MATTERS ARISING
https://doi.org/10.1038/s42256-020-00251-5

nature machine intelligence

Li Yan et al. reply

Jorge Goncalves[1,2], Li Yan[3], Hai-Tao Zhang[4], Yang Xiao[4], Maolin Wang[4], Yuqi Guo[4], Chuan Sun[4], Xiuchuan Tang[5], Zhiguo Cao[4], Shusheng Li[3], Hui Xu[6], Cheng Cheng[4], Junyang Jin[7] and Ye Yuan[4]✉

REPLYING TO Marian Quanjel et al. *Nature Machine Intelligence* https://doi.org/10.1038/s42256-020-00253-3 (2020)
REPLYING TO Claire Dupuis et al. *Nature Machine Intelligence* https://doi.org/10.1038/s42256-020-00252-4 (2020)
REPLYING TO Matthew Barish et al. *Nature Machine Intelligence* https://doi.org/10.1038/s42256-020-00254-2 (2020)

We would like to thank the authors for their interest in our paper[1], the application of the model to new datasets and for sharing the data. Differences in hospital and laboratory protocols can lead to significant changes in blood sample distributions. In addition, it is possible that genetic heterogeneity between Asians and Caucasians also impacts blood samples. We are very interested in understanding the differences between the data and would welcome a collaboration between our groups.

To understand the differences in the performance of the model among these datasets and those in ref. [1], we directly compared the distributions of the three key biomarkers (Figs. 1 and 2). Figure 1 shows the distributions of the three biomarkers in all blood samples, while Fig. 2 separates the blood samples according to patient outcome. What is clear from Fig. 1 is that the distributions from Tongji Hospital (top row) are very different from those from the St Antonius Hospital (Nieuwegein, the Netherlands) (AH, second row), French Outcomerea (FO, third row) and Northwell Health (US) (NH, bottom row) datasets. We examined pairwise Kolmogorov–Smirnov tests between Tongji Hospital and AH, FO and NH for each of the three biomarkers. These show that the data distributions for all three biomarkers, for Tongji and the other hospitals, are statistically different.

In Fig. 2, all three biomarkers in the data from Tongji Hospital (training and external test data are combined) have a clear separation between survival and death. However, this is not the case with the datasets from AH, FO and NH, as there are considerable overlaps between surviving and deceased patients, thus making it difficult to predict a patient's outcome.

The reasons for these changes in distributions are unknown and require further investigation.

One possible explanation for the differences in the distributions in Figs. 1 and 2 is the differences in hospital protocols. In particular, the discharging protocols seem very different. Table 1 looks at the last blood samples before outcome. On average, surviving patients from AH, FO and NH were discharged with the three biomarkers considerably outside normal ranges (for LDH, ~80–250 $U\,l^{-1}$ (ref. [2]); for CRP, <10 $mg\,l^{-1}$ (ref. [3]); for lymphocytes (%), ~20–40% (ref. [4]))[5–7]. In other words, surviving patients from AH, FO and NH seem to be released earlier than those at Tongji Hospital. Hence, patients with relatively high values of LDH are assigned with a survival outcome. It is possible that if these patients had remained in the hospital longer, their 10 days to outcome blood samples distributions would have been closer to those of Tongji Hospital.

All hospitals in China follow the following strict discharge protocol set by the China National Health Commission[8]:

- the patient's temperature has remained normal (<37.3 °C) for more than three days
- respiratory symptoms have been relieved
- COVID-19 nucleic acid in respiratory tract specimens has tested negative twice in a row (sampling interval of at least 24 h)
- the chest image shows absorption in the lungs

It would be interesting to compare the discharge protocols for patients in the AH, FO and NH datasets with those of Tongji Hospital.

A second explanation may be related to the different laboratory protocols used in the hospitals. For example, Tongji Hospital uses the 'lactate dehydrogenase acc.to IFCC ver.2' kit made by Roche Diagnostics GmbH to measure LDH. It would be interesting to review the literature to compare protocols between hospitals. Also, Tongji Hospital measures hs-CRP, while some hospitals measure CRP. As shown in ref. [9], these two measurements are not equivalent. Finally, it would be interesting to compare the rates of haemolysis and details of the laboratory protocols overall.

Third, as mentioned by the authors, LDH expression seems to have substantial genetic heterogeneity between Asians and Caucasians[10,11]. Assuming that NH hospitals have a wide range of ethnic patients, further data separated by ethnicity may provide new clues.

Fourth, different hospital treatments or baseline characteristics of patients can influence outcomes.

Fifth, mortality in intensive care and in non-critical care settings has been dropping by 2–5% every week since April 2020[12]. This could create discrepancies between the data used in ref. [1] and the data from the AH, FO and NH datasets. One solution would be to retrain the model as new data become available (see below).

Sixth, it has been reported in refs. [13,14] that there are at least two lineages of SARS-CoV-2 virus. As yet, the implications of these evolutionary changes for disease aetiology remain unclear. It is possible that patients may have different expressions of these biomarkers because they have been infected with different strains.

[1]Luxembourg Centre for System Biomedicine, Belvaux, Luxembourg. [2]Department of Plant Sciences, University of Cambridge, Cambridge, UK. [3]Department of Emergency, Tongji Hospital of Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China. [4]School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, China. [5]School of Mechanical Science and Engineering, Huazhong University of Science and Technology, Wuhan, China. [6]Department of Anesthesiology, Tongji Hospital of Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China. [7]Huazhong University of Science and Technology–Wuxi Research Institute, Wuxi, China. ✉e-mail: yye@hust.edu.cn

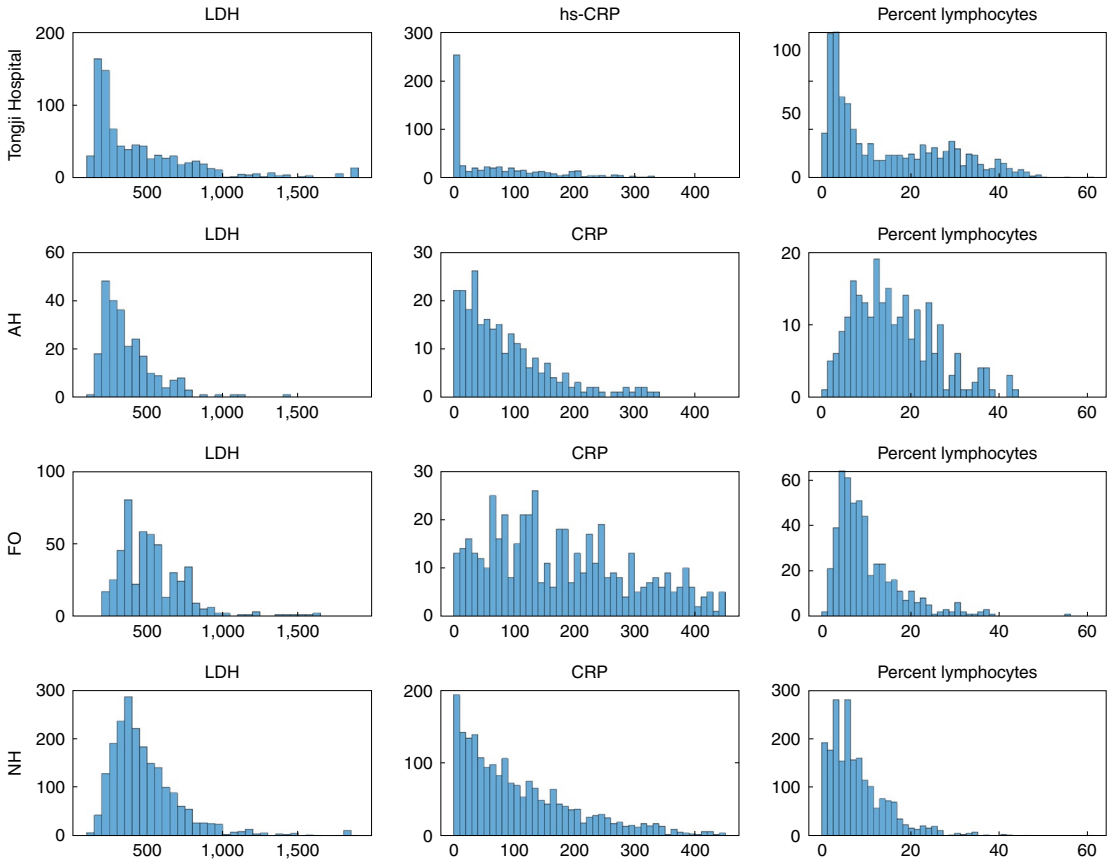


Fig. 1 | Distribution of all three biomarkers for the different datasets. From left to right: distributions of lactate dehydrogenase (LDH/LD), high-sensitivity C reactive protein (hs-CRP) or C reactive protein (CRP) and percent lymphocytes in blood samples from the different datasets: Tongji Hospital (last 10 days), AH (last 10 days), FO (all data) and NH (all data).

In particular, China, Europe and the United States seem to be classified in distinct clusters.

Seventh, patient selection for the FO dataset did not follow the complete patient selection process that was used in ref. [1] and hence did not serve as unbiased validation of the model in ref. [1]. According to ref. [7], the following patients were excluded from the FO dataset: 'very low LDH and CRP serum levels and high lymphocyte counts (these patients have good outcomes)' and 'some of the most severely ill patients with high CRP and LDH serum levels and low lymphocyte counts, who are not admitted to ICU because of therapeutic limitation (these patients have the worst outcomes)'. In essence, patients that would have been correctly classified by our model were removed, leaving only a selection of intermediate patients that are harder to classify, thus reducing the overall accuracy of our model. This is confirmed by their statement, 'Thus, it is not surprising that the predictive rule of Yan et al. was not accurate in our cohort'[7].

Model retraining

Recall that the model in ref. [1] was trained only with the last samples taken, although it could then be applied to other blood samples, including at hospital admission (see below)[15]. For the AH dataset, we retrained the model using data within 10 days from outcome because there was no information as to which samples were the last before outcome (there is only one sample per patient available in the data from AH). The retrained model followed exactly the same single-tree XGBoost method specified in ref. [1]: max depth equal to 3, learning rate equal to 0.1, number of tree estimators set to 1, regularization parameter α set to 0, and 'subsample' and 'colsample_bytree' both set to 1. We achieved an averaged accuracy of 0.83 (0.76, 0.82, 0.9, 0.82, 0.84) using fivefold cross-validation. The fivefold cross-validation was necessary because there were no further available data to test the model. Moreover, we could not test the model on admission data and over time, given that this information was not available.

Retraining was not possible for the FO dataset due to the use of a patient selection process different from that of ref. [1].

For the NH dataset we retrained our decision tree model with the same single-tree XGBoost method and used the last blood samples (exactly the same as in ref. [1]). We achieved a training accuracy of 0.78. Moreover, the retrained model achieved 0.72 accuracy using only the first blood sample (admission samples), showing that the retrained model is useful in triaging patients upon admission. Finally, the mortality arm considerably improved its performance: 75% (65%) on the retrained model versus 41% (40%) on the original model in ref. [1] using the last (first) blood samples. We could not test the model over time as in ref. [1], because the dates of blood samples were not available.

Further validation

Since the publication of the paper[1], we have obtained further data besides that from Tongji Hospital, including from two new hospitals in China. We applied the decision tree in ref. [1] to new patient data from Jinyintan Hospital in Wuhan and No. 3 People's Hospital in Shenzhen[16]. The datasets from Jinyintan and Shenzhen include all patients with COVID-19 for whom there were values for all three biomarkers until 31 March 2020 and 13 April 2020, respectively. The results are presented in Fig. 3. Overall, both hospitals show a performance similar to that of Tongji, with accuracies

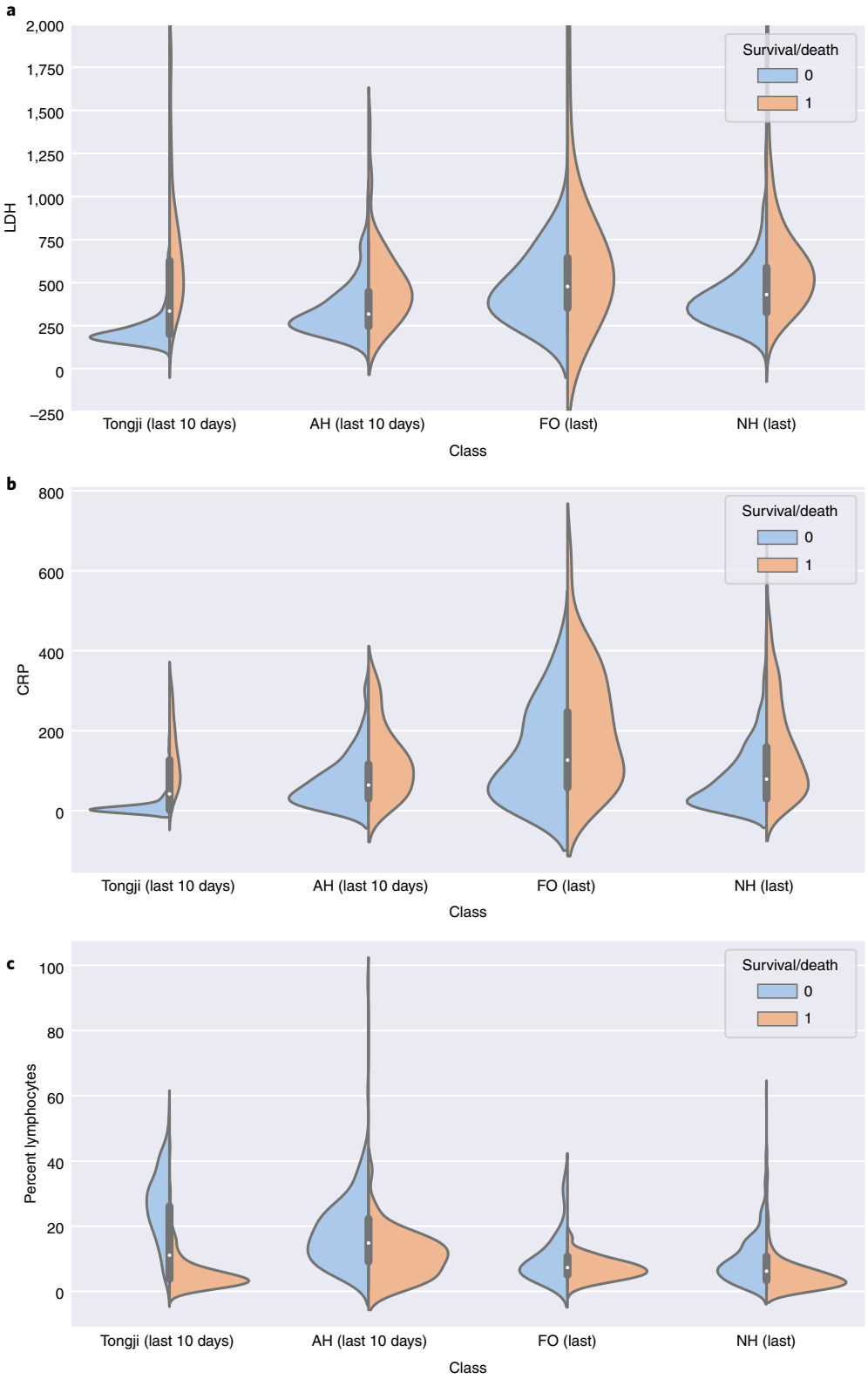


Fig. 2 | Fitted distributions of the three biomarkers on the different datasets, separated by outcome. a–c, Distributions of LDH/LD (**a**), hs-CRP/CRP (**b**) and percent lymphocytes (**c**) in blood samples from the Tongji Hospital (training, test), AH, FO and NH datasets.

of 94% and 90%, respectively. This demonstrates that exactly the same model can predict the mortality of individual patients more than 10 days in advance with more than 90% accuracy in different centres in China.

Another independent study (the WHO COVID-19 database[17]) identified the same top three biomarkers (LDH, lymphocyte, CRP) using data from five Chinese centres (Guizhou Provincial People's Hospital, Affiliated Hospital of Zunyi Medical University,

Table 1 | LDH, CRP and percent lymphocytes results

	Total ($N=454$)	Alive ($n=287$)	Deceased ($n=167$)
LDH	420 (370)	224 (79)	756 (427)
CRP	55 (77)	12 (24)	131 (80)
Lymphocytes (%)	19 (13)	26 (11)	6 (6)
	Total ($N=135$)	**Alive ($n=121$)**	**Deceased ($n=14$)**
LDH	343 (175)	319 (139)	551 (290)
CRP	69 (71)	63 (64)	128 (101)
Lymphocytes (%)	19 (10)	20 (10)	11 (9)
	Total ($N=84$)	**Alive ($n=46$)**	**Deceased ($n=38$)**
LDH	574 (441)	496 (301)	708 (592)
CRP	166 (132)	140 (118)	210 (146)
Lymphocytes (%)	9 (6)	10 (8)	7 (3)
	Total ($N=1,038$)	**Alive ($n=678$)**	**Deceased ($n=360$)**
LDH	500 (285)	426 (214)	640 (344)
CRP	109 (105)	83 (79)	159 (126)
Lymphocytes (%)	8 (7)	10 (7)	5 (6)

Mean (s.d.) for Tongji dataset (last measurements before outcome). Mean (s.d.) for AH dataset (within three days from outcomes). Mean (s.d.) for FO dataset (last measurements before outcome). Mean (s.d.) for NH dataset (last measurements before outcome).

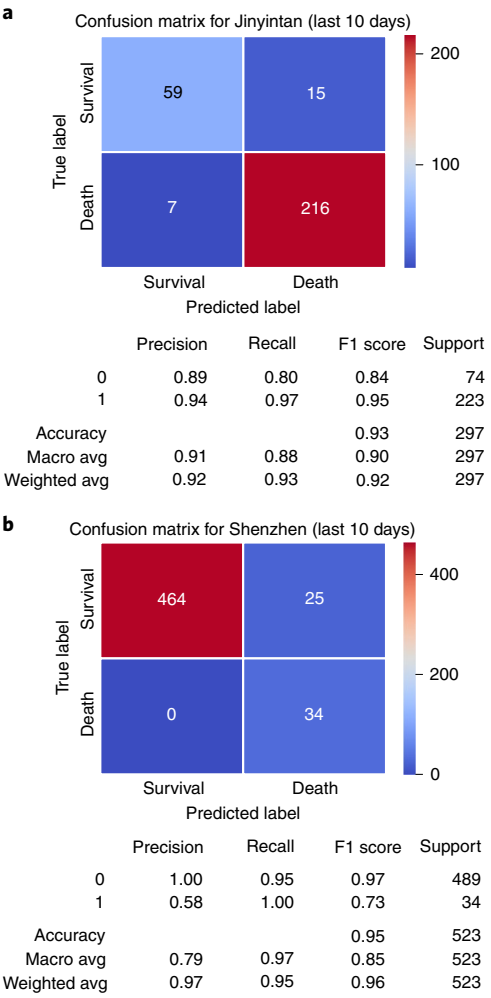


Fig. 3 | Results with datasets from two further hospitals in China. **a**,**b**, Performance of the discovered clinical route on the external test dataset from Jinyintan Hospital (**a**) and No. 3 People's Hospital Shenzhen (**b**).

Jiangjunshan Hospital of Guizhou Province, Zhongnan Hospital of Wuhan University and the Radiology Quality Control Center database of Hunan Province). In addition, there are many other publications that have independently identified similar biomarkers. For example, lymphocytes were identified as a risk factor in refs. [18–22], CRP in refs. [19–21,23,24] and LDH in refs. [18,19,24,25]. This further validates the importance of these three biomarkers.

Admission samples

The comment in ref. [26] suggested presenting the performance of the model from ref. [1] on blood samples taken at admission. Indeed, this analysis should have been in ref. [1] and is now shown in Fig. 4.

The overall accuracy at admission is 88%, with a survival (death) accuracy of 98.8% (48%). More importantly, at admission, of the 110 patients, the model would have stratified 85 as low risk (with only one wrong) and 25 as high risk (of which 12 died). Hence, and as expected, the model was more conservative with high-risk patients. Overall, 85 patients out of a total of 110 (77%) would have been classified at admission as low risk and relieved hospital resources. This shows that the model provides useful triage information at admission.

Discussion

As stated in the last paragraph of the discussion in ref. [1], the model was developed and tested with high accuracy with data from a single hospital in Wuhan, China. All statements in the paper are based on data from Tongji Hospital, as this was the only data we had available at the time of publication. Since the publication of ref. [1], we have further validated the model on data from two additional hospitals in China.

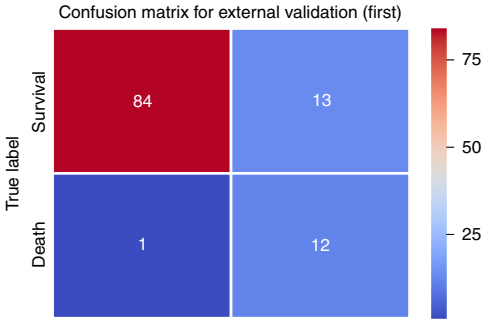
Reference [1] and the comments in relation to AH, FO and NH have opened interesting discussions and research questions that we hope to pursue together. Moreover, we call on the participation of hospitals around the world to share their data and we welcome an opportunity to collaborate. For example, the comments in relation to FO list a number of very interesting extensions for the model and we would welcome collaborations to tackle them.

Finally, it is clear that, at any given time, we do not know how many days are left until the outcome. Nevertheless, fig. 3d,e in ref. [1] shows that the accuracy of prediction improves as new blood samples become available. This remains true even when the date of outcome is unknown (that is, in a practical clinical situation). Note that, even with 18 days until the outcome, the overall cumulative prediction accuracy is still above 90%, and at admission is at 88%.

Summary

We tested the model with three new datasets from St Antonius Hospital in Nieuwegein, the Netherlands (referred to as AH), French Outcomerea (FO) and Northwell Health, United States (NH). The key messages are as follows:

- We retrained our model with data from AH, and the overall cross-validation accuracy increased from 53% to 83%.
- Patient selection for FO followed very different criteria from ref. [1] and hence it does not serve as unbiased validation of the model in ref. [1].
- We retrained our model for NH, and the accuracy increased from 50% to 78% on last blood samples and the testing accuracy increased from 48% to 72% on first blood samples (admission).



	Precision	Recall	F1 score	Support
0	0.99	0.87	0.92	97
1	0.48	0.92	0.63	13
Accuracy			0.87	110
Macro avg	0.73	0.89	0.78	110
Weighted avg	0.93	0.87	0.89	110

Fig. 4 | Results using first blood samples taken at admission from patients at Tongji Hospital. Performance of the discovered clinical route on the external test dataset from Tongji Hospital using patients' first blood samples taken at admission.

- The same model as in ref. [1] (without retraining) was applied to new data from Jinyintan Hospital in Wuhan and No. 3 People's Hospital in Shenzhen, with accuracies of 94% and 90%, respectively.

In some cases, no retraining is needed, while in others, retraining significantly improves model performance.

Data availability

The original data from Tongji Hospital are available in ref. [1]. We do not have permission from Jinyintan and Shenzhen hospitals to release these datasets as they were shared under strict confidentiality agreements.

Code availability

The code implementation is available at https://github.com/HAIRLAB/Pre_Surv_COVID_19.

Received: 22 June 2020; Accepted: 8 October 2020;
Published online: 12 November 2020

References

1. Yan, L. et al. An interpretable mortality prediction model for COVID-19 patients. *Nat. Mach. Intell.* **2**, 283–288 (2020).
2. Quist, J. & Hill, A. R. Serum lactate dehydrogenase (LDH) in *Pneumocystis carinii* pneumonia, tuberculosis and bacterial pneumonia. *Chest* **108**, 415–418 (1995).
3. Chew, K. S. What's new in emergencies trauma and shock? C-reactive protein as a potential clinical biomarker for influenza infection: more questions than answers. *J. Emerg. Trauma Shock* **5**, 115–117 (2012).
4. An, X. et al. Elevated neutrophil to lymphocyte ratio predicts survival in advanced pancreatic cancer. *Biomarkers* **15**, 516–522 (2010).
5. Yuan Y. et al. Development and validation of a prognostic risk score system for COVID-19 inpatients: a multi-center retrospective study in China. Preprint at https://doi.org/10.21203/rs.3.rs-41151/v1 (2020).
6. Quanjel M. et al. Replication of a mortality prediction model in Dutch patients with COVID-19. *Nat. Mach. Intell.* https://doi.org/10.1038/s42256-020-0180-7 (2020).
7. Dupuis C. et al. Limited applicability of a COVID-19 specific mortality prediction rule to the intensive care setting. *Nat. Mach. Intell.* https://doi.org/10.1038/s42256-020-00252-4 (2020).
8. China National Health Commission *Diagnosis and Treatment of 2019-nCoV Pneumonia in China* (in Chinese) (2020); http://www.nhc.gov.cn/yzygj/s7653p/202002/d4b895337e19445f8d728fcaf1e3e13a.shtml.
9. Helal, I. et al. Comparison of C-reactive protein and high-sensitivity C-reactive protein levels in patients on hemodialysis. *Saudi J. Kidney Dis. Transpl.* **23**, 477–483 (2012).
10. Lv, J. et al. Prognostic value of lactate dehydrogenase expression in different cancers: a meta-analysis. *Am. J. Med. Sci.* **358**, 412–421 (2019).
11. Wang, D. et al. Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus-infected pneumonia in Wuhan, China. *JAMA* **323**, 1061–1069 (2020).
12. Armstrong R. A., Kane A. D. & Cook T. M. Outcomes from intensive care in patients with COVID-19: a systematic review and meta-analysis of observational studies. *Anaesthesia* (2020); https://doi.org/10.1111/anae.15201
13. Tang, X. et al. On the origin and continuing evolution of SARS-CoV-2. *Natl Sci. Rev.* **7**, 1012–1023 (2020).
14. Forster, P. et al. Phylogenetic network analysis of SARS-CoV-2 genomes. *Proc. Natl Acad. Sci. USA* **117**, 9241–9243 (2020).
15. Vickers, A. J. et al. Net benefit approaches to the evaluation of prediction models, molecular markers and diagnostic tests. *BMJ* **352**, i6 (2016).
16. Chen, C. et al. Predicting illness severity and short-term outcomes of COVID-19: a retrospective cohort study in China. *Innovation* **1**, 1 (2020).
17. Zheng Y. et al. A learning-based model to evaluate hospitalization priority in COVID-19 pandemics. *Patterns* (2020); https://doi.org/10.1016/j.patter.2020.100092
18. Ji, D. et al. Prediction for progression risk in patients with COVID-19 pneumonia: the CALL Score. *Clin. Infect. Dis.* (2020); https://doi.org/10.1093/cid/ciaa414
19. Xie, J. et al. Development and external validation of a prognostic multivariable model on admission for hospitalized patients with COVID-19. Preprint at https://doi.org/10.1101/2020.03.28.20045997 (2020).
20. Zhang, H. et al. Risk prediction for poor outcome and death in hospital in-patients with COVID-19: derivation in Wuhan, China and external validation in London, UK. Preprint at https://doi.org/10.1101/2020.04.28.20082222 (2020).
21. Guo, Y. et al. Development and validation of an early warning score (EWAS) for predicting clinical deterioration in patients with coronavirus disease 2019. Preprint at https://doi.org/10.1101/2020.04.17.20064691 (2020).
22. Cambridge Clinical Trials Unit. *TACTIC trial* (accessed 1 July 2020); https://cctu.org.uk/portfolio/COVID-19/TACTIC.
23. Lu, J. et al. ACP risk grade: a simple mortality index for patients with confirmed or suspected severe acute respiratory syndrome coronavirus 2 disease (COVID-19) during the early stage of outbreak in Wuhan, China. Preprint at https://doi.org/10.1101/2020.02.20.20025510 (2020).
24. Colombi, D. et al. Well-aerated lung on admitting chest CT to predict adverse outcome in COVID-19 pneumonia. *Radiology* **296**, E86–E96 (2020).
25. Huang, H. et al. Prognostic factors for covid-19 pneumonia progression to severe symptoms based on earlier clinical features: A retrospective analysis. *Front. Med.* **7**, 643 (2020).
26. Barish M. et al. External validation demonstrates limited clinical utility of the interpretable mortality prediction model for patients with COVID-19. *Nat. Mach. Intell.* https://doi.org/10.1038/s42256-020-00254-2 (2020).

Author contributions

J.G. and Y.Y. drafted the manuscript. All authors provided critical review of the manuscript and approved the final draft for publication.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Y.Y.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.