indicates this measure may better reflect renal function than urea or creatinine (which were also included in our analysis) [66, 67], and may serve as a superior predictor for COVID-19 severity.

Other potential risk factors briefly highlighted below were less reported. As some were listed only once or twice among the top 10, and for some their ShapVal were close to other risk factors, further replications are required. For example, testosterone was top-ranked by XGboost (for mortality), with higher levels associated with increased risk. This may partially reflect that males are at a higher risk of fatal infections, but it remains to be studied whether testosterone itself is involved in the pathophysiology of severe COVID-19, as the ML model chose this variable instead of sex. Studies have suggested elevated or reduced testosterone levels may be both associated with a more severe clinical course[68]. Also, interestingly, 5-alpha-reductase inhibitors or androgen-deprivation therapy have been shown to be associated with a lower risk or severity of disease [69, 70]. We also found a few hematological indices that may be potential risk factors. High red cell distribution width (RDW) was associated with mortality in our study and was also identified in a recent meta-analysis of three studies as a risk factor [71]. Low lymphocyte percentage was a top-10 risk factor in cohort B, which may be related to immune functioning and response to infections. Lymphopenia was reported as a main hematological finding in those with severe illnesses [35, 72]. Most previous studies considered hematological indices at admission or during hospitalization. Slightly surprisingly, this study suggested that high RDW or reduced lymphocyte percentage *prior to the diagnosis* may also be predictive of worse outcomes.

Comorbid diseases associated with severity as highlighted by PermImp

Among the diseases being included as covariates, T2DM is most consistently ranked among the top, no matter in the full or lite models, and regardless of ranking by ShapVal or PermImp (p-value). T2DM has been shown in numerous studies to be associated with higher risk and severity of infection[73, 74]. We noted some discrepancy between the ranked results based on ShapVal and those based on PermImp. In general, the latter measure favors binary variable while ShapVal alone tends to rank continuous variables higher. We are unsure about the exact reason but it may be an interesting topic for further methodology studies. If we employed a composite ranking criteria based on PermImp then by ShapVal (if equal PermImp), then a few more diseases were ranked among the top 10, such as hypertension and COPD. For cohort D, T2DM, dementia, COPD, AF, heart failure and CAD were also top-ranked, suggesting that a range of chronic cardiovascular, respiratory and neuropsychiatric conditions may be associated with increased mortality.

Full and lite prediction models

We note that the simplified ('lite') prediction model has very similar predictive performance (as assessed by AUC) to the 'full' model with a larger panel of predictors. However, it is important to note that features associated with the outcome may not always improve predictive power. AUC is relatively insensitive to detecting changes in predictive performance when additional risk factors are added [75-77].

For example, Pencina et al. [75] showed that in the prediction of CVD risk in the Women's Health Study, adding extra established risk factors often result in minimal improvements in AUC. For instance, in a model with age, SBP and smoking, adding any lipid measures result in only an increase of 0.01 in AUC from the

baseline of 0.76. In the same vein, starting from a full prediction model [containing Ln(age), Ln(SBP), smoking, Ln(Total cholesterol), Ln(HDL)], deleting any one of these established risk factors (except age) resulted in a very small reduction of AUC of <0.02. In general, for a model with high baseline AUC from existing predictors (e.g. age, sex and obesity in the case of COVID-19), including additional predictors may not result in much improvement in discriminative power or AUC[78].

Nevertheless, it is still valuable to study variable importance (e.g. ShapVal) from ML model as they may shed light on the pathophysiology of the disease. For example, many factors such as age and T2DM may lead to poorer renal function (and higher cystatin C), which in turn may increase the severity of infection. Given that age, T2DM and other main comorbidities are already modeled, adding cystatin C may not improve discriminative power of the model. However, its inclusion may still change the predicted probability of outcome, which will be reflected in ShapVal. The high ranking of cystatin C (based on ShapVal) may shed light on renal impairment as a potential mechanism associated with clinical deterioration.

Some limitations have been discussed above, for example the use of hospitalization as a proxy for severity, and that the predictors were recorded prior to the pandemic. We briefly discuss other limitations here. The UK biobank is a very large-scale study with detailed phenotypic data, but still the number of fatal cases is relatively small. Also, the UKBB is not entirely representative of the UK population, as participants tend to be healthier and wealthier overall[79]. Also, it remains to be studied whether the findings are generalizable to other populations. Symptom measures and lung imaging features were not available at the time of analysis. Despite adjusting for a rich set of predictors and that all predictors were recorded prior to the outbreak, causality cannot be confirmed from this study, due to risk of residual confounding by unknown factors. The current study was performed on a cohort with age>50 and generalizability to younger individuals remains to be studied. In cohorts C and D, the population with no known infection was regarded as controls. It is expected that some may become infected in the future, and some may have been infected but not tested; however, the chance of missing cases of severe infection is probably not high. Since the UKBB represents a relatively healthy population with low rate of severe COVID-19 cases so far (~0.5%), we expect the use of ‘unscreened’ controls is unlikely to result in substantial bias.

Regarding the ML model, XGboost is a state-of-the-art method that has been consistently shown to the best or one of the best ML methods in supervised learning tasks/competitions [80] (especially for tasks not involving computer vision or natural language processing). Nevertheless, other ML methods may still be useful or may uncover novel risk factors. Assessing variable importance is a long-standing problem in ML; here we mainly employed ShapVal which is both computationally fast and was shown to have good theoretical properties[41, 42].

Conclusions

In conclusion, we identified a number of baseline risk factors for severe/fatal infection by an ML approach. Shapley dependence plots revealed possible non-linear and ‘threshold’ effects of risk factors on the risks of infection or severity. To summarize, age, central obesity, impaired renal function, multi-comorbidities, cardiometabolic abnormalities or disorders (especially T2DM) and low socioeconomic status may predispose to poorer outcomes, among other risk factors. The prediction models (of cohorts C/D) may be useful at a population level to identify those susceptible to developing severe/fatal infections, hence facilitating targeted prevention strategies. Further replication and validation in independent cohorts are required to confirm our findings.

Supplementary Materials are available at https://drive.google.com/drive/folders/17O1PUd1tE5gk_lFiXUyrn2vz7KUb_fRM?usp=sharing or https://doi.org/10.6084/m9.figshare.14833104

An online risk calculation tool is available at https://labsocuhk.ddns.net:8890/covid19/

Author Contributions

Conception and design: HCS. Analytic methodology: HCS and KCYW. Data analysis: KCYW (main), YX, LY, HCS. Interpretation: HCS and KCYW. Supervision of study: HCS. Drafting of manuscript: HCS, with input from KCYW.

Acknowledgements

This work was supported partially by the Lo Kwee Seong Biomedical Research Fund from The Chinese University of Hong Kong, and the KIZ-CUHK Joint Laboratory of Bioresources and Molecular Research of Common Diseases, Kunming Institute of Zoology and The Chinese University of Hong Kong. We thank Prof. Pak Sham for support on data access and analyses, and Ms Qiu Jinghong for formatting and editing of the manuscript.

Conflicts of interest

The authors declare no conflict of interest.

References

1. Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, et al. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia. The New England journal of medicine. 2020 Jan 29. PMID: 31995857. doi: 10.1056/NEJMoa2001316.
2. Novel-Coronavirus-Pneumonia-Emergency-Response-Epidemiology-Team. [The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (COVID-19) in China]. Zhonghua liu xing bing xue za zhi = Zhonghua liuxingbingxue zazhi. 2020 Feb 17;41(2):145-51. PMID: 32064853. doi: 10.3760/cma.j.issn.0254-6450.2020.02.003.
3. Guan W-j, Ni Z-y, Hu Y, Liang W-h, Ou C-q, He J-x, et al. Clinical Characteristics of Coronavirus Disease 2019 in China. New England Journal of Medicine. 2020. doi: 10.1056/NEJMoa2002032.
4. 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 (Clinical research ed). 2020;370:m3339-m. PMID: 32907855. doi: 10.1136/bmj.m3339.
5. Chung H, Ko H, Kang WS, Kim KW, Lee H, Park C, et al. Prediction and Feature Importance Analysis for Severity of COVID-19 in South Korea Using Artificial Intelligence: Model Development and Validation. Journal of medical Internet research. 2021;23(4):e27060.
6. Wynants L, Van Calster B, 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. doi: 10.1136/bmj.m1328.
7. Alballa N, Al-Turaiki I. Machine learning approaches in COVID-19 diagnosis, mortality, and severity risk prediction: A review. Informatics in Medicine Unlocked. 2021 2021/01/01/;24:100564. doi: https://doi.org/10.1016/j.imu.2021.100564.
8. Adamidi ES, Mitsis K, Nikita KS. Artificial intelligence in clinical care amidst COVID-19 pandemic: A systematic review. Comput Struct Biotechnol J. 2021;19:2833-50. PMID: 34025952. doi: 10.1016/j.csbj.2021.05.010.

9. Vaid A, Somani S, Russak AJ, De Freitas JK, Chaudhry FF, Paranjpe I, et al. Machine Learning to Predict Mortality and Critical Events in a Cohort of Patients With COVID-19 in New York City: Model Development and Validation. J Med Internet Res. 2020 Nov 6;22(11):e24018. PMID: 33027032. doi: 10.2196/24018.
10. Atkins JL, Masoli JAH, Delgado J, Pilling LC, Kuo C-L, Kuchel GA, et al. Preexisting Comorbidities Predicting COVID-19 and Mortality in the UK Biobank Community Cohort. The Journals of Gerontology: Series A. 2020. doi: 10.1093/gerona/glaa183.
11. McQueenie R, Foster HME, Jani BD, Katikireddi SV, Sattar N, Pell JP, et al. Multimorbidity, polypharmacy, and COVID-19 infection within the UK Biobank cohort. PLOS ONE. 2020;15(8):e0238091. doi: 10.1371/journal.pone.0238091.
12. Yates T, Razieh C, Zaccardi F, Davies MJ, Khunti K. Obesity and risk of COVID-19: analysis of UK biobank. Prim Care Diabetes. 2020;14(5):566-7. PMID: 32493608. doi: 10.1016/j.pcd.2020.05.011.
13. Rod JE, Oviedo-Trespalacios O, Cortes-Ramirez J. A brief-review of the risk factors for covid-19 severity. Rev Saude Publica. 2020;54:60. PMID: 32491116. doi: 10.11606/s1518-8787.2020054002481.
14. Romero Starke K, Petereit-Haack G, Schubert M, Kampf D, Schliebner A, Hegewald J, et al. The Age-Related Risk of Severe Outcomes Due to COVID-19 Infection: A Rapid Review, Meta-Analysis, and Meta-Regression. Int J Environ Res Public Health. 2020 Aug 17;17(16). PMID: 32824596. doi: 10.3390/ijerph17165974.
15. Wingert A, Pillay J, Gates M, Guitard S, Rahman S, Beck A, et al. Risk factors for severe outcomes of COVID-19: a rapid review. medRxiv. 2020:2020.08.27.20183434. doi: 10.1101/2020.08.27.20183434.
16. Wolff D, Nee S, Hickey NS, Marschollek M. Risk factors for Covid-19 severity and fatality: a structured literature review. Infection. 2020 2020/08/28. doi: 10.1007/s15010-020-01509-1.
17. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, 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 Medicine. 2015;12(3):e1001779. doi: 10.1371/journal.pmed.1001779.
18. Lopez Bernal J, Andrews N, Gower C, Robertson C, Stowe J, Tessier E, et al. Effectiveness of the Pfizer-BioNTech and Oxford-AstraZeneca vaccines on covid-19 related symptoms, hospital admissions, and mortality in older adults in England: test negative case-control study. BMJ. 2021;373:n1088. doi: 10.1136/bmj.n1088.
19. Tenforde MW. Effectiveness of Pfizer-BioNTech and Moderna Vaccines Against COVID-19 Among Hospitalized Adults Aged≥ 65 Years—United States, January–March 2021. MMWR Morbidity and mortality weekly report. 2021;70.
20. Dagan N, Barda N, Kepten E, Miron O, Perchik S, Katz MA, et al. BNT162b2 mRNA Covid-19 Vaccine in a Nationwide Mass Vaccination Setting. New England Journal of Medicine. 2021 2021/04/15;384(15):1412-23. doi: 10.1056/NEJMoa2101765.
21. Haas EJ, Angulo FJ, McLaughlin JM, Anis E, Singer SR, Khan F, et al. Impact and effectiveness of mRNA BNT162b2 vaccine against SARS-CoV-2 infections and COVID-19 cases, hospitalisations, and deaths following a nationwide vaccination campaign in Israel: an observational study using national surveillance data. The Lancet. 2021;397(10287):1819-29. doi: 10.1016/S0140-6736(21)00947-8.
22. Dashti H, Roche EC, Bates DW, Mora S, Demler O. SARS2 simplified scores to estimate risk of hospitalization and death among patients with COVID-19. Scientific Reports. 2021 2021/03/02;11(1):4945. doi: 10.1038/s41598-021-84603-0.
23. Karaismailoglu E, Karaismailoglu S. Two novel nomograms for predicting the risk of hospitalization or mortality due to COVID-19 by the naïve Bayesian classifier method. Journal of Medical Virology. 2021 2021/05/01;93(5):3194-201. doi: https://doi.org/10.1002/jmv.26890.
24. Hamer M, Kivimäki M, Gale CR, Batty GD. Lifestyle risk factors, inflammatory

mechanisms, and COVID-19 hospitalization: A community-based cohort study of 387,109 adults in UK. Brain, Behavior, and Immunity. 2020 2020/07/01/;87:184-7. doi: https://doi.org/10.1016/j.bbi.2020.05.059.
25. Pairo-Castineira E, Clohisey S, Klaric L, Bretherick AD, Rawlik K, Pasko D, et al. Genetic mechanisms of critical illness in COVID-19. Nature. 2021 2021/03/01;591(7848):92-8. doi: 10.1038/s41586-020-03065-y.
26. Walsh KA, Jordan K, Clyne B, Rohde D, Drummond L, Byrne P, et al. SARS-CoV-2 detection, viral load and infectivity over the course of an infection. Journal of Infection. 2020 2020/09/01/;81(3):357-71. doi: https://doi.org/10.1016/j.jinf.2020.06.067.
27. Lever J, Krzywinski M, Altman N. Points of significance: model selection and overfitting. Nature Publishing Group; 2016.
28. Stekhoven DJ, Bühlmann P. MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics. 2012;28(1):112-8.
29. Waljee AK, Mukherjee A, Singal AG, Zhang Y, Warren J, Balis U, et al. Comparison of imputation methods for missing laboratory data in medicine. BMJ Open. 2013;3(8):e002847. doi: 10.1136/bmjopen-2013-002847.
30. Shah AD, Bartlett JW, Carpenter J, Nicholas O, Hemingway H. Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study. American journal of epidemiology. 2014;179(6):764-74.
31. Penone C, Davidson AD, Shoemaker KT, Di Marco M, Rondinini C, Brooks TM, et al. Imputation of missing data in life-history trait datasets: which approach performs the best? Methods in Ecology and Evolution. 2014 2014/09/01;5(9):961-70. doi: https://doi.org/10.1111/2041-210X.12232.
32. Crowson CS, Atkinson EJ, Therneau TM. Assessing calibration of prognostic risk scores. Stat Methods Med Res. 2016 Aug;25(4):1692-706. PMID: 23907781. doi: 10.1177/0962280213497434.
33. Nixon J, Dusenberry MW, Zhang L, Jerfel G, Tran D, editors. Measuring Calibration in Deep Learning. CVPR Workshops; 2019.
34. Demler OV, Paynter NP, Cook NR. Tests of calibration and goodness-of-fit in the survival setting. Stat Med. 2015 May 10;34(10):1659-80. PMID: 25684707. doi: 10.1002/sim.6428.
35. Tan L, Wang Q, Zhang D, Ding J, Huang Q, Tang Y-Q, et al. Lymphopenia predicts disease severity of COVID-19: a descriptive and predictive study. Signal Transduction and Targeted Therapy. 2020 2020/03/27;5(1):33. doi: 10.1038/s41392-020-0148-4.
36. Song H, Diethe T, Kull M, Flach P. Distribution calibration for regression. In: Kamalika C, Ruslan S, editors. Proceedings of the 36th International Conference on Machine Learning; Proceedings of Machine Learning Research: PMLR; 2019. p. 5897--906.
37. Niculescu-Mizil A, Caruana R. Predicting good probabilities with supervised learning. Proceedings of the 22nd international conference on Machine learning; Bonn, Germany: Association for Computing Machinery; 2005. p. 625–32.
38. Jiang X, Osl M, Kim J, Ohno-Machado L. Smooth isotonic regression: a new method to calibrate predictive models. AMIA Jt Summits Transl Sci Proc. 2011;2011:16-20. PMID: 22211175.
39. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. Eur Heart J. 2014 Aug 1;35(29):1925-31. PMID: 24898551. doi: 10.1093/eurheartj/ehu207.
40. Niculescu-Mizil A, Caruana R, editors. Obtaining Calibrated Probabilities from Boosting. UAI; 2005.
41. Lundberg SM, Lee S-I, editors. A unified approach to interpreting model predictions. Advances in neural information processing systems; 2017.
42. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From Local Explanations to Global Understanding with Explainable AI for Trees. Nature machine intelligence. 2020 Jan;2(1):56-67. PMID: 32607472. doi: 10.1038/s42256-019-0138-9.

43. Lundberg S, Erion G, Lee S-I. Consistent Individualized Feature Attribution for Tree Ensembles. 2018 02/11.
44. Altmann A, Toloşi L, Sander O, Lengauer T. Permutation importance: a corrected feature importance measure. Bioinformatics. 2010;26(10):1340-7. doi: 10.1093/bioinformatics/btq134.
45. Lundberg SM, Erion GG, Lee S-I. Consistent individualized feature attribution for tree ensembles. arXiv preprint arXiv:180203888. 2018.
46. Witten DM, Tibshirani R. A framework for feature selection in clustering. Journal of the American Statistical Association. 2010 Jun 1;105(490):713-26. PMID: 20811510. doi: 10.1198/jasa.2010.tm09415.
47. Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2001;63(2):411-23.
48. Boyd K, Eng KH, Page CD, editors. Area under the Precision-Recall Curve: Point Estimates and Confidence Intervals. 2013; Berlin, Heidelberg: Springer Berlin Heidelberg.
49. Subudhi S, Verma A, Patel AB, Hardin CC, Khandekar MJ, Lee H, et al. Comparing machine learning algorithms for predicting ICU admission and mortality in COVID-19. npj Digital Medicine. 2021 2021/05/21;4(1):87. doi: 10.1038/s41746-021-00456-x.
50. Patrício A, Costa RS, Henriques R. Predictability of COVID-19 Hospitalizations, Intensive Care Unit Admissions, and Respiratory Assistance in Portugal: Longitudinal Cohort Study. J Med Internet Res. 2021 2021/4/28;23(4):e26075. doi: 10.2196/26075.
51. Yun K, Lee JS, Kim EY, Chandra H, Oh B-L, Oh J. Severe COVID-19 Illness: Risk Factors and Its Burden on Critical Care Resources. Frontiers in Medicine. 2020 2020-November-19;7(767). doi: 10.3389/fmed.2020.583060.
52. DeCaprio D, Gartner JA, Burgess T, Kothari S, Sayed S, McCall CJ. Building a COVID-19 Vulnerability Index. medRxiv. 2020:2020.03.16.20036723. doi: 10.1101/2020.03.16.20036723.
53. Clift AK, Coupland CAC, Keogh RH, Diaz-Ordaz K, Williamson E, Harrison EM, 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. 2020;371:m3731. doi: 10.1136/bmj.m3731.
54. Barda N, Riesel D, Akriv A, Levy J, Finkel U, Yona G, et al. Developing a COVID-19 mortality risk prediction model when individual-level data are not available. Nature Communications. 2020 2020/09/07;11(1):4439. doi: 10.1038/s41467-020-18297-9.
55. Williamson EJ, Walker AJ, Bhaskaran K, Bacon S, Bates C, Morton CE, et al. Factors associated with COVID-19-related death using OpenSAFELY. Nature. 2020 2020/08/01;584(7821):430-6. doi: 10.1038/s41586-020-2521-4.
56. Noor FM, Islam MM. Prevalence and Associated Risk Factors of Mortality Among COVID-19 Patients: A Meta-Analysis. Journal of community health. 2020 Sep 12. PMID: 32918645. doi: 10.1007/s10900-020-00920-x.
57. Rahman A, Sathi NJ. Risk Factors of the Severity of COVID-19: a Meta-Analysis. medRxiv. 2020:2020.04.30.20086744. doi: 10.1101/2020.04.30.20086744.
58. Zhou Y, Chi J, Lv W, Wang Y. Obesity and diabetes as high-risk factors for severe coronavirus disease 2019 (Covid-19). Diabetes/Metabolism Research and Reviews. 2020;n/a(n/a):e3377. doi: 10.1002/dmrr.3377.
59. Harrison SL, Fazio-Eynullayeva E, Lane DA, Underhill P, Lip GYH. Comorbidities associated with mortality in 31,461 adults with COVID-19 in the United States: A federated electronic medical record analysis. PLoS Med. 2020 Sep;17(9):e1003321. PMID: 32911500. doi: 10.1371/journal.pmed.1003321.
60. Chen T, He T, Benesty M, Tang Y. Understand Your Dataset with Xgboost. CRAN. Retrieved June; 2020.
61. Hamer M, Gale CR, Kivimäki M, Batty GD. Overweight, obesity, and risk of hospitalization

for COVID-19: A community-based cohort study of adults in the United Kingdom. Proceedings of the National Academy of Sciences. 2020;117(35):21011-3. doi: 10.1073/pnas.2011086117.
62. Popkin BM, Du S, Green WD, Beck MA, Algaith T, Herbst CH, et al. Individuals with obesity and COVID-19: A global perspective on the epidemiology and biological relationships. Obesity reviews : an official journal of the International Association for the Study of Obesity. 2020 Aug 26. PMID: 32845580. doi: 10.1111/obr.13128.
63. Tamara A, Tahapary DL. Obesity as a predictor for a poor prognosis of COVID-19: A systematic review. Diabetes & metabolic syndrome. 2020 Jul - Aug;14(4):655-9. PMID: 32438328. doi: 10.1016/j.dsx.2020.05.020.
64. Di Castelnuovo A, Bonaccio M, Costanzo S, Gialluisi A, Antinori A, Berselli N, et al. Common cardiovascular risk factors and in-hospital mortality in 3,894 patients with COVID-19: survival analysis and machine learning-based findings from the multicentre Italian CORIST Study. Nutrition, metabolism, and cardiovascular diseases : NMCD. 2020 Jul 31. PMID: 32912793. doi: 10.1016/j.numecd.2020.07.031.
65. Uribarri A, Núñez-Gil IJ, Aparisi A, Becerra-Muñoz VM, Feltes G, Trabattoni D, et al. Impact of renal function on admission in COVID-19 patients: an analysis of the international HOPE COVID-19 (Health Outcome Predictive Evaluation for COVID 19) Registry. Journal of nephrology. 2020 Aug;33(4):737-45. PMID: 32602006. doi: 10.1007/s40620-020-00790-5.
66. Hojs R, Bevc S, Ekart R, Gorenjak M, Puklavec L. Serum cystatin C as an endogenous marker of renal function in patients with mild to moderate impairment of kidney function. Nephrology Dialysis Transplantation. 2006;21(7):1855-62. doi: 10.1093/ndt/gfl073.
67. Shlipak MG, Mattes MD, Peralta CA. Update on Cystatin C: Incorporation Into Clinical Practice. Am J Kidney Dis. 2013 Sep;62(3):595-603. PMID: WOS:000324023700022. doi: 10.1053/j.ajkd.2013.03.027.
68. Giagulli VA, Guastamacchia E, Magrone T, Jirillo E, Lisco G, De Pergola G, et al. Worse progression of COVID-19 in men: Is testosterone a key factor? Andrology. 2020:10.1111/andr.12836. PMID: 32524732. doi: 10.1111/andr.12836.
69. Cadegiani FA, McCoy J, Wambier CG, Goren A. 5-Alpha-Reductase Inhibitors Reduce Remission Time of COVID-19: Results From a Randomized Double Blind Placebo Controlled Interventional Trial in 130 SARS-CoV-2 Positive Men. medRxiv. 2020:2020.11.16.20232512. doi: 10.1101/2020.11.16.20232512.
70. Montopoli M, Zumerle S, Vettor R, Rugge M, Zorzi M, Catapano CV, et al. Androgen-deprivation therapies for prostate cancer and risk of infection by SARS-CoV-2: a population-based study (N = 4532). Annals of Oncology. 2020 2020/08/01/;31(8):1040-5. doi: https://doi.org/10.1016/j.annonc.2020.04.479.
71. Lippi G, Henry BM, Sanchis-Gomar F. Red Blood Cell Distribution Is a Significant Predictor of Severe Illness in Coronavirus Disease 2019. Acta Haematologica. 2020. doi: 10.1159/000510914.
72. Terpos E, Ntanasis-Stathopoulos I, Elalamy I, Kastritis E, Sergentanis TN, Politou M, et al. Hematological findings and complications of COVID-19. American Journal of Hematology. 2020;95(7):834-47. doi: 10.1002/ajh.25829.
73. Gupta R, Hussain A, Misra A. Diabetes and COVID-19: evidence, current status and unanswered research questions. European journal of clinical nutrition. 2020 Jun;74(6):864-70. PMID: 32404898. doi: 10.1038/s41430-020-0652-1.
74. Apicella M, Campopiano MC, Mantuano M, Mazoni L, Coppelli A, Del Prato S. COVID-19 in people with diabetes: understanding the reasons for worse outcomes. The lancet Diabetes & endocrinology. 2020 Sep;8(9):782-92. PMID: 32687793. doi: 10.1016/S2213-8587(20)30238-2.
75. Pencina MJ, D' Agostino Sr RB, D' Agostino Jr RB, Vasan RS. Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. Statistics in Medicine. 2008 2008/01/30;27(2):157-72. doi: https://doi.org/10.1002/sim.2929.
76. Cook Nancy R. Use and Misuse of the Receiver Operating Characteristic Curve in Risk

Prediction. Circulation. 2007 2007/02/20;115(7):928-35. doi: 10.1161/CIRCULATIONAHA.106.672402.
77. Ware JH. The limitations of risk factors as prognostic tools. The New England journal of medicine. 2006 Dec 21;355(25):2615-7. PMID: 17182986. doi: 10.1056/NEJMp068249.
78. Janssens ACJW, Martens FK. Reflection on modern methods: Revisiting the area under the ROC Curve. International Journal of Epidemiology. 2020;49(4):1397-403. doi: 10.1093/ije/dyz274.
79. Fry A, Littlejohns TJ, Sudlow C, Doherty N, Adamska L, Sprosen T, et al. Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. American Journal of Epidemiology. 2017;186(9):1026-34. doi: 10.1093/aje/kwx246.
80. Nielsen D. Tree Boosting With XGBoost. Master thesis, Norwegian University of Science and Technology. 2016.

Supplementary Files

Multimedia Appendixes

Main Figures and Main Tables.
URL: http://asset.jmir.pub/assets/68d0d88fe38dd5f20c547c42f669746b.docx

List of supplementary Tables and Figures.
URL: http://asset.jmir.pub/assets/2de5e31abcd7ad6c7fe3f5dd3dfcbb85.docx

Supplementary Figures.
URL: http://asset.jmir.pub/assets/371fdb4853d9bb5d0ba8bfbcce8c77a8.xlsx

Supplementary Tables.
URL: http://asset.jmir.pub/assets/03402fb6e1d9c986be2a9d0bea19843b.xlsx

Supplementary Text and Table S8.
URL: http://asset.jmir.pub/assets/28922584b5ca2fa7b86638ba4ed0a2eb.docx