Prospective predictive performance comparison between clinical gestalt and validated COVID-19 mortality scores

Adrian Soto-Mota ,[1,2] Braulio Alejandro Marfil-Garza,[2,3] Santiago Castiello-de Obeso,[4,5] Erick Jose Martinez Rodriguez,[2] Daniel Alberto Carrillo Vazquez,[2] Hiram Tadeo-Espinoza,[2] Jessica Paola Guerrero Cabrera,[2] Francisco Eduardo Dardon-Fierro,[2] Juan Manuel Escobar-Valderrama,[2] Jorge Alanis-Mendizabal,[2] Juan Gutierrez-Mejia[2]

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/jim-2021-002037).

[1]Metabolic Diseases Research Unit, National Institute of Medical Sciences and Nutrition Salvador Zubirán, Mexico City, Mexico
[2]Internal Medicine, National Institute of Medical Sciences and Nutrition Salvador Zubiran, Mexico
[3]CHRISTUS-LatAm Hub – Excellence and Innovation Center, Monterrey, Mexico
[4]Experimental Psychology, University of Oxford, Oxford, UK
[5]Universidad de Guadalajara, Guadalajara, Jalisco, Mexico

Correspondence to
Dr Adrian Soto-Mota, University of Oxford, Oxford OX1 2JD, UK; adrian.sotom@incmnsz.mx

Accepted 15 September 2021

To cite: Soto-Mota A, Marfil-Garza BA, Castiello-de Obeso S, *et al*. *J Investig Med* Epub ahead of print: [*please include* Day Month Year]. doi:10.1136/jim-2021-002037

ABSTRACT

Most COVID-19 mortality scores were developed at the beginning of the pandemic and clinicians now have more experience and evidence-based interventions. Therefore, we hypothesized that the predictive performance of COVID-19 mortality scores is now lower than originally reported. We aimed to prospectively evaluate the current predictive accuracy of six COVID-19 scores and compared it with the accuracy of clinical gestalt predictions. 200 patients with COVID-19 were enrolled in a tertiary hospital in Mexico City between September and December 2020. The area under the curve (AUC) of the LOW-HARM, qSOFA, MSL-COVID-19, NUTRI-CoV, and NEWS2 scores and the AUC of clinical gestalt predictions of death (as a percentage) were determined. In total, 166 patients (106 men and 60 women aged 56±9 years) with confirmed COVID-19 were included in the analysis. The AUC of all scores was significantly lower than originally reported: LOW-HARM 0.76 (95% CI 0.69 to 0.84) vs 0.96 (95% CI 0.94 to 0.98), qSOFA 0.61 (95% CI 0.53 to 0.69) vs 0.74 (95% CI 0.65 to 0.81), MSL-COVID-19 0.64 (95% CI 0.55 to 0.73) vs 0.72 (95% CI 0.69 to 0.75), NUTRI-CoV 0.60 (95% CI 0.51 to 0.69) vs 0.79 (95% CI 0.76 to 0.82), NEWS2 0.65 (95% CI 0.56 to 0.75) vs 0.84 (95% CI 0.79 to 0.90), and neutrophil to lymphocyte ratio 0.65 (95% CI 0.57 to 0.73) vs 0.74 (95% CI 0.62 to 0.85). Clinical gestalt predictions were non-inferior to mortality scores, with an AUC of 0.68 (95% CI 0.59 to 0.77). Adjusting scores with locally derived likelihood ratios did not improve their performance; however, some scores outperformed clinical gestalt predictions when clinicians' confidence of prediction was <80%. Despite its subjective nature, clinical gestalt has relevant advantages in predicting COVID-19 clinical outcomes. The need and performance of most COVID-19 mortality scores need to be evaluated regularly.

Significance of this study

What is already known about this subject?

- Multiple scores have been designed or repurposed to predict survival in patients with COVID-19; however, all of them were designed or validated during the early days of the pandemic and COVID-19 healthcare has greatly improved since then.
- Clinical gestalt has been proven to accurately predict survival in other clinical contexts.

What are the new findings?

- The observed area under the curve of all scores was significantly lower than originally reported.
- No score was significantly better than clinical gestalt predictions.

How might these results change the focus of research or clinical practice?

- The need and performance of most COVID-19 mortality scores need to be re-evaluated with regularity.

INTRODUCTION

Background

Many prediction models have been developed for COVID-19[1–5] and their applications in healthcare range from bedside counseling to triage systems.[6] However, most have been developed within specific clinical contexts[1 2] or validated with data from the early months of the pandemic.[4 5] Since then, health systems have implemented protocols and adaptations to cope with surge in hospitalization rates,[7] and now clinicians have more knowledge and experience in managing these patients. Additionally, other non-biological factors such as critical care availability have been found to strongly influence the prognosis of patients with COVID-19.[8 9] These frequently intangible factors (eg, the experience of the staff with specific healthcare tasks) impact prognosis but are ignored by mortality scores.

Prediction models are context-sensitive[10]; therefore, to preserve their accuracy, they must be applied in contexts as similar as possible to the ones where they were derived from. Considering that healthcare systems and settings are quite different around the world, there are many examples of scores requiring adjustments or local adaptations.[11 12]

Predicting is an everyday activity in most medical fields, and in other scenarios clinicians' subjective predictions have been observed to be as accurate as mathematically derived models.[13–15] However, the opposite has been observed as well; for example, clinicians tend to overestimate the long-term survival of oncological patients.[16]

This work aimed to compare the predictive performance of different mortality prediction models for COVID-19 (some of them in the same hospital they were developed) against their original performance and clinical gestalt predictions.

METHODS

Study design

This observational prospective study was carried out in a tertiary hospital in Mexico City, fully dedicated to providing COVID-19 healthcare, between October and December 2020.

Selection of subjects

Data from 200 consecutive hospital admissions (for RT-PCR-confirmed COVID-19 infection) were obtained between October and December 2020. We excluded from the analysis all patients without a documented clinical outcome (eg, if they had not been discharged at the moment of data collection, transferred to another hospital, voluntarily discharged). A total of 166 patients were included in the analysis because 34 patients were either transferred to other hospitals or voluntarily discharged. The most frequent criteria for hospital admission were requiring supplemental oxygen to reach oxygen saturation >90%, respiratory rate >20, need for ventilation (non-invasive or invasive), severity of pneumonia based on CT, hemodynamic instability, and impossibility of home isolation.

A total of 24 internal medicine residents with more than 6 months of experience (all residency programs in Mexico start every year on March 1) in COVID-19 healthcare participated in the study. Their median years of hospital experience was 2 (IQR 1–3).

Measurements

Clinical gestalt predictions and all necessary data to calculate prognostic scores were obtained at hospital admission from October to December 2020. Internal medicine residents in charge of collecting clinical history, physical examination, and initial imaging and laboratory work-up were asked the following questions once all initial imaging and laboratory reports were available:

- How likely do you think this patient will die from COVID-19? (as a percentage).
- How confident are you of that prediction? (as a percentage).

To obtain the earliest and best informed clinical gestalt prediction available, we asked only the resident in charge of each patient's hospital admission. While it is likely that clinical gestalt scores vary between evaluators, inviting more evaluators would require evaluating the same patient at different times (giving a 'predictive advantage' to later scorers who would be able to see if a patient is improving after their initial therapeutic interventions) and would allow predictions with different levels of information (from evaluators who did not spend the same amount of time directly examining the patient).

To test the hypothesis that updating the statistical weights of a score with local data could help preserve its original accuracy, we developed a second version of the LOW-HARM score (LOW-HARM V.2 score) using positive and negative likelihood ratios derived from cohorts of Mexican patients[4 8] (instead of only positive likelihood ratios from Chinese patients[17 18] as in the original version).

The likelihood ratios (LR+/LR−) used to calculate the LOW-HARM V.2 score were as follows: oxygen saturation <88%=2.61/0.07; previous diagnosis of hypertension=2.37/0.65; elevated troponin (>20 pg/mL)=15.6/0.62; elevated creatine phosphokinase (>223 U/L)=2.37/0.88; leukocyte count >10.0 cells $\times 10^9$/L=5.6/0.48; lymphocyte count <800 cells/µL (<0.8 cells/mm^3)=2.24/0.48; and serum creatinine >1.5 mg/dL=19.1/0.6.

All previously validated scores were calculated by the research team.

Outcomes

The primary outcome of this study was the area under the curve (AUC) of each COVID-19 mortality prediction method. To test the hypothesis that the predictive performance of already validated scores declined over time, we chose the LOW-HARM,[4] MSL-COVID-19, and NUTRI-CoV[5] scores because all these three were validated with data from Mexican patients with COVID-19. To rule out that this was a phenomenon exclusive of scores developed with Mexican data, we re-evaluated the accuracy of NEWS2[1] and qSOFA[2] scores and the neutrophil to lymphocyte ratio to predict mortality from COVID-19.[19]

To test the hypothesis that scores outperformed clinical gestalt predictions when their confidence was 'low' (below or equal to the median perceived confidence; ie, <80%), we conducted a comparative AUC analysis of cases below or above this threshold.

Analysis

Clinical and demographic data were analyzed using mean or median (depending on their distribution) and SD or IQR as dispersion measures. Shapiro-Wilk tests were used to assess if variables were normally distributed.

R V.4.0.3 packages 'caret' for confusion matrix calculations and 'pROC' for receiver operating characteristic curve (ROC) analysis and STATA V.12 software were used for statistical analysis. AUC differences were analyzed using DeLong's method with the STATA function 'roccomp'.[20] A p value of <0.05 for inferring statistical significance was used in all statistical tests. Missing data were handled by mean substitution.

Table 1 Patient demographics and clinical data

	Total (N=166)	Died (n=47)	Survived (n=119)	P value*
Female, n (%)	60 (36.1)	18 (38.3)	42 (35.3)	0.717
Age, years (IQR)	56 (45–64)	61 (54–69)	52 (42–63)	0.0002
Weight, kg (IQR)	78 (70–90)	78 (65–96)	79 (72–90)	0.659
Height, cm (IQR)	165 (158–170)	165 (160–172)	164 (580–170)	0.578
BMI (IQR)	29 (25.4–33)	28 (24–33)	29 (27–32)	0.302
Obesity, n (%)	77 (46.4)	21 (44.7)	56 (47.1)	0.782
Diabetes mellitus, n (%)	42 (25.3)	12 (25.5)	30 (25.2)	0.966
Hypertension, n (%)	49 (29.5)	17 (36.2)	32 (26.9)	0.238
Smoking, n (%)	37 (22.3)	10 (21.3)	27 (22.7)	0.844
Immunosuppression, n (%)	25 (15.1)	6 (12.8)	19 (16.0)	0.603
COPD, n (%)	7 (4.2)	4 (8.5)	3 (2.5)	0.084
CKD, n (%)	9 (5.4)	2 (4.3)	7 (5.9)	0.677
CAD, n (%)	8 (4.8)	4 (8.5)	4 (3.4)	0.163
SpO_2 % <88% with supplemental oxygen, n (%)	156 (94.0)	47 (100)	109 (91.6)	0.040
IMV/CPAP, n (%)	96 (57.8)	33 (70.2)	63 (52.9)	0.042
Positive troponin/CPK, n (%)	77 (46.4)	31 (66)	46 (38.7)	0.001
Creatinine >1.5 mg/dL, n (%)	25 (15.1)	14 (29.8)	11 (9.2)	0.001
WCC >10.0 cells × 10^9 / L, n (%)	94 (56.6)	35 (74.5)	59 (50)	0.004
Lymphocytes <800 cells/µL, n (%)	113 (68.1)	38 (80.9)	75 (63)	0.026
Neutrophil to lymphocyte ratio >9.8, n (%)	60 (36.1)	27 (57.5)	33 (27.7)	<0.0001
Length of stay, days (IQR)	15.5 (9–27)	17 (11–27)	13 (8–27)	0.5408

*Comparisons were done between deaths and survivors. X^2 was used to compare categorical variables. Mann-Whitney U test was used to compare continuous variables.
BMI, body mass index; CAD, coronary artery disease; CKD, chronic kidney disease; COPD, chronic obstructive pulmonary disease; CPAP, continuous positive airway pressure; CPK, creatine phosphokinase; IMV, invasive mechanical ventilation; SpO_2, oxygen saturation; WCC, white cell count.

Sample size rationale

We calculated sample size using 'easyROC',[21] an open R-based web tool used to estimate sample sizes for direct and non-inferior AUC comparisons using Obuchowski's method[22]; to detect non-inferiority with >0.05 maximal AUC difference with the reported AUC of LOW-HARM (0.96, 95% CI 0.94 to 0.98), a case allocation ratio of 0.7 (because the mortality at our center is ~0.3), a power of 0.8, and a significance cut-off level of 0.05, 159 patients would be needed. To detect >0.1 difference between AUCs, 99 patients would be needed with the rest of the parameters held constant. To allow a patient loss rate of ~25%, we obtained data from 200 consecutive hospital admissions.

Patient and public involvement

Patients or the public were not involved in the design, or conduct, or reporting, or dissemination plans of our research.

RESULTS

Characteristics of study subjects

We included 166 patients in our study. Of these, 47 (28.3%) died, while 119 (71.7%) survived. The general demographics and clinical characteristics of these populations are shown in table 1. As expected, decreased peripheral saturation, ventilatory support, cardiac injury, renal injury, leukocytosis, and lymphocytosis were more prevalent in the group of patients who died during their hospitalization.

Main results

Table 2 shows the median scores and their IQR for each prediction tool. As expected, there was a more pronounced mean difference between groups in scores that were based on a 100-point scale (clinical gestalt, LOW-HARM scores). Table 2 shows the originally reported AUC versus the AUC we observed in our data.

Performance characteristics of selected predictive models and AUC comparisons

Figure 1 shows the performance characteristics of the selected predictive models. Overall, we found a statistically significant difference between predictive models (p=0.002). However, we did not find statistically significant differences between clinical gestalt and other prediction tools.

As expected, we found that the confidence of prediction increased in cases in which the predicted probability of death was clearly high or clearly low (figure 2).

We found a moderate-strong, bimodal correlation between the confidence of prediction and the predicted probability of death at <50% predicted probability of death (Pearson's r=0.60, p<0.0001) and at >50% predicted probability of death (Pearson's r=0.50, p=0.0002).

We further explored the performance characteristics of the selected predictive models in specific contexts (online supplemental appendix table 1). Figure 3 shows the results of the analysis including cases in which the certainty of prediction was below and above 80%. Overall, we found a statistically significant difference between predictive models in both settings. In cases in which the confidence of prediction was ≤80%, both versions of the LOW-HARM scores showed a larger AUC compared with clinical gestalt (figure 3B and online supplemental appendix table 1).

An additional analysis restricted to cases in which the certainty of prediction was ≤80% and the predicted probability of death was ≤30% (ie, median value for all cases) found a statistically significant difference between

Table 2 Distribution and accuracy of selected mortality prediction tools

Prediction tool	Total (N=166)	Died (n=47)	Survived (n=119)	Original AUC (95% CI)	Observed AUC (95% CI)
Clinical gestalt, confidence (IQR)	30 (20–50) 80 (70–90)	40 (30–70) 80 (60–90)	30 (15–40) 80 (70–90)	–	0.68 (0.59 to 0.77)
LOW-HARM (IQR)	46 (8.4–83.8)	86 (37.5–99.3)	37.5 (6.4–69)	0.96 (0.94 to 0.98)	0.76 (0.69 to 0.84)
LOW-HARM V.2 (IQR)	9.7 (0.9–52.7)	49 (9.7–96.3)	3.2 (0.5–28.1)	–	0.78 (0.71 to 0.86)
NUTRI-CoV (IQR)	9 (7–12)	10 (8–12)	9 (7–11)	0.79 (0.76 to 0.82)	0.60 (0.51 to 0.69)
MSL-COVID-19 (IQR)	8 (7–10)	8 (8–10)	8 (7–9)	0.72 (0.69 to 0.75)	0.64 (0.55 to 0.73)
qSOFA (IQR)	1 (1–1)	1 (1–2)	1 (1–1)	0.74 (0.65 to 0.81)	0.61 (0.53 to 0.69)
NEWS2 (IQR)	7.5 (6–9)	9 (7–10)	7 (5–9)	0.84 (0.79 to 0.90)	0.65 (0.56 to 0.75)
Neutrophil to lymphocyte ratio >9.8 (%)	64.4	55.3	27.7	0.74 (0.62 to 0.85)	0.65 (0.57 to 0.73)

Overall comparison test for observed AUC=0.002.
Individual AUROC comparisons: clinical gestalt vs all scores, p>0.05.
To calculate the relative mean difference, some scores (those not based on 100 points) were converted to a percentage in the following manner: (patient score/maximum possible score)×100.
AUC, area under the curve.

predictive models (p=0.0005). Similarly, individual comparisons showed a larger statistically significant AUC differences between clinical gestalt and both versions of the LOW-HARM score (online supplemental appendix table 1).

DISCUSSION

Outcome prediction plays an important role in everyday clinical practice. This work highlights the inherent limitations of statistically derived scores and some of the advantages of clinical gestalt predictions. In other scenarios where using predictive scores is frequent, more experienced clinicians can always ponder their sometimes subjective, yet quite valuable insight. However, with the COVID-19 pandemic, clinicians of all levels of training started their learning curve at the same time. In this study, we had the unique opportunity of re-evaluating more than one score (two of them in the same setting and for the same purpose they were designed for), while testing the accuracy of clinical gestalt, in a group of clinicians who started their learning curve for managing a disease at the same time (experience and training within healthcare teams are usually mixed for other diseases).

Additionally, we explored the accuracy of clinical gestalt across different degrees of prediction confidence. To our knowledge, this is the first time that this type of analysis is done for subjective clinical predictions and proved to be quite insightful. The fact that clinical gestalt's accuracy correlates with confidence in prediction suggests that while there is value in subjective predictions, it is also important to ask ourselves about how confident we are about our predictions. Interestingly, our results suggest clinical gestalt predictions are particularly prone to being positively biased,

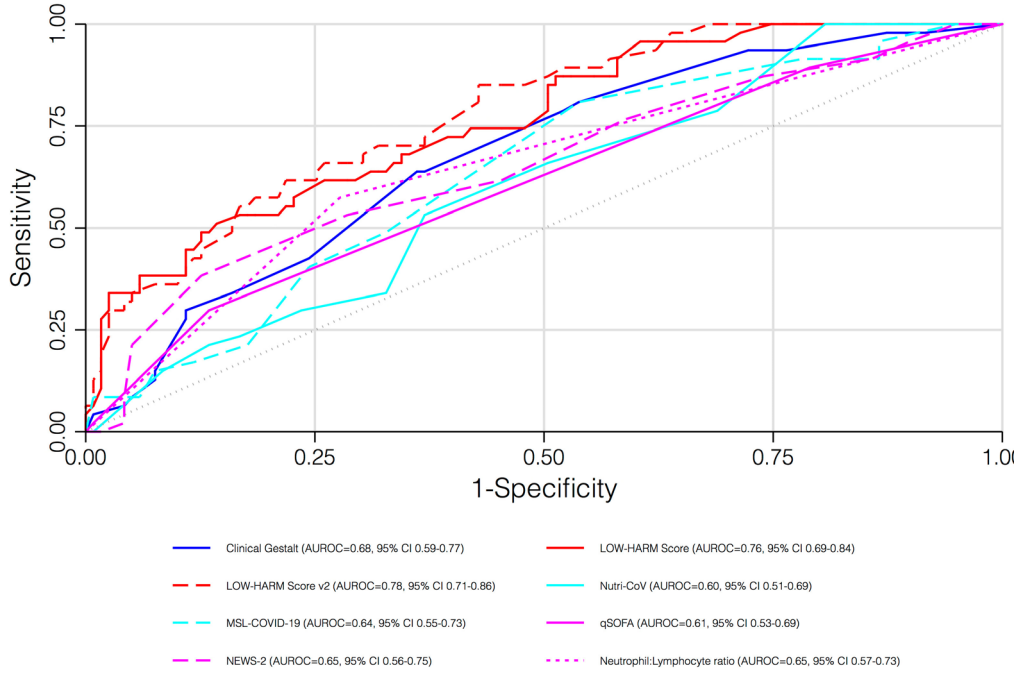

Figure 1 AUC comparison of selected mortality prediction tools. AUC, area under the curve.

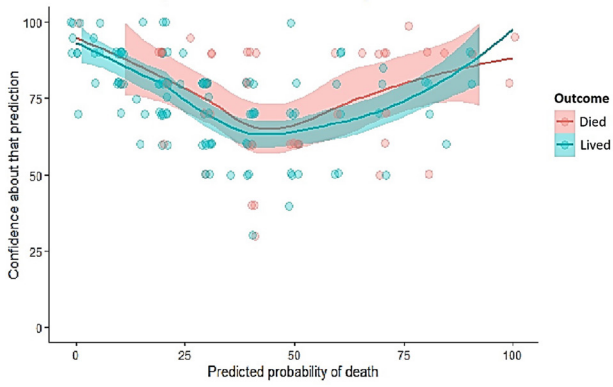

Figure 2 Clinical gestalt prediction and confidence of prediction.

and that clinicians were more likely to correctly predict which patients would survive than which patients would die (figure 2 and online supplemental figure 1). This is consistent with other studies that have found that clinicians tend to overestimate the effectiveness of their treatments and therefore patient survival.[16]

Since it is expected that scores will lose at least some of their predictive accuracy when used outside the context they were developed in, it has already been reported that local adaptations improve or help retain their predictive performance. In this work, we tried to evaluate if by updating the likelihood ratio values used in the calculation of the LOW-HARM score with data from Mexican patients we could mitigate its loss of accuracy. However, despite the AUC of the LOW-HARM V.2 score being slightly larger than the AUC of the original LOW-HARM score, the difference was not statistically significant nor significantly more accurate than clinical gestalt predictions. This highlights the fact that scores are far from being final or perfect tools even after implementing local adjustments.

Limitations

Even when some of the results in this study can prove insightful for other clinical settings and challenges, our results cannot be widely extrapolated due to the local setting of our work and the highly heterogeneous nature of COVID-19 healthcare systems. Additionally, it is likely that emerging variants, vaccination, or the seasonality of contagion waves[23] will continue to influence the predictive capabilities of all predictive models. Additionally, our sample size was calculated to detect non-inferiority between prediction methods. On the other hand, it is possible that, despite having comparable experience with COVID-19, overall clinical experience still influences the accuracy of clinical gestalt predictions. We were not able to account for this source of variability because of how our hospital's patient admission workflows are designed (senior attendings usually meet patients after their initial work-up is complete and their prediction would also be informed by the success or failure of the early therapeutic interventions).

Furthermore, individual consistency cannot be accurately estimated as, on average, each clinician evaluated seven patients only. Nonetheless, 87.5% of the residents (21 of 24) provided at least one prediction per quartile, and we did not observe any of them consistently registering high nor low clinical gestalt scores.

Specifically designed studies are needed to better investigate the relationship between subjective confidence, accuracy, and positive bias. Clinical predictions will always be challenging because all medical fields are in constant development and clinical challenges are highly dynamic phenomena.

All scores had lower predictive accuracy than in their original publications and none of them showed better predictive performance than clinical gestalt predictions; however, scores could still outperform clinical gestalt when confidence in clinical gestalt predictions is perceived to be low. These results remind us that prognostic scores require constant re-evaluation even after being properly validated and adjusted and that no score can or should ever substitute careful medical assessments and thoughtful clinical judgment. Despite its inherent subjectivity, clinical gestalt immediately incorporates context-specific factors, and in contrast

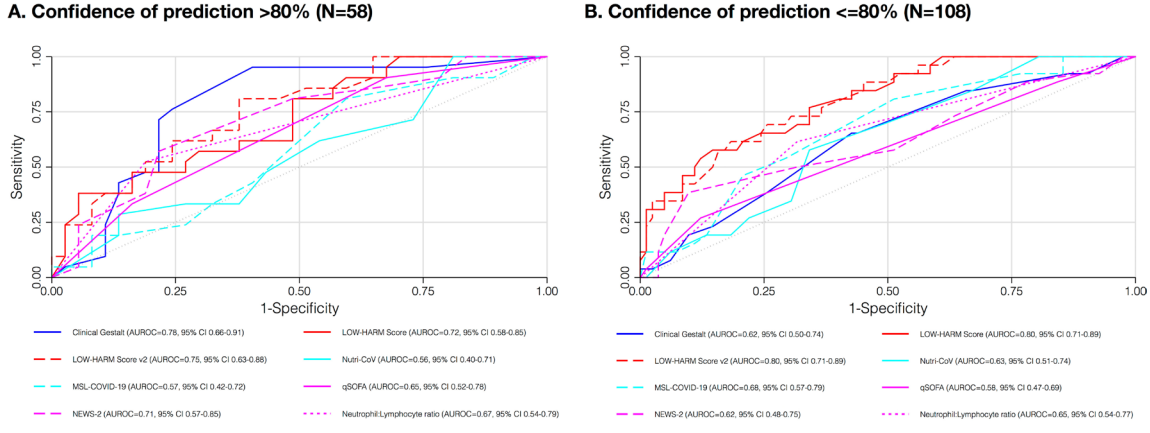

Figure 3 AUC comparison of selected mortality prediction tools according to confidence of prediction. (A) AUC comparison of selected mortality prediction tools in cases where the confidence of prediction was >80%. (B) AUC comparison of selected mortality prediction tools in cases where the confidence of prediction was ≤80%. AUC, area under the curve.

to statistically derived models it is likely to improve its accuracy over time.

Acknowledgements All authors wish to thank the invaluable support of the National Institute of Medical Sciences and Nutrition Salvador Zubirán Emergency Department staff.

Contributors AS-M led, designed, collected, and analyzed research data. BAM-G, SCdO, and JG-M designed and analyzed research data. EMR, DACV, HT-E, JPGC, FED-F, JME-V, and JA-M collected and analyzed research data. All authors contributed to elaborating this manuscript and approved this version.

Funding BAMG is currently supported by the patronage of the National Institute of Medical Sciences and Nutrition Salvador Zubirán and by the Foundation for Health and Education Dr Salvador Zubirán (FunSaEd), and the CHRISTUS Excellence and Innovation Center.

Competing interests None declared.

Patient consent for publication Not required.

Ethics approval This study was approved by the Ethics Committee for Research on Humans of the National Institute of Medical Sciences and Nutrition Salvador Zubirán on August 25, 2020 (reg no DMC-3369-20-20-1-1a).

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request.

ORCID iD

Adrian Soto-Mota http://orcid.org/0000-0002-9173-7440

REFERENCES

1 Rigoni M, Torri E, Nollo G, *et al*. NEWS2 is a valuable tool for appropriate clinical management of COVID-19 patients. *Eur J Intern Med* 2021;85:118–20.
2 Liu S, Yao N, Qiu Y, *et al*. Predictive performance of SOFA and qSOFA for in-hospital mortality in severe novel coronavirus disease. *Am J Emerg Med* 2020;38:2074–80.
3 Ma A, Cheng J, Yang J, *et al*. Neutrophil-to-lymphocyte ratio as a predictive biomarker for moderate-severe ARDS in severe COVID-19 patients. *Crit Care* 2020;24.
4 Soto-Mota A, Marfil-Garza BA, Martínez Rodríguez E, *et al*. The low-harm score for predicting mortality in patients diagnosed with COVID-19: a multicentric validation study. *J Am Coll Emerg Physicians Open* 2020;1:1436–43.
5 Bello-Chavolla OY, Antonio-Villa NE, Ortiz-Brizuela E, *et al*. Validation and repurposing of the MSL-COVID-19 score for prediction of severe COVID-19 using simple clinical predictors in a triage setting: the Nutri-CoV score. *PLoS One* 2020;15:e0244051.
6 White DB, Lo B. A framework for rationing ventilators and critical care beds during the COVID-19 pandemic. *JAMA* 2020;323:1773.
7 OECD/European Union. How resilient have European health systems been to the COVID-19 crisis? In: *Health at a glance: Europe 2020: state of health in the EU cycle*, 2020: 23–81.
8 Olivas-Martínez A, Cárdenas-Fragoso JL, Jiménez JV, *et al*. In-Hospital mortality from severe COVID-19 in a tertiary care center in Mexico City; causes of death, risk factors and the impact of hospital saturation. *PLoS One* 2021;16:e0245772.
9 Najera H, Ortega-Avila AG. Health and institutional risk factors of COVID-19 mortality in Mexico, 2020. *Am J Prev Med* 2021;60:471–7.
10 Khan Z, Hulme J, Sherwood N. An assessment of the validity of SOFA score based triage in H1N1 critically ill patients during an influenza pandemic. *Anaesthesia* 2009;64:1283–8.
11 Fronczek J, Polok K, Devereaux PJ, *et al*. External validation of the revised cardiac risk index and national surgical quality improvement program myocardial infarction and cardiac arrest calculator in noncardiac vascular surgery. *Br J Anaesth* 2019;123:421–9.
12 Carr E, Bendayan R, Bean D, *et al*. Evaluation and improvement of the National Early Warning Score (NEWS2) for COVID-19: a multi-hospital study. *BMC Med* 2021;19:23.
13 Ros MM, van der Zaag-Loonen HJ, Hofhuis JGM, *et al*. Survival prediction in severely ill patients study—the prediction of survival in critically ill patients by ICU physicians. *Crit Care Explor* 2021;3:e0317.
14 Donzé J, Rodondi N, Waeber G, *et al*. Scores to predict major bleeding risk during oral anticoagulation therapy: a prospective validation study. *Am J Med* 2012;125:1095–102.
15 Nazerian P, Morello F, Prota A, *et al*. Diagnostic accuracy of physician's gestalt in suspected COVID-19: prospective bicentric study. *Acad Emerg Med* 2021;28:404–11.
16 Cheon S, Agarwal A, Popovic M, *et al*. The accuracy of clinicians' predictions of survival in advanced cancer: a review. *Ann Palliat Med* 2016;5:22–9.
17 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.
18 Yan L, Zhang H-T, Goncalves J, *et al*. An interpretable mortality prediction model for COVID-19 patients. *Nat Mach Intell* 2020;2:283–8.
19 Liu J, Liu Y, Xiang P. Neutrophil-To-Lymphocyte ratio predicts severe illness patients with 2019 novel coronavirus in the early stage. *medR* 2020.
20 DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988;44:837.
21 Goksuluk D, Korkmaz S, Zararsiz G, *et al*. EasyROC: an interactive web-tool for ROC curve analysis using R language environment. *R J* 2016;8:213–30.
22 Obuchowski NA. Roc analysis. *AJR Am J Roentgenol* 2005;184:364–72.
23 Birkmeyer JD, Barnato A, Birkmeyer N, *et al*. The impact of the COVID-19 pandemic on hospital admissions in the United States. *Health Aff* 2020;39:2010–7.