External validation of risk scores to predict in-hospital mortality in patients hospitalized due to coronavirus disease 2019.

Running head: Validation of COVID-19 risk scores

Shermarke Hassan[1,2], Chava L. Ramspek[2], Barbara Ferrari[3], Merel van Diepen[2], Raffaella Rossio[3], Rachel Knevel[4], Vincenzo la Mura[1,3], Andrea Artoni[5], Ida Martinelli[5], Alessandra Bandera[1,6], Alessandro Nobili[7], Andrea Gori[1,6], Francesco Blasi[1,8], Ciro Canetta[9], Nicola Montano[10], Frits R. Rosendaal[2], Flora Peyvandi[1,3], LUMC-COVID-19 Research Group, COVID-19 Network working group

1. Università degli Studi di Milano, Dipartimento di Fisiopatologia medico-chirurgica e dei trapianti, Milan, Italy
2. Leiden University Medical Center, Department of Clinical Epidemiology, Leiden, the Netherlands
3. Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, U.O.C. Medicina Generale Emostasi e Trombosi, Milan, Italy
4. Leiden University Medical Center, Department of Rheumatology, Leiden, the Netherlands
5. Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Angelo Bianchi Bonomi Hemophilia and Thrombosis Centre, Milan, Italy
6. Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Infectious Disease Unit, Milan, Italy
7. Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Department of Health Policy, Milan, Italy
8. Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Respiratory Unit and Cystic Fibrosis Adult Center, Milan, Italy.
9. Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Department of Medicine, High Care Internal Medicine Unit, Milan, Italy
10. Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Medicina Generale Immunologia e Allergologia, Milan, Italy.

Correspondence address

Dr. F. Peyvandi

Università degli Studi di Milano

Department of Pathophysiology and Transplantation

Via Francesco Sforza 35, 20122, Milan, Italy

Tel: +39 0250320288

E-mail: flora.peyvandi@unimi.it

Manuscript information

Word count text:	3654
Figure/table count:	2 tables, 4 figures
Word count abstract:	249
Reference count:	36

Keywords: SARS-CoV-2, COVID-19, prediction, mortality

Abstract

Background: The coronavirus disease 2019 (COVID-19) presents an urgent threat to global health. Prediction models that accurately estimate mortality risk in hospitalized patients could assist medical staff in treatment and allocating limited resources.

Aims: To externally validate two promising previously published risk scores that predict in-hospital mortality among hospitalized COVID-19 patients.

Methods: Two cohorts were available; a cohort of 1028 patients admitted to one of nine hospitals in Lombardy, Italy (the Lombardy cohort) and a cohort of 432 patients admitted to a hospital in Leiden, the Netherlands (the Leiden cohort). The primary endpoint was in-hospital mortality. All patients were adult and tested COVID-19 PCR-positive. Model discrimination and calibration were assessed.

Results: The C-statistic of the 4C mortality score was good in the Lombardy cohort (0.85, 95CI: 0.82-0.89) and in the Leiden cohort (0.87, 95CI: 0.80-0.94). Model calibration was acceptable in the Lombardy cohort but poor in the Leiden cohort due to the model systematically overpredicting the mortality risk for all patients. The C-statistic of the CURB-65 score was good in the Lombardy cohort (0.80, 95CI: 0.75-0.85) and in the Leiden cohort (0.82, 95CI: 0.76-0.88). The mortality rate in the CURB-65 development cohort was much lower than the mortality rate in the Lombardy cohort. A similar but less pronounced trend was found for patients in the Leiden cohort.

Conclusion: Although performances did not differ greatly, the 4C mortality score showed the best performance. However, because of quickly changing circumstances, model recalibration may be necessary before using the 4C mortality score.

Introduction

The coronavirus disease 2019 (COVID-19) epidemic, which started in early December 2019, presents an important threat to global health. As of October 18th 2021, the number of patients confirmed to have the disease has exceeded 240 million and more than 4,900,537 people have died from COVID-19 infection. [1]

The outbreak overwhelmed the healthcare system in several countries, leading to shortages in hospital beds and medical equipment. [2–4] Prediction models that estimate the risk of hospitalized patients experiencing a poor outcome could assist medical staff in triaging patients when allocating limited healthcare resources.

A systematic review of prognostic prediction models for poor outcomes in hospitalized COVID-19 patients [5] reported that most models showed good predictive performance, but almost all models had a high risk of bias owing to a combination of poor reporting and poor methodological conduct for participant selection, predictor description, and/or statistical methods used. Of the 107 reviewed prognostic scores, only one, the 4C mortality score [6], was identified as being of good methodological quality. [5] The 4C mortality score development cohort consisted of 35,463 patients enrolled between February 6th 2020 and May 20th 2020. Several external validation studies reported C-statistics between 0.78 and 0.84. [7–11]

The systematic review by Wynants et al. [5] also identified another prediction model of interest [12]. In addition to reporting on the quality of newly developed COVID-19 specific prediction models, the review also reported on several previously published "general-purpose" mortality risk prediction models that were externally validated in the COVID-19 population. Among these models, the CURB-65 score [13], was also found to be a promising candidate model. The CURB-65 development cohort consisted of 718 patients with community-acquired pneumonia enrolled between October 1998 and December 2000. The CURB-65 score has been externally validated in COVID-19 patients in several studies that reported C-statistics ranging from 0.58-84. [12, 14–17] The CURB-65 score was also directly compared with the 4C mortality score. This was done in the external validation cohort (N = 15,560) used in the same publication that also reported on the development of the 4C mortality score. The C-statistic was 0.72 (0.71-0.73). [6]

Unfortunately, few studies have properly evaluated the 4C mortality score and the CURB-65. Many studies lacked the sample size to properly externally validate these scores. Furthermore, most studies only assessed model discrimination (by calculating a C-statistic) but not model calibration (which refers to the degree to which the mortality risk predicted by a given model is in agreement with the observed risk). In addition, none of these studies considered recalibrating the 4C mortality score to increase the performance of this score in the local population. Poorly calibrated models can lead to wrong clinical decisions, as the predicted mortality risk for a given patient may be widely different from the actual mortality risk. [18]

Therefore, the aim of the study was to externally validate two promising risk tools that predict mortality among hospitalized COVID-19 positive patients; one COVID-19-specific score (the 4C mortality score) and one general-purpose score (the CURB-65 score) in a cohort of 1028 COVID-19 positive patients admitted to one of nine hospitals in Lombardy, Italy (the Lombardy cohort) and a cohort of 432 COVID-19 positive patients admitted to a hospital in Leiden, the Netherlands (the Leiden cohort).

Methods

Study design

Two cohorts of adult patients diagnosed with COVID-19 were available for external validation. The first cohort consisted of adult patients hospitalized at one of nine hospitals in the province of Lombardy, Italy (the Lombardy cohort). The second cohort consisted of adult patients hospitalized at Leiden University Medical Center, Leiden, the Netherlands (the Leiden cohort). Patients were included at hospital admission. For both cohorts, patients that were transferred from other hospitals, and patients that were directly admitted to the ICU were excluded. This study was approved by the Medical Ethics Committee of the Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico and the Institutional Review Board of the LUMC for observational studies.

Data collection

Clinical data were collected using a case report form (which was based on the ISARIC-WHO case report form). Only data that were available during the first 24 hours of admission were used. If multiple values were present, the earliest recorded value was used.

Predictor variables for the 4C mortality score

The 4C mortality score was developed in a cohort of COVID-19 PCR-positive adult patients that were admitted to one of 260 hospitals in England, Scotland, or Wales between 6 February and 20 May 2020. The outcome was in-hospital mortality. The 4C mortality score includes the following eight predictors, collected on the day of admission: age in years (categorical variable: <50, 50-59, 60-69, 70-79, ≥80); sex at birth (dichotomous variable: male, female); respiratory rate in breaths/min (categorical variable: >20, 20-29, ≥30); oxygen saturation on room air (dichotomous variable: ≥92%, <92%); Glasgow coma scale (dichotomous variable: 15 points, <15 points); urea (categorical variable: <7 mmol/L, ≥7 to ≤14 mmol/L, >14 mmol/L); CRP (categorical variable: <50 mg/L, 50-99 mg/L, ≥100 mg/L) and the number of comorbidities. The list of comorbidities is based on the Charlson comorbidity index [19] with the addition of clinician-defined obesity.

Predictor variables for the CURB-65 score

The CURB-65 score was developed in a cohort of adult patients admitted as medical emergencies with community acquired pneumonia (CAP) to hospitals in the UK, New Zealand, and the Netherlands between October 1998 and December 2000. The outcome was 30-day mortality. The CURB-65 score consisted of the following 5 predictors that were to be measured at the emergency department: mental confusion, defined as a score of ≤ 8 on the Abbreviated Mental Test score; urea (categorical variable: ≤7 mmol/L, >7 mmol/L); respiratory rate in breaths/min (categorical variable: <30, ≥30); a systolic blood pressure of < 90 mmHg or a diastolic blood pressure of ≤ 60 mmHg (dichotomous variable; no, yes) and age (dichotomous variable; <65 years, ≥65 years). If information on mental confusion was missing, a Glasgow Coma Scale score of ≤ 14 was used as an indicator of mental confusion. (this method was also used previous studies [20])

Outcome

The outcome used in this study for all models was 30-day in-hospital mortality.

External validation of all risk scores

For the 4C mortality score, the predicted risk of 30-day in-hospital mortality was calculated for each individual in the validation cohort. The model formula and coefficients for the 4C mortality score were obtained from the supplement of the original paper. [6] In addition, the number of points that each individual scored on the 4C mortality score was also calculated. The original publication of the CURB-65 score did not provide the model formula underlying the score, only the score itself. [13] Therefore, it was not possible to calculate the individual predicted risk. Instead, only the individual CURB-65 score was calculated for each patient.

Model discrimination for all models was assessed using the C-statistic. The C-statistic reflects the degree to which a model can distinguish between patients with and patients without the outcome and ranges from 0.5 (no discrimination) to 1.0 (perfect discrimination). [21] For the 4C mortality score, the C-statistic was calculated in two ways; via the predicted risk of 30-day in-hospital mortality (which was calculated from the regression model underlying the risk tool), and via the simplified points score (which was constructed to facilitate usage of the risk tool by clinicians). For the CURB-65, only the points score was available and we therefore only used the points score to calculate model discrimination.

Model calibration of the 4C mortality score was assessed visually by plotting calibration curves. [21] To construct the calibration curve, the predicted outcome probabilities were plotted against observed outcome frequencies, for each quintile of predicted risk. To examine calibration across the whole range, a LOESS (Locally Estimated Scatterplot Smoothing) line was estimated. As it was not possible to obtain the

individual predicted mortality risk using the CURB-65 score, plotting a calibration curve was not possible. Instead, an alternative method, also implemented in other studies [22, 23], was used. Patients were divided into groups, based on the number of points on the CURB-65 score. Next, a bar chart was constructed where the proportion of deaths of each group in the external validation cohort was compared against the proportion of deaths in the same group in the CURB-65 development cohort.

The 4C mortality score was also recalibrated to better reflect the mortality incidence in the validation cohorts. Recalibration methods range from conservative to very extensive. Two recalibration methods were used. The first recalibration method was the most conservative and consisted of fitting a logistic regression model to the validation cohort dataset, with the intercept as a free parameter and the linear predictor of the 4C mortality score as an offset variable. [24] This method corrects for systematic over- or underprediction by the model. The second recalibration method was more extensive and consisted of fitting the same logistic regression model as described above, but leaving both the intercept and the coefficient for the linear predictor to be freely estimated. [24] A more detailed explanation of this methodology, as well as the details of the various recalibrated models presented in the results section, are given in the Appendix. As the predicted mortality risk for each patient was not available for the CURB-65 score, model recalibration was not done for this score.

Handling missing values

Missing values in the Lombardy cohort and Leiden cohort datasets were imputed using multivariate imputation by chained equations. [25] The C-statistic was pooled using Rubin's rules. [26] All imputed datasets were combined to create one dataset, which was then used to create the calibration plots.

Statistical packages

All analyses were performed using R version 3.6.2.

Data sharing

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

Results

General characteristics

The Lombardy cohort consisted of 1028 patients enrolled between February 25th 2020 and August 1st 2020. The Leiden cohort consisted of 432 patients enrolled between March 7th 2020 and March 5th 2021.

The mortality rate was 21% in the Lombardy cohort, 10% in the Leiden cohort, 32% in the 4C mortality score development cohort and 10% in the CURB-65 score development cohort. (Table 1) The median age for all patients was 66 years in the Lombardy cohort, 65 years in the Leiden cohort, 73 years in the 4C mortality score development cohort. (Table 1) The portion of patients with ≥ 2 comorbidities was 24% in the Lombardy cohort, 35% in the Leiden cohort and 48% in the 4C mortality score development cohort. (Table 1)

External validation of 4C mortality score

The C-statistic of the 4C mortality score in the Lombardy cohort, which was calculated based on the predicted risk of 30-day in-hospital mortality obtained from the underlying regression model, was 0.85 (95CI: 0.82-0.89). (Table 2, Figure 1) The C-statistic of the 4C mortality score in the Lombardy cohort, calculated based on the simplified points score, was exactly the same (C-statistic: 0.85, 95CI: 0.82-0.89). Calibration was acceptable. However, the 4C mortality score did overpredict the risk of mortality in the 0-50% risk range. (Figure 2A) After applying the first recalibration method, the 4C mortality score showed better calibration in the lower risk range but underpredicted the risk in the >50% risk range. (Figure 2B) Applying the second recalibration method resulted in almost perfect calibration. (Figure 2C)

The C-statistic of the 4C mortality score in the Leiden cohort, which was calculated based on the predicted risk of 30-day in-hospital mortality obtained from the underlying regression model, was 0.87 (95CI: 0.80-0.94). (Table 2, Figure 1) The C-statistic of the 4C mortality score in the Leiden cohort, calculated based on the simplified points score, was exactly the same (C-statistic: 0.87, 95CI: 0.80-0.94). Calibration was poor as the 4C mortality score overpredicted the mortality rate across the entire risk range. (Figure 3A) After applying the first recalibration method, the 4C mortality score showed slightly improved calibration but now underpredicted the risk across most of the risk range. (Figure 3B) Applying the second recalibration method resulted in a model with very good calibration, which was almost perfect for the 0-50% predicted risk range. (Figure 3C) Calibration in the higher risk ranges was less accurate, but few patients had a predicted risk higher than 50%.

External validation of CURB-65 score

The C-statistic of the CURB-65 score in the Lombardy cohort was 0.80 (95CI: 0.75-0.85). (Table 2, Figure 1) The observed mortality risk for patients in the CURB-65 development cohort was much lower than the observed mortality risk for patients in the Leiden cohort, for every group of patients with a specific CURB-65 score. (Figure 4) The C-statistic of the CURB-65 score in the Leiden cohort was 0.82 (95CI: 0.76-0.88). (Table 2, Figure 1) The observed mortality risk for patients in the CURB-65 development cohort was slightly lower than the observed mortality risk for patients in the Leiden cohort, especially for patients with ≥ 2

points on the CURB-65 score. (Figure 4)

Discussion

Main findings

Two promising prognostic scores that are used to predict in-hospital mortality in patients hospitalized with COVID-19 were selected for external validation based on the results of a previous systematic review. [5]

Although both models showed good discrimination, the 4C mortality score performed the best, both in the Lombardy cohort (C-statistic: 0.85) as well as the Leiden cohort (C-statistic: 0.87). The 4C model performed less well in terms of model calibration, as the standard model overpredicted the risk in most patients, in both the Lombardy cohort and the Leiden cohort. In the Lombardy cohort, the degree of overprediction was still acceptable but it was unacceptably high in the Leiden cohort. Updating the 4C mortality score to the local setting by applying a conservative recalibration method strongly improved model calibration in the Leiden cohort in the 0-30% risk range (which contained most of the patients in the cohort). Excellent calibration in both cohorts was obtained by applying a second, slightly more extensive, recalibration method.

The CURB-65 score also showed good discrimination. However, direct assessment of model calibration was not possible. We indirectly assessed model calibration by comparing the mortality rate in the development cohort with the mortality rate in the external validation cohort.

The differences in model performance between cohorts can be explained by two factors. [27] Firstly, part of the reduction in model performance is due to overfitting. Secondly, differences in the type of patients admitted to the hospital as well as differences in treatment protocol between the development cohorts and the external validation cohorts could also have influenced model performance.

For example, almost all patients (91%) in the Lombardy cohort were enrolled during the first two months of the first COVID-19 wave in Italy (March and April 2020) while patients in the Leiden cohort were uniformly enrolled from March 7th 2020 and March 5th 2021. This means that most patients in the Leiden cohort would have had access to better treatment, as much more was known about the effectiveness of different COVID-19 treatment options. This will have impacted the performance of the model in the external validation cohort.

Another issue is that the 4C mortality score development cohort was older (73 years) than the Leiden cohort (65 years) and the Lombardy cohort (66 years). This also explains the difference in mortality rate, which was considerably higher in the 4C mortality score development cohort (32.2%) than in the Leiden cohort (8.3%) or the Lombardy cohort (21.7%). Due to this, it was to be expected that the 4C mortality score would systematically overpredict the mortality risk in the external validation cohorts, given that it

was developed in a cohort of patients with a much higher average mortality rate. Recalibrating was an adequate way to solve this issue.

Comparison with other studies

The development of the 4C mortality score was in line with the most recent guidelines for prediction model development. [28] Furthermore, the patient cohorts used to develop and validate the 4C mortality score were extremely large, minimizing the chance of overfitting. The model has since been validated in a number of cohorts from other populations. In a large cohort of 14,343 patients from hospitals in the greater Paris area, the 4C mortality score had a C-statistic of 0.79 (95% CI: 0.78–0.80). Calibration was acceptable, although the model somewhat overpredicted the risk in most patients. [29] in a cohort of 925 Brazilian and 438 Spanish patients, a C-statistic of 0.78 (95%CI: 0.75-0.81) was found. Overall calibration was good in this cohort. [30] Furthermore, the 4C mortality score had a C-statistic of 0.78 (0.70–0.85) in a cohort of 1027 Canadian patients from Toronto. Calibration was not assessed formally but by plotting model scores against observed probabilities, which makes calibration difficult to interpret. [11] Lastly, a Japanese study of 693 patients reported good discrimination (0.84, 95%CI: 0.80-0.88) and calibration. [31] None of the aforementioned studies recalibrated their models to better fit the local population.

The CURB-65 score has also been previously validated in different populations of COVID-19 positive patients. The CURB-65 showed a C-statistic of 0.83 (0.82–0.84) in a very large cohort of 10,328 patients from Spain. [32] Furthermore, a preprint manuscript reported that the CURB-65 score was tested in patients hospitalized in one of thirteen acute care hospitals in the New York City area. The score had a C-statistic of 0.80 in cohort of 2229 patients and 0.72 in another cohort of 3328 patients. [17] In a cohort of 1717 COVID-19 positive patients admitted to a hospital in Shanghai, China, the CURB-65 score had a fairly low C-statistic of 0.70 (95CI: 0.66-0.73). [33] Lastly, in a cohort of 1181 patients from Qatar, a C-statistic of 0.78 (95% CI 0.70-0.86) was reported. [34] Overall, the discriminative performance of the CURB-65 reported by these studies varied from moderate to good. Similar to our study, none of the aforementioned studies assessed model calibration of the CURB-65 score. (as this was not possible)

Limitations

It has been suggested that the minimum sample size for external validation should be at least 100 events and 100 non-events. [35] The Leiden cohort had 41 deaths, falling below the number suggested by this rule of thumb.

The sample size of the Lombardy cohort was acceptable for external validation. However, the Lombardy cohort consisted of patients that were enrolled in the first months of 2020. After this period, the incidence COVID-19 related mortality has changed a lot due to many treatment changes. This limits the applicability of the recalibrated 4C risk score as the population in which the scores were recalibrated may not be

representative of the current patient population in 2022.

Furthermore, some patients who were already very ill before being hospitalized for COVID-19 would have chosen to receive end-of-life care at home. Despite not dying in the hospital (in-hospital mortality was the study outcome), the 4C mortality score would have assigned these patients a very high predicted in-hospital mortality risk. This will have reduced the model performance (especially model discrimination), depending on the proportion of patients that received end-of-life care at home.

Lastly, missing data may have influenced model performance. For example, the oxygen saturation on room air (a predictor in the 4C mortality score) was missing in more than half of all patients in both the Lombardy cohort and the Leiden cohort. This is most likely because these patients were already receiving oxygen therapy at admission.

Clinical use

Given the gradual uptake of the COVID-19 vaccine, the chances of the COVID-19 virus overburdening hospital resources at the national level in developed countries is becoming smaller, although localized outbreaks might still occur, especially in places with high rates of vaccine hesitancy. On the other hand, vaccine uptake in developing countries is still extremely low [1] and a viral outbreak could severely strain local resources. Furthermore, viral evolution could lead to a novel variant that current vaccines are not or only partially effective against. [36]

Risk scores could be used to identify patients with a high mortality risk. These patients could be candidates for early escalation to critical care while low-risk patients could be safely managed outside the hospital. For this purpose, an accurate estimate of the absolute risk of mortality for a given patient is essential. A risk score is sometimes used in a patient population that is very different (in terms of patient- and treatment characteristics) than the patient population in which the risk score was originally developed. In this situation, it is highly likely that the score will be miscalibrated and model recalibration might be necessary before this score can be used to obtain an accurate estimate of the absolute risk of mortality.

If recalibration of the risk tool is necessary, we would suggest using a more conservative recalibration method that changes the original model as little as possible to minimize overfitting the model to the new setting. Depending on the sample size, more extensive recalibration models can be considered. Recalibrating a model before use might be too complicated for end-users (i.e. clinicians). An app or web tool that automatically produces an adjusted risk score based on a user-specified mean mortality rate would simplify the process significantly and increase usage of these scores among clinicians. Lastly, as vaccination rates increase, an updated model that also includes information on vaccination status might yield better predictions.

In situations where hospital ICUs are overburdened, a risk score could be used to only admit patients with a low predicted mortality risk and transfer high-risk patients to other centers with more ICU capacity. For this purpose, the clinician only needs to know if a given patient has a lower or higher risk of mortality relative to other patients. In this situation, a risk score only needs to have good model discrimination. The 4C mortality score showed good model discrimination (as measured with the C-statistic) across different populations, and could therefore be used in clinical practice for this purpose, without recalibrating the model.

Conclusion

Two previously published risk scores were externally validated in two different settings. Although performances did not differ greatly, the 4C mortality score showed the best model performance. However, if the reason for using this risk tool is to obtain accurate absolute mortality risks, recalibration of the model to the local patient population might be necessary before use.

Funding

The authors received no financial support for this study.

Authorship

Author contributions

S. Hassan analyzed the data and wrote the manuscript. C.L. Ramspek, M. van Diepen, F. Peyvandi and F.R. Rosendaal interpreted the data and reviewed the manuscript. All other authors were involved in either coordination, patient enrollment, data collection and/or reviewing the manuscript.

Conflict of interest disclosure

Barbara Ferrari has received consulting fees and travel support from Sanofi Genzyme. Vincenzo la Mura has received consulting fees and travel support from Gore and Takeda. Ida Martinelli reports personal and non-financial support from Bayer, Roche, Rovi and Novo Nordisk outside of the submitted work. Alessandra Bandera has received speaker's honoraria and fees for attending advisory boards from Janssen-Cilag, Pfizer, Nordic Pharma, Qiagen and received research grants from Gilead Sciences. Andrea Gori has received speaker's honoraria and fees for attending advisory boards from ViiV Healthcare, Gilead, Janssen-Cilag, Merck Sharp & Dohme, Bristol-Myers Squibb, Pfizer and Novartis and has received research grants from ViiV, Bristol-Myers Squibb, and Gilead. Francesco Blasi reports grants and personal fees from Astrazeneca, Chiesi, GlaxoSmithKline, Sanofi Genzyme, Insmed and Menarini and personal fees from Grifols, Guidotti, Novartis, Zambon, Vertex and Viatris, outside the submitted work. Ciro Canetta has received honoraria for participating as a speaker at educational meetings from Boehringer Ingelheim and Novartis. Nicola Montano has received honoraria as speaker for educational meetings and advisory boards by Novartis,

Novo Nordisk, Gilead and Philips. Flora Peyvandi has participated in advisory boards of Sanofi, Sobi, Takeda, Roche and Biomarin and educational meetings of Grifols and Roche. All other authors have no relevant financial or non-financial interests to disclose.

Acknowledgments

We are indebted to all the patients with COVID-19 who participated in this research. We would also like to thank the COVID-19 Network working group and the LUMC-COVID-19 Research Group members. Their names are listed below:

Silvano Bosari, Luigia Scudeller, Giuliana Fusetti, Laura Rusconi, Silvia Dell'Orto, Daniele Prati, Luca Valenti, Silvia Giovannelli, Maria Manunta, Giuseppe Lamorte, Francesca Ferarri, Andrea Gori, Alessandra Bandera, Antonio Muscatello, Davide Mangioni, Laura Alagna, Giorgio Bozzi, Andrea Lombardi, Riccardo Ungaro, Giuseppe Ancona, Gianluca Zuglian, Matteo Bolis, Nathalie Iannotti, Serena Ludovisi, Agnese Comelli, Giulia Renisi, Simona Biscarini, Valeria Castelli, Emanuele Palomba, Marco Fava, Valeria Fortina, Carlo Alberto Peri, Paola Saltini, Giulia Viero, Teresa Itri, Valentina Ferroni, Valeria Pastore, Roberta Massafra, Arianna Liparoti, Toussaint Muheberimana, Alessandro Giommi, Rosaria Bianco, Rafaela Montalvao De Azevedo, Grazia Eliana Chitani, Flora Peyvandi, Roberta Gualtierotti, Barbara Ferrari, Raffaella Rossio, Nadia Boasi, Erica Pagliaro, Costanza Massimo, Michele De Caro, Andrea Giachi, Nicola Montano, Barbara Vigone, Chiara Bellocchi, Angelica Carandina, Elisa Fiorelli, Valerie Melli, Eleonora Tobaldini, Francesco Blasi, Stefano Aliberti, Maura Spotti, Leonardo Terranova, Sofia Misuraca, Alice D'Adda, Silvia Della Fiore, Marta Di Pasquale, Marco Mantero, Martina Contarini, Margherita Ori, Letizia Morlacchi, Valeria Rossetti, Andrea Gramegna, Maria Pappalettera, Mirta Cavallini, Agata Buscemi, Marco Vicenzi, Irena Rota, Giorgio Costantino, Monica Solbiati, Ludovico Furlan, Marta Mancarella, Giulia Colombo, Giorgio Colombo, Alice Fanin, Mariele Passarella, Valter Monzani, Ciro Canetta, Angelo Rovellini, Laura Barbetta, Filippo Billi, Christian Folli, Silvia Accordino, Diletta Maira, Cinzia Maria Hu, Irene Motta, Natalia Scaramellini, Anna Ludovica Fracanzani, Rosa Lombardi, Annalisa Cespiati, Matteo Cesari, Tiziano Lucchi, Marco Proietti, Laura Calcaterra, Clara Mandelli, Carlotta Coppola, Arturo Cerizza, Antonio Maria Pesenti, Giacomo Grasselli, Alessandro Galazzi, Alessandro Nobili, Mauro Tettamanti, Igor Monti, Alessia Antonella Galbussera, Ernesto Crisafulli, Domenico Girelli, Alessio Maroccia, Daniele Gabbiani, Fabiana Busti, Alice Vianello, Marta Biondan, Filippo Sartori, Paola Faverio, Alberto Pesci, Stefano Zucchetti, Paolo Bonfanti, Marianna Rossi, Ilaria Beretta, Anna Spolti, Sergio Harari, Davide Elia, Roberto Cassandro, Antonella Caminati, Francesco Cipollone, Maria Teresa Guagnano, Damiano D'Ardes, Ilaria Rossi, Francesca Vezzani, Antonio Spanevello, Francesca Cherubino, Dina Visca, Marco Contoli, Alberto Papi, Luca Morandi, Nicholas Battistini, Guido Luigi Moreo, Pasqualina Iannuzzi, Daniele Fumagalli, Sara Leone, Josine A. Oud, Meryem Baysan, Jeanette Wigbers, Lieke J. van Heurn, Susan B. ter Haar, Alexandra G.L. Toppenberg, Laura Heerdink, Annekee A. van IJlzinga Veenstra, Anna M. Eikenboom, Julia Wubbolts, Jonathan Uzorka, Willem Lijferink, Romy Meier,

Ingeborg de Jonge, Sesmu M. Arbou, Mark G.J. de Boer, Anske G. van der Bom, Olaf M. Dekkers and Frits Rosendaal.

References

1. Dong E, Du H, Gardner L (2020) An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis 20:533–534. https://doi.org/10.1016/s1473-3099(20)30120-1

2. Arabi YM, Murthy S, Webb S (2020) COVID-19: a novel coronavirus and a novel challenge for critical care. Intensive Care Med 46:833–836. https://doi.org/10.1007/s00134-020-05955-1

3. Grasselli G, Pesenti A, Cecconi M (2020) Critical Care Utilization for the COVID-19 Outbreak in Lombardy, Italy: Early Experience and Forecast During an Emergency Response. Jama. https://doi.org/10.1001/jama.2020.4031

4. Xie J, Tong Z, Guan X, et al (2020) Critical care crisis and some recommendations during the COVID-19 epidemic in China. Intensive Care Med 46:837–840. https://doi.org/10.1007/s00134-020-05979-7

5. Wynants L, Van Calster B, Collins GS, et al (2020) Prediction models for diagnosis and prognosis of covid-19: Systematic review and critical appraisal. BMJ 369:m1328. https://doi.org/10.1136/bmj.m1328

6. Knight SR, Ho A, Pius R, et al (2020) Risk stratification of patients admitted to hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: Development and validation of the 4C Mortality Score. BMJ 370:. https://doi.org/10.1136/bmj.m3339

7. Lassau N, Ammari S, Chouzenoux E, et al (2021) Integrating deep learning CT-scan model, biological and clinical variables to predict severity of COVID-19 patients. Nat Commun 12:. https://doi.org/10.1038/s41467-020-20657-4

8. van Dam PMEL, Zelis N, van Kuijk SMJ, et al (2021) Performance of prediction models for short-term outcome in COVID-19 patients in the emergency department: a retrospective study. Ann Med 53:402–409. https://doi.org/10.1080/07853890.2021.1891453

9. Çınar T, Hayıroğlu Mİ, Çiçek V, et al (2021) Is prognostic nutritional index a predictive marker for estimating all-cause in-hospital mortality in COVID-19 patients with cardiovascular risk factors? Hear Lung 50:307–312. https://doi.org/10.1016/j.hrtlng.2021.01.006

10. Covino M, De Matteis G, Burzo ML, et al (2021) Predicting In-Hospital Mortality in COVID-19 Older Patients with Specifically Developed Scores. J Am Geriatr Soc 69:37–43. https://doi.org/10.1111/jgs.16956

11. Verma AA, Hora T, Jung HY, et al (2021) Characteristics and outcomes of hospital admissions for COVID-19 and influenza in the Toronto area. Cmaj 193:E410–E418.

https://doi.org/10.1503/cmaj.202795

12. Luo M, Liu J, Jiang W, et al (2020) IL-6 and CD8+ T cell counts combined are an early predictor of in-hospital mortality of patients with COVID-19. JCI Insight 5:. https://doi.org/10.1172/jci.insight.139024

13. Lim WS, Van Der Eerden MM, Laing R, et al (2003) Defining community acquired pneumonia severity on presentation to hospital: An international derivation and validation study. Thorax 58:377–382. https://doi.org/10.1136/thorax.58.5.377

14. Liang W, Liang H, Ou L, et al (2020) Development and validation of a clinical risk score to predict the occurrence of critical illness in hospitalized patients with COVID-19. JAMA Intern Med 180:1081–1089. https://doi.org/10.1001/jamainternmed.2020.2033

15. Choi MH, Ahn H, Ryu HS, et al (2020) Clinical Characteristics and Disease Progression in Early-Stage COVID-19 Patients in South Korea. J Clin Med 9:1959. https://doi.org/10.3390/jcm9061959

16. Satici C, Demirkol MA, Sargin Altunok E, et al (2020) Performance of pneumonia severity index and CURB-65 in predicting 30-day mortality in patients with COVID-19. Int J Infect Dis 98:84–89. https://doi.org/10.1016/j.ijid.2020.06.038

17. Levy TJ, Richardson S, Coppa K, et al (2020) Development and validation of a survival calculator for hospitalized patients with COVID-19. medRxiv Prepr Serv Heal Sci 2020.04.22.20075416. https://doi.org/10.1101/2020.04.22.20075416

18. Van Calster B, McLernon DJ, Van Smeden M, et al (2019) Calibration: The Achilles heel of predictive analytics. BMC Med 17:. https://doi.org/10.1186/s12916-019-1466-7

19. Charlson ME, Pompei P, Ales KL, MacKenzie CR (1987) A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation. J Chronic Dis 40:373–383. https://doi.org/10.1016/0021-9681(87)90171-8

20. Shin YM, Lee N, Ip M, et al (2007) Prospective comparison of three predictive rules for assessing severity of community-acquired pneumonia in Hong Kong. Thorax 62:348–353. https://doi.org/10.1136/thx.2006.069740

21. Steyerberg EW, Vickers AJ, Cook NR, et al (2010) Assessing the performance of prediction models: A framework for traditional and novel measures. Epidemiology 21:128–138

22. Ocak G, Ramspek C, Rookmaaker MB, et al (2019) Performance of bleeding risk scores in dialysis patients. Nephrol Dial Transplant 34:1223–1231. https://doi.org/10.1093/ndt/gfy387

23. O'Brien EC, Simon DN, Thomas LE, et al (2015) The ORBIT bleeding score: a simple bedside score to assess bleeding risk in atrial fibrillation. Eur Heart J 36:ehv476. https://doi.org/10.1093/eurheartj/ehv476

24. Steyerberg EW (2009) Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating

25. van Buuren S, Groothuis-Oudshoorn K (2011) mice: Multivariate imputation by chained equations in R. J Stat Softw 45:1–67. https://doi.org/10.18637/jss.v045.i03

26. Marshall A, Altman DG, Holder RL, Royston P (2009) Combining estimates of interest in prognostic modelling studies after multiple imputation: Current practice and guidelines. BMC Med Res Methodol 9:. https://doi.org/10.1186/1471-2288-9-57

27. Steyerberg EW, Vergouwe Y (2014) Towards better clinical prediction models: seven steps for development and an ABCD for validation. Eur Heart J 35:1925–1931. https://doi.org/10.1093/eurheartj/ehu207

28. Collins GS, Reitsma JB, Altman DG, Moons KGM (2015) Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. BMJ 350:. https://doi.org/10.1136/bmj.g7594

29. Lombardi Y, Azoyan L, Szychowiak P, et al (2021) External validation of prognostic scores for COVID-19: a multicenter cohort study of patients hospitalized in Greater Paris University Hospitals. Intensive Care Med. https://doi.org/10.1007/s00134-021-06524-w

30. Lazar Neto F, Marino LO, Torres A, et al (2021) Community-acquired pneumonia severity assessment tools in patients hospitalized with COVID-19: a validation and clinical applicability study. Clin Microbiol Infect 27:1037.e1-1037.e8. https://doi.org/10.1016/j.cmi.2021.03.002

31. Kuroda S, Matsumoto S, Sano T, et al (2021) External validation of the 4C Mortality Score for patients with COVID-19 and pre-existing cardiovascular diseases/risk factors. BMJ Open 11:. https://doi.org/10.1136/bmjopen-2021-052708

32. Artero A, Madrazo M, Fernández-Garcés M, et al (2021) Severity Scores in COVID-19 Pneumonia: a Multicenter, Retrospective, Cohort Study. J Gen Intern Med 36:1338–1345. https://doi.org/10.1007/s11606-021-06626-7

33. Jiang M, Li C, Zheng L, et al (2021) A biomarker-based age, biomarkers, clinical history, sex (ABCS)-mortality risk score for patients with coronavirus disease 2019. Ann Transl Med 9:230–230. https://doi.org/10.21037/atm-20-6205

34. Elmoheen A, Abdelhafez I, Salem W, et al (2021) External Validation and Recalibration of the CURB-65 and PSI for Predicting 30-Day Mortality and Critical Care Intervention in Multiethnic Patients with COVID-19. Int J Infect Dis 111:108–116. https://doi.org/10.1016/j.ijid.2021.08.027

35. Riley RD, Debray TPA, Collins GS, et al (2021) Minimum sample size for external validation of a clinical prediction model with a binary outcome. Stat Med 40:4230–4251. https://doi.org/10.1002/sim.9025

36. Telenti A, Arvin A, Corey L, et al (2021) After the pandemic: perspectives on the future trajectory of COVID-19. Nature

Tables & Figures

Table 1: General characteristics

Characteristics	Lombardy cohort (N = 1028)		Leiden cohort (N = 432)		4C mortality score *development cohort (N = 35463)		CURB-65 ‡development cohort §(N = 718)
	No of patients (%) or median (IQR)	% of missing values	No of patients (%) or median (IQR)	% of missing values	No of patients (%) or median (IQR)	% of missing values	No of patients (%) or median (IQR)
In-hospital mortality	216 (21%)	0%	41 (10%)	0%	11426 (32%)	0%	69 (9.6%)
Age (years)	66 (26)	0.1%	65 (21)	0%	73 (24)	0.5%	NR
Sex at birth							
Female	383 (37%)	0.2%	168 (39%)	0%	14741 (42%)	0.3%	NR
Male	643 (63%)		264 (61%)		20615 (58%)		NR
Clinician-defined obesity (%)							
No	885 (89%)	3.2%	265 (68%)	10%	26415 (89%)	15.9%	NR
Yes	110 (11%)		124 (32%)		3414 (11%)		NR
No of comorbidities							
0	496 (51%)	5%	102 (26%)	10%	8497 (24%)	0%	NR
1	245 (25%)		148 (38%)		†9941 (28%)		NR
≥2	234 (24%)		137 (35%)		†17025 (48%)		NR
Respiratory rate (breaths/min)	22 (8)	25%	20 (7)	2%	22 (9)	6.0%	NR
Oxygen saturation, room air (%)	96 (4)	61.5%	94 (7)	53.7%	94 (6)	5.0%	NR
Oxygen therapy at admission							
No	394 (39%)	1.3%	200 (46%)	0%	NR	NR	NR

Yes	621 (61%)		232 (54%)		NR		NR
Urea (mmol/L)	12 (11)	16.2%	6 (4)	4.2%	7 (6)	26.3%	NR
C-reactive protein (mg/L)	71 (127)	5.9%	68 (75)	6.2%	85 (122)	21.5%	NR
Temperature (°C)	36.9 (1.3)	6.5%	37.6 (1.6)	0%	37.3 (1.5)	5.6%	NR
Lymphocytes (10^9/L)	1.0 (0.7)	5.5%	0.8 (0.6)	14%	0.9 (0.7)	16.7%	NR
Lactate dehydrogenase (U/L)	285 (165)	29.7%	311 (169)	14.1%	NR	NR	NR
Mental confusion[¶]							
No	464 (88%)	48.8%	316 (86%)	14.8%	NR	NR	NR
Yes	62 (12%)		52 (14%)		NR		NR
Systolic blood pressure (mmHg)	130 (22)	3.3%	134 (27)	0%	124 (33)	5.1%	NR
Diastolic blood pressure (mmHg)	75 (14)	5.4%	76 (18)	0%	70 (19)	5.3%	NR

*: The cohort that was used to develop the 4C mortality score, as reported by S.R. Knight et al. in the original publication.[6] †Obesity was also counted as a comorbidity. ‡The cohort that was used to develop the CURB-65 score, as reported in the original publication.[13] §No missing data. ¶: defined as a score of ≤ 8 on the Abbreviated Mental Test score or a score of ≤ 14 on the Glasgow Coma Scale. NR: not reported.

Table 2: C-statistic of the 4C mortality score and the CURB-65 score during external validation

Models	Lombardy cohort	Leiden cohort
4C mortality score	0.85 (95CI: 0.82-0.89)	0.87 (95CI: 0.80-0.94)
CURB-65 score	0.80 (95CI: 0.75-0.85)	0.82 (95CI: 0.76-0.88)

*: Not applicable, model was developed in this cohort.

Figure 1: Receiver operating characteristic (ROC) curve

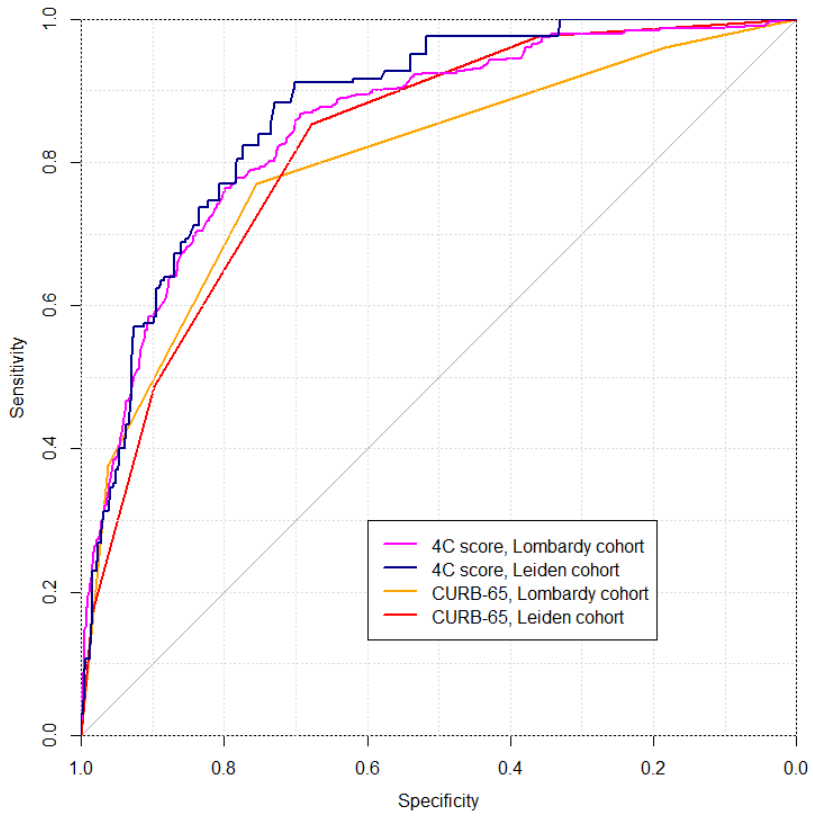

Figure legend: The figure shows the ROC curve for the 4C mortality score and the CURB-65 score for both the Lombardy cohort and the Leiden cohort. The area under the curve is equal to the C-statistic, and varies from 0.5 (poor discrimination, basically the same as flipping a coin to determine the outcome) to 1 (the model can discriminate perfectly between patients that died and patients that survived). The C-statistics and corresponding confidence intervals are reported in Table 2.

Figure 2: Calibration of the standard 4C mortality score and the recalibrated models in the Lombardy cohort

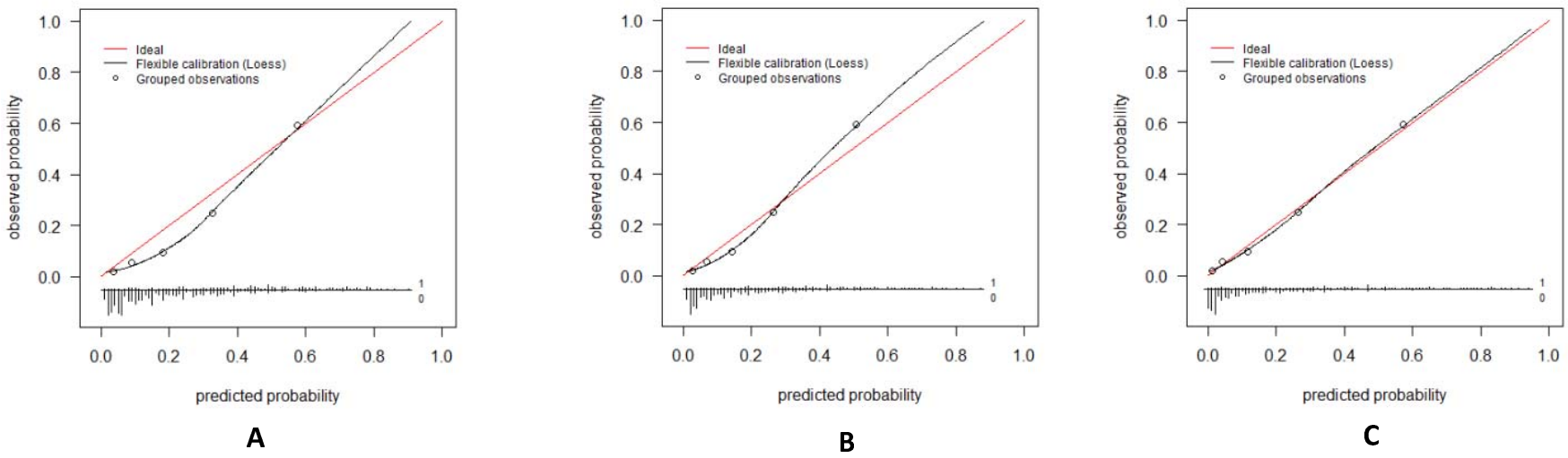

Figure legend: Model calibration was assessed visually by plotting calibration curves. Figure **2A** represents the calibration plot for the standard model in the Lombardy cohort. Figure **2B** represents the calibration plot for the first recalibrated model in the Lombardy cohort. (only intercept was re-estimated) Figure **2C** represents the calibration plot for the second recalibrated model in the Lombardy cohort. (intercept and slope were re-estimated) To construct the calibration curve, the cohort was divided into quintiles based on the predicted mortality risk. Next, the predicted mortality risk of each group was plotted against the observed mortality rate for that group. (these points are represented as circles) Ideally, the predicted and observed mortality risk should be equal, for each group. With ideal model calibration, all points should fall on the diagonal red line. To examine model calibration across the entire risk range, a LOESS (Locally Estimated Scatterplot Smoothing) line was estimated. Below the calibration curves are histograms showing the distribution of predicted probabilities for patients that died during follow-up (all bars above the horizontal line, labeled '1') and for patients that survived (all bars below the horizontal line, labeled '0'). The length of a bar corresponds to the number of patients with that predicted probability.

Figure 3: Calibration of the standard 4C mortality score and the recalibrated models in the Leiden cohort

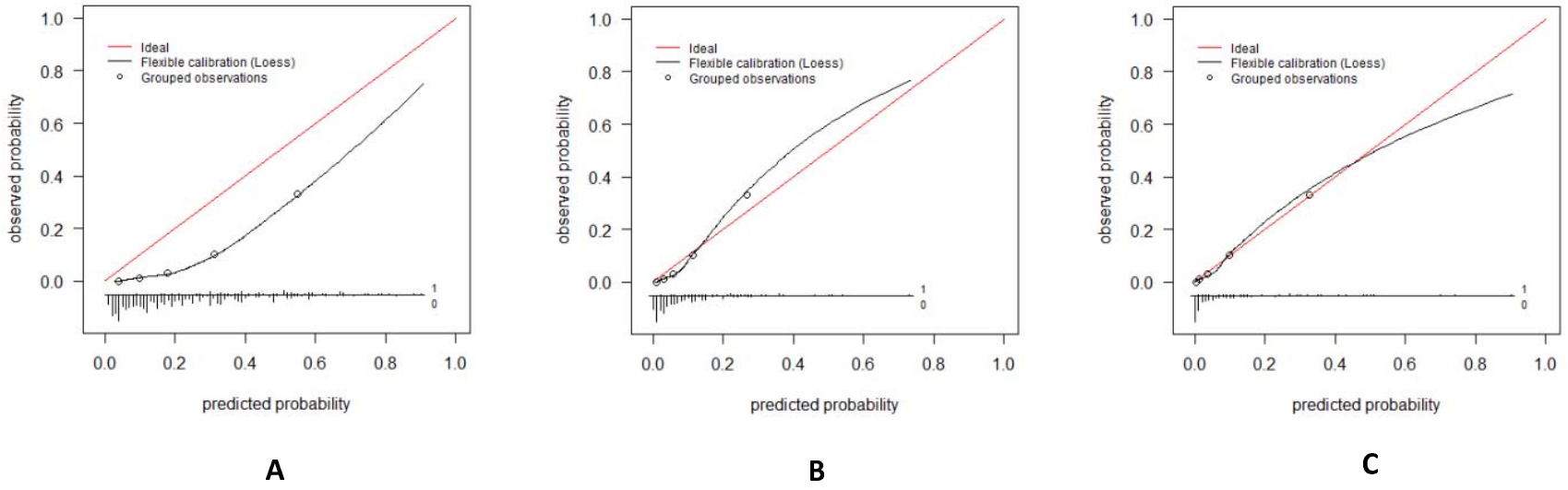

Figure legend: Model calibration was assessed visually by plotting calibration curves. Figure **3A** represents the calibration plot for the standard model in the Leiden cohort. Figure **3B** represents the calibration plot for the first recalibrated model in the Leiden cohort. (only intercept was re-estimated) Figure **3C** represents the calibration plot for the second recalibrated model in the Leiden cohort. (intercept and slope were re-estimated) To construct the calibration curve, the cohort was divided into quintiles based on the predicted mortality risk. Next, the predicted mortality risk of each group was plotted against the observed mortality rate for that group. (these points are represented as circles) Ideally, the predicted and observed mortality risk should be equal, for each group. With ideal model calibration, all points should fall on the diagonal red line. To examine model calibration across the entire risk range, a LOESS (Locally Estimated Scatterplot Smoothing) line was estimated. Below the calibration curves are histograms showing the distribution of predicted probabilities for patients that died during follow-up (all bars above the horizontal line, labeled '1') and for patients that survived (all bars below the horizontal line, labeled '0'). The length of a bar corresponds to the number of patients with that predicted probability.

Figure 4: Calibration of the CURB-65 score in the Lombardy cohort and Leiden cohort

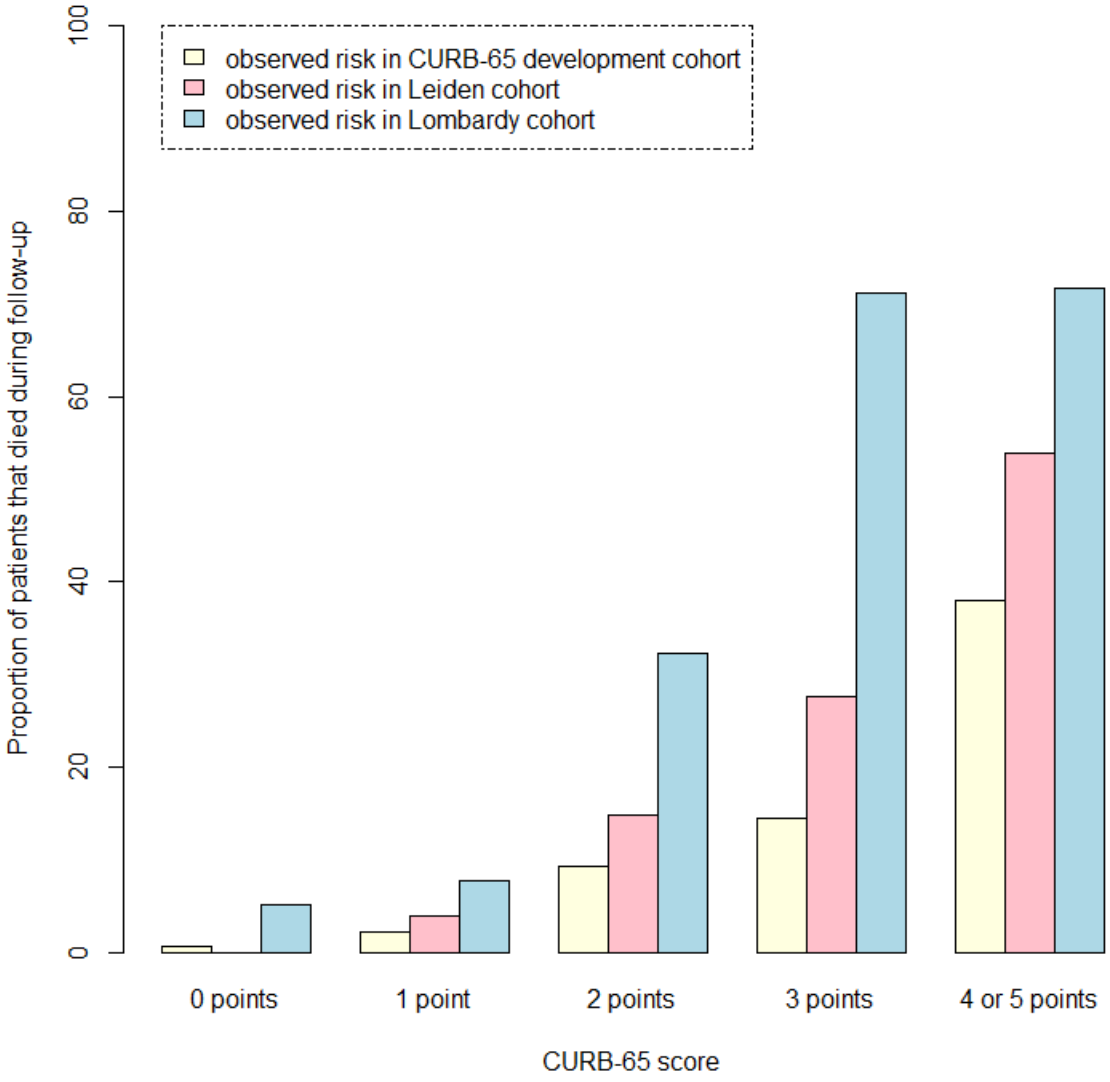

Figure legend: Model calibration was assessed visually in bar chart. Each cohort was divided into groups, based on the number of points on the CURB-65 score. For each group, the observed proportion of deaths in the CURB-65 development cohort (represented by the light-yellow bars) was compared against the observed proportion of deaths in the Leiden cohort (represented by the pink bars) and the Lombardy cohort (represented by the light-blue bars). Ideally, all bars for a given group should have the same height.