nature machine intelligence

MATTERS ARISING

https://doi.org/10.1038/s42256-020-00254-2

Check for updates

External validation demonstrates limited clinical utility of the interpretable mortality prediction model for patients with COVID-19

Matthew Barish[1,2,8], Siavash Bolourani[3,4,5,6,8], Lawrence F. Lau[1,6,8], Sareen Shah[1,7,8] and Theodoros P. Zanos[1,3,8] ✉

ARISING FROM Yan et al. *Nature Machine Intelligence* https://doi.org/10.1038/s42256-020-0180-7 (2020)

Typical artificial intelligence or machine learning algorithm deployment begins with model training followed by testing under varying circumstances to enable model adaptation and verification of a model's performance. Despite the urgency that the coronavirus disease 2019 (COVID-19) pandemic has posed for deploying predictive models, proper validation is critical before any claims of utility or generalizability are made[1]. In the recent publication by Yan et al.[2], the authors claimed to have developed a novel simplified mortality prediction model that can be applied in clinical practice. We validated their proposed decision tree against a large database of patients with COVID-19[3]. Yan et al. attempted to determine a subset of biomarkers from blood samples taken throughout a patient's hospital course that could be used to predict mortality. The authors described the problem as a classification task, where the inputs were the results of the last set of laboratory tests taken from patients with variable severity and where the outcomes were either discharge or death. Following algorithm generation, the authors assessed the feature importance of each parameter. This generated a set of three key features: lactate dehydrogenase (LDH), lymphocyte proportion and high-sensitivity C-reactive protein (hs-CRP). The simplicity of the three-branch model based on these three values is enticing for rapid, wide-scale adoption in clinical practice. Although the authors reported success in their attempt to determine markers of imminent mortality in their dataset, they used a small validation sample size and did not externally validate their model[3]. We have attempted, under various use cases, to validate their model using data from patients with COVID-19 treated in Northwell Health hospitals. We also attempted to recalibrate the primary branch of the proposed tree-based model based on our data to account for differences between our populations and health systems[1].

Model performance as a triage tool

The model is purported to be applicable to any blood sample, far ahead of the primary clinical outcome, thereby suggesting its use as an admission triage tool. We tested the performance of their model on their validation data at the first time point that all three blood tests were collected, similar to how physicians would risk-stratify new patients (Fig. 1a). This critical time point was not included in the original paper, which is important as the authors suggest that their model can be used to prioritize care. Predictive clinical models are used prospectively where the time of outcome is unknown, unlike this model, which retrospectively utilizes data based on the known date of outcome. Given that the clinician cannot know when the date of discharge or death is to occur, the fact that the performance of the proposed model improves closer to the time of outcome is not clinically useful information. Therefore, it is important to show that the model has sufficient performance to justify changes to clinical care (as the authors suggest) at admission.

Our analysis was performed on Python and R, using code available at https://github.com/siabolourani/YIN_reply. The precision was 0.48 for predicting mortality, meaning that over half of the patients that the model predicted would die actually survived. The accuracy was 0.88 and the F1 score was 0.41. In interpreting these results, one must take care in accounting for imbalanced data—their validation set had a survival rate of 0.88, meaning that the null model of always predicting survival had a similar accuracy as the proposed full model.

Model performance on external data

To test the clinical portability of the mortality prediction model, we validated it externally using the Northwell Health electronic health record database. Northwell Health is the largest academic health system in New York, comprising 12 acute care hospitals that serve ~11 million people in the North American epicentre of the COVID-19 pandemic[4]. The data used for this validation were collected from the Enterprise Electronic Health record (Sunrise Clinical Manager, Allscripts, Chicago) and included patients who had had COVID-19 and had been discharged from Northwell hospitals between 1 March and 31 May 2020. All patients with a final outcome (death or discharged alive) and LDH, hs-CRP and lymphocyte values measured at least once during their hospitalization were considered. Thus,

[1]Donald and Barbara Zucker School of Medicine at Hofstra/Northwell, Northwell Health, Hempstead, NY, USA. [2]Institute of Health Innovations and Outcomes Research, Feinstein Institutes for Medical Research, Northwell Health, Manhasset, NY, USA. [3]Institute of Bioelectronic Medicine, Feinstein Institutes for Medical Research, Northwell Health, Manhasset, NY, USA. [4]Center for Immunology and Inflammation, Feinstein Institutes for Medical Research, Northwell Health, Manhasset, NY, USA. [5]Elmezzi Graduate School of Molecular Medicine, Manhasset, NY, USA. [6]Department of Surgery, Donald and Barbara Zucker School of Medicine at Hofstra/Northwell, Manhasset, NY, USA. [7]Pediatric Critical Care, Cohen Children's Medical Center, New Hyde Park, NY, USA. [8]These authors contributed equally: Matthew Barish, Siavash Bolourani, Lawrence F. Lau, Sareen Shah, Theodoros P. Zanos.
✉e-mail: tzanos@northwell.edu

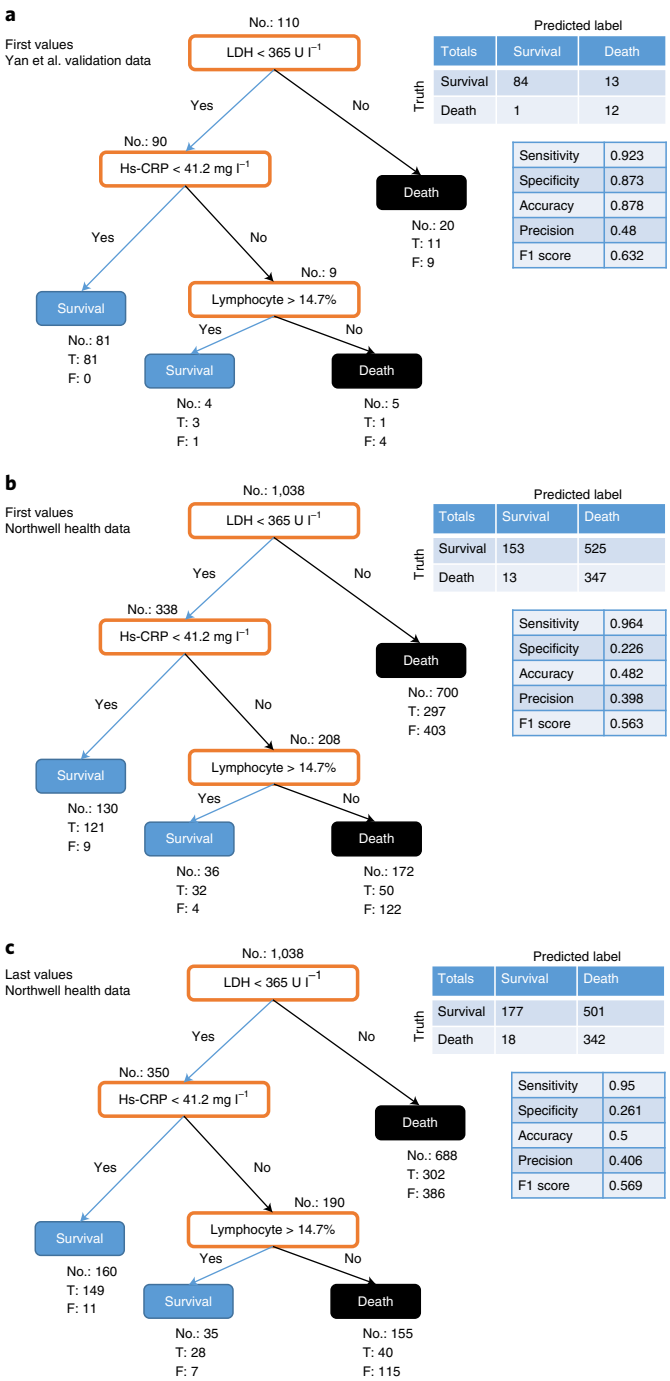

Fig. 1 | Performance of decision rule. a–c, Performance of the decision rule for three settings: first values from the Yan et al. validation data (**a**) and first values (**b**) and last values (**c**) from the Northwell patient data ($n = 1{,}038$). T, true; F, false.

from a total of 13,106 patients, 1,038 patients were included for the validation of the model.

We initially tested the model performance using the first time point when all three laboratory values were available (Fig. 1b). Simulating the operation of the model at this initial triage point, the precision was 0.40 for death (F1 score of 0.56), with an overall accuracy of 0.48.

The model's accuracy is reported to increase with laboratory values drawn closer to the patient's outcome. As previously stated, a clinical model that is contingent on knowing, in advance, the date of outcome, is of dubious use. Nevertheless, we externally validated the model using the final (pre-death or discharge) laboratory values in our dataset. The precision for death remained low at 0.41, with an overall model accuracy of 0.50 (Fig. 1c).

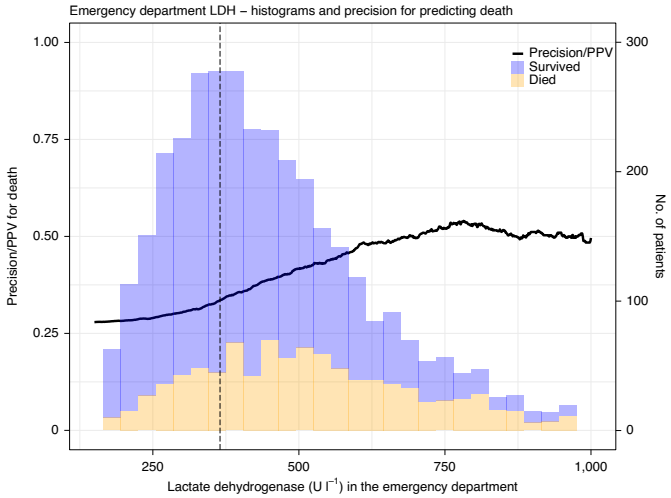

Fig. 2 | LDH as a mortality predictor. Histogram of emergency department LDH values and precision for mortality of an LDH rule for different thresholds based on Northwell COVID-19 patient data ($n = 3{,}595$). The dashed line represents the LDH threshold of 365 U l^{-1} of Yan et al. PPV, positive predictive value.

Recalibrated model performance on external data

LDH alone was the primary driver of the decision tree and an LDH value of >365 U l^{-1} led to a terminal node accounting for 93.0% (146/157) of the true positive predictions of mortality in their dataset. Therefore, it could be argued that LDH alone is a sufficiently robust mortality predictor and naturally lends itself to serve as a triage tool. To test this hypothesis, from all our patients ($n = 13{,}106$), we included those with at least one LDH value from the emergency department ($n = 3{,}595$). With the proposed threshold of LDH > 365 U l^{-1}, the precision for mortality was 0.34 (Fig. 2). We then varied the LDH threshold and found that the maximal precision achieved for this branch was only 0.54, revealing its lack of prognostic utility at our institution as part of a mortality prediction model upon admission.

Conclusion

An interpretable mortality prediction model for patients with COVID-19 is a worthwhile pursuit to help inform clinicians in the battle against this pandemic. We have shown that the recently published model by Yan et al. does not perform as a triage tool based on the internal validation dataset provided by the original authors. Furthermore, we have demonstrated that the decision algorithm was not portable to our large external validation dataset, both with unmodified and optimized parameters. We have thus demonstrated the importance of externally validating this model before its widespread adoption in actual clinical practice, especially given the rapid and widespread dissemination of this model post-publication[5]. Furthermore, our findings, consistent with other studies[6], confirm that the proposed model cannot be recommended for routine clinical implementation.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The datasets analysed and generated during the current study are available from the corresponding author on reasonable request.

Code availability

The analysis code is available at the following repository: https://github.com/siabolourani/YIN_reply.

Received: 4 June 2020; Accepted: 2 October 2020;
Published online: 12 November 2020

References

1. Wynants, L. et al. Prediction models for diagnosis and prognosis of Covid-19 infection: systematic review and critical appraisal. *BMJ* **369**, m1328 (2020).
2. Yan, L. et al. An interpretable mortality prediction model for COVID-19 patients. *Nat. Mach. Intell.* **2**, 283–288 (2020).
3. Riley, R. D. et al. Calculating the sample size required for developing a clinical prediction model. *BMJ* **368**, m441 (2020).
4. Richardson, S. et al. Presenting characteristics, comorbidities and outcomes among 5,700 patients hospitalized with COVID-19 in the New York City area. *JAMA* **323**, 2052–2059 (2020).
5. An interpretable mortality prediction model for COVID-19 patients. *Altmetric* https://www.altmetric.com/details/82019437/ (2020).
6. Gupta, R. K. et al. Systematic evaluation and external validation of 22 prognostic models among hospitalised adults with COVID-19: an observational cohort study. *Eur. Resp. J.* https://doi.org/10.1183/13993003.03498-2020 (2020).

Acknowledgements

The authors would like to thank J. Hirsh and K. Coppa for providing the queries and data that enabled this study, as well as the Northwell Machine Learning in Medicine group, whose discussions inspired this paper. We acknowledge and honor all our Northwell team members who consistently put themselves in harm's way during the COVID-19 pandemic. We dedicate this article to them, as their vital contribution to knowledge about COVID-19 and sacrifices on the behalf of patients made it possible.

Author contributions

S.B., S.S. contributed in the data analysis. S.S. and T.P.Z. created the figures. M.B., S.B., S.S., L.F.L. and T.P.Z. contributed in the writing, editing and overall concept of the manuscript. T.P.Z. coordinated the project and led the submission.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s42256-020-00254-2.

Correspondence and requests for materials should be addressed to T.P.Z.

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.

nature research

Corresponding author(s): Theodoros Zanos

Last updated by author(s): Jun 11, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Data were collected from the enterprise electronic health record (EHR; Sunrise Clinical Manager, Allscripts, Chicago, IL). Transfers from 1 in-system hospital to another were merged and considered 1 hospital visit. All patients with a final outcome (death or discharged alive) were considered. Data collected for the validation of the tool included the specific laboratory values (LDH, Hs-CRP and Lymphocyte) in the sequence they were acquired, and the patient final outcomes (i.e., death, discharge).
Data analysis	Data analysis was performed on Python and R, using code available at the following repository: https://github.com/siabolourani/YIN_reply

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Provide your data availability statement here.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We used two patient cohorts. For the first cohort (FIgure 1B & Figure 1C), we included all COVID-19 positive hospitalized patients that had the three required values (LDH, Hs-CRP and Lympocyte) available, yielding N=1038. For the second cohort, we used all the COVID-19 positive hospitalized patients that had LDH measured, yielding N=3595.
Data exclusions	No patients were excluded from our analysis.
Replication	Two team members (SB & SS) have run independently all analyses and confirmed all values reported.
Randomization	Not relevant since this is a retrospective
Blinding	Blinding was not possible or needed in this retrospective validation analysis.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☐	☒ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	This is not a clinical trial but a retrospective minimal-risk research project.
Study protocol	This study was approved by the Institutional Review Boards at Northwell Health as minimal-risk research that used data collected for routine clinical practice, and as such, waived the requirement for informed consent.
Data collection	Data were collected from the enterprise electronic health record (EHR; Sunrise Clinical Manager, Allscripts, Chicago, IL).
Outcomes	All patients with a final outcome (death or discharged alive) were considered. Data collected for the validation of the tool included the specific laboratory values (LDH, Hs-CRP and Lymphocyte) in the sequence they were acquired, and the patient final outcomes (i.e., death, discharge).