BMJ Open

Prospective validation study of prognostic biomarkers to predict adverse outcomes in patients with COVID-19: a study protocol

Benjamin Tang,[1,2] Maryam Shojaei,[1,3] Ya Wang ,[1,2] Marek Nalos,[1] Anthony Mclean,[1] Ali Afrasiabi,[2,4] Tim N Kwan,[1] Win Sen Kuan ,[5,6] Yoann Zerbib,[7] Velma Herwanto,[2,8] Gunawan Gunawan,[9] Davide Bedognetti,[10,11] Gabriele Zoppoli,[11,12] Alberto Ballestrero,[11,12] Darawan Rinchai,[10] Paolo Cremonesi,[13] Michele Bedognetti,[14] Martin Matejovic,[15] Thomas Karvunidis,[15] Stephen P J Macdonald,[16] Amanda J Cox,[17] Nicholas P West,[17] Allan William Cripps,[17] Klaus Schughart,[18,19] Andrea de Maria,[20,21] Damien Chaussabel,[22] Jonathan Iredell,[23] Stephen Weng,[24] for the PREDICT-19 consortium

To cite: Tang B, Shojaei M, Wang Y, *et al*. Prospective validation study of prognostic biomarkers to predict adverse outcomes in patients with COVID-19: a study protocol. *BMJ Open* 2021;**11**:e044497. doi:10.1136/bmjopen-2020-044497

► Prepublication history for this paper is available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2020-044497).

Received 04 September 2020
Revised 17 November 2020
Accepted 15 December 2020

For numbered affiliations see end of article.

Correspondence to
Dr Benjamin Tang;
benjamin.tang@sydney.edu.au

ABSTRACT

Introduction Accurate triage is an important first step to effectively manage the clinical treatment of severe cases in a pandemic outbreak. In the current COVID-19 global pandemic, there is a lack of reliable clinical tools to assist clinicians to perform accurate triage. Host response biomarkers have recently shown promise in risk stratification of disease progression; however, the role of these biomarkers in predicting disease progression in patients with COVID-19 is unknown. Here, we present a protocol outlining a prospective validation study to evaluate the biomarkers' performance in predicting clinical outcomes of patients with COVID-19.

Methods and analysis This prospective validation study assesses patients infected with COVID-19, in whom blood samples are prospectively collected. Recruited patients include a range of infection severity from asymptomatic to critically ill patients, recruited from the community, outpatient clinics, emergency departments and hospitals. Study samples consist of peripheral blood samples collected into RNA-preserving (PAXgene/Tempus) tubes on patient presentation or immediately on study enrolment. Real-time PCR (RT-PCR) will be performed on total RNA extracted from collected blood samples using primers specific to host response gene expression biomarkers that have been previously identified in studies of respiratory viral infections. The RT-PCR data will be analysed to assess the diagnostic performance of individual biomarkers in predicting COVID-19-related outcomes, such as viral pneumonia, acute respiratory distress syndrome or bacterial pneumonia. Biomarker performance will be evaluated using sensitivity, specificity, positive and negative predictive values, likelihood ratios and area under the receiver operating characteristic curve.

Ethics and dissemination This research protocol aims to study the host response gene expression biomarkers in severe respiratory viral infections with a pandemic potential (COVID-19). It has been approved by the local ethics committee with approval number 2020/ETH00886. The results of this project will be disseminated in international peer-reviewed scientific journals.

Strengths and limitations of this study

- ► The study has a prospective study design, optimised to evaluate the performance of predictive biomarkers.
- ► Data generated from this study will enhance triage across a diverse range of clinical settings during COVID-19 pandemic.
- ► All outcomes are prespecified and have high clinical relevance to the management of patients with COVID-19.
- ► Study limitations include potential heterogeneity in management protocols of patients with COVID-19 across different countries (eg, medications administered to patients will vary depending on clinician preference or local institutional protocol).

INTRODUCTION

An enhanced ability to predict disease progression is central to the management of the current COVID-19 crisis. As the COVID-19 crisis escalates, health services can be overwhelmed by the rapid rise in infected cases. In some locations, such as northern Italy, Mexico, Brazil and some US states, hospital beds and ventilator requirements exceeded the maximum capacity at the peak of the outbreak. In these circumstances, clinicians are often confronted with difficult triage questions: (1) Which patients should be hospitalised? (2) Which patients will need

an intensive care bed? (3) Which patients can be safely sent home for self-isolation?

There is currently a lack of clinical tools to address these questions, especially in the early phase of COVID-19 illness when symptoms are often very non-specific (eg, fever, cough, anosmia), and thus not informative. Furthermore, initial symptoms may correlate poorly with clinical trajectory (eg, recovery vs deterioration). Other clinical observations such as radiographic evidence (eg, chest CT/CT) suggestive of pneumonia or signs of hypoxaemia (eg, cyanosis, desaturation) are helpful. However, those are often late signs of impending respiratory failure. Therefore, there is an urgent need to develop new biomarkers to help identify high-risk patients with COVID-19 in the early stage of the disease.

A number of biomarkers have been identified as potential candidates for risk stratification in patients with COVID-19. These biomarkers fall mainly into three broad categories: (1) routine laboratory parameters (eg, D-dimers, lactate dehydrogenase (LDH), lymphopenia),[1 2] (2) inflammatory biomarkers (eg, interleukin 6 (IL-6), C-reactive protein (CRP)),[3 4] and (3) research-only immune assays (eg, surface markers on CD4 and CD8 lymphocytes).[5] However, there are major limitations to the use of these biomarkers for risk stratification. First, they are not specific to viral illness. For example, CRP, D-dimers or IL-6 can be elevated in many non-viral conditions, such as trauma, thromboembolism and sepsis. Second, these biomarkers do not correlate well with the immune response to SARS-CoV-2, the key driver of progression in viral pneumonia. Therefore, there is a need to develop and validate biomarkers that (1) are specific to viral pneumonia, and (2) predict the risk of COVID-19 complications, such as acute respiratory distress syndrome (ARDS).

Recent studies have identified a number of immune response biomarkers that are specific to viral pneumonia.[6–8] In this study, we hypothesise that virally induced gene expression biomarkers can predict outcomes of patients with COVID-19. Here, we present a protocol for a prospective validation study to evaluate the prognostic performance of these biomarkers in patients with COVID-19.

METHOD

Study design

This is a prospective validation study of previously identified gene expression biomarkers in respiratory viral infections (figure 1). All biomarkers have been identified or validated in previous studies in respiratory virus infections (except in COVID-19).[6–8] The primary objective of the study is to prospectively validate these biomarkers in a new cohort of patients with COVID-19.

Reporting of study findings

The reporting of the study will follow international best practice guidelines for prognostic studies, according to the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis guideline.[9] This guideline includes a checklist to ensure completeness and transparency in reporting findings of the study.

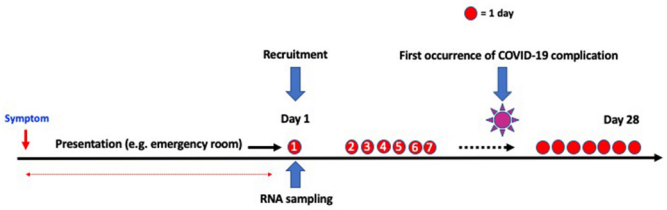

Figure 1 Overall schema including recruitment and study period. Examples of COVID-19 complication include acute respiratory distress syndrome or intensive care unit admission.

Candidate biomarkers

Real-time PCR will be performed to measure the gene expression levels of selected biomarkers in peripheral blood samples of study participants. A list of representative biomarkers is shown in table 1, which are selected because they represent key molecular pathways in the host response to respiratory viral infection, and their diagnostic or prognostic performance in viral respiratory infection has already been demonstrated in recent studies.[6–8] To account for multiplicity issues, some biomarkers will be summarised as a single genomic score in data analysis. An example of a genomic score is given in table 1, where five of the selected biomarkers (*TGFBI*, *DEFA4*, *LY86*, *BATF* and *HK3*) are condensed into a single genomic score, based on recent analyses performed on patients with respiratory viral infection.[8] Additional novel biomarkers in respiratory infection or COVID-19 may be added pending new evidence emerging from the most recent literature.

Patient recruitment and sampling

Clinical settings

To capture the full spectrum of disease severity, we will recruit a heterogeneous patient population using convenience sampling. The target populations will be drawn from preselected sites representing six clinical settings, including:

- The community.
- Outpatient clinics (eg, 'Fever clinic').
- Hospital wards.
- Emergency departments.
- Field clinics (eg, 'Cruise ship clinic').
- Intensive care units (ICU).

Eligible participants

Eligible participants include patients whose respiratory samples, such as nasopharyngeal swab, sputum or bronchoalveolar lavage (BAL), are positive for SARS-CoV-2, as confirmed by standard virological testing.

Virological testing

Respiratory samples (nasal/throat swab/sputum/BAL) collected from patients are tested for SARS-CoV-2 virus.

Tang B, *et al*. *BMJ Open* 2021;**11**:e044497. doi:10.1136/bmjopen-2020-044497

Table 1 Representative examples of gene expression biomarkers selected from literature

Biomarker	Biological role	Genomic score
IFI27	Interferon pathway	Not applicable
CD177	Neutrophil migration	Not applicable
HLADPB1	Antigen presentation	A summary genomic score will be used to represent these five genes collectively with a single readout
TGFBI	Cell adhesion	
DEFA4	Host defence	
LY86	Lymphocyte response	
BATF	Transcription factor	
HK3	Bioenergetics	

Nucleic acid amplification testing using established WHO primers will be performed and, where appropriate, other tests (eg, rapid antigen assay, viral culture or serology) are used.

Sample collection

PAXgene/Tempus blood samples are collected from study participants at presentation to the hospital/clinic or, in the case of asymptomatic patients, as soon as virus testing is performed. Follow-up blood sample collection is performed during hospitalisation or follow-up visits. We anticipate that the number of follow-up samples from each patient and the interval between follow-up samples will vary across the entire cohort. This is because the study is undertaken in multiple hospitals/institutions, with different management protocols implemented at each institution. We estimate, on average, that between one and three follow-up samples will be available for collection from each study participant. To account for this variability, we will undertake the following approaches during the analysis phase of the study, including (1) performing analysis in cohorts with matched time points (eg, days 1, 3 and 5); (2) stratifying patients into different disease stages (eg, mild vs severe); and (3) using an unbalanced linear mixed effects model to account for sampling variability.

Sample processing

Blood samples will be collected from individuals using the PAXgene/Tempus blood RNA system. Collected tubes will be incubated at room temperature for 4 hours following blood collection and then stored at −80°C. Prior to RNA isolation, tubes will be removed from −80°C and allowed to thaw at room temperature overnight. Total RNA will be isolated following the manufacturer's recommended protocol (PAXgene Blood RNA Kit; QIAGEN/Tempus Spin RNA Isolation; Thermo Fisher). Quality of the resulting RNA samples will be verified on an Agilent 2100 Bioanalyzer (Agilent Technologies, Palo Alto, California); RNA concentrations will be determined using a NanoDrop ND-1000 spectrophotometer (NanoDrop Technologies, Wilmington, Delaware). Reverse transcription will be performed using the qScript cDNA SuperMix per the manufacturer's protocol (Gene Target Solutions, Australia). Quantitative PCR will be carried out using TaqMan Gene Expression Master Mix (Thermo Fisher Scientific, Australia) on a CFX384 Real-Time PCR Detection System (Bio-Rad Laboratories, Hercules, California). CFX Maestro Software will be used for gene expression analysis.

Study outcomes

This study aims to assess the performance of an individual biomarker or a single genomic score in predicting complications of COVID-19. The types of complications are predefined in accordance with international guidelines (see below) and are included as part of the definitions for primary and secondary outcomes.

Primary outcomes

A composite outcome is used as defined by the first occurrence of (1) any International Severe Acute Respiratory and Emerging Infection Consortium (ISARIC)-defined complications, such as viral pneumonia, ARDS or bacterial pneumonia (table 2); or (2) prolonged virus shedding; or (3) ICU admission; or (4) hospital stay >7 days. The study period to record the first occurrence of an outcome will be 28 days.

Secondary outcomes

(1) Hospital mortality; (2) ICU length of stay; and (3) ventilation-free days (if admitted to ICU).

Definition of COVID-19 complications

COVID-19 complications are defined by ISARIC (https://isaric.tghn.org/) and represent internationally agreed consensus definitions that are endorsed by the WHO. A list of the predefined COVID-19-related complications is found in table 2.

Ascertainment of outcome

Study participants are followed up for up to 28 days after their enrolment into the study to ascertain the occurrence of outcomes (outlined above). The outcome data will be independently assessed by a clinician who is blinded to the biomarker data and has not been involved in the design of this project or the analysis of the biomarker data.

Clinical data

An electronic case report form (eCRF) is designed and developed based on best practice guidelines.[10] Clinical data to be collected are prespecified, again using ISARIC guidelines. These clinical data include ISARIC-defined outcomes (outlined above), study participants' demographics (age and gender), common COVID-19 symptoms, routine laboratory results and treatment. Examples of items to be included in the eCRF are provided in table 2.

Data security and confidentiality

Study-related information is stored securely at designated study sites in accordance with ethics and research governance guidelines of each participating institution. To maintain study participant confidentiality, all laboratory

Table 2 Prespecified COVID-19 symptoms, complications and risk factors

COVID-19-related symptoms	COVID-19-related complications	Known risk factors for disease progression
Fever	Viral pneumonia	Age
Cough	Bacterial pneumonia	Number of symptoms
Sore throat	ARDS	Number of comorbidities
Runny nose	Cardiac complications	History of malignancy
Myalgia	Bacteraemia	C-reactive protein (CRP)
Arthralgia	DIC	Neutrophilia
Fatigue	Liver dysfunction	Lymphopenia
Chest pain	Stroke/CVA	Neutrophil:lymphocyte ratio
Shortness of breath	Seizure	Lactate dehydrogenase (LDH)
Headache	Acute kidney injury	Direct bilirubin
GIT symptoms (eg, diarrhoea)	Hyperglycaemia	D-dimer
Loss of smell	Pulmonary embolism	CT scan abnormalities
Loss of taste	Deep venous thrombosis	CXR abnormalities

ARDS, acute respiratory distress syndrome; CVA, cerebrovascular accident; CXR, chest X-ray; DIC, disseminated intravascular coagulation; GIT, gastrointestinal tract.

specimens, reports, collected data and administrative forms are identified by coded study identification number only. The study participants are assigned unique study IDs, administered by each participating institution.

STATISTICAL ANALYSIS

The statistical analysis aims to generate data to evaluate the prognostic performance of an individual biomarker (or a genomic score) in predicting clinical outcomes using a prospective validation cohort. In evaluating predictive performance (see below), a prevalidated cut-off threshold (for each biomarker) or a prevalidated summary genomic score from previous studies will be used.

Variable selection

The main predictor variable is biomarker expression level or genomic score. The main analysis (eg, sensitivity, specificity, positive and negative predictive values) is based on predetermined cut-off values for each biomarker or genomic score. Additional risk factors (eg, age, comorbidities, CRP level, lymphocyte count) will be included as predictor variables in a multivariate analysis. These additional risk factors are prespecified prior to the analyses (table 2), based on a published systematic review of COVID-19 prediction studies.[11]

Other predictor variables

A multivariate analysis will be performed to assess the incremental value of gene expression biomarkers compared with conventional laboratory parameters (eg, lymphocyte count) or clinical factors in predicting COVID-19 outcomes. The clinical factors and biomarkers have been preselected based on a recently published systematic review of prognostic studies relating to COVID-19 (27 studies in total), which showed that clinical factors and conventional biomarkers may have predictive value.[11] These variables include CRP, LDH, lymphocyte counts and procalcitonin (PCT), age, comorbidities and symptoms. Examples of prespecified conventional biomarkers and clinical factors are listed in table 2.

Data imputation

Multiple imputation with chained questions will be used to estimate missing values for continuous variables (eg, CRP, LDH, PCT, lymphocyte counts).[12] All other available data will be included in imputation models to create 10 imputed data sets. We will compare the distributions of each imputation data set with that of the original data to explore their stability and convergence. Using Robin's rules, we will pool the regression equations estimated from each imputation data set to generate a combined model.[13]

Performance of prediction accuracy

We will assess the prognostic performance of each biomarker using an established method by Metz and Zhou, as implemented in the NCSS statistical software (Utah, USA).[14] Sensitivity, specificity, positive predictive values, negative predictive values, positive likelihood ratio and negative likelihood ratio will be generated using the previously established cut-off values. For positive and negative predictive values, a range of prevalence will be provided in 5% increments (20%–40%). For all performance metrics, 95% CIs will be calculated based on the Exact (Clopper-Pearson) method.[15]

Net reclassification improvement

Net reclassification improvement provides an additional measure of prediction accuracy. It is performed by adding the sum total of 'true positive' and 'true negative' values

for each predictor variable (biomarker, age, comorbidity, symptom scores, CRP). We will also calculate the concordant and disconcordant cases. Concordant cases are defined as cases where biomarker and clinical variables (eg, age, comorbidity or symptoms) are both correct in predicting outcome. Disconcordant cases are defined as cases where either biomarker variables are correct but clinical variables are incorrect, or biomarker variables are incorrect but clinical variables are correct. Based on these data, we will calculate the net reclassification improvement index.

Decision curve analysis

Decision curve analysis will be performed to provide additional insight into the possible consequence of misclassification error. In this analysis, we assume that the harm arising from false positive is relatively limited, for example, unnecessarily admitting a patient to hospital for further monitoring but the patient does not develop any adverse outcome. We also assume that the harm of false negative is relatively serious, for example, discharging a high-risk patient to home quarantine where the patient subsequently deteriorates and dies at home. We will assign an appropriate harm to benefit ratio to determine an optimal decision threshold for clinicians. This is done by calculating the net benefit of one or more models (for instance, with and without predictive biomarker), in comparison to a default strategy of treating all or no patients. Net benefit can be defined as: the sensitivity × prevalence − (1-specificity) × (1-prevalence) × w, where w is the odds at threshold probability. By defining a low threshold (treat many), an assumption can be made that harms arising from false positive are relatively limited but may increase costs and capacity, whereas defining a high threshold (treat few) could result in fewer false positives but introduce more false negatives, individuals who are still at high risk.

Sample size calculation

Assuming an event rate of 0.2 (eg, 20% of recruited subjects develop COVID-19 complications such as respiratory failure) and a sample sensitivity of 0.80, the sample size needed for a two-sided 95% sensitivity CI with a width of at most 0.2 is 350. Assuming an event rate of 0.2 and a sample specificity of 0.8, the sample size needed for a two-sided 95% specificity CI with a width of at most 0.2 is 88. The whole table sample size required is 350, so that both CIs have widths less than 0.2, the larger of the two sample sizes. Therefore, the appropriate sample size for this study is 350 patients. The sample size calculation was estimated using PASS V.15 (NCSS, Kaysville, Utah, USA).

PATIENT AND PUBLIC INVOLVEMENT

Study participants are not directly involved in the study design or the formulation of hypotheses, nor will they be involved in conducting the study. However, the research protocol has been reviewed by local ethics committees, which include representatives from the local community who act as a bridge between the researchers and the local population. Study results that can positively change clinical practice will be disseminated through local meetings and, when appropriate, via selected media outlets (as determined by each participating institution). Individual data will not be reported back to study participants.

ETHICS AND DISSEMINATION

Informed consents are obtained from study participants or next of kin (if study participants are unable to give consent). The participant information sheet contains details of the study, the implications and constraints of the study protocols and the known side effects or any risks involved in participating in the study. It also states that a participant is free to withdraw himself/herself from the study at any time for any reason without prejudice to future care and with no obligation to give the reason to withdraw. All data will be deidentified and no patient-related information will be revealed during analysis. The WSLHD Ethics Committee has approved the study (reference number: 2020/ETH00886). The result of this study will be published in international peer-reviewed journals. Only anonymised data will be presented for publication of the study findings.

DISCUSSION

In the current COVID-19 literature, many studies that evaluate the performance of prognostic biomarkers (or clinical scores) have a high risk of bias.[11 11] The sources of biases include: (1) studies often being retrospective in design; (2) the lack of independent, external validation for identified biomarkers; (3) overfitting (this is a common problem); and (4) the lack of transparent reporting procedure in information relating to study methodologies. In this study, we have incorporated several study design features to reduce potential sources of bias. These features include: (1) using a prospective study design; (2) transparency in the study protocol and analysis workflow; and (3) all study endpoints being prespecified prior to analysis.

The most common problem in prognostic studies is model overfitting. This occurs when the sample size is relatively small compared with the number of candidate biomarkers and is frequently seen in high-dimensional data sets (eg, RNA sequencing data) where hundreds of candidate biomarkers are tested against a small number of patient samples (eg, less than 30 patients). Here, we limit the number of predictor variables by preselecting a small number of candidate biomarkers that have already been validated in previous studies. To further minimise overfitting, five of these candidate biomarkers are condensed into a composite genomic score. In total, three candidate biomarkers and one composite genomic score will be validated in this prospective study.

Another common problem in prognostic biomarker study design is the lack of separation between internal and external validation. Internal validation, involving training-testing splits of the available data or cross-validation, is usually performed during the discovery phase of the biomarkers. This is done by using a 'training' cohort (to build a prediction model) and a 'test' cohort (to test model performance). External validation, on the other hand, is usually performed using an independent data set that is distinct from the 'training' and 'test' cohorts. Many biomarker studies either do not perform external validation or do not separate internal and external validation (ie, the same data set is used for 'training', 'testing' and external validation). In this study, all selected biomarkers have already undergone internal validation in previous studies.[6–8] The new COVID-19 cohort, assembled for this study, represents a completely independent cohort that is separate from the original studies (from which these biomarkers were first discovered). This clear separation of the internal and external validation cohorts strengthens the external validity of our study.

In conclusion, we have presented a detailed protocol that aims to independently evaluate the prognostic performance of several previously published gene expression biomarkers using a new cohort of patients with COVID-19 assembled for the purpose of this study. To our knowledge, this study represents the first systematic evaluation of gene expression biomarkers to predict clinical outcomes in patients with COVID-19.

Author affiliations
[1]Nepean Clinical School, University of Sydney, Sydney, New South Wales, Australia
[2]Centre for Immunology and Allergy Research, Westmead Institute for Medical Research, Westmead, New South Wales, Australia
[3]Westmead Institute for Medical Research, Westmead, New South Wales, Australia
[4]System Biology and Health Data Analytic Lab, The Graduate School of Biomedical Engineering, University of New South Wales, Sydney, New South Wales, Australia
[5]Emergency Medicine, National University Hospital, Singapore
[6]Department of Surgery, National University Singapore Yong Loo Lin School of Medicine, Singapore
[7]Department of Intensive Care Medicine, Amiens University Hospital, Amiens, France
[8]Tarumanagara University Faculty of Medicine, Jakarta Barat, Jakarta, Indonesia
[9]Department of Internal Medicine, Medistra Hospital, Jakarta, Indonesia
[10]Cancer Research Department, Sidra Medicine, Doha, Qatar
[11]Department of Internal Medicine, University of Genoa, Genova, Italy
[12]Division of Internal Medicine and Oncology, IRCCS Ospedale Policlinico San Martino, Genova, Italy
[13]Ente Ospedaliero Ospedali Galliera, Genova, Liguria, Italy
[14]Azienda Sanitaria Locale 3 Genovese, Genova, Liguria, Italy
[15]First Medical Department, Faculty of Medicine in Pilsen, Charles University Medical School and Teaching Hospital Plzen, Pilsen, Czech Republic
[16]Centre for Clinical Research in Emergency Medicine, Royal Perth Hospital, Perth, Western Australia, Australia
[17]Mucosal Immunology Research Group, Griffith University Menzies Health Institute Queensland, Southport, Queensland, Australia
[18]Department of Infection Genetics, Helmholtz Centre for Infection Research, Braunschweig, Niedersachsen, Germany
[19]University of Veterinary Medicine Hannover, Hannover, Niedersachsen, Germany
[20]Division of Infectious and Tropical Diseases, IRCCS Ospedale Policlinico San Martino, Genova, Liguria, Italy
[21]Department of Health Sciences, University of Genoa, Genova, Liguria, Italy
[22]Immunology Department, Sidra Medical and Research Center, Doha, Ad Dawhah, Qatar
[23]Centre for Infectious Diseases and Microbiology, Westmead Hospital, Westmead, New South Wales, Australia
[24]Division of Primary Care, University of Nottingham, Nottingham, UK

Acknowledgements We would like to thank Derek Blankenship for his helpful review on the statistical section of the manuscript.

Collaborators The "Predicting disease progression in severe viral respiratory infections and COVID-19" (PREDICT-19) Consortium is an international consortium formed by a group of researchers who share common interests in identifying, developing and validating clinical and/or bioinformatic tools to improve patient triage in a pandemic such as COVID-19.

Contributors BT conceived the study and wrote the manuscript. SW provided significant input in the Statistical Analysis section of the manuscript. DB, MS, TNK, WSK, VH, SPJM and KS provided significant input in the Method section of the manuscript. YW, MN, AM, AA, YZ, GG, DB, GZ, AB, DR, PC, MB, MM, TK, AJC, NPW, AWC, KS, AdM, DC and JI participated in the critical evaluation of the initial manuscript draft and contributed significant intellectual input in the revision of the final draft.

Funding SW has received honorarium from AMGEN for providing education and training to clinical stakeholders. KS is supported by intramural grants from the Helmholtz Association (Program Infection and Immunity) and NIAID Research Grants 2-U19-AI100625-06 REVISED and 5U19AI100625-07. MM and TK are supported by project number CZ.02.1.01/0.0/0.0/16_019/0000787 'Fighting Infectious Diseases' awarded by the MEYS CR, financed by EFRR. BT is supported by two NHMRC Centers of Research Excellence: (1) Australia Partnership for Preparedness Research on Infectious Disease Emergencies (APPRISE); and (2) Centre of Research Excellence in Emerging Infectious Diseases (CREID). AJC, NPW and AWC are supported by grants from the Menzies Health Research Institute Queensland (MHIQ) and Griffith University. GZ is supported by a Fondazione AIRC per la Ricerca sul Cancro IG 2018 ID 21761.

Competing interests AM, BT and MS hold a patent on IFI27.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

ORCID iDs
Ya Wang http://orcid.org/0000-0002-0297-8227
Win Sen Kuan http://orcid.org/0000-0002-2134-7842

REFERENCES

1 Wu C, Chen X, Cai Y, *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;180:934.
2 Zhou F, Yu T, Du R, *et al*. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *Lancet* 2020;395:1054–62.
3 Chen X, Zhao B, Qu Y. Detectable serum SARS-CoV-2 viral load (RNAaemia) is closely associated with drastically elevated interleukin 6 (IL-6) level in critically ill COVID-19 patients. *Medrxiv*.
4 Ling W. C-Reactive protein levels in the early stage of COVID-19. *Med Mal Infect* 2020.
5 Wang F, Nie J, Wang H, *et al*. Characteristics of peripheral lymphocyte subset alteration in COVID-19 pneumonia. *J Infect Dis* 2020;221:1762–9.
6 Tang BM, Shojaei M, Parnell GP, *et al*. A novel immune biomarker *IFI27* discriminates between influenza and bacteria in patients with suspected respiratory infection. *Eur Respir J* 2017;49:1602098.
7 Tang BM, Shojaei M, Teoh S, *et al*. Neutrophils-related host factors associated with severe disease and fatality in patients with influenza infection. *Nat Commun* 2019;10:3422.

8 Mayhew MB, Buturovic L, Luethy R, *et al*. A generalizable 29-mRNA neural-network classifier for acute bacterial and viral infections. *Nat Commun* 2020;11:1177.
9 Collins GS, Reitsma JB, Altman DG, *et al*. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD). *Circulation* 2015;131:211–9.
10 Bellary S, Krishnankutty B, Latha MS. Basics of case report form designing in clinical research. *Perspect Clin Res* 2014;5:159–66.
11 Wynants L, Van Calster B, Collins GS, *et al*. Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal. *BMJ* 2020;369:m1328.
12 Royston P. Multiple imputation of missing values: update. *Stata J* 2005;5:188–201.
13 Marshall A, Altman DG, Holder RL, *et al*. Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines. *BMC Med Res Methodol* 2009;9:57.
14 Metz CE. Basic principles of ROC analysis. *Semin Nucl Med* 1978;8:283–98.
15 Brown LD, Cai TT, science ADS. Interval estimation for a binomial proportion. *Statist Sci* 2001;16:101–33.