Title: Development and validation of age-specific predictive model on the risk of post-acute mortality within one year of COVID-19 infection

Ivan Chun Hang Lam, PhD[1,2†], Jiayi Zhou, MPH[3,4†], Wenlong Liu, MSc[1†], Kenneth Keng Cheung Man, PhD[1,5,6,7], Qingpeng Zhang, PhD[1,8], Hao Luo, PhD[4,9,10,11], Carlos King Ho Wong, PhD[1,3,7], Celine Sze Ling Chui, PhD[7,12,13], Francisco Tsz Tsun Lai, PhD[1,3,7], Xue Li, PhD[1,7,14], Esther Wai Yin Chan, PhD[1,7,15,16], Ian Chi Kei Wong, PhD[1,7,17], Eric Yuk Fai Wan, PhD[1,3,7*]

† Co-first author. ICHL, JZ and WL contributed equally to this article.

* Corresponding author

1 Centre for Safe Medication Practice and Research, Department of Pharmacology and Pharmacy, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong SAR, China.
2 Pharmaco- and Device Epidemiology, Centre for Statistics in Medicines, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK
3Department of Family Medicine and Primary Care, School of Clinical Medicine, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong SAR, China.
4 Department of Social Work and Social Administration, The University of Hong Kong, Hong Kong SAR, China
5 Research Department of Practice and Policy, School of Pharmacy, University College London, London, United Kingdom.
6 Centre for Medicines Optimisation Research and Education, University College London Hospitals NHS Foundation Trust, London, United Kingdom.
7 Laboratory of Data Discovery for Health ($D^2$4H), Hong Kong Science and Technology Park, Sha Tin, Hong Kong SAR, China.
8 Musketeers Foundation Institute of Data Science, The University of Hong Kong, Hong Kong SAR, China.
9 School of Public Health Sciences, University of Waterloo, Waterloo, ON, Canada
10 The Hong Kong Jockey Club Centre for Suicide Research and Prevention, The University of Hong Kong, Hong Kong SAR, China
11 Sau Po Centre on Ageing, The University of Hong Kong, Hong Kong, China
12 School of Nursing, Li Ka Shing Faculty of Medicine, The University of Hong Kong SAR, Hong Kong, China.
13 School of Public Health, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong SAR, China.
14 Department of Medicine, School of Clinical Medicine, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong SAR, China.
15 Department of Medicine, The University of Hong Kong-Shenzhen Hospital, Shenzhen, China
16 The University of Hong Kong Shenzhen Institute of Research and Innovation, Hong Kong SAR, China

17 Aston Pharmacy School, Aston University, Birmingham, United Kingdom.

Full professors: Esther Wai Yin Chan and Ian Chi Kei Wong.

Correspondence to:

Dr. Eric Yuk Fai Wan
Postal Address: 3/F, Ap Lei Chau Clinic, 161 Ap Lei Chau Main Street, Ap Lei Chau, Hong Kong SAR, China
Email: yfwan@hku.hk
Tel: +852 2552 5756

Word Count: 3,730 words

Abstract

Background

The existing risk prediction models for COVID-19 associated mortality have not considered the difference in risk factors in patients across an aging population.

Aim

To develop age-specific prediction models to forecast the risk of all-cause mortality in patients recovering from COVID-19 infection

Design

Population-based, retrospective cohort study

Methods

Patients with COVID-19 between 1 April 2020 and 31 July 2022 survived beyond the acute phase of infection were stratified into separate age cohorts (<45, 45-64, ≥65) and followed-up for one year. Backward stepwise logistic regression and four statistical and machine learning algorithms were employed to develop age-specific models on the risk of post-acute mortality following COVID-19 infection, based on a comprehensive set of clinical parameters including demographics, COVID-19 vaccination status, pre-existing comorbidities and laboratory-test findings.

Results

Of the 891,246 patients with COVID-19 identified, 13,578 (1.05%) died within one year of the index date. Age, COVID-19 vaccination status and history of acute respiratory syndrome prior infection were identified as predictors in the models for separate age groups. The model for patients aged ≥65 exhibited excellent prediction performance with an AUROC of 0.87 (95% CI: 0.87, 0.88), followed by the model

for patients aged 45-64 [AUROC=0.83 (95% CI: 0.81, 0.85)] and those aged <45 [AUROC=0.79 (95% CI: 0.72, 0.86)].

Conclusion

The age-specific models reported accurately predicted the risk of post-acute mortality in their corresponding age-group of patients, providing valuable asset in optimising clinical strategies and resource allocation in the management of the global burden of Long COVID.

Keywords

COVID-19; SARS-CoV-2 infection; Post-acute sequelae of SARS-CoV-2; Prediction modelling; Machine-learning; All-cause mortality

Introduction

Since the outbreak of the COVID-19 pandemic in 2019, there have been over seven billion confirmed cases and seven million deaths reported worldwide, with most of these recorded among older adults.(1) Whilst many patients recover from the acute phase of the illness, as much as 80% of patients continue to experience persistent and newly developed symptoms or adverse clinical outcomes beyond the acute infection. The current literature referred post-acute sequelae of SARS-CoV-2 (PASC), also known as Long COVID, as clinical presentation which persists or develops 30 days following the initial COVID-19 infection.(2-5) The constellation of signs and symptoms of PASC could range from mild symptoms including fatigue, shortness of breath, and cognitive impairment to severe complications including acute respiratory distress syndrome (ARDS), sepsis, and multiorgan failure, which could lead to mortality in severe cases.(3, 6) Despite the milder disease severity associated with the dominant Omicron variant and the gradual reduction in the spread COVID-19, over 16,000 new cases of infection continued to be reported worldwide every week.(1) Previous studies have reported an approximately four-fold increase in risk of post-acute all-cause mortality amongst patients with COVID-19 compared to their matched non-COVID-19 controls.(7) The considerable risk increase and the large population of patients with COVID-19 at risk of developing post-acute sequelae following infection raises the need for more effective strategies to identify individuals at high risk of poor clinical prognosis and subsequent death following their infection.

Emerging evidence throughout the course of the pandemic has demonstrated the association of age, certain comorbidities and severe COVID-19 conditions with various post-acute adverse outcomes of COVID-19 including hospital admission or death.(8-10) Nevertheless, physiological changes and dysregulation in the immune system and inflammatory response associated with aging has led to the speculation on the variations in the associated risk factors across patients of different age.(11-13) Building on the current knowledge of the risk factors associated with poorer prognosis of COVID-19 infection, various prognosis predictive models with the aim of predicting the risk of Long COVID,

hospital admission and mortality amongst patients with COVID-19 have been devised and implemented to support clinical decisions in hospital settings. (14-17) Nonetheless, existing prediction models have focused mainly on the prediction of the risk of adverse clinical outcomes after COVID-19 during the earlier acute phase of infection. Besides, the risk prediction based on the same factors across patients of different age could compromise the accuracy of previous models. Lastly, laboratory measurements, such as elevated levels of inflammatory markers and lower Cycle Threshold (CT) value, which have been identified as important predictors of disease severity of infection, were not included in previous prediction models for the post-acute outcomes of patients with COVID-19.

This study aims to develop age-specific prediction models on the risk of post-acute mortality following COVID-19 infection drawing on a diverse range of clinical parameters, including demographic information, pre-existing medical conditions, COVID-19 vaccination history and laboratory measurements. Robust statistical models devised based on key predictive risk factors serves as a valuable tool for identifying high-risk patients in clinical settings.

Methods

Data Source

The study extracted patient's electronic medical records from the Hong Kong Hospital Authority (HKHA) database. As a statutory body, HKHA is responsible for managing all public hospitals and ambulatory clinics in Hong Kong. The healthcare service is accessible to more than 7.3 million HK residents, covering around 80% of all routine hospital admissions and all patients diagnosed with COVID-19. The electronic medical records in the HKHA database comprised disease diagnoses recorded during planned or unplanned doctor consultations, hospital visits, and emergency visits. This facilitates the prompt recording of medical records for all users of public health services in HK. Vaccination records were provided by the Department of Health, the Government of Hong Kong Special Administrative

Region. Death records were obtained from the Hong Kong Deaths Registry. The database has been used in previous studies on the long-term sequelae of COVID-19 infection, COVID-19 vaccines, oral antiviral safety surveillance and effectiveness.(7, 18-24)

Study Design and Population

The base population for this study was defined as patients aged 18 years or above with a SARS-CoV-2 infection (confirmed by either rapid antigen test [RAT] adopted for self-testing or polymerase chain reaction [PCR] test in throat swab, nasopharyngeal aspirate, or deep throat sputum specimens conducted in hospital and community testing centres) between 1 April 2020 and 31 July 2022. For patients with multiple records of positive COVID-19 screening test, the date of the first record of positive result from either PCR or RAT was taken as the date of their COVID-19 infection. Patients were stratified by age into under 45, 45-64 and 65 or over representing young adults, middle-age and older adults, respectively.(25, 26) Patients who survived the acute-phase of infection (30-days post-infection) were eligible for prediction model development. The index date of patients was defined as 30 days after COVID-19 infection. Each patient was followed up from the index date until their death or 12 months from the index date (the last record was on 31 August 2023), whichever occurs the earliest. This study followed the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) reporting guideline for prognostic studies.(27)

Outcome

The outcome of this study was post-acute all-cause mortality within 1 year of index date.

Predictors

A comprehensive range of predicting variables were selected based on the risk factors for poor prognosis following COVID-19 infection including patients' characteristics, comorbidities and laboratory test parameters. Specific patient's characteristics included age, sex, history of COVID-19 vaccination (2/+ doses), history of disease diagnoses involving multiple organ systems defined by the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM; Supplementary table 1) were selected as potential predictors. These variables were measured on index date whilst laboratory testing results were obtained within 14 days infection to capture the rapid onset of changes following COVID-19 infection. The specific disease diagnoses and laboratory test results selected for model development were further described in Supplementary method 1.

Statistical Analysis

Before model construction, univariate analysis adjusting for age and sex only was conducted to assess the association between each candidate predictor and all-cause mortality. The significant predictors were then included in a multivariable regression model and were examined by variance inflation factors (VIF). To avoid multicollinearity issues, predictors with a VIF above 3 were excluded.(28) Descriptive characteristics were summarised using counts and percentages for categorical variables and mean and standard deviation for continuous variables.

The cohort of individuals from each age strata were divided randomly into training set and validation set in the ratio of 7:3. To develop age-specific prediction models for post-acute 1-year all-cause mortality, logistic regression with backward selection and four statistical and machine learning algorithms were performed in different age groups using training sets. The predictors selected by backward logistic regression were determined by Bayesian Information Criterion (BIC).(29, 30) In the developed model, we further tested the multicollinearity assumption by VIF and linearity assumption by the Box-Tidwell test and plotting the numerical parameters (i.e. age) and the log odds of the outcome (Supplementary figure

1).(31) The squared term of age was included in the prediction model for patients over 65 ensure linearity assumption of the model. Stepwise Cox proportional hazards regression with backward selection by BIC was also conducted, however, due to the violation of the proportional hazard assumption and the relatively short followed-up of one-year, logistic regression with backward selection was chosen for the current analysis (Supplementary table 2). Four additional statistical and machine learning algorithms represent a spectrum of modelling strategies ranging from penalised regression, Least Absolute Shrinkage and Selection Operator (LASSO) to tree-based ensemble methods, Random Forest, eXtreme Gradient Boosting (XGBoost), and Light Gradient-Boosting Machine (LightGBM) were employed to cross-validate the findings in ensuring the robustness of findings and independence on the choice of a single algorithm.(32-34) These models, which do not demand adherence to statistical assumptions typically required by conventional models, are merited for their capability to consider complex, nonlinear interactions in clinical data.(35) Detailed description on the development and hyperparameters tuning of the statistical and machine learning models were provided in Supplementary method 2 and Supplementary table 3.

Model performance in the testing sets was evaluated by examining discrimination and calibration. Discrimination was assessed by the area under receiver operating characteristic curve (AUROC) and the area under precision recall curve (AUPRC). The precision recall curve can be more informative in imbalanced datasets.(36) The reference value for AUPRC was the prevalence of all-cause mortality in each age group (0.06% in those under 45; 0.47% in patients aged 45-64; 4.56% in patients over 65), whereas the reference value for AUROC was 0.5. Sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and F1 score of the models were also reported at different predicted probability cut-offs. Calibration was graphically depicted in calibration plot between the predicted probability and observed probability, and measured by calibration slope, calibration-in-the-large (i.e. intercept in the calibration plot), and the Hosmer-Lemeshow goodness-of-fit test. The 95% confidence interval (CI) of performance measures were estimated by bootstrapping using the percentile method.

Further sensitivity analyses were conducted by including only 1) patients hospitalised during the acute phase of COVID-19 and 2) individuals screened positive by PCR testing. Data cleaning and logistic regression analyses were conducted by R version 3.6.3, and statistical and machine learning was conducted by Python (version 3.9.13) and PyCharm 2022 (Professional Edition).

Results

A total of 901,801 patients with COVID-19 between 1 April 2020 and 31 July 2022 were identified, in which 891,246 patients survived the acute phase and were included in prediction model development.. Of the 891,246 COVID-19 survivors, 13,578 (1.05%) died within 1 year of the index date. These patients were stratified into three age groups: <45 (n=306,338), 45-64 (n=324,085), ⩾65 (n=260,823). Patients aged ⩾65 had a highest rate of all-cause mortality within one year of 4.56%, followed by patients aged 45-64 and <45 with 0.47% and 0.06%, respectively (Table 1). The cumulative number of deaths after COVID-19 infection was plotted in Supplementary figure 2. The univariate analysis adjusting for age and sex showed that older patients, males, with less than 2 doses of vaccination, existing chronic diseases (except for psychotic disorder, Bell's palsy, hypertension and anxiety), developed ARDS during the acute phase, and abnormal levels of biomarkers shortly after the infection were associated with increased risk of mortality within one year after COVID-19 infection (Supplementary table 4). The statistically significant factors were included in a multivariable logistic model and tested for multicollinearity, where five correlated factors with high collinearity were excluded in feature selection process. A total of 27 candidate predictors were eventually selected, taking into account the VIF estimated (Supplementary table 5).

Figure 1 illustrates the performance of different prediction models in each age group against the testing sets. Models developed by separate statistical and machine learning model exhibits comparably high predictive performance. Given the parsimonious and greater clinical interpretability of model developed

by the logistic regression, the model based on logistic regression was identified as the preferred model for clinical application (Supplementary tables 6-8). Models developed for patients aged over 65 achieved the highest predictive performance [AUROC: 0.87 (95% CI 0.87, 0.88); AUPRC: 0.29 (95% CI 0.27, 0.30)], followed by the model for patients aged 45-64 [AUROC: 0.83 (95% CI 0.81, 0.85); AUPRC: 0.10 (95% CI 0.07, 0.13)] and those aged <45 [AUROC: 0.79 (95% CI 0.72, 0.86); AUPRC: 0.02 (95% CI 0.01, 0.04)]. Table 2 and Supplementary table 9 present the performance measures at different probabilities cut offs for age-specific models. All models were well calibrated with intercepts and slopes of the calibration plots close to 0 and 1, respectively (Figure 2 and Supplementary figure 3).

The odds ratios of selected predictors estimated were summarised in Table 3. The prediction model for 1-year all-cause mortality varied across different age groups. Age, sex, receipt of two doses of vaccination, ARDS before infection and lymphocyte were found to be significant predictors across all the models, while the model for patients aged 45-64 years additionally included ARDS during acute phase of COVID-19, heart failure, ESKD, type 2 diabetes, seizure, DVT and CT; the model for those aged ⩾ 65 years included the addition of ARDS during acute phase of COVID-19, heart failure, ESKD, seizure, myocardial infarction, stroke, atrial fibrillation, DVT, CPD, CT and serum ferritin. Increased age and fewer than 2 doses of vaccination showed similar adverse effects on 1-year all-cause mortality across age groups, while the adverse effects of ARDS before infection decreased with age from OR of 21.63 (95% CI: 8.82, 53.02) for patients aged <45, 6.72 (95 %CI: 5.15, 8.75) for patients aged 45-64, to 2.61 (95% CI: 2.32, 2.94) for patients aged ⩾65. For patients aged between 45 and 64, the adverse effect of ARDS during acute phase of COVID-19 [OR 4.70 (95% CI 2.82, 7.83)] was lower than ARDS before infection, whereas in patients aged over 65, ARDS during acute phase became to exhibit higher risk [OR 3.72 (95% CI 3.04, 4.55)] than ARDS before infection at the same age group. In addition, all morbidities included in the models were associated with higher risk of all-cause mortality. Of the laboratory parameters, lymphocyte $<1.0 \times 10^9$/L and serum ferritin > 400ng/mL was associated with higher risk of all-cause mortality, while the absence of laboratory variables constituted to a lower risk of all-cause mortality.

Patients requiring hospital admission following SARS-CoV-2 infection incurred a higher rate of mortality (N=6,753 [12.91%]), with a more complete record of laboratory parameters (Supplementary table 10) Models developed among hospitalised patients identified comparable predictors and consistent direction for risk factors associated with mortality, with the absence of laboratory variables shown to constitute to a greater risk of all-cause mortality (Supplementary tables 11-12). As routine blood tests are typically performed for hospitalised patients, missing laboratory data may reflect distinct comorbidity profiles or early mortality before testing. Given the uncertainty surrounding absence values in this group of patients, the findings should be interpreted cautiously and not over-interpreted. The characteristics of patients with COVID-19 identified through either a positive PCR and RAT screening test were presented and compared in Supplementary table 13. Prediction models for patients with PCR developed demonstrated largely consistent findings in the predictors identified in the main analysis (Supplementary tables 14-15 and Supplementary figures 4-5).

Discussion

This study reported age-specific risk prediction models for all-cause mortality within one year of post-acute COVID-19 infection accounting for variations in risk factors associated with poor disease prognosis among patients as they age. Advancing age, incomplete COVID-19 vaccination, reduced lymphocyte count and history of ARDS were significant predictors contributing to an increased risk of mortality across patients of all age groups. The model for middle-age and older adults further included existing co-morbidities including heart failure, end-stage renal diseases, diabetes and other laboratory variables including CT value, and absolute neutrophil count as predictors. All models demonstrated excellent predictive capability with AUROC of 0.8 or above; in particular, the model for patients aged ⩾65 achieved an AUROC of 0.9. The variation in the predictors identified along with the potential physiological differences associated with aging highlighted the clinical benefit in the enhanced predictive

accuracy of age-specific models reported compared to existing models. Through adopting robust statistical modelling techniques, the findings provided further understanding on the various risk factors associated with post-acute mortality identified through the current data-driven process. The evidence generated serve as crucial step in the research roadmap address the intricate challenges posed by Long COVID.(37)

Our risk prediction model identified several key clinical parameters which could influence the risk of all-cause mortality within one year of acute COVID-19 infection across all age groups of patients. Firstly, the associated increased risk with advanced age highlighted the greater susceptibility for severe COVID-19 infection and clinical manifestations in older adult and emphasised the need for focused care amongst these patients.(38, 39) Secondly, complete vaccination with two or more doses of COVID-19 vaccines was observed to confer a protective effect against mortality within one year of infection, supporting the effectiveness of COVID-19 vaccines in reducing the risk and persistence in risk of PASC including all-cause mortality.(2, 40)

The development of ARDS, whether prior to or during acute COVID-19 infection, was associated with increased post-acute mortality across all age groups. Given that the respiratory system serves as the primary target for SARS-CoV-2 infection, viral pneumonia caused by severe ARDS could impair alveolar gas exchange and perfusion, resulting in subsequent irreversible damage to the vital organs.(39) (41-43) Notably, this study identified an age-dependent pattern where ARDS during the acute phase posed a greater risk in older adults, likely due to diminished recovery capacity and pulmonary reserve, suggesting the acute respiratory deterioration from acute infection may be a more critical prognostic factor than chronic respiratory condition in this age-group of patients. Furthering to the existing evidence indicating the association between cardiovascular and renal disorder with poorer prognosis following COVID-19 infection, this study identified specific patient groups with certain conditions that may face an elevated risk of mortality during the post-acute phase of COVID-19.(10, 44, 45) These insights provide essential guidance for clinical decision-making.

Beyond the common parameters affecting the risk of mortality in patients across all age range, several clinical and haemological abnormalies were found to associate with the risk of mortality in patients from different age groups. Firstly, a reduced lymphocyte counts is identified as a significant predictor across all the age groups whilst an elevated level of ferritin was associated increased risk of mortality among older patients aged 65 or above. Given the correlation between lymphopenia with COVID-19 severity, our current findings provided further insight into the pathophysiology of PASC attributed to the irreversible organ damage associated with severe condition of COVID-19, resulting in subsequent mortality beyond the acute infection.(46, 47) Secondly, an increased CT values indicating a low viral load was not associated with a greater risk of post-acute mortality. The absence of the biological markers and CT value measurements amongst the general population of patients could indicate cases of milder disease severity, where testing was deemed unnecessary by clinical judgment or patients managed in community settings where RT-PCR testing is less routinely performed. This lower severity constituted to a lower risk of mortality over the long-term. Whilst abnormal levels of laboratory parameters are often observed in severe COVID-19 disease, physiological changes in the immune systems levels of inflammatory cytokines and biomarkers associated with aging could lead to variation in the composition of cells and soluble mediators involved in both innate and adaptive immune responses within lymphoid and non-lymphoid peripheral tissues. These changes determine not only the susceptibility to infections, but also the disease progression and subsequent clinical outcomes, thus implicating the need for considering specific parameters in forecasting the risk of post-acute mortality in different age-groups of patients as well as emphasising the need for the age-stratified models reported in clinical settings. (11, 12, 48, 49) By leveraging the extensive clinical features present in our electronic health record (EHR) data, models reported in our current study, especially that developed based on hospitalised patients with a lower proportion of missing measurements of laboratory-based predictors, is able to comprehensively account for the variations in physiology and risk factors across patients of different ages as well as enhance the accuracy of mortality prediction compared to existing models.

While prior research has aimed to facilitate patient-level risk assessment for adverse outcomes of COVID-19, existing prediction models are susceptible to biases arising from studies conducted on highly-selective cohorts and limited transparency.(50) As illustrated by the Supplementary Figure 1, the non-linear relationship between age and post-acute mortality could undermine the accuracy of the existing prediction models developed for patients of all age. The extensive clinical features, including laboratory-based measurements and coverage of patient's medical records within the territory-wide hospital database provided a comprehensive and reliable source for the development of individualised prediction models taking into consideration of the age-specific physiology to enhance the accuracy in the risk prediction generated through our current models. The stepwise logistic regression adopted in the development of models for separate age-groups also ensures the accuracy and selection of the most relevant predictors for the separate patient cohorts compared to existing models developed for the general population. The highly consistent findings in the predictors identified through different statistical modelling approaches further emphasised the robustness of our prediction model.

Nevertheless, this study is subjected to several limitations. Firstly, the current logistic model did not include complex terms such as non-linear relationships and interactions between variables. However, statistical and machine learning models which took these complex relationships into account achieved comparable predictive performance, suggesting a low degree of interactions between predictors and the lack of significant nonlinear association between predictors and outcomes. To facilitate the ease of clinical application, such complex terms were not included to maintain simplicity of the prediction model. Secondly, the variant of COVID-19 was not considered in the models reported owing to data availability. However, owing to the success in containing the Alpha and Delta variants through its zero COVID policy, it is expected that the majority of cases of COVID-19 included in our current study were caused by the Omicron variant.(51) Thirdly, the co-morbidity profile of patients within our study population could change over the course of the follow-up. However, given the relatively short follow-up period of one year, any significant change in patient's chronic health conditions affecting its survivability is considered

unlikely. Fourthly, mild diseases and symptoms such as fatigue and shortness of breath were not included as parameters in our prediction model. However, such mild symptoms presented following COVID-19 infection were unlikely to result in a severe COVID-19 condition or subsequent mortality, thus should not greatly affect the accuracy of this model. Furthermore, the inclusion of laboratory-based parameters such as inflammatory markers and biomarkers provided a more accurate data indicating the severity of COVID-19 of individual patients in predicting patient's prognosis. Lastly, lifestyle factors including smoking, drinking and exercise habits were not considered due to the lack of relevant data. Nevertheless, the comprehensive laboratory parameters and the disease diagnosis incorporated would provide reflection on the health status of individual patients.

Conclusion

This study reported age-specific risk prediction models for all-cause mortality within one year beyond the acute phase of COVID-19 infection based on a comprehensive range of patient demographics, clinical diagnosis and laboratory-based parameters. Models developed for separate age-group demonstrated high prediction performance. Given the vast number of individuals with a history of COVID-19 infection and the limited healthcare resources available, the findings from this study may assist clinicians in identifying patients at higher risk of death following COVID-19 infection, optimising clinical strategies and resource allocation to alleviate the global burden of Long COVID.

Acknowledgment

The authors thank the Hospital Authority for the generous provision of data for this study. This work was supported by HMRF Research on COVID-19, The Hong Kong Special Administrative Region (HKSAR) Government (Principal Investigator: EWYC; Ref No. COVID1903011); Collaborative Research Fund, University Grants Committee, the HKSAR Government (Principal Investigator: ICKW; Ref. No. C7154-

20GF); and Research Grant from the Health Bureau, the HKSAR Government (Principal Investigator: ICKW; Ref. No. COVID19F01). The funders did not have any role in design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication. ICKW and FTTL are partially supported by the Laboratory of Data Discovery for Health ($D^2$4H) funded by the AIR@InnoHK administered by Innovation and Technology Commission.

Data Access

EYFW had full access to all the data in the study and took responsibility for the integrity of the data and the accuracy of the data analysis.

Data Sharing Statement

The data contains confidential information and hence cannot be shared with the public due to third-party use restrictions.

Codes availability

The codes used to derive the current findings are made available in https://github.com/Jiayiz2222/LongCovid_prediction to ensure transparency and reproducibility of the findings reported.

Ethical approval

Ethical approval for this study was granted by the Institutional Review Board of the University of HK/HA HK West Cluster (UW20-556 and UW21-149) and Department of Health, HK (L/M21/2021 and L/M175/2022) with an exemption for informed consent from participants as patients' confidentiality was maintained in this retrospective cohort study.

Authors Contribution

ICHL, JZ, WL and EYFW had the original idea for the study, contributed to the development of the study, extracted data from the source database, constructed the study design and the statistical model, reviewed the literature, and act as guarantors for the study. ICHL, JZ, WL and EYFW accessed and verified the data, performed statistical analysis. ICHL, JZ, WL and EYFW wrote the first draft of the manuscript. ICKW is the principal investigator and provided oversight for all aspects of this project. KKCM, QZ, HL, CKHW, CSLC, FTTL, XL, EWYC, EYFW and ICKW provided critical input to the analyses, study design, and discussion. ICHL, RZ, EYFW, and ICKW had full access to and accessed all underlying data in this study. All authors contributed to the interpretation of the analysis, critically reviewed and revised the manuscript, and approved the final manuscript to be submitted. All authors had final responsibility for the decision to submit for publication.

Conflict of Interest

EYFW has received research grants from the Excellent Young Scientists Fund (Hong Kong and Macau), National Natural Science Foundation of China (NSFC), Health Bureau of the Government of the Hong Kong Special Administrative Region, and the Hong Kong Research Grants Council, outside the submitted work. FTTL has been supported by the RGC Postdoctoral Fellowship under the Hong Kong Research Grants Council and has received research grants from the Health Bureau of the Government of the Hong Kong Special Administrative Region, outside the submitted work. CSLC has received grants from the

Health Bureau of the Hong Kong Government, Hong Kong Research Grant Council, Hong Kong Innovation and Technology Commission, Pfizer, IQVIA, and Amgen; and personal fees from PrimeVigilance; outside the submitted work. XL has received research grants from the Health Bureau of the Government of the Hong Kong Special Administrative Region; research and educational grants from Janssen and Pfizer; internal funding from the University of Hong Kong; and consultancy fees from Merck Sharp & Dohme, unrelated to this work. CKHW. reports the receipt of General Research Fund, Research Grant Council, Government of Hong Kong SAR; EuroQol Research Foundation; AstraZeneca and Boehringer Ingelheim, all outside the submitted work. ICKW reports grants from Amgen, Bristol-Myers Squibb, Pfizer, Janssen, Bayer, GSK and Novartis, the Hong Kong RGC, and the Hong Kong Health and Medical Research Fund in Hong Kong, National Institute for Health Research in England, European Commission, National Health and Medical Research Council in Australia, consulting fees from IQVIA and World Health Organization, payment for expert testimony for Appeal Court of Hong Kong and is a non-executive director of Jacobson Medical in Hong Kong and Therakind in England, outside of the submitted work; no other relationships or activities that could appear to have influenced the submitted work. EWC reports grants from Research Grants Council (RGC, Hong Kong), Research Fund Secretariat of the Food and Health Bureau, National Natural Science Fund of China, Wellcome Trust, Bayer, Bristol-Myers Squibb, Pfizer, Janssen, Amgen, Takeda, and Narcotics Division of the Security Bureau of the Hong Kong Special Administrative Region; honorarium from Hospital Authority; outside the submitted work.

References

1. WHO Coronavirus (COVID-19) Dashboard 2024 [Available from: https://data.who.int/dashboards/covid19/data.
2. Lam ICH, Zhang R, Man KKC, Wong CKH, Chui CSL, Lai FTT, et al. Persistence in risk and effect of COVID-19 vaccination on long-term health consequences after SARS-CoV-2 infection. Nature Communications. 2024;15(1):1716.
3. Al-Aly Z, Xie Y, Bowe B. High-dimensional characterization of post-acute sequelae of COVID-19. Nature. 2021;594(7862):259-64.
4. Cohen K, Ren S, Heath K, Dasmariñas MC, Jubilo KG, Guo Y, et al. Risk of persistent and new clinical sequelae among adults aged 65 years and older during the post-acute phase of SARS-CoV-2 infection: retrospective cohort study. BMJ. 2022;376:e068414.
5. Lam ICH, Chai Y, Man KKC, Lau WCY, Luo H, Lin X, et al. The short-, medium- and long-term risk and the multi-organ involvement of clinical sequelae after COVID-19 infection: a multinational network cohort study. Journal of the Royal Society of Medicine. 2025;118(7):213-29.
6. Nalbandian A, Sehgal K, Gupta A, Madhavan MV, McGroder C, Stevens JS, et al. Post-acute COVID-19 syndrome. Nat Med. 2021;27(4):601-15.
7. Lam ICH, Wong CKH, Zhang R, Chui CSL, Lai FTT, Li X, et al. Long-term post-acute sequelae of COVID-19 infection: a retrospective, multi-database cohort study in Hong Kong and the UK. eClinicalMedicine. 2023;60.
8. Tsampasian V, Elghazaly H, Chattopadhyay R, Debski M, Naing TKP, Garg P, et al. Risk Factors Associated With Post−COVID-19 Condition: A Systematic Review and Meta-analysis. JAMA Internal Medicine. 2023;183(6):566-80.
9. 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;584(7821):430-6.
10. Subramanian A, Nirantharakumar K, Hughes S, Myles P, Williams T, Gokhale KM, et al. Symptoms and risk factors for long COVID in non-hospitalized adults. Nature Medicine. 2022;28(8):1706-14.
11. Farshbafnadi M, Kamali Zonouzi S, Sabahi M, Dolatshahi M, Aarabi MH. Aging & COVID-19 susceptibility, disease severity, and clinical outcomes: The role of entangled risk factors. Exp Gerontol. 2021;154:111507.
12. Hu Y, Liu Y, Zheng H, Liu L. Risk Factors for Long COVID in Older Adults. Biomedicines. 2023;11(11).
13. Team C-F. Variation in the COVID-19 infection-fatality ratio by age, time, and geography during the pre-vaccine era: a systematic analysis. Lancet. 2022;399(10334):1469-88.
14. 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.
15. Knight SR, Ho A, Pius R, Buchan I, Carson G, Drake TM, et al. Risk stratification of patients admitted to hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: development and validation of the 4C Mortality Score. BMJ. 2020;370:m3339.
16. Sperrin M, McMillan B. Prediction models for covid-19 outcomes. BMJ. 2020;371:m3777.
17. Pfaff ER, Girvin AT, Bennett TD, Bhatia A, Brooks IM, Deer RR, et al. Identifying who has long COVID in the USA: a machine learning approach using N3C data. The Lancet Digital Health. 2022;4(7):e532-e41.

18. Lai FTT, Li X, Peng K, Huang L, Ip P, Tong X, et al. Carditis After COVID-19 Vaccination With a Messenger RNA Vaccine and an Inactivated Virus Vaccine. Annals of Internal Medicine. 2022;175(3):362-70.
19. Wan EYF, Wang Y, Chui CSL, Mok AHY, Xu W, Yan VKC, et al. Safety of an inactivated, whole-virion COVID-19 vaccine (CoronaVac) in people aged 60 years or older in Hong Kong: a modified self-controlled case series. The Lancet Healthy Longevity. 2022;3(7):e491-e500.
20. Wan EYF, Mathur S, Zhang R, Yan VKC, Lai FTT, Chui CSL, et al. Association of COVID-19 with short- and long-term risk of cardiovascular disease and mortality: a prospective cohort in UK Biobank. Cardiovascular Research. 2023.
21. Yan VKC, Wan EYF, Ye X, Mok AHY, Lai FTT, Chui CSL, et al. Effectiveness of BNT162b2 and CoronaVac vaccinations against mortality and severe complications after SARS-CoV-2 Omicron BA.2 infection: a case-control study. Emerg Microbes Infect. 2022;11(1):2304-14.
22. Yan X, Huang H, Wang C, Jin Z, Zhang Z, He J, et al. Follow-up study of pulmonary function among COVID-19 survivors 1 year after recovery. J Infect. 2021;83(3):381-412.
23. Wan EYF, Chui CSL, Lai FTT, Chan EWY, Li X, Yan VKC, et al. Bell's palsy following vaccination with mRNA (BNT162b2) and inactivated (CoronaVac) SARS-CoV-2 vaccines: a case series and nested case-control study. Lancet Infect Dis. 2022;22(1):64-72.
24. Wan EYF, Yan VKC, Mok AHY, Wang B, Xu W, Cheng FWT, et al. Effectiveness of Molnupiravir and Nirmatrelvir-Ritonavir in Hospitalized Patients With COVID-19 : A Target Trial Emulation Study. Ann Intern Med. 2023;176(4):505-14.
25. Wei C, Liu Y, Liu Y, Zhang K, Su D, Zhong M, et al. Clinical characteristics and manifestations in older patients with COVID-19. BMC Geriatrics. 2020;20(1):395.
26. Zhang J, Hao Y, Ou W, Ming F, Liang G, Qian Y, et al. Serum interleukin-6 is an indicator for severity in 901 patients with SARS-CoV-2 infection: a cohort study. Journal of Translational Medicine. 2020;18(1):406.
27. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ. 2015;350:g7594.
28. Sun C, Gong X, Hou L, Yang D, Li Q, Li L, et al. A nomogram based on conventional and contrast-enhanced ultrasound radiomics for the noninvasively prediction of axillary lymph node metastasis in breast cancer patients. Frontiers in Oncology. 2024;14.
29. Karim MN, Reid CM, Tran L, Cochrane A, Billah B. Variable selection methods for multiple regressions influence the parsimony of risk prediction models for cardiac surgery. The Journal of Thoracic and Cardiovascular Surgery. 2017;153(5):1128-35.e3.
30. Vrieze SI. Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). Psychological Methods. 2012;17(2):228-43.
31. Harris JK. Primer on binary logistic regression. Family Medicine and Community Health. 2021;9(Suppl 1):e001290.
32. Osborne MR, Presnell B, Turlach BA. On the LASSO and its Dual. Journal of Computational and Graphical Statistics. 2000;9(2):319-37.
33. Breiman L. Random Forests. Machine Learning. 2001;45(1):5-32.
34. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; San Francisco, California, USA: Association for Computing Machinery; 2016. p. 785–94.
35. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: a highly efficient gradient boosting decision tree. Proceedings of the 31st International Conference on Neural Information Processing Systems; Long Beach, California, USA: Curran Associates Inc.; 2017. p. 3149–57.

36. Saito T, Rehmsmeier M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. PLOS ONE. 2015;10(3):e0118432.
37. Al-Aly Z, Davis H, McCorkell L, Soares L, Wulf-Hanson S, Iwasaki A, et al. Long COVID science, research and policy. Nature Medicine. 2024;30(8):2148-64.
38. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. The Lancet. 2020;395(10223):497-506.
39. Zhou F, Yu T, Du R, Fan G, Liu Y, Liu Z, et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. The Lancet. 2020;395(10229):1054-62.
40. Bowe B, Xie Y, Al-Aly Z. Acute and postacute sequelae associated with SARS-CoV-2 reinfection. Nature Medicine. 2022;28(11):2398-405.
41. Chiumello D, Modafferi L, Fratti I. Risk Factors and Mortality in Elderly ARDS COVID-19 Compared to Patients without COVID-19. J Clin Med. 2022;11(17).
42. Parada-Gereda HM, Avendaño JM, Melo JE, Ruiz CI, Castañeda MI, Medina-Parra J, et al. Association between ventilatory ratio and mortality in patients with acute respiratory distress syndrome and COVID 19: A multicenter, retrospective cohort study. BMC Pulmonary Medicine. 2023;23(1):425.
43. Wu C, Chen X, Cai Y, Xia J, Zhou X, Xu S, et al. Risk Factors Associated With Acute Respiratory Distress Syndrome and Death in Patients With Coronavirus Disease 2019 Pneumonia in Wuhan, China. JAMA Intern Med. 2020;180(7):934-43.
44. Bruchfeld A. The COVID-19 pandemic: consequences for nephrology. Nature Reviews Nephrology. 2021;17(2):81-2.
45. Zheng Y-Y, Ma Y-T, Zhang J-Y, Xie X. COVID-19 and the cardiovascular system. Nature Reviews Cardiology. 2020;17(5):259-60.
46. Xiong S, Liu L, Lin F, Shi J, Han L, Liu H, et al. Clinical characteristics of 116 hospitalized patients with COVID-19 in Wuhan, China: a single-centered, retrospective, observational study. BMC Infectious Diseases. 2020;20(1):787.
47. Huang I, Pranata R. Lymphopenia in severe coronavirus disease-2019 (COVID-19): systematic review and meta-analysis. Journal of Intensive Care. 2020;8(1):36.
48. Bajaj V, Gadi N, Spihlman AP, Wu SC, Choi CH, Moulton VR. Aging, Immunity, and COVID-19: How Age Influences the Host Immune Response to Coronavirus Infections? Front Physiol. 2020;11:571416.
49. Weyand CM, Goronzy JJ. Aging of the Immune System. Mechanisms and Therapeutic Targets. Ann Am Thorac Soc. 2016;13 Suppl 5(Suppl 5):S422-S8.
50. 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.
51. Burki T. Hong Kong's fifth COVID-19 wave-the worst yet. Lancet Infect Dis. 2022;22(4):455-6.

569 **Table 1. Demographics, medical history and clinical laboratory parameters of study population**

	Age <45			Age 45-64			Age >=65		
Variable	**Alive**	**Dead within 1 year**	**p†**	**Alive**	**Dead within 1 year**	**p†**	**Alive**	**Dead within 1 year**	**p†**
Demographic characteristics									
Number of patients	306155	183		322571	1514		248942	11881	
Male (n, %)	138932 (45.4)	110 (60.1)	<0.001	136998 (42.5)	938 (62.0)	<0.001	122442 (49.2)	6684 (56.3)	<0.001
Age in years (Mean, SD)	33.26 (6.94)	36.90 (6.11)	<0.001	55.18 (5.75)	57.63 (5.26)	<0.001	74.39 (8.10)	84.45 (9.60)	<0.001
Vaccinate 2+ doses (n, %)	247653 (80.9)	108 (59.0)	<0.001	274374 (85.1)	768 (50.7)	<0.001	164390 (66.0)	3342 (28.1)	<0.001
Diagnosis (n, %)									
Myocardial infarction	126 (<0.1)	2 (1.1)	<0.001	1843 (0.6)	56 (3.7)	<0.001	4169 (1.7)	842 (7.1)	<0.001
Heart Failure	172 (0.1)	4 (2.2)	<0.001	1223 (0.4)	105 (6.9)	<0.001	7484 (3.0)	1996 (16.8)	<0.001
Stroke	589 (0.2)	1 (0.5)	0.803	6186 (1.9)	143 (9.4)	<0.001	24039 (9.7)	2967 (25.0)	<0.001
Atrial fibrillation	158 (0.1)	0 (0.0)	-	1789 (0.6)	47 (3.1)	<0.001	11632 (4.7)	1933 (16.3)	<0.001
Deep vein thrombosis	110 (<0.1)	4 (2.2)	<0.001	380 (0.1)	28 (1.8)	<0.001	781 (0.3)	178 (1.5)	<0.001
Psychotic disorder	199 (0.1)	0 (0.0)	-	103 (<0.1)	1 (0.1)	0.984	25 (<0.1)	0 (0.0)	-
Encephalitis and Encephalopathy	89 (<0.1)	1 (0.5)	0.054	73 (<0.1)	3 (0.2)	<0.001	78 (<0.1)	9 (0.1)	0.020
Bell's Palsy	46 (<0.1)	0 (0.0)	-	135 (<0.1)	0 (0.0)	-	167 (0.1)	8 (0.1)	1.000
Interstitial lung disease	1 (<0.1)	0 (0.0)	-	21 (<0.1)	2 (0.1)	<0.001	96 (<0.1)	24 (0.2)	<0.001
Chronic pulmonary disease	2416 (0.8)	8 (4.4)	<0.001	3923 (1.2)	48 (3.2)	<0.001	10899 (4.4)	1534 (12.9)	<0.001
ARDS before infection	264 (0.1)	11 (6.0)	<0.001	1081 (0.3)	133 (8.8)	<0.001	2629 (1.1)	733 (6.2)	<0.001
ARDS during acute phase	19 (<0.1)	3 (1.6)	<0.001	110 (0.0)	32 (2.1)	<0.001	405 (0.2)	331 (2.8)	<0.001
Pancreatitis	167 (0.1)	3 (1.6)	<0.001	454 (0.1)	7 (0.5)	0.003	925 (0.4)	121 (1.0)	<0.001
Liver injury	91 (<0.1)	1 (0.5)	0.058	208 (0.1)	5 (0.3)	<0.001	277 (0.1)	32 (0.3)	<0.001
End stage kidney disease	41 (<0.1)	2 (1.1)	<0.001	188 (0.1)	37 (2.4)	<0.001	502 (0.2)	168 (1.4)	<0.001
Acute kidney injury and failure	205 (0.1)	6 (3.3)	<0.001	727 (0.2)	65 (4.3)	<0.001	3055 (1.2)	1014 (8.5)	<0.001
Type 1 diabetes	239 (0.1)	1 (0.5)	0.346	325 (0.1)	12 (0.8)	<0.001	389 (0.2)	35 (0.3)	<0.001
Type 2 diabetes	3776 (1.2)	12 (6.6)	<0.001	38962 (12.1)	382 (25.2)	<0.001	73724 (29.6)	4245 (35.7)	<0.001
Hypertension	6650 (2.2)	21 (11.5)	<0.001	71338 (22.1)	454 (30.0)	<0.001	132081 (53.1)	7532 (63.4)	<0.001
Anxiety	1046 (0.3)	4 (2.2)	<0.001	1833 (0.6)	11 (0.7)	0.518	1884 (0.8)	77 (0.6)	0.159
Post-traumatic stress disorder	2838 (0.9)	8 (4.4)	<0.001	2662 (0.8)	21 (1.4)	0.024	1608 (0.6)	140 (1.2)	<0.001
Seizure	1907 (0.6)	7 (3.8)	<0.001	1896 (0.6)	80 (5.3)	<0.001	1914 (0.8)	351 (3.0)	<0.001
Clinical laboratory parameters (n, %)									
COVID CT			<0.001			<0.001			<0.001

Missing	268554 (87.7)	118 (64.5)		283162 (87.8)	762 (50.3)		203080 (81.6)	4659 (39.2)	
< 20	16653 (5.4)	25 (13.7)		19666 (6.1)	371 (24.5)		22681 (9.1)	3576 (30.1)	
≥ 20	20948 (6.8)	40 (21.9)		19743 (6.1)	381 (25.2)		23181 (9.3)	3646 (30.7)	
C-reactive protein			<0.001			<0.001			<0.001
Missing	297688 (97.2)	150 (82.0)		313493 (97.2)	1073 (70.9)		228152 (91.6)	6475 (54.5)	
≤ 15mg/L	6919 (2.3)	12 (6.6)		6177 (1.9)	123 (8.1)		9469 (3.8)	1353 (11.4)	
> 15mg/L	1548 (0.5)	21 (11.5)		2901 (0.9)	318 (21.0)		11321 (4.5)	4053 (34.1)	
WBC (TLC) count			<0.001			<0.001			<0.001
Missing	294843 (96.3)	130 (71.0)		308513 (95.6)	860 (56.8)		219513 (88.2)	5104 (43.0)	
≤ 10×10^9/L	10199 (3.3)	38 (20.8)		12721 (3.9)	501 (33.1)		24651 (9.9)	4962 (41.8)	
> 10×10^9/L	1113 (0.4)	15 (8.2)		1337 (0.4)	153 (10.1)		4778 (1.9)	1815 (15.3)	
Absolute Neutrophil Count			<0.001			<0.001			<0.001
Missing	295673 (96.6)	132 (72.1)		309545 (96.0)	885 (58.5)		221501 (89.0)	5383 (45.3)	
≤ 7.5×10^9/L	9572 (3.1)	38 (20.8)		11832 (3.7)	467 (30.8)		22485 (9.0)	4481 (37.7)	
> 7.5×10^9/L	910 (0.3)	13 (7.1)		1194 (0.4)	162 (10.7)		4956 (2.0)	2017 (17.0)	
Platelet count			<0.001			<0.001			<0.001
Missing	294846 (96.3)	130 (71.0)		308528 (95.6)	860 (56.8)		219538 (88.2)	5109 (43.0)	
< 150×10^9/L	864 (0.3)	16 (8.7)		1613 (0.5)	130 (8.6)		5794 (2.3)	1480 (12.5)	
≥ 150×10^9/L	10445 (3.4)	37 (20.2)		12430 (3.9)	524 (34.6)		23610 (9.5)	5292 (44.5)	
Procalcitonin			<0.001			<0.001			<0.001
Missing	305678 (99.8)	179 (97.8)		321602 (99.7)	1418 (93.7)		246639 (99.1)	11151 (93.9)	
≤ 0.25 ng/mL	385 (0.1)	0 (0.0)		715 (0.2)	39 (2.6)		1389 (0.6)	351 (3.0)	
> 0.25 ng/mL	92 (<0.1)	4 (2.2)		254 (0.1)	57 (3.8)		914 (0.4)	379 (3.2)	
Serum Ferritin			<0.001			<0.001			<0.001
Missing	304183 (99.4)	170 (92.9)		319971 (99.2)	1355 (89.5)		242699 (97.5)	10157 (85.5)	
≤ 400ng/mL	1573 (0.5)	5 (2.7)		1421 (0.4)	60 (4.0)		3060 (1.2)	712 (6.0)	
> 400ng/mL	399 (0.1)	8 (4.4)		1179 (0.4)	99 (6.5)		3183 (1.3)	1012 (8.5)	
Lactate Dehydrogenase			<0.001			<0.001			<0.001
Missing	297960 (97.3)	148 (80.9)		313915 (97.3)	1126 (74.4)		230515 (92.6)	7329 (61.7)	
< 245 U/L	7345 (2.4)	16 (8.7)		6647 (2.1)	182 (12.0)		11452 (4.6)	2137 (18.0)	
≥ 245 U/L	850 (0.3)	19 (10.4)		2009 (0.6)	206 (13.6)		6975 (2.8)	2415 (20.3)	

Lymphocyte			<0.001		<0.001		<0.001
Missing	295673 (96.6)	132 (72.1)	309545 (96.0)	885 (58.5)	221501 (89.0)	5383 (45.3)	
< 1.0 x10^9/L	2666 (0.9)	31 (16.9)	3938 (1.2)	339 (22.4)	11654 (4.7)	3587 (30.2)	
≥ 1.0 x10^9/L	7816 (2.6)	20 (10.9)	9088 (2.8)	290 (19.2)	15787 (6.3)	2911 (24.5)	
Neutrophil-lymphocyte ratio (NLR)			<0.001		<0.001		<0.001
Missing	295673 (96.6)	132 (72.1)	309545 (96.0)	885 (58.5)	221501 (89.0)	5383 (45.3)	
≤ 3	9647 (3.2)	45 (24.6)	11638 (3.6)	560 (37.0)	23688 (9.5)	5553 (46.7)	
> 3	835 (0.3)	6 (3.3)	1388 (0.4)	69 (4.6)	3753 (1.5)	945 (8.0)	
Erythrocyte sedimentation rate (ESR)			<0.001		<0.001		<0.001
Missing	295673 (96.6)	132 (72.1)	309545 (96.0)	885 (58.5)	221501 (89.0)	5383 (45.3)	
Male: ≤ age/2 Female: ≤ (age+10)/2	9647 (3.2)	45 (24.6)	11638 (3.6)	560 (37.0)	23688 (9.5)	5553 (46.7)	
Male: > age/2 Female: > (age+10)/2	835 (0.3)	6 (3.3)	1388 (0.4)	69 (4.6)	3753 (1.5)	945 (8.0)	

Note: SD, standard deviation.

† The characteristics between alive persons and dead persons were compared using the Chi-squared test for categorical variables, and t test for continuous variables.

Table 2. Performance of logistic regression model against testing sets in each age groups at different predicted probability cut-offs

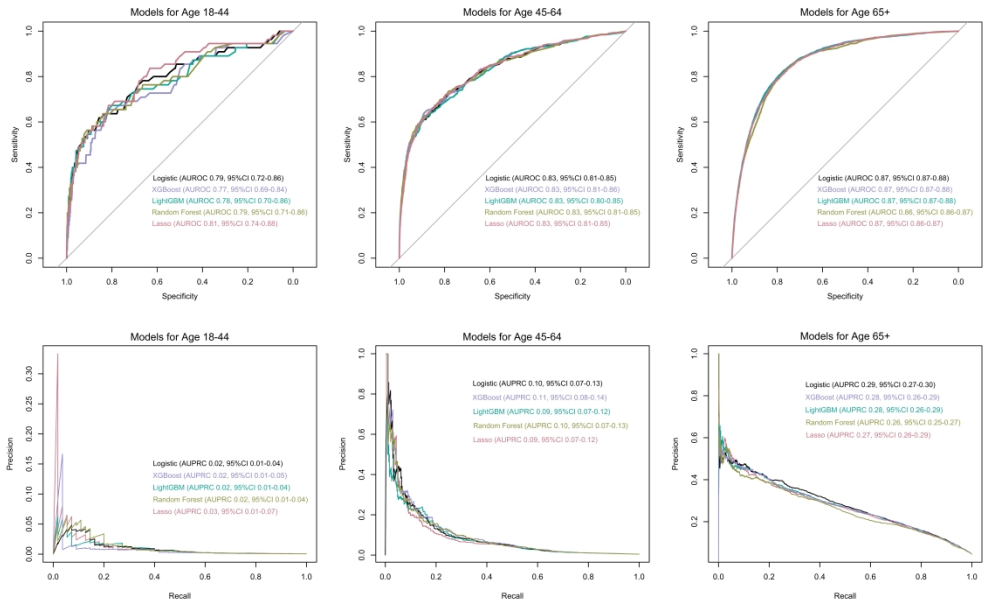
Risk cut-offs	Sensitivity, %	Specificity, %	PPV, %	NPV, %	F1, %
Model 18-44					
0.05%	78.18 (67.27,88.89)	66.28 (65.98,66.56)	0.14 (0.10,0.19)	99.98 (99.97,99.99)	0.28 (0.20,0.37)
0.1%	54.55 (42.01,67.21)	90.58 (90.38,90.76)	0.35 (0.23,0.48)	99.97 (99.96,99.98)	0.69 (0.46,0.95)
0.2%	34.55 (22.58,47.30)	97.66 (97.56,97.75)	0.88 (0.54,1.27)	99.96 (99.95,99.97)	1.71 (1.05,2.46)
Model 45-64					
0.2%	88.33 (85.28,91.26)	48.90 (48.58,49.22)	0.77 (0.70,0.85)	99.89 (99.86,99.92)	1.54 (1.39,1.68)
0.5%	64.99 (60.72,69.34)	85.54 (85.32,85.76)	1.99 (1.76,2.22)	99.82 (99.79,99.84)	3.86 (3.43,4.30)
1%	53.32 (48.73,58.08)	94.06 (93.92,94.20)	3.89 (3.38,4.41)	99.78 (99.75,99.80)	7.26 (6.35,8.18)
Model 65+					
2%	91.10 (90.18,92.09)	63.95 (63.61,64.29)	10.66 (10.31,11.03)	99.35 (99.28,99.42)	19.09 (18.52,19.68)
4%	81.18 (79.85,82.48)	79.17 (78.87,79.45)	15.55 (15.02,16.05)	98.89 (98.81,98.97)	26.10 (25.31,26.83)
5%	76.23 (74.82,77.56)	82.95 (82.68,83.22)	17.43 (16.81,18.01)	98.66 (98.57,98.75)	28.38 (27.49,29.21)

Note: PPV: Positive Predictive Value; NPV: Negative Predictive Value; The values denote the estimates (and 95% Confidence Interval)

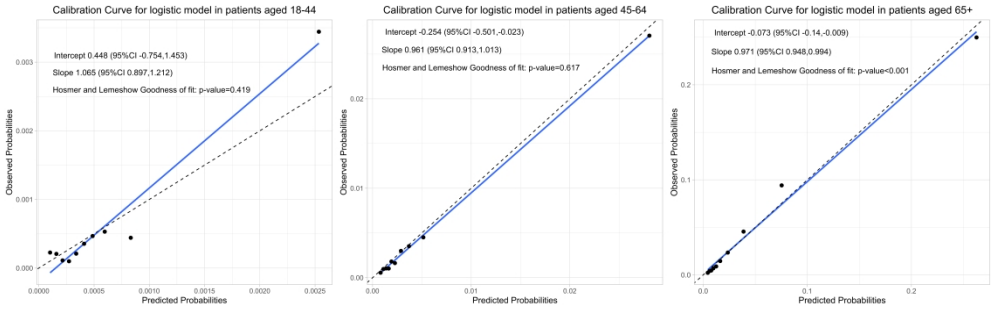
576 **Table 3. Summary of predictors selection from logistic regression against the training sets in each age groups.**

Variable	Age < 45 model			45 ≤ Age ≤ 64 model			Age ≥ 65 model		
	Coefficient	OR (95% CI)	P-value	Coefficient	OR (95% CI)	P-value	Coefficient	OR (95% CI)	P-value
(Intercept)	-9.390		<0.001	-7.101		<0.001	-5.520		<0.001
Age	0.087	1.09 (1.06, 1.12)	<0.001	0.064	1.07 (1.05, 1.08)	<0.001			
Age^2							0.001	1.001 (1.000, 1.001)	<0.001
Male	0.702	2.02 (1.41, 2.88)	<0.001	0.603	1.83 (1.61, 2.08)	<0.001	0.469	1.60 (1.52, 1.68)	<0.001
Vaccinate 2+ doses	-0.729	0.48 (0.33, 0.71)	<0.001	-1.029	0.36 (0.31, 0.41)	<0.001	-0.735	0.48 (0.45, 0.51)	<0.001
ARDS before infection	3.074	21.63 (8.82, 53.02)	<0.001	1.904	6.72 (5.15, 8.75)	<0.001	0.963	2.62 (2.33, 2.95)	<0.001
ARDS during acute phase				1.547	4.70 (2.82, 7.83)	<0.001	1.312	3.71 (3.04, 4.54)	<0.001
Heart failure				1.427	4.16 (3.11, 5.57)	<0.001	0.505	1.66 (1.53, 1.79)	<0.001
End stage kidney disease				1.651	5.21 (3.19, 8.51)	<0.001	0.636	1.89 (1.46, 2.45)	<0.001
Type 2 diabetes				0.359	1.43 (1.23, 1.67)	<0.001			
Seizure				0.710	2.03 (1.47, 2.82)	<0.001	0.643	1.90 (1.63, 2.23)	<0.001
Myocardial infarction							0.367	1.44 (1.30, 1.61)	<0.001
Stroke							0.284	1.33 (1.25, 1.41)	<0.001
Atrial fibrillation							0.160	1.17 (1.09, 1.27)	<0.001
Deep vein thrombosis				1.416	4.12 (2.37, 7.15)	<0.001	0.630	1.88 (1.49, 2.36)	<0.001
Chronic pulmonary disease							0.238	1.27 (1.17, 1.37)	<0.001
Acute kidney injury and failure							0.624	1.87 (1.69, 2.07)	<0.001
COVID CT < 20				Ref			Ref		
COVID CT - missing				-0.840	0.43 (0.35, 0.53)	<0.001	-0.534	0.59 (0.54, 0.64)	<0.001
COVID CT ≥ 20				-0.148	0.86 (0.72, 1.04)	0.119	-0.055	0.95 (0.89, 1.01)	0.107
Lymphocyte ≥1.0 × 10^9/L	Ref			Ref			Ref		
Lymphocyte - Missing	-1.208	0.30 (0.16, 0.56)	<0.001	-1.237	0.29 (0.23, 0.36)	<0.001	-0.802	0.45 (0.41, 0.49)	<0.001
Lymphocyte < 1.0 × 10^9/L	1.640	5.16 (2.54, 10.45)	<0.001	0.559	1.75 (1.42, 2.15)	<0.001	0.338	1.40 (1.31, 1.51)	<0.001
Serum Ferritin ≤ 400ng/mL							Ref		
Serum Ferritin - Missing							0.014	1.01 (0.91, 1.14)	0.761
Serum Ferritin > 400ng/mL							0.314	1.37 (1.19, 1.58)	<0.001

577 Note: OR: odds ratio; CI: confidence interval.



4649x2830mm (38 x 38 DPI)



6400x2021mm (38 x 38 DPI)