Machine-Learning Model for Mortality Prediction in Patients with Community-acquired Pneumonia: Development and validation study

Journal:	*CHEST*
Manuscript ID	CHEST-22-0131.R3
Article Type:	Original Research
Date Submitted by the Author:	n/a
Complete List of Authors:	cilloniz, catia; Biomedical Research Networking Centres in Respiratory Diseases (Ciberes) Barcelona, Department of Pneumology Mogensen, Mads Lause; Treat Systems ApS Ward, Logan; Treat Systems Pericas, Juan Manuel; Hospital Clínic de Barcelona, Infectious Diseases; Vall d'Hebron Institut de Recerca, Méndez, Raúl; Hospital Universitari i Politecnic La Fe, Pneumology Gabarrús , Albert ; Hospital Clinic of Barcelona Ferrer, Miquel; Department of Pneumology, Institut Clinic de Respiratori - Hospital Clinic of Barcelona,Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), University of Barcelona (UB), ICREA Academia award, Ciber de Enfermedades Respiratorias (Ciberes, CB06/06/0028) GARCIA-VIDAL, CAROLINA; Hospital Clinic de Barcelona, Infectious diseases Menéndez, Rosario; IIS/Hospital Universitario y Politécnico La Fe. CIBER Enfermedades Respiratorias (CIBERES), Pneumology Torres, Antoni; Biomedical Research Networking Centres in Respiratory Diseases (Ciberes) Barcelona, Department of Pneumology; ICREA
Keywords:	PNEUMONIA, artificial intelligence, machine learning, COMMUNITY-ACQUIRED PNEUMONIA

Word counts abstract: 275
Word counts text: 3006

Machine-Learning Model for Mortality Prediction in Patients with Community-acquired Pneumonia: Development and validation study

Running title: Machine-Learning for Mortality Prediction in CAP

Catia Cilloniz, PhD[1*], Logan Ward, PhD[2*], Mads Lause Mogensen, PhD[2*], Juan M Pericàs, MD, PhD[3,4*], Raúl Méndez, MD,PhD[5], Albert Gabarrús, MSc[1], Miquel Ferrer, MD, PhD[1], Carolina Garcia-Vidal, MD, PhD[6], Rosario Menendez, MD, PhD[5], Antoni Torres, MD, PhD[1.6]

[1]Department of Pneumology, Hospital Clinic of Barcelona; August Pi i Sunyer Biomedical Research Institute–IDIBAPS, University of Barcelona; Biomedical Research Networking Centres in Respiratory Diseases (Ciberes) Barcelona, Spain.

[2] Treat Systems ApS, Aalborg, Denmark

[3] Department of Infectious Diseases, Hospital Clinic of Barcelona, Barcelona, Spain.

[4] Liver Unit, Vall d'Hebron University Hospital, Vall d'Hebron Institute for Research (VHIR), Barcelona, Spain.

[5] Department of Pneumology, Hospital La Fe de Valencia, Valencia, Spain

[6] ICREA, Barcelona, Spain

*Equal contribution

Corresponding author: Prof. Antoni Torres (primary) or Dr. Catia Cillóniz (alternative)

Department of Pulmonary Medicine, Hospital Clinic of Barcelona

C/ Villarroel 170, 08036 Barcelona, Spain

Tel: (+34) 93-227-5779, fax: (+ 34) 93-227-9813

Email: atorres@clinic.cat (primary) or catiacilloniz@yahoo.com (alternative)

Summary conflict of interest statements:

The authors declare that they have no conflicts of interest with the study. Dr Cillóniz is the recipient of the 2018 SEPAR fellowship and a grant from the *Fondo de Investigación Sanitaria* (PI19/00207). RM has received honoraria for lectures from Pfizer.

Funding information:

This study was supported by *CIBER de Enfermedades Respiratorias* (CIBERES CB06/06/0028), and by 2009 Support to Research Groups of Catalonia 911, IDIBAPS. The founders of the study had no role in the study design, data collection, analysis, or interpretation, writing of the report, or decision to submit for publication. SEPAR integrated respiratory infections program.

Notation of prior abstract publication/presentation:

None

Key words list:

Community-acquired pneumonia

Machine learning

Pneumonia

Artificial intelligence

Mortality prediction

Abbreviation list:

ML: machine learning

AI: artificial intelligence

CAP: community-acquired pneumonia

CPN: causal probabilistic network

SeF: SepsisFinder

SeF-ML: SeF adapted model

ATS/IDSA: American Thoracic Society/Infectious Diseases Society of America

SOFA: Sequential Organ Failure Assessment

qSOFA: quick Sequential Organ Failure Assessment

ED: emergency department

MAP: median arterial pressure

SBP: systolic blood pressure

DBP: diastolic blood pressure

CRP: C-reactive protein

ARDS indicates acute respiratory distress syndrome

AUC-ROC: area under the receiver operating characteristic curve

CI: confidence intervals

Q1: first quartile

Q3: third quartile

FiO_2, fraction of inspired oxygen

IDSA/ATS: Infectious Diseases Society of America/American Thoracic Society

PaO_2: partial pressure of arterial oxygen

PSI, pneumonia severity index

Q1: first quartile

Q3: third quartile

Abstract

Background: Artificial intelligence tools and techniques such as machine learning (ML) are increasingly seen as a suitable manner to increase the prediction capacity of currently available clinical tools, including prognostic scores. However, studies evaluating the efficacy of ML methods in enhancing the predictive capacity of existing scores for community-acquired pneumonia (CAP) are limited. We aimed to apply and validate a causal probabilistic network (CPN) model to predict mortality in patients with CAP.

Research question: Is a CPN model able to predict mortality in patients with CAP better than the commonly-used severity scores?

Study design and methods: Derivation-validation retrospective study conducted in two Spanish University hospitals. The ability to predict 30-day mortality of a CPN designed to predict mortality in sepsis (SeF) and adapted for CAP (SeF-ML) was assessed and compared to other scoring systems (PSI, SOFA, qSOFA and CURB-65). The SeF models are proprietary software. Differences between ROC curves were assessed using De Long's method for correlated ROC curves.

Results: The derivation cohort comprised 4,531 patients whilst the validation cohort had 1,034 patients. In the derivation cohort, the AUC of SeF-ML, CURB-65, SOFA, PSI and qSOFA were 0.801, 0.759, 0.671, 0.799 and 0.642, respectively, for 30-day mortality prediction. In the validation study, the AUC of SeF-ML was 0.826, concordantly with the AUC (0.801) in the derivation data (p=0.51). The AUC of SeF-ML was significantly higher than those of CURB-65 (0.764, p=0.03) and qSOFA (0.729, p=0.005). However, it did not differ significantly from PSI (0.830, p=0.92) and SOFA (0.771, p=0.14).

Interpretation: SeF-ML shows potential for improving mortality prediction amongst patients with CAP using structured health data. Additional external validation studies should be conducted to support generalisability.

Introduction

Community-acquired pneumonia (CAP) remains a major cause of morbidity and mortality worldwide, with children and elderly being the most affected population[1,2]. Early-risk stratification of severity and adequate antimicrobial therapy are critical to improving CAP outcomes[1,3]. Scoring systems such as Pneumonia Severity Index (PSI)[4] and CURB-65 score[5] have been widely used to stratify patients with CAP and support clinical decision-making processes[3]. However, these scores have several limitations that hinder their capacity as clinical decision-making supporting tools, e.g., low positive and negative predictive values in predicting the need of intensive care, or underestimated severity in certain age groups.

Artificial intelligence (AI) tools and techniques such as machine learning (ML) are increasingly seen as a suitable manner to increase the prediction capacity of currently available tools in infectious diseases, e.g., sepsis, antimicrobial resistance and COVID-19[6,7]. Within respiratory medicine, the main applications of AI and ML have included the interpretation of thoracic imaging, lung pathology slides and physiologic data such as pulmonary function tests[8]. Nonetheless, studies evaluating the efficacy of ML methods in enhancing the predictive capacity of existing scores for CAP are limited[9,10].

We aimed to apply a causal probabilistic network (CPN) model previously used in sepsis (SeF) [11–13] to predict 30-day mortality in patients with CAP, comparing the accuracy of this model to that of the established clinical scores. In addition, we pursued validating the ML model in CAP using a large cohort of patients with CAP.

Methods

Study design and participants

A derivation-validation retrospective study was performed when using an innovative CPN model[12,13] to predict mortality in adult patients hospitalised with CAP. The rationale for using this new statistical approach is trying to diminish the effect of correlations and make better use of the variables, thus avoiding losing patients with missing information. We applied the guidelines provided by Leisman et al [14] for reporting of prediction models.

Institutional approval was provided by the IRBs of both University Hospital La Fe of Valencia (EC2011/2019) and Hospital Clínic of Barcelona (HCP2009/5451), which waived the need for informed consent due to the retrospective nature of the study.

Data were collected within the first 24 hours of admission. Collected data had similar definitions for both cohorts, and harmonization between cohorts was elementary. All data set were anonymously analyzed, and the study was performed following current recommendation of the Declaration of Helsinki.

Definitions

CAP was diagnosed if new acute respiratory symptoms, signs, and compatible infiltrate(s) on chest x-ray were present. Severe CAP was defined according to the ATS/IDSA guidelines[3]. Prior antibiotic treatment was defined as the intake of antibiotics during the week before hospital admission. The appropriateness of empiric antibiotic treatment was determined according to multidisciplinary guidelines for the management of CAP[15]. Sepsis was defined as the presence of pneumonia and an increase of ≥2 points in the Sequential Organ Failure Assessment (SOFA) score[16]. We also calculated median arterial pressure (MAP) from systolic blood pressure (SBP) and

diastolic blood pressure (DBP) as 1/3*SBP + 2/3*DBP. The ML technique used is based on CPN [11,12,17]. TREAT-Lab (Treat Systems ApS, Aalborg, Denmark) is a medical device software program that provides a risk assessment of patients with suspected infection (proprietary software). The aim of such software is to inform the use of additional or adjunct diagnostics. For example, the clinician (or clinical microbiologist) can use the risk score to identify high-risk patients, i.e., classified as those with high predicted probability of mortality, who may benefit from rapid diagnostics. Conversely, it may be used to identify low-risk patients who only receive standard of care. Customisable risk thresholds can be set for individual clinical installations depending on resources available and local practice requirements. The risk assessment model used within TREAT-Lab is the SeF CPN.

Patient selection, inclusion and exclusion criteria

Derivation cohort

We enrolled all consecutive adult patients with a CAP diagnosis in the Emergency Department of Hospital Clinic of Barcelona between January 2003 and December 2016. We excluded patients with severe immunosuppression due but not limited to human immunodeficiency virus (HIV) infection, active solid or haematologic malignancy treated with chemotherapy, oral corticosteroid treatment with at least 20 mg of prednisone (or equivalent) per day for at least two weeks, and treatment with other immunosuppressive drugs. We also excluded those with active tuberculosis or a confirmed alternative diagnosis.

Validation cohort

All consecutive patients admitted with CAP to Hospital Universitario y Politécnico La Fe (Valencia, Spain) between January 2012 and December 2018 were included. The

inclusion criteria were CAP diagnosis based on a new radiologic infiltrate with at least two compatible clinical symptoms and age ≥18 years. Exclusion criteria were hospital admission within the previous 15 days, immunosuppressive treatments and HIV infection.

Data collection and evaluation

Derivation cohort

Demographic variables, comorbidities, and physiologic parameters were collected at the emergency department within 24 hours of admission. The PSI, CURB-65, qSOFA and SOFA score were calculated at admission[4,5,18]. We recorded whether patients had specific complications, including multilobar infiltration, pleural effusions, acute respiratory distress syndrome (ARDS)[19], septic shock[20] and acute renal failure[21] during hospitalisation. All surviving patients were visited or contacted by telephone within 30 days of discharge; hospital records and the Catalunya Health Department database were reviewed at the 1-year mark. We included all available patients in this analysis. We also calculated MAP from SBP and DBP as 1/3*SBP + 2/3*DBP. We discretised PO2 and FiO2 as required by the model. Finally, we transformed creatinine, C-reactive protein (CRP), lactate, bilirubin and platelets through the natural logarithm.

Validation cohort

Demographic characteristics and comorbidities (diabetes mellitus, respiratory, heart, liver, neurological, and renal diseases) were collected at time of admission. The severity of disease at presentation was assessed with the PSI. Antibiotic treatment before CAP diagnosis in the current episode was recorded. We included all available patients in this analysis.

Outcomes

Primary outcome: death within 30 days of admission.

Statistical analysis

We report the number and percentage of patients for categorical variables and the median and interquartile range for continuous variables (non-normal distribution confirmed by the Kolmogorov-Smirnov test). Categorical variables were compared using the χ^2 test. Continuous variables were compared using the nonparametric Mann-Whitney test.

Derivation cohort

We used an updated version of the published SepsisFinder (SeF) model (see supplementary material for a description of the modelling techniques used including variable selection and parameterization). The main differences in input variables between SeF and the adapted model (SeF-ML) are shown in Table 1.

We adapted the respiratory component of SeF, which comprises acid-base balance (pH, HCO3-), respiratory rate, and oxygen perfusion (measured through PaO2, SaO2 and FiO2 [FiO2 was discretized based on a conversion from oxygen flow rate in L/min. PaO2 was discretized into 8 bins]). We retrained this portion of the model using data included in the study. The remainder of the model was deemed invariant. To evaluate the degree of overfitting to the derivation data, we performed a 10-fold cross-validation, shuffling the data and stratifying to ensure a similar proportion of outcomes in each cross-validation fold. We then retrained the model using the full derivation dataset.

No explicit steps to handle missing data were required. CPNs are inherently tolerant of missing information and are able to perform inference with partial evidence:

combining the model's structure and conditional probability tables with the axioms of probability theory allows the marginal probabilities of all nodes in a CPN to be calculated when only some of the nodes are observed (supplemental material and e-Table 1)[22,23]. We then used the SeF-ML model to calculate the probability of death within 30 days (Pmort) for all patients with >2 input variables recorded.

We assessed predictive performance of SeF and SeF-ML by using the area under the receiver operating characteristic curve (AUC-ROC). As a general rule, the relation between AUC and diagnostic accuracy is as follows: AUC between 0.90 and 1.00 has outstanding discrimination ability; 0.80 - 0.90, excellent; 0.70 - 0.80, acceptable; 0.60 - 0.70, poor, and 0.50 - 0.60 fails to accurately diagnose a certain disease or condition[24,25]. Kaplan-Meier survival curves were also constructed.

In our preliminary analysis of this data, we noted that SeF-ML performance was not negatively impacted by the greater degree of missing data amongst patients included from 2003 to 2006 (e-Table 2).

We compared SeF and SeF-ML performance (Pmort as a predictor of death within 30 days) with other scoring systems (PSI, SOFA, qSOFA and CURB-65). Differences between ROC curves were assessed using De Long's method for correlated ROC curves as implemented in the pROC package of R.

In addition to assessing the model's performance, we computed an example for a potential use-case for TREAT Lab, showing patients being stratified into groups of 40% low-risk, 40% medium-risk and 20% high-risk.

Validation cohort

We used the SeF-ML model (Table 1), where the respiratory components were adapted through learning from 4,531 patients with CAP at the Hospital Clinic of

Barcelona. We used the adapted model to calculate the probability of death within 30 days (Pmort) for all patients.

We assessed predictive performance using the AUC. We compared SeF-ML's performance (Pmort as a predictor of death within 30 days) with other scoring systems (i.e., PSI, CURB-65, SOFA and qSOFA). Differences between ROC curves were assessed using DeLong's method for correlated ROC curves as implemented in the pROC package of R. We also compared mortality for a range of risk cut-offs to assess potential operating points for clinical implementation of risk stratification. Model calibration was assessed using the Hosmer-Lemeshow goodness of fit test and by calculating the Brier-score loss.

More details about the selection of variables and the ML model are displayed in the supplementary material (e-Figure 1, e-Figure 2 and e-Figure 3).

Results

Patients' characteristics

The derivation cohort comprised 4,531 (15% outpatients and 85% inpatients) patients and the validation cohort 1,034 patients. Clinical characteristics of the studied population are shown in Table 2.

Predictive performance of risk assessment

Derivation cohort

We used the SeF-ML model to calculate Pmort for all patients with >2 input variables – 4,500/4,531=99.3% of patients. The AUC for 30-day mortality prediction was 0.801 for SeF-ML, being significantly better than CURB-65 (0.759, p<0.001), SOFA (0.671, p<0.001), and qSOFA (0.642, p<0.001) (Table 3). The mean cross-validation performance, measured by AUC, was 0.800 (range: 0.749-0.832) which did not differ from the AUC for the full dataset. Details of the cross-validation assessment are included in the supplemental material.

PSI was only available for 58% of patients and had an AUC of 0.799; it was not significantly different from SeF-ML (p=0.58). CURB-65 provided a "fair" prediction of mortality, while SeF-ML and PSI provided a "good" prediction of 30-day mortality (Figure 1). The calibration of SeF-ML was measured using the Hosmer-Lemeshow statistic and the Brier-score loss. The Hosmer-Lemeshow statistic was 15.62 (p=0.048) which suggests the model may not be well calibrated. However, the Hosmer-Lemeshow statistic is known to be very sensitive to sample size. The model appears visually well-calibrated. The Brier-score loss for the model was 0.056.

Survival curves are shown in Figure 2 for patients stratified according to Pmort; the low-risk group comprises 40% of patients with the lowest Pmort; the medium-risk

group represents the next highest 40%; and the high-risk group accounts for the top 20% according to Pmort. The additional value provided by SeF-ML with respect to CURB-65 and PSI scores is shown in Figure 3: SeF-ML provides additional discrimination for mortality, e.g., for patients with PSI ≥4.

Validation cohort

For the validation cohort, the AUC for SeF-ML was 0.826. It was not significantly different from performance in the derivation cohort (p=0.51) (Figure 4). According to the Hosmer-Lemeshow statistic, SeF-ML was well-calibrated (HL statistic = 11.93, p=0.15). The Brier-score loss for SeF-ML was 0.036. Notably, only 23.1% patients had complete information to calculate the SOFA score, being not significantly different from SeF-ML (p=0.85). On the contrary, SeF-ML had a significantly higher AUC than both CURB-65 and qSOFA (p=0.03 and p=0.005, respectively) (Table 4).

When analysing the imputed SOFA score (assuming missing components were normal), the AUC of SOFA improved from 0.728 to 0.771; the difference from SeF-ML remained non-significant.

Supplementary e-Table 3 shows the number of patients in each PSI risk class and associated 30-day mortality. Also, 30-day mortality within quantile-matched risk classes for SeF-ML is also shown. SeF-ML risk groups were defined by choosing cut-offs that resulted in the same number of patients assigned to each risk group like in the corresponding PSI risk class.

Potential use of risk assessment

Patients can be stratified by their probability of death within 30 days. An odds ratio (OR) for death can also be calculated for high-risk patients vs. others. The OR represents the degree of separation between high-risk and low-risk patients. For this

analysis, we considered two strategies for stratification: high-risk patients receive rapid diagnostics and low-risk patients receive minimum standard of care. For example, the top 20% had a probability of death within 30 days of 18.4%, whilst OR for the high-risk group vs. remaining patients was 7.6 (e-Figure 4). One advantage of SeF-ML is the continuous probability output that allows custom thresholds to be set depending on the care environment versus the five potential operating points for CURB-65, for example. For comparison, three potential scenarios are shown below: 1) *PSI=5* (N=98) 30-day mortality = 21.9% (PSI), 24% (SeF-ML); OR for high-risk vs. low risk = 11.5 (PSI), 14.3 (SeF-ML); 2) PSI ≥4 (N=428) 30-day mortality = 8.6% (PSI), 8.4% (SeF-ML); OR for high-risk vs. low risk = 9.3 (PSI), 7.7 (SeF-ML); 3) CURB-65 ≥3 (N=153) 30-day mortality = 14.4% (CURB-65), 18.3% (SeF-ML); OR for high-risk vs. low risk = 6.8 (CURB-65), 12.8 (SeF-ML). A fuller picture of the effect of choosing different operating points is shown in Supplementary e-Figure 4 and e-Figure 5. It shows the 30-day mortality as functions of the size of the high-risk group along with operating points for PSI (state) and CURB-65. The smaller the high-risk group, the higher the 30-day mortality.

Discussion

In our study, we applied a ML approach to develop and validate a 30-day mortality prediction clinical model in patients with CAP. Although not pneumonia-specific, the SeF-ML model shows potential for improving mortality prediction amongst patients with CAP. Remarkably, SeF-ML's performance in the validation set matched that of the training data in the derivation cohort, as did those of PSI and CURB-65, whereas the performance of SOFA and qSOFA scores did not match. SeF-ML not only enhances the mortality prediction ability of currently available tools but optimizes the use and quality of available electronic health records (EHR) data. Hence, although the advantages conferred by SeF-ML need further refining and interpretation, this ML model might be applied in clinical practice in the near future, i.e., patients can be stratified by their probability of death and such stratification might then be used to determine a patient's care trajectory and diagnostic workup.

This study is one of the first to use a large cohort of hospitalised patients with CAP to generate a ML model validated through an external cohort of patients with CAP. A barrier for physicians in using ML is its potential "black box" opacity. However, studies like ours show that results obtained with ML predictions are consistent with other severity scores that we have used so far[26,27]. In the near future, ML techniques will allow us to analyse a large volume of data that current techniques cannot do, facilitating the possibility amongst investigators to directly collect data from EHR [7,26–28]. The potential of SeF-ML to improve the current ability of available clinical scores for CAP primarily relies on the findings suggesting that SeF-ML better predicts 30-day mortality than qSOFA and CURB-65 according to our data, which nonetheless require further clinical validation. Compounding this is also the fact that SeF-ML had a higher

AUC than SOFA and PSI, albeit non-significant. Furthermore, the AUC value of SeF-ML was consistent between the predictive and validation model. There is strong evidence supporting the importance of accurately assessing the severity of CAP and stratifying patients based on their mortality risk to improve clinical outcomes[1,3,29]. Despite that, PSI and CURB-65 remain the most widely used CAP severity scores and recommended by international guidelines[3,29,30]. The ability of these scores to predict mortality has some limitations[31]. Moreover, although qSOFA score ≥2 is strongly associated with mortality in patients with pneumonia, its use in early identification of patients with CAP and mortality risk is hindered[32,33]. SeF-ML seems to provide better discriminative capacity to discern between high- and low-risk patients, which is key to adapting the intensity of care and resources per foreseen prognosis. However, this still needs further validation to prove its actual clinical validity. In addition, SeF-ML increases efficiency in exploiting available data. The lower requirements set for minimum data and the ability to handle missing data mean that 99% of patients, on average, would have sufficient data for predictions to be established. This would facilitate optimization of EHR use, with fewer investments on data collection and curation when compared to other scores. Also, the continuous output achieved with SeF-ML allows for adaptive fine-tuning of patient classification. Cut-offs for defining risk can be smoothly and accurately adjusted. In particular, the ability of SeF-ML to identify high-risk patients with low CURB65 and PSI scores, as well as low-risk patients with high CURB65 and PSI scores paramount as these CAP-specific clinical scores are the two most widely used. It is worth noting, however, that enhanced mortality prediction through continuous appraisals is not unique to ML but is also found in other models with continuous

output, e.g. the eCURB model, which achieved greater predictive accuracy using CURB-65 variables using regression splines[34].

The ability of SeF-ML to reclassify risk of patients with CAP might be more beneficial for certain patient subgroups, e.g., those with low CURB-65/PSI yet high SeF and mortality, or those with high CURB-65/PSI yet low SeF and mortality.

A foremost strength of our study is the use of a large set of clinical data that are representative of the population in a defined area of influence. This study feature enabled the integration of surveillance data into direct clinical care of individual patients and can be helpful in making decisions by applying ML models like SeF-ML. When developing and validating the algorithm, we used two large patient slices to ensure that our model can be implemented using real-time patient data.

However, some limitations need to be highlighted. This score validation against mortality is only the first step toward clinical utility. Remarkably, the ability of SeF-ML to predict ICU admission remains unproven. In particular, our study lacked information on post-admission disposition, and therefore we were not able to test SeF-ML against potentially different clinical pathways other than direct admission to the ICU. Furthermore, the use of closed databases does not incorporate new information and therefore does not allow "learning" of ML models. Besides, despite the large sample sizes of both cohorts, some baseline variables and clinical features, including CAP severity and outcomes, differed. However, this does not hamper the training and validation of the CPN model. There are components of the current SeF that were not available in the dataset due to them not generally being measured in the ED for CAP. However, SeF performs well despite not including these variables, which helps to demonstrate its robustness to missing values.

Interpretation

SeF-ML performance at predicting 30-day mortality appears to be overall superior than that of existing CAP-specific scores, with the exception of PSI in the validation model. SeF-ML offers some advantages over current scoring methods, eg., calculations easily made on routinely collected data and based upon; structured EHR data (vs. subjective criteria and arterial blood gas sampling needed for PSI for instance); and tunable performance so as to allow risk cut-offs to be tailored to workflow requirements and capacities of the individual institution (compared with fewer states in CURB-65). In addition, SeF-ML performance seems to not be dependent upon data availability, therefore allowing for more effective calculation of risk scores for CAP based upon data sources with limited access to or completeness of certain variables. Our findings need further validation in other cohorts from different settings to assess the actual clinical utility of SeF-ML in predicting CAP prognosis.

Acknowledgements

Guarantor:

Dr. Antoni Torres is the guarantor of the article

Author contributions:

Conception and design of the study: CC, LW, ML, MF, JP, RM, RM, AT. Data acquisition: all authors. Data analysis and interpretation: LW, ML, CC, AG. Manuscript drafting or critical revision for important intellectual content: all authors. Final approval of the submitted version: all authors.

Financial/nonfinancial disclosures:

The authors have reported to Chest the following: This study was supported by *CIBER de Enfermedades Respiratorias* (CIBERES CB06/06/0028), and by 2009 Support to Research Groups of Catalonia 911, IDIBAPS. The founders of the study had no role in the study design, data collection, analysis, or interpretation, writing of the report, or decision to submit for publication. SEPAR integrated respiratory infections program.

Role of sponsors:

The funders did not have a role in the conduct of the study; in the collection, management, analysis, or interpretation of data; or in the preparation of the manuscript.

References

1. Torres A, Cilloniz C, Niederman MS, et al. Pneumonia. *Nature Reviews Disease Primers* 2021;7(1):1–28.

2. GBD 2019 Diseases and Injuries Collaborators. Global burden of 369 diseases and injuries in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet* 2020;396(10258):1204–1222.

3. Metlay JP, Waterer GW, Long AC, et al. Diagnosis and Treatment of Adults with Community-acquired Pneumonia. An Official Clinical Practice Guideline of the American Thoracic Society and Infectious Diseases Society of America. *Am J Respir Crit Care Med* 2019;200(7):e45–e67.

4. Fine MJ, Auble TE, Yealy DM, et al. A prediction rule to identify low-risk patients with community-acquired pneumonia. *N Engl J Med* 1997;336(4):243–250.

5. Lim WS, Lewis S, Macfarlane JT. Severity prediction rules in community acquired pneumonia: a validation study. *Thorax* 2000;55(3):219–223.

6. Peiffer-Smadja N, Rawson TM, Ahmad R, et al. Machine learning for clinical decision support in infectious diseases: a narrative review of current applications. *Clin Microbiol Infect* 2020;26(5):584–595.

7. Garcia-Vidal C, Sanjuan G, Puerta-Alcalde P, Moreno-García E, Soriano A. Artificial intelligence to support clinical decision-making processes. *EBioMedicine* 2019;46:27–29.

8. Gonem S, Janssens W, Das N, Topalovic M. Applications of artificial intelligence and machine learning in respiratory medicine. *Thorax* 2020;75(8):695–701.

9. Kang SY, Cha WC, Yoo J, et al. Predicting 30-day mortality of patients with pneumonia in an emergency department setting using machine-learning models. *Clin Exp Emerg Med* 2020;7(3):197–205.

10. Chumbita M, Cillóniz C, Puerta-Alcalde P, et al. Can Artificial Intelligence Improve the Management of Pneumonia. *J Clin Med* 2020;9(1):248.

11. Ward L, Andreassen S, Astrup JJ, Rahmani Z, Fantini M, Sambri V. Clinical- vs. model-based selection of patients suspected of sepsis for direct-from-blood rapid diagnostics in the emergency department: a retrospective study. *Eur J Clin Microbiol Infect Dis* 2019;38(8):1515–1522.

12. Ward L, Møller JK, Eliakim-Raz N, Andreassen S. Prediction of Bacteraemia and of 30-day Mortality Among Patients with Suspected Infection using a CPN Model of Systemic Inflammation. *IFAC-PapersOnLine* 2018;116–121.

13. Andreassen S, Møller JK, Eliakim-Raz N, Lisby G, Ward L. A comparison of predictors for mortality and bacteraemia in patients suspected of infection. *BMC Infect Dis* 2021;21(1):864.

14. Leisman DE, Harhay MO, Lederer DJ, et al. Development and Reporting of Prediction Models: Guidance for Authors From Editors of Respiratory, Sleep, and Critical Care Journals. *Crit Care Med* 2020;48(5):623–633.

15. Torres A, Barberán J, Falguera M, et al. [Multidisciplinary guidelines for the management of community-acquired pneumonia]. *Med Clin (Barc)* 2013;140(5):223.e1-223.e19.

16. Singer M, Deutschman CS, Seymour CW, et al. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA* 2016;315(8):801–810.

17. Logan Morgan Ward. Gradation of the Severity of Sepsis: Learning in a Causal Probabilistic Network. 2016;55.

18. Vincent JL, Moreno R, Takala J, et al. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. On behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine. *Intensive Care Med* 1996;22(7):707–710.

19. ARDS Definition Task Force, Ranieri VM, Rubenfeld GD, et al. Acute respiratory distress syndrome: the Berlin Definition. *JAMA* 2012;307(23):2526–2533.

20. Rhodes A, Evans LE, Alhazzani W, et al. Surviving Sepsis Campaign: International Guidelines for Management of Sepsis and Septic Shock: 2016. *Intensive Care Med* 2017;43(3):304–377.

21. Bellomo R, Ronco C, Kellum JA, Mehta RL, Palevsky P, Acute Dialysis Quality Initiative workgroup. Acute renal failure - definition, outcome measures, animal models, fluid therapy and information technology needs: the Second International Consensus Conference of the Acute Dialysis Quality Initiative (ADQI) Group. *Crit Care* 2004;8(4):R204-212.

22. Steen Andreassen, Leonard Leibovici, Mical Paul, Anders D. Nielsen, Alina Zalounina, Leif E. Kristensen, Karsten Falborg, Brian Kristensen, Uwe Frank, Henrik C. Schønheyder. A Probabilistic Network for Fusion of Data and Knowledge in Clinical Microbiology [Internet]. springerprofessional.de. 2005 [cited 2022 Mar 22];Available from: https://www.springerprofessional.de/en/a-probabilistic-network-for-fusion-of-data-and-knowledge-in-clin/1020738

23. Lauritzen SL, Spiegelhalter DJ. Local Computations with Probabilities on Graphical Structures and Their Application to Expert Systems. *Journal of the Royal Statistical Society: Series B (Methodological)* 1988;50(2):157–194.

24. Cole TJ. Applied logistic regression. D. W. Hosmer and S. Lemeshow, Wiley, New York, 1989. *Statistics in Medicine* 1989;10(7):1162–1163.

25. Altman DG, Bland JM. Diagnostic tests 3: receiver operating characteristic plots. *BMJ* 1994;309(6948):188.

26. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. *Lancet* 2019;393(10181):1577–1579.

27. Mekov E, Miravitlles M, Petkov R. Artificial intelligence and machine learning in respiratory medicine. *Expert Rev Respir Med* 2020;14(6):559–564.

28. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. *Nat Med* 2019;25(1):30–36.

29. Menéndez R, Cilloniz C, España PP, et al. Community-Acquired Pneumonia. Spanish Society of Pulmonology and Thoracic Surgery (SEPAR) Guidelines. 2020 Update. *Arch Bronconeumol* 2020;56 Suppl 1:1–10.

30. Lim WS, Baudouin SV, George RC, et al. BTS guidelines for the management of community acquired pneumonia in adults: update 2009. *Thorax* 2009;64 Suppl 3:iii1-55.

31. Loke YK, Kwok CS, Niruban A, Myint PK. Value of severity scales in predicting mortality from community-acquired pneumonia: systematic review and meta-analysis. *Thorax* 2010;65(10):884–890.

32. Jiang J, Yang J, Jin Y, Cao J, Lu Y. Role of qSOFA in predicting mortality of pneumonia: A systematic review and meta-analysis. *Medicine (Baltimore)* 2018;97(40):e12634.

33. Ranzani OT, Prina E, Menéndez R, et al. New Sepsis Definition (Sepsis-3) and Community-acquired Pneumonia Mortality. A Validation and Clinical Decision-Making Study. *Am J Respir Crit Care Med* 2017;196(10):1287–1297.

34. Jones BE, Jones J, Bewick T, et al. CURB-65 pneumonia severity assessment adapted for electronic decision support. *Chest* 2011;140(1):156–163.

Table 1. Variables used in scoring systems

Variable	SeF	SeF-ML	PSI	SOFA	qSOFA	CURB-65
Demography						
Age	Yes	Yes	Yes			Yes
Sex			Yes			
Nursing home resident			Yes			
Comorbidities						
Neoplastic disease			Yes			
Liver disease			Yes			
Congestive heart failure			Yes			
Cerebrovascular disease			Yes			
Renal disease			Yes			
Vitals						
Temperature	Yes	Yes	Yes			
Chills	Yes	Yes				
Systolic Blood Pressure			Yes		Yes	Yes
Diastolic Blood Pressure						Yes
Mean Arterial Pressure	Yes	Yes		Yes		
Heart Rate	Yes	Yes	Yes			
Respiratory Rate		Yes	Yes		Yes	Yes
Mental Status	Yes	Yes	Yes	Yes, GCS	Yes	Yes
SpO2		Yes				
Labs						
CRP	Yes	Yes				
Creatinine	Yes	Yes		Yes		
Albumin	Yes*	Yes*				
Lactate	Yes*	Yes*				
Platelets	Yes	Yes		Yes		
Neutrophils%	Yes	Yes				
Bilirubin		Yes		Yes		
BUN or Urea		Yes	Yes			Yes
Sodium			Yes			
Glucose			Yes			
pH		Yes	Yes			

Variable	SeF	SeF-ML	PSI	SOFA	qSOFA	CURB-65
Hematocrit		Yes	Yes			
PaO2		Yes	Yes	Yes		
FiO2		Yes		Yes		
Bicarbonate		Yes				
Leukocytes		Yes				
Radiology						
Chest x-ray			Yes			
Treatment/devices						
Mechanical ventilation				Yes		
Pressors				Yes		

Abbreviations: Glasgow coma scale (GCS); * not part of standard testing for CAP patients at the emergency department (ED)

534 **Table 2. Clinical characteristics of the studied cohorts**

Variable	Derivation cohort (N = 4,531)	Validation cohort (N = 1,034)	P-value
Age, median (Q1; Q3), years	73 (55; 82)	72 (59; 81)	0.650
Male sex, n (%)	2,708 (60)	642 (62)	0.176
Previous antibiotic, n (%)	1,057 (25)	338 (33)	<0.001
Influenza vaccine, n (%)	1,770 (44)	438 (42)	0.350
Pneumococcal vaccine, n (%)	695 (17)	79 (8)	<0.001
Previous inhaled corticosteroids, n (%)	818 (18)	-	-
Previous systemic corticosteroids, n (%)	180 (4)	-	-
Previous episode of pneumonia (last year), n (%)	630 (14)	-	-
Comorbidities, n (%) [a]	2,982 (66)	731 (71)	0.005
Chronic respiratory disease	1,729 (39)	325 (31)	<0.001
Chronic cardiovascular disease	635 (14)	347 (34)	<0.001
Diabetes mellitus	914 (20)	288 (28)	<0.001
Neurologic disease	836 (19)	176 (17)	0.137
Chronic renal disease	318 (7)	139 (13)	<0.001
Chronic liver disease	212 (5)	38 (4)	0.146
Nursing home, n (%)	352 (8)	43 (4)	<0.001
Confusion, n (%)	820 (18)	77 (7)	<0.001
Respiratory rate, median (Q1; Q3), breaths/min	24 (20; 30)	18 (16; 24)	<0.001
Heart rate, median (Q1; Q3), beats/min	97 (84; 110)	97 (85; 110)	0.685
Systolic blood pressure, median (Q1; Q3), mmHg	129 (112; 148)	134 (118; 152)	<0.001
Diastolic blood pressure, median (Q1; Q3), mmHg	72 (64; 80)	70 (61; 80)	0.007
Creatinine, median (Q1; Q3), mg/dL	1.0 (0.8; 1.4)	1.0 (0.8; 1.4)	0.003
Glucose level, median (Q1; Q3), mg/dL	124 (105; 157)	-	-

Variable	Derivation cohort (N = 4,531)	Validation cohort (N = 1,034)	P-value
PaO_2/FiO_2, median (Q1; Q3)	281 (238; 327)	271 (238; 311)	0.012
PSI score, median (Q1; Q3)	98 (74; 123)	86 (66; 105)	<0.001
Severe CAP, n (%)	868 (26)	144 (14)	<0.001
Bacteraemia, n (%) [b]	390 (12)	54 (8)	0.001
Appropriate empiric treatment, n (%)	2,844 (96)	-	-
Length of hospital stay, median (Q1; Q3), days	7 (4; 11)	6 (5; 9)	0.261
Mechanical ventilation, n (%)	375 (10)	26 (3)	<0.001
Non-invasive	169 (4)	-	-
Invasive	206 (5) [c]	-	-
In-hospital mortality, n (%)	272 (6)	35 (3)	<0.001
30-day mortality, n (%)	293 (7)	43 (4)	0.004

Abbreviations: CAP indicates community-acquired pneumonia; PSI, pneumonia severity index; Q1, first quartile; Q3, third quartile. Percentages calculated on non-missing data. [a] May have >1 comorbid condition. [b] Calculated only for patients with blood samples (3,206 in the derivation cohort and 696 in the validation cohort). [c] Patients who initially received non-invasive ventilation yet needed intubation subsequently were included in the invasive mechanical ventilation group.

Table 3. Discriminatory ability of SeF-ML vs. other scoring systems for 30-day mortality in the derivation cohort

Model/Score	Availability	AUC (95% CI)	P-value (vs. SeF-ML)
SeF-ML	99.3%	0.801 (0.777-0.825)	-
(limited to those where PSI was available)		0.808 (0.776-0.839)	-
(limited to those where SOFA was available)		0.804 (0.780-0.829)	-
(limited to those where CURB-65 was available)		0.814 (0.788-0.839)	-
PSI (raw score)	57.7%	0.799 (0.768-0.830)	0.58
SOFA	91.4%	0.671 (0.638-0.704)	<0.001
qSOFA [a]	100%	0.642 (0.611-0.673)	<0.001
CURB-65	82.5%	0.759 (0.732-0.786)	<0.001

Abbreviations: AUC indicates area under the receiver operating characteristic curve; CI, confidence interval; PSI, pneumonia severity index. [a] Calculated assuming missing values were normal.

Table 4. Discriminatory ability of SeF-ML vs. other scoring systems for 30-day mortality in the validation cohort

Model/Score	Availability	AUC (95% CI)	P-value (vs. SeF-ML)
SeF-ML	100%	0.826 (0.753-0.899)	-
PSI (raw score)	100%	0.830 (0.753-0.90)	0.92
CURB-65	100%	0.764 (0.694-0.834)	0.03
SOFA	23.1%	0.728 (0.588-0.869)	0.85 [a]
SOFA-imputed [b]	100%	0.771 (0.706-0.836)	0.14
qSOFA	98.3%	0.729 (0.653-0.804)	0.005 [a]

Abbreviations: AUC indicates area under the receiver operating characteristic curve; CI, confidence interval; PSI, pneumonia severity index. [a] Compared only for patients with complete SOFA/qSOFA score. [b] Calculated assuming missing values were normal.

Figure Legends

Figure 1. Model performance for 30-day mortality prediction in the derivation cohort. A: ROC curves B: Calibration curves for SeF models

Figure 2. Survival curves for patients stratified according to Pmort in the derivation cohort

Figure 3. Survival curves stratified by SF risk group, set together according to CURB-65/PSI score in the derivation cohort

Figure 4. Model performance for 30-day mortality prediction in the validation cohort. A: ROC curves B: Calibration curves for SeF models

Take-home Points:

Study Questions: Did a CPN model could predict mortality in patients with CAP better than the frequent severity scores?

Results: SeF-ML performance at predicting 30-day mortality was overall significantly better than that of existing CAP-specific scores.

Interpretation: Our results showed that SeF-ML shows potential for improving mortality prediction amongst patients with CAP using structured health data.

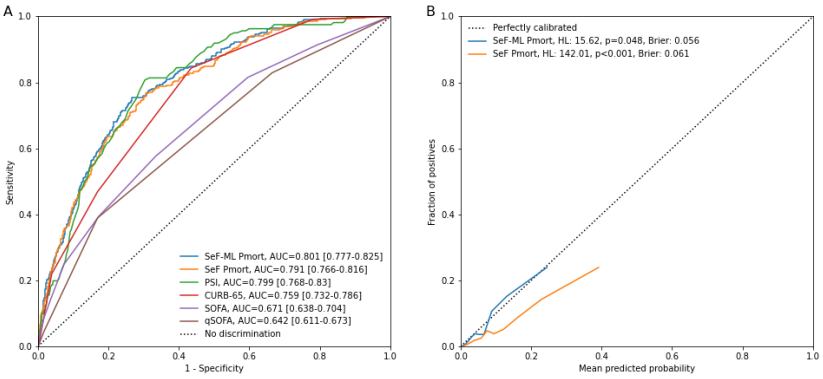


Figure 1. Model performance for 30-day mortality prediction in the derivation cohort. A: ROC curves B: Calibration curves for SeF models

406x177mm (72 x 72 DPI)

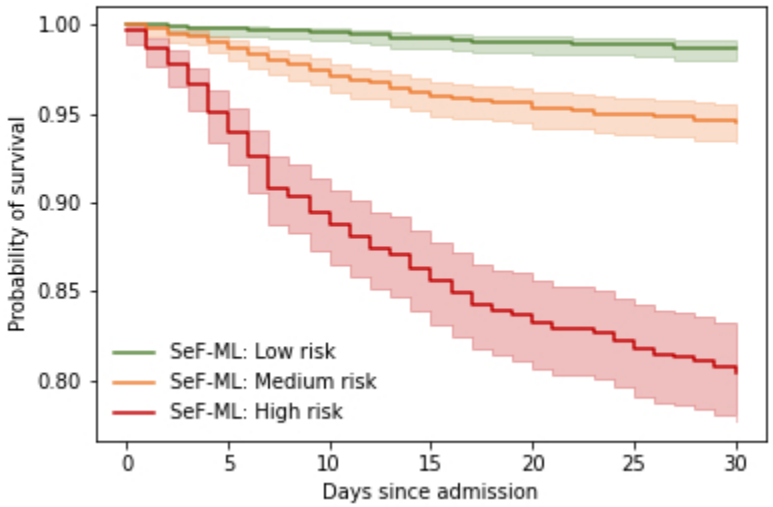


Figure 2. Survival curves for patients stratified according to Pmort in the derivation cohort

152x101mm (72 x 72 DPI)

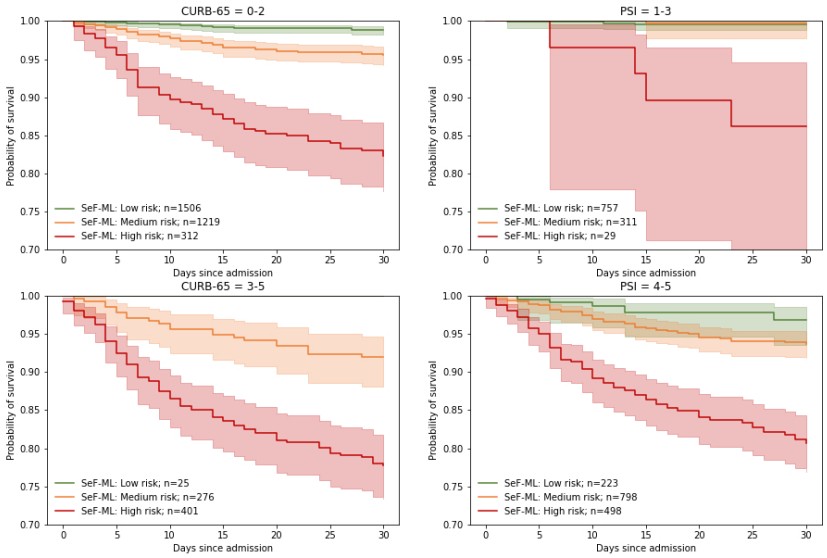


Figure 3. Survival curves stratified by SF risk group, set together according to CURB-65/PSI score in the derivation cohort

381x254mm (72 x 72 DPI)

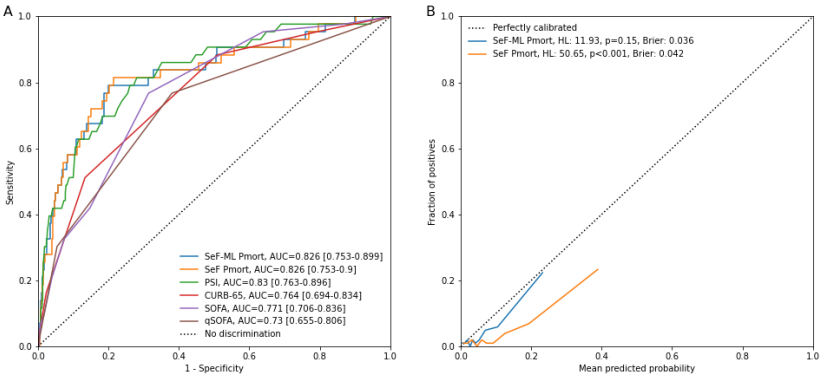
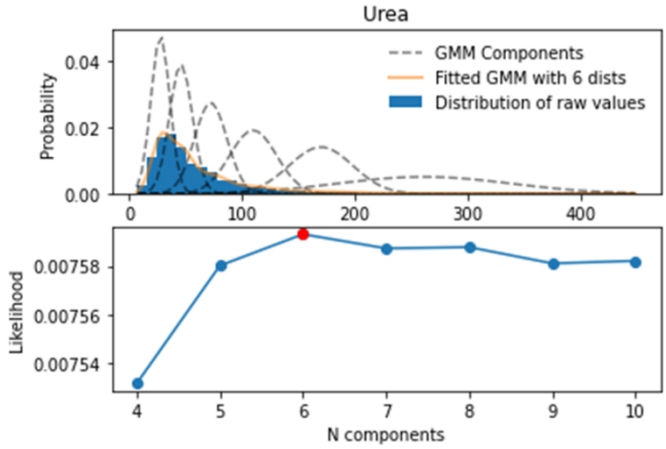


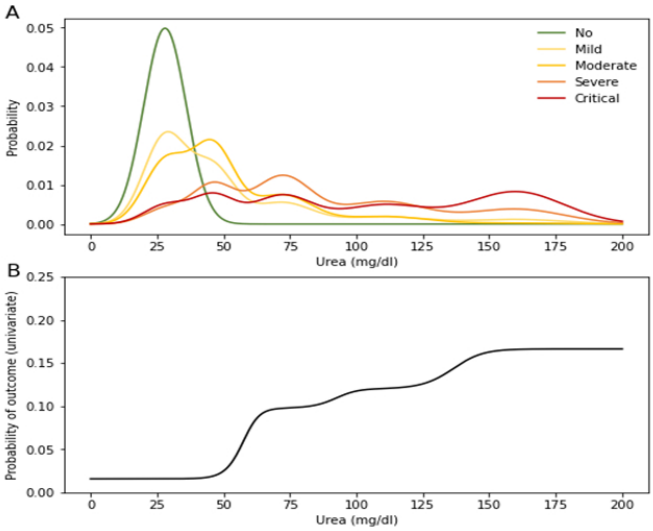
Figure 4. Model performance for 30-day mortality prediction in the validation cohort. A: ROC curves B: Calibration curves for SeF models

406x177mm (72 x 72 DPI)



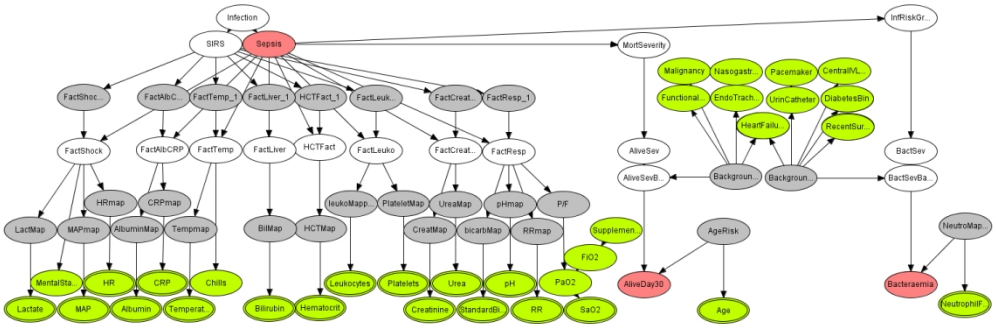
e-Figure 1. Example of continuous variable parameterization using Gaussian Mixture

338x190mm (96 x 96 DPI)



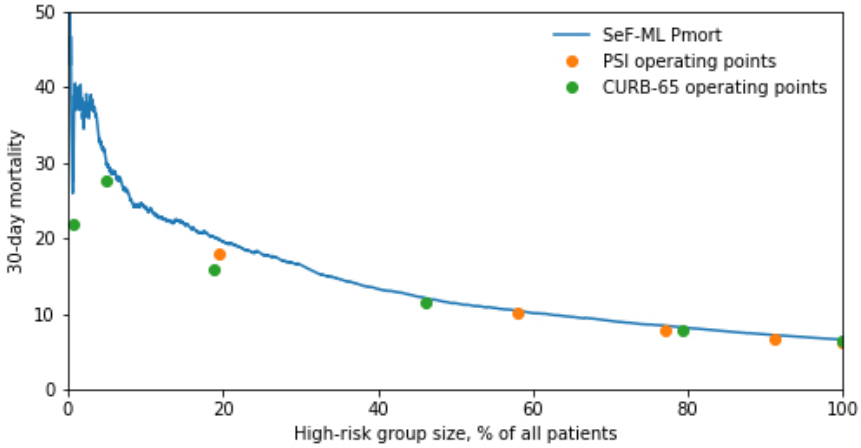
e-Figure 2. Example of composite Gaussian distributions created by mapping across severity states (conditional probabilities learned in the CPN). A: learned distributions B: univariate outcome prediction across the predictor domain

254x190mm (96 x 96 DPI)



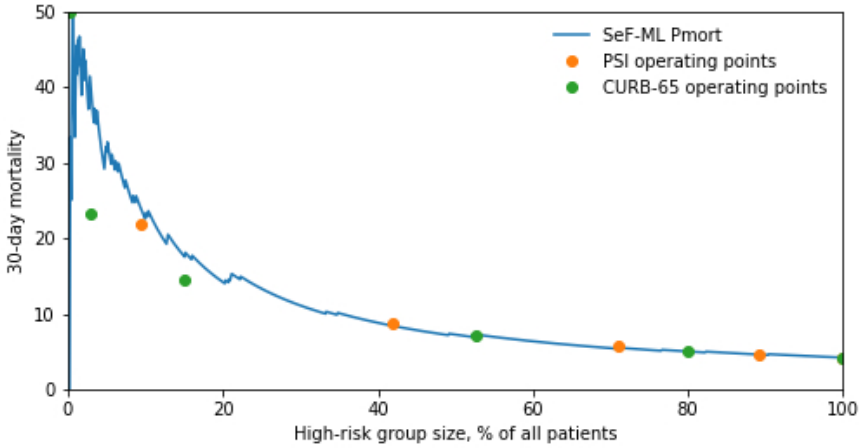
e- Figure 3. Structure of the SeF CPN

536x177mm (72 x 72 DPI)



e-Figure 4. 30-day mortality as a function of the size of the high-risk group, compared with PSI and CURB-65 operating points in the derivation cohort

203x101mm (72 x 72 DPI)



e-Figure 5. 30-day mortality as a function of the size of the high-risk group in the validation cohort

203x101mm (72 x 72 DPI)

Supplementary Material

Machine Learning Model for Mortality Prediction in Patients with Community-acquired Pneumonia: Development and validation study

Catia Cilloniz, PhD[1*], Logan Ward, PhD[2*], Mads Lause Mogensen, PhD[2*], Juan M Pericàs, PhD[3,4*], Raúl Méndez, PhD[5], Albert Gabarrús, MSc[1], Miquel Ferrer, PhD[1], Carolina Garcia-Vidal, PhD[6], Rosario Menendez, PhD[5], Antoni Torres, PhD[1.6]

Changes made to the published SepsisFinder model

Addition of input variables

The published version of the SepsisFinder model [1] was updated to include a wider set of input variables. Variables which are to include or extend modelling of dysfunction associated with the hepatic system (bilirubin), renal system (urea), haematology (WBC, hematocrit) and respiratory system (respiratory rate, SpO2, PaO2, FiO2, pH and HCO3). Background models were added for both bacteraemia (pacemaker, urinary catheter, central IV line, diabetes, recent surgery, heart failure) and mortality (malignancy, functional capacity, nasogastric/endotracheal tubes, heart failure). However, background variables were not used as inputs for analysis in this manuscript. Placement of the neutrophil fraction variable in the model was changed to make it an independent predictor of bacteraemia unrelated to severity.

Relearning conditional probability distributions

Conditional probabilities in the CPN model were learned using the same derivation data and learning technique described in [1–4].

Variable selection

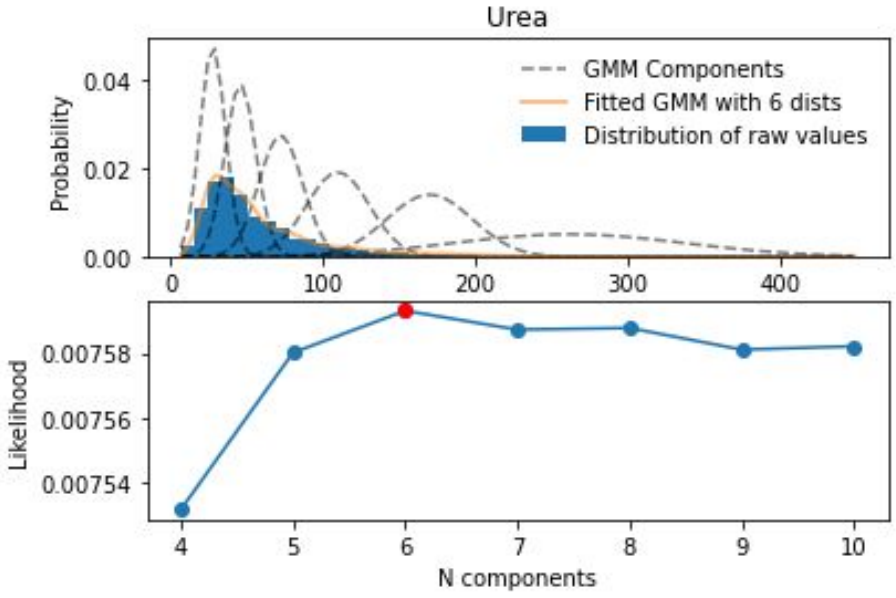
Variable selection for the published version of the SepsisFinder model is described in the literature. Variables were selected if they showed utility in predicting bacteraemia, 30-day mortality or both, as well as a documented link to some underlying physiological process which suggests clinical relevance regarding linkages to infection/organ dysfunction/critical illness. Stepwise selection of variables was not performed.

Variable parameterization

The measured lab/vital variables are input as continuous variables. The SepsisFinder model models a set of severity states, for each input variable this is modeled as a Gaussian mixture. Design of the individual Gaussian components is described in the literature for earlier versions of the SepsisFinder model[3,4], while components for variables added to the latest version were generated by fitting a Gaussian Mixture Model (GMM). GMM modelling was performed in python (scikit-learn).

To determine the optimal number of Gaussian components and their mean/variance, a grid search across the number of components (min=4, max=10) was performed to assess the 10-fold cross-validation performance. The number of components with the highest likelihood (best overall fit of the input data) was selected and the GMM was then relearned with the complete training set. Manual adjustments were then made to avoid distributions with very high variance as this can create undesired behavior near the edges of other distributions. An example is shown in e-Figure 1 for Urea.

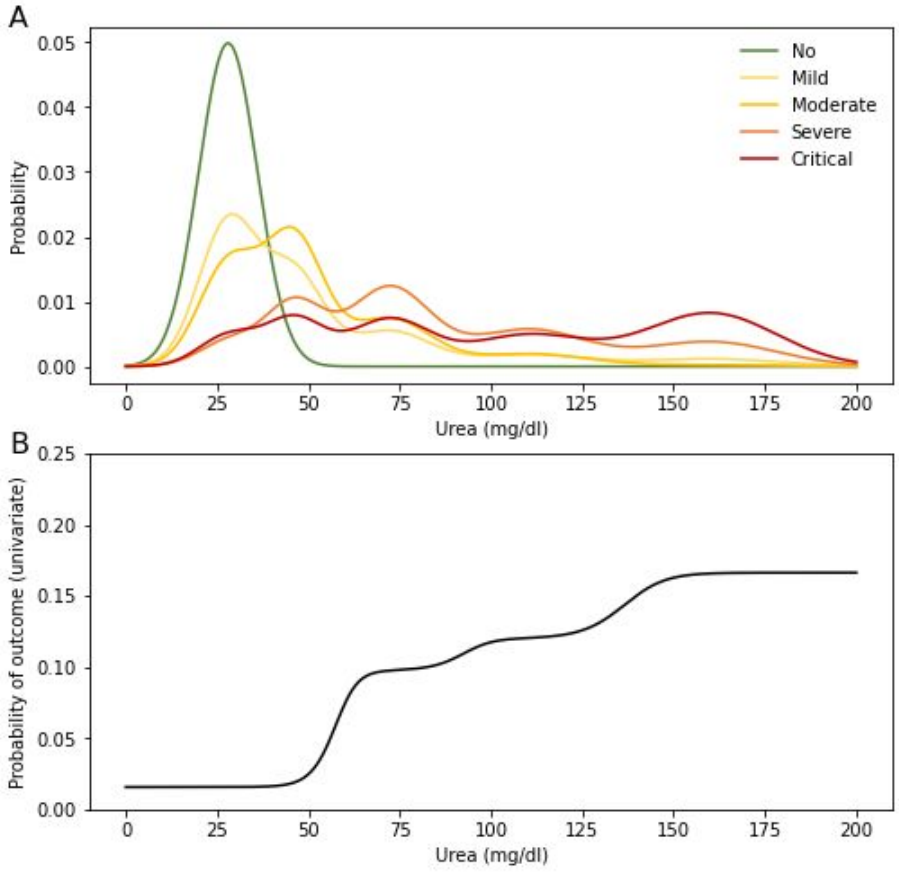
e-Figure 1. Example of continuous variable parameterization using Gaussian Mixture



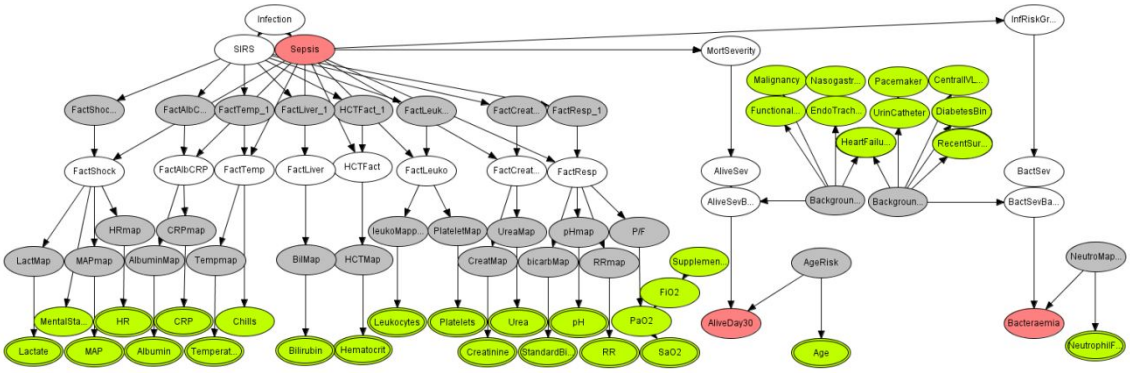
Models

The mapping of severity states across the individual Gaussian components occurs when learning the conditional probability distributions in the model. Observation of a certain value propagates an odds ratio equivalent to the relative probability of each state given the observed value, which is used to adjust the model's *a priori* beliefs. An example of such a mapping is shown for urea in e-Figure 2. This concept is also described in the literature [1,3].

e-Figure 2. Example of composite Gaussian distributions created by mapping across severity states (conditional probabilities learned in the CPN). A: learned distributions B: univariate outcome prediction across the predictor domain



e- Figure 3. Structure of the SeF CPN



Evaluation of overfitting to the derivation dataset

To ensure the model did not significantly overfit to the derivation data, a ten-fold cross-validation was performed to assess the internal performance. The derivation data were shuffled and split into ten cross-validation folds, stratifying by outcome to ensure a similar proportion of outcomes in each fold. Ten models were trained and evaluated, in each case holding out one of the cross-validation folds as a validation set and training using the remaining nine. The ROC AUC was then computed for each model on its respective held-out validation set. The mean cross validation AUC was 0.800 (range: 0.749-0.832). The final model was then learned using the entire derivation cohort. The internal performance of the final model on the derivation cohort was AUC = 0.801 (95% CI 0.777-0.825) which did not differ significantly from the cross-validation performance.

Handling of missing data

The CPN model does not require any explicit handling of missing data. Ability to handle missing data is an inherent property of CPNs. e-Table 1 shows the measurement rates of the CPN input variables for the derivation and validation datasets.

For three of the variables (CRP, platelets, neutrophil fraction) where there was a discrepancy in measurement rates between the cohorts, this was a temporal feature of the derivation data. Prior to 2007, measurement of these parameters was less common. However, the lower availability of these parameters did not adversely affect model performance, as shown in e-Table 2.

e-Table 1. Measurement rates

Measurement rates	Derivation n (%)	Validation n (%)
Temperature	4235 (93.5)	1012 (97.9)
Chills	4388 (96.8)	na
MAP	4241 (93.6)	1000 (96.7)
HR	4184 (92.3)	1032 (99.8)
RR	3907 (86.2)	1016 (98.3)
Mental status	4531 (100.0)	1031 (99.7)
SpO2	3098 (68.4)	992 (95.9)
CRP	3971 (87.6)	1017 (98.4)
Creatinine	4457 (98.4)	1031 (99.7)
Platelets	3388 (74.8)	1027 (99.3)
Neutrophil fraction	3884 (85.7)	1032 (99.8)
Urea	744 (16.4)	1021 (98.7)
pH	3317 (73.2)	668 (64.6)
PaO2	3334 (73.6)	645 (62.4)
FiO2	3167 (69.9)	983 (95.1)
HCO3	2998 (66.2)	653 (63.2)
Leukocytes	4452 (98.3)	1031 (99.7)
Lactate	na	181 (17.5)
Bilirubin	na	253 (24.5)
Hematocrit	na	1031 (99.7)

na: not available

e-Table 2. Impact of missing data

Years	Patients, n (%)	Mortality, n/N (%)	Availability of selected model inputs			AUC 30-day mortality
			CRP	Platelets	Neut.Frac	
2003-2006	1687	90/1679 (5.4)	76.5%	41.6%	67.2%	0.856 [0.824-0.888]
2007-2016	2844	202/2744 (7.4)	94.6%	94.5%	97.0%	0.771 [0.739-0.804]
All	4531	292/4423 (6.6)	87.6%	74.8%	85.7%	0.801 [0.777-0.825]

Abbreviations: CRP indicates C-reactive protein

e-Table 3. Mortality in PSI risk classes and quantile-matched SF risk classes in the validation cohort

PSI Risk Class	Patients, n (%)	30-day Mortality	SeF-ML – matched, 30-day mortality
1	112 (10.9)	1 (0.9)	1 (0.9)
2	185 (18.1)	0 (0.0)	2 (1.1)
3	299 (29.2)	5 (1.7)	4 (1.3)
4	332 (32.4)	16 (4.8)	13 (3.9)
5	96 (9.4)	21 (21.9)	23 (24.0)

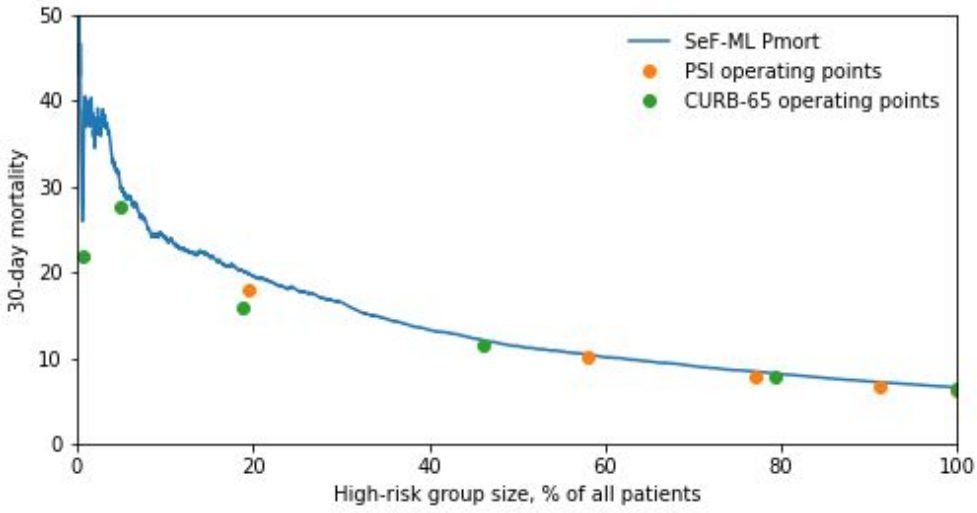
Abbreviations: PSI indicates pneumonia severity index; SeF-ML, SepsisFinder-Machine Learning.

e-Table 4. Mortality in CURB65 risk classes and quantile-matched SF risk classes in the validation cohort

CURB65 Risk Class	Patients, n (%)	30-day Mortality	SeF-ML – matched, 30-day mortality
0	207 (20.0)	2 (1.0)	2 (1.0)
1	282 (27.3)	3 (1.1)	2 (0.7)
2	390 (37.7)	16 (4.1)	12 (3.1)
3	124 (12.0)	15 (12.1)	16 (12.9)
4	29 (2.8)	6 (20.7)	12 (41.4)
5	2 (0.2)	1 (50.0)	0 (0.0)

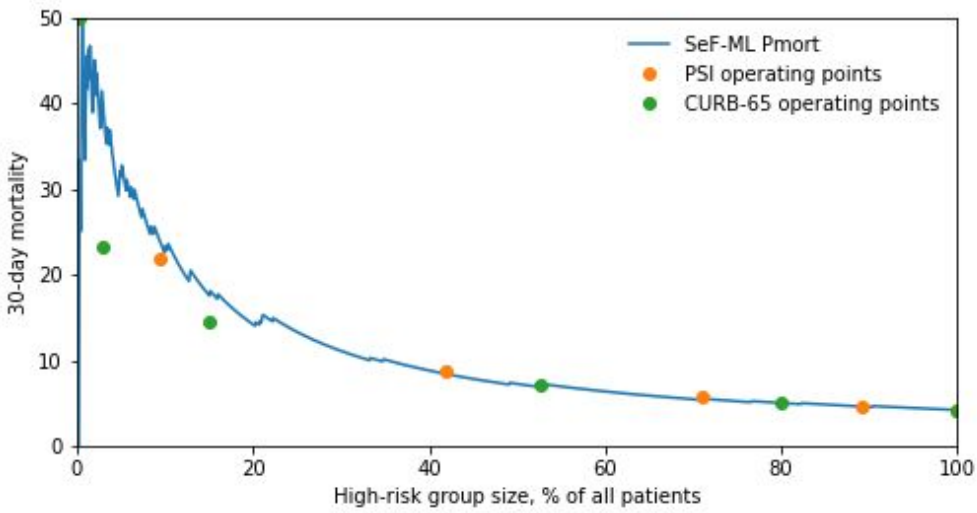
Abbreviations: CURB65 indicates Confusion, Urea, Respiratory Rate, Blood pressure and Age 65; SeF-ML, SepsisFinder-Machine Learning.

e-Figure 4. 30-day mortality as a function of the size of the high-risk group, compared with PSI and CURB-65 operating points in the derivation cohort



The coloured circles show possible operating points for the PSI and CURB-65 scores.

e-Figure 5. 30-day mortality as a function of the size of the high-risk group in the validation cohort



The coloured circles show possible operating points for the PSI and CURB-65 scores.

References.

1. Ward L, Andreassen S. A Bayesian Approach to Model-Development: Automatic Learning for Tuning Predictive Performance. IFAC-PapersOnLine [Internet] 2015;48:481–6. doi: http://linkinghub.elsevier.com/retrieve/pii/S2405896315020789doi: 10.1016/j.ifacol.2015.10.187
2. Ward LM. Gradation of the severity of sepsis - Learning in a causal probabilistic network. 2016;doi: 10.5278/VBN.PHD.MED.00064
3. Ward L, Paul M, Andreassen S. Automatic Learning of mortality in a CPN model of the Systemic Inflammatory Response Syndrome. Math. Biosci. 2017;284:12–20. doi: http://dx.doi.org/10.1016/j.mbs.2016.11.004
4. Ward L, Møller JK, Eliakim-Raz N, Andreassen S. Prediction of Bacteraemia and of 30-day Mortality Among Patients with Suspected Infection using a CPN Model of Systemic Inflammation. IFAC-PapersOnLine [Internet] 2018;51:116–21. doi: https://linkinghub.elsevier.com/retrieve/pii/S2405896318333731doi: 10.1016/j.ifacol.2018.11.657

STROBE Statement—Checklist of items that should be included in reports of ***cohort studies***

	Item No	Recommendation	Page No
Title and abstract	1	(*a*) Indicate the study's design with a commonly used term in the title or the abstract (*b*) Provide in the abstract an informative and balanced summary of what was done and what was found	1-3
Introduction			
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported	4
Objectives	3	State specific objectives, including any prespecified hypotheses	4
Methods			
Study design	4	Present key elements of study design early in the paper	5
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection	5-7
Participants	6	(*a*) Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up (*b*) For matched studies, give matching criteria and number of exposed and unexposed	5-7
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable	5-7
Data sources/ measurement	8*	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group	5-7
Bias	9	Describe any efforts to address potential sources of bias	5-7
Study size	10	Explain how the study size was arrived at	5-7
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why	7
Statistical methods	12	(*a*) Describe all statistical methods, including those used to control for confounding (*b*) Describe any methods used to examine subgroups and interactions (*c*) Explain how missing data were addressed (*d*) If applicable, explain how loss to follow-up was addressed (*e*) Describe any sensitivity analyses	7-11
Results			
Participants	13*	(a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed (b) Give reasons for non-participation at each stage (c) Consider use of a flow diagram	12
Descriptive data	14*	(a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders (b) Indicate number of participants with missing data for each variable of interest (c) Summarise follow-up time (eg, average and total amount)	12-13
Outcome data	15*	Report numbers of outcome events or summary measures over time	14

Main results	16	(*a*) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included (*b*) Report category boundaries when continuous variables were categorized (*c*) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period	7-17
Other analyses	17	Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses	Not performed
Discussion			
Key results	18	Summarise key results with reference to study objectives	18
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias	20
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence	18-21
Generalisability	21	Discuss the generalisability (external validity) of the study results	18-21
Other information			
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based	22

*Give information separately for exposed and unexposed groups.

Note: An Explanation and Elaboration article discusses each checklist item and gives methodological background and published examples of transparent reporting. The STROBE checklist is best used in conjunction with this article (freely available on the Web sites of PLoS Medicine at http://www.plosmedicine.org/, Annals of Internal Medicine at http://www.annals.org/, and Epidemiology at http://www.epidem.com/). Information on the STROBE Initiative is available at http://www.strobe-statement.org.

Word counts abstract: 275
Word counts text: 3006

Machine-Learning Model for Mortality Prediction in Patients with Community-acquired Pneumonia: Development and validation study

Running title: Machine-Learning for Mortality Prediction in CAP

Catia Cilloniz, PhD[1*], Logan Ward, PhD[2*], Mads Lause Mogensen, PhD[2*], Juan M Pericàs, MD, PhD[3,4*], Raúl Méndez, MD,PhD[5], Albert Gabarrús, MSc[1], Miquel Ferrer, MD, PhD[1], Carolina Garcia-Vidal, MD, PhD[6], Rosario Menendez, MD, PhD[5], Antoni Torres, MD, PhD[1.6]

[1]Department of Pneumology, Hospital Clinic of Barcelona; August Pi i Sunyer Biomedical Research Institute–IDIBAPS, University of Barcelona; Biomedical Research Networking Centres in Respiratory Diseases (Ciberes) Barcelona, Spain.

[2] Treat Systems ApS, Aalborg, Denmark

[3] Department of Infectious Diseases, Hospital Clinic of Barcelona, Barcelona, Spain.

[4] Liver Unit, Vall d'Hebron University Hospital, Vall d'Hebron Institute for Research (VHIR), Barcelona, Spain.

[5] Department of Pneumology, Hospital La Fe de Valencia, Valencia, Spain

[6] ICREA, Barcelona, Spain

*Equal contribution

Corresponding author: Prof. Antoni Torres (primary) or Dr. Catia Cillóniz (alternative)

Department of Pulmonary Medicine, Hospital Clinic of Barcelona

C/ Villarroel 170, 08036 Barcelona, Spain

Tel: (+34) 93-227-5779, fax: (+ 34) 93-227-9813

Email: atorres@clinic.cat (primary) or catiacilloniz@yahoo.com (alternative)

Summary conflict of interest statements:

The authors declare that they have no conflicts of interest with the study. Dr Cillóniz is the recipient of the 2018 SEPAR fellowship and a grant from the *Fondo de Investigación Sanitaria* (PI19/00207). RM has received honoraria for lectures from Pfizer.

Funding information:

This study was supported by *CIBER de Enfermedades Respiratorias* (CIBERES CB06/06/0028), and by 2009 Support to Research Groups of Catalonia 911, IDIBAPS. The founders of the study had no role in the study design, data collection, analysis, or interpretation, writing of the report, or decision to submit for publication. SEPAR integrated respiratory infections program.

Notation of prior abstract publication/presentation:

None

Key words list:

Community-acquired pneumonia

Machine learning

Pneumonia

Artificial intelligence

Mortality prediction

Abbreviation list:

ML: machine learning

AI: artificial intelligence

CAP: community-acquired pneumonia

CPN: causal probabilistic network

SeF: SepsisFinder

SeF-ML: SeF adapted model

ATS/IDSA: American Thoracic Society/Infectious Diseases Society of America

SOFA: Sequential Organ Failure Assessment

qSOFA: quick Sequential Organ Failure Assessment

ED: emergency department

MAP: median arterial pressure

SBP: systolic blood pressure

DBP: diastolic blood pressure

CRP: C-reactive protein

ARDS indicates acute respiratory distress syndrome

AUC-ROC: area under the receiver operating characteristic curve

CI: confidence intervals

Q1: first quartile

Q3: third quartile

FiO_2, fraction of inspired oxygen

IDSA/ATS: Infectious Diseases Society of America/American Thoracic Society

PaO_2: partial pressure of arterial oxygen

PSI, pneumonia severity index

Q1: first quartile

Q3: third quartile

Abstract

Background: Artificial intelligence tools and techniques such as machine learning (ML) are increasingly seen as a suitable manner to increase the prediction capacity of currently available clinical tools, including prognostic scores. However, studies evaluating the efficacy of ML methods in enhancing the predictive capacity of existing scores for community-acquired pneumonia (CAP) are limited. We aimed to apply and validate a causal probabilistic network (CPN) model to predict mortality in patients with CAP.

Research question: Is a CPN model able to predict mortality in patients with CAP better than the commonly-used severity scores?

Study design and methods: Derivation-validation retrospective study conducted in two Spanish University hospitals. The ability to predict 30-day mortality of a CPN designed to predict mortality in sepsis (SeF) and adapted for CAP (SeF-ML) was assessed and compared to other scoring systems (PSI, SOFA, qSOFA and CURB-65). The SeF models are proprietary software. Differences between ROC curves were assessed using De Long's method for correlated ROC curves.

Results: The derivation cohort comprised 4,531 patients whilst the validation cohort had 1,034 patients. In the derivation cohort, the AUC of SeF-ML, CURB-65, SOFA, PSI and qSOFA were 0.801, 0.759, 0.671, 0.799 and 0.642, respectively, for 30-day mortality prediction. In the validation study, the AUC of SeF-ML was 0.826, concordantly with the AUC (0.801) in the derivation data (p=0.51). The AUC of SeF-ML was significantly higher than those of CURB-65 (0.764, p=0.03) and qSOFA (0.729, p=0.005). However, it did not differ significantly from PSI (0.830, p=0.92) and SOFA (0.771, p=0.14).

Interpretation: SeF-ML shows potential for improving mortality prediction amongst patients with CAP using structured health data. Additional external validation studies should be conducted to support generalisability.

Introduction

Community-acquired pneumonia (CAP) remains a major cause of morbidity and mortality worldwide, with children and elderly being the most affected population[1,2]. Early-risk stratification of severity and adequate antimicrobial therapy are critical to improving CAP outcomes[1,3]. Scoring systems such as Pneumonia Severity Index (PSI)[4] and CURB-65 score[5] have been widely used to stratify patients with CAP and support clinical decision-making processes[3]. However, these scores have several limitations that hinder their capacity as clinical decision-making supporting tools, e.g., low positive and negative predictive values in predicting the need of intensive care, or underestimated severity in certain age groups.

Artificial intelligence (AI) tools and techniques such as machine learning (ML) are increasingly seen as a suitable manner to increase the prediction capacity of currently available tools in infectious diseases, e.g., sepsis, antimicrobial resistance and COVID-19[6,7]. Within respiratory medicine, the main applications of AI and ML have included the interpretation of thoracic imaging, lung pathology slides and physiologic data such as pulmonary function tests[8]. Nonetheless, studies evaluating the efficacy of ML methods in enhancing the predictive capacity of existing scores for CAP are limited[9,10].

We aimed to apply a causal probabilistic network (CPN) model previously used in sepsis (SeF) [11–13] to predict 30-day mortality in patients with CAP, comparing the accuracy of this model to that of the established clinical scores. In addition, we pursued validating the ML model in CAP using a large cohort of patients with CAP.

Methods

Study design and participants

A derivation-validation retrospective study was performed when using an innovative CPN model[12,13] to predict mortality in adult patients hospitalised with CAP. The rationale for using this new statistical approach is trying to diminish the effect of correlations and make better use of the variables, thus avoiding losing patients with missing information. We applied the guidelines provided by Leisman et al [14] for reporting of prediction models.

Institutional approval was provided by the IRBs of both University Hospital La Fe of Valencia (EC2011/2019) and Hospital Clínic of Barcelona (HCP2009/5451), which waived the need for informed consent due to the retrospective nature of the study.

Data were collected within the first 24 hours of admission. Collected data had similar definitions for both cohorts, and harmonization between cohorts was elementary. All data set were anonymously analyzed, and the study was performed following current recommendation of the Declaration of Helsinki.

Definitions

CAP was diagnosed if new acute respiratory symptoms, signs, and compatible infiltrate(s) on chest x-ray were present. Severe CAP was defined according to the ATS/IDSA guidelines[3]. Prior antibiotic treatment was defined as the intake of antibiotics during the week before hospital admission. The appropriateness of empiric antibiotic treatment was determined according to multidisciplinary guidelines for the management of CAP[15]. Sepsis was defined as the presence of pneumonia and an increase of ≥2 points in the Sequential Organ Failure Assessment (SOFA) score[16]. We also calculated median arterial pressure (MAP) from systolic blood pressure (SBP) and

diastolic blood pressure (DBP) as 1/3*SBP + 2/3*DBP. The ML technique used is based on CPN [11,12,17]. TREAT-Lab (Treat Systems ApS, Aalborg, Denmark) is a medical device software program that provides a risk assessment of patients with suspected infection (proprietary software). The aim of such software is to inform the use of additional or adjunct diagnostics. For example, the clinician (or clinical microbiologist) can use the risk score to identify high-risk patients, i.e., classified as those with high predicted probability of mortality, who may benefit from rapid diagnostics. Conversely, it may be used to identify low-risk patients who only receive standard of care. Customisable risk thresholds can be set for individual clinical installations depending on resources available and local practice requirements. The risk assessment model used within TREAT-Lab is the SeF CPN.

Patient selection, inclusion and exclusion criteria

Derivation cohort

We enrolled all consecutive adult patients with a CAP diagnosis in the Emergency Department of Hospital Clinic of Barcelona between January 2003 and December 2016. We excluded patients with severe immunosuppression due but not limited to human immunodeficiency virus (HIV) infection, active solid or haematologic malignancy treated with chemotherapy, oral corticosteroid treatment with at least 20 mg of prednisone (or equivalent) per day for at least two weeks, and treatment with other immunosuppressive drugs. We also excluded those with active tuberculosis or a confirmed alternative diagnosis.

Validation cohort

All consecutive patients admitted with CAP to Hospital Universitario y Politécnico La Fe (Valencia, Spain) between January 2012 and December 2018 were included. The

inclusion criteria were CAP diagnosis based on a new radiologic infiltrate with at least two compatible clinical symptoms and age ≥18 years. Exclusion criteria were hospital admission within the previous 15 days, immunosuppressive treatments and HIV infection.

Data collection and evaluation

Derivation cohort

Demographic variables, comorbidities, and physiologic parameters were collected at the emergency department within 24 hours of admission. The PSI, CURB-65, qSOFA and SOFA score were calculated at admission[4,5,18]. We recorded whether patients had specific complications, including multilobar infiltration, pleural effusions, acute respiratory distress syndrome (ARDS)[19], septic shock[20] and acute renal failure[21] during hospitalisation. All surviving patients were visited or contacted by telephone within 30 days of discharge; hospital records and the Catalunya Health Department database were reviewed at the 1-year mark. We included all available patients in this analysis. We also calculated MAP from SBP and DBP as 1/3*SBP + 2/3*DBP. We discretised PO2 and FiO2 as required by the model. Finally, we transformed creatinine, C-reactive protein (CRP), lactate, bilirubin and platelets through the natural logarithm.

Validation cohort

Demographic characteristics and comorbidities (diabetes mellitus, respiratory, heart, liver, neurological, and renal diseases) were collected at time of admission. The severity of disease at presentation was assessed with the PSI. Antibiotic treatment before CAP diagnosis in the current episode was recorded. We included all available patients in this analysis.

Outcomes

Primary outcome: death within 30 days of admission.

Statistical analysis

We report the number and percentage of patients for categorical variables and the median and interquartile range for continuous variables (non-normal distribution confirmed by the Kolmogorov-Smirnov test). Categorical variables were compared using the χ^2 test. Continuous variables were compared using the nonparametric Mann-Whitney test.

Derivation cohort

We used an updated version of the published SepsisFinder (SeF) model (see supplementary material for a description of the modelling techniques used including variable selection and parameterization). The main differences in input variables between SeF and the adapted model (SeF-ML) are shown in Table 1.

We adapted the respiratory component of SeF, which comprises acid-base balance (pH, HCO3-), respiratory rate, and oxygen perfusion (measured through PaO2, SaO2 and FiO2 [FiO2 was discretized based on a conversion from oxygen flow rate in L/min. PaO2 was discretized into 8 bins]). We retrained this portion of the model using data included in the study. The remainder of the model was deemed invariant. To evaluate the degree of overfitting to the derivation data, we performed a 10-fold cross-validation, shuffling the data and stratifying to ensure a similar proportion of outcomes in each cross-validation fold. We then retrained the model using the full derivation dataset.

No explicit steps to handle missing data were required. CPNs are inherently tolerant of missing information and are able to perform inference with partial evidence:

combining the model's structure and conditional probability tables with the axioms of probability theory allows the marginal probabilities of all nodes in a CPN to be calculated when only some of the nodes are observed (supplemental material and e-Table 1)[22,23]. We then used the SeF-ML model to calculate the probability of death within 30 days (Pmort) for all patients with >2 input variables recorded.

We assessed predictive performance of SeF and SeF-ML by using the area under the receiver operating characteristic curve (AUC-ROC). As a general rule, the relation between AUC and diagnostic accuracy is as follows: AUC between 0.90 and 1.00 has outstanding discrimination ability; 0.80 - 0.90, excellent; 0.70 - 0.80, acceptable; 0.60 - 0.70, poor, and 0.50 - 0.60 fails to accurately diagnose a certain disease or condition[24,25]. Kaplan-Meier survival curves were also constructed.

In our preliminary analysis of this data, we noted that SeF-ML performance was not negatively impacted by the greater degree of missing data amongst patients included from 2003 to 2006 (e-Table 2).

We compared SeF and SeF-ML performance (Pmort as a predictor of death within 30 days) with other scoring systems (PSI, SOFA, qSOFA and CURB-65). Differences between ROC curves were assessed using De Long's method for correlated ROC curves as implemented in the pROC package of R.

In addition to assessing the model's performance, we computed an example for a potential use-case for TREAT Lab, showing patients being stratified into groups of 40% low-risk, 40% medium-risk and 20% high-risk.

Validation cohort

We used the SeF-ML model (Table 1), where the respiratory components were adapted through learning from 4,531 patients with CAP at the Hospital Clinic of

Barcelona. We used the adapted model to calculate the probability of death within 30 days (Pmort) for all patients.

We assessed predictive performance using the AUC. We compared SeF-ML's performance (Pmort as a predictor of death within 30 days) with other scoring systems (i.e., PSI, CURB-65, SOFA and qSOFA). Differences between ROC curves were assessed using DeLong's method for correlated ROC curves as implemented in the pROC package of R. We also compared mortality for a range of risk cut-offs to assess potential operating points for clinical implementation of risk stratification. Model calibration was assessed using the Hosmer-Lemeshow goodness of fit test and by calculating the Brier-score loss.

More details about the selection of variables and the ML model are displayed in the supplementary material (e-Figure 1, e-Figure 2 and e-Figure 3).

Results

Patients' characteristics

The derivation cohort comprised 4,531 (15% outpatients and 85% inpatients) patients and the validation cohort 1,034 patients. Clinical characteristics of the studied population are shown in Table 2.

Predictive performance of risk assessment

Derivation cohort

We used the SeF-ML model to calculate Pmort for all patients with >2 input variables – 4,500/4,531=99.3% of patients. The AUC for 30-day mortality prediction was 0.801 for SeF-ML, being significantly better than CURB-65 (0.759, p<0.001), SOFA (0.671, p<0.001), and qSOFA (0.642, p<0.001) (Table 3). The mean cross-validation performance, measured by AUC, was 0.800 (range: 0.749-0.832) which did not differ from the AUC for the full dataset. Details of the cross-validation assessment are included in the supplemental material.

PSI was only available for 58% of patients and had an AUC of 0.799; it was not significantly different from SeF-ML (p=0.58). CURB-65 provided a "fair" prediction of mortality, while SeF-ML and PSI provided a "good" prediction of 30-day mortality (Figure 1). The calibration of SeF-ML was measured using the Hosmer-Lemeshow statistic and the Brier-score loss. The Hosmer-Lemeshow statistic was 15.62 (p=0.048) which suggests the model may not be well calibrated. However, the Hosmer-Lemeshow statistic is known to be very sensitive to sample size. The model appears visually well-calibrated. The Brier-score loss for the model was 0.056.

Survival curves are shown in Figure 2 for patients stratified according to Pmort; the low-risk group comprises 40% of patients with the lowest Pmort; the medium-risk

group represents the next highest 40%; and the high-risk group accounts for the top 20% according to Pmort. The additional value provided by SeF-ML with respect to CURB-65 and PSI scores is shown in Figure 3: SeF-ML provides additional discrimination for mortality, e.g., for patients with PSI ≥4.

Validation cohort

For the validation cohort, the AUC for SeF-ML was 0.826. It was not significantly different from performance in the derivation cohort (p=0.51) (Figure 4). According to the Hosmer-Lemeshow statistic, SeF-ML was well-calibrated (HL statistic = 11.93, p=0.15). The Brier-score loss for SeF-ML was 0.036. Notably, only 23.1% patients had complete information to calculate the SOFA score, being not significantly different from SeF-ML (p=0.85). On the contrary, SeF-ML had a significantly higher AUC than both CURB-65 and qSOFA (p=0.03 and p=0.005, respectively) (Table 4).

When analysing the imputed SOFA score (assuming missing components were normal), the AUC of SOFA improved from 0.728 to 0.771; the difference from SeF-ML remained non-significant.

Supplementary e-Table 3 shows the number of patients in each PSI risk class and associated 30-day mortality. Also, 30-day mortality within quantile-matched risk classes for SeF-ML is also shown. SeF-ML risk groups were defined by choosing cut-offs that resulted in the same number of patients assigned to each risk group like in the corresponding PSI risk class.

Potential use of risk assessment

Patients can be stratified by their probability of death within 30 days. An odds ratio (OR) for death can also be calculated for high-risk patients vs. others. The OR represents the degree of separation between high-risk and low-risk patients. For this

analysis, we considered two strategies for stratification: high-risk patients receive rapid diagnostics and low-risk patients receive minimum standard of care. For example, the top 20% had a probability of death within 30 days of 18.4%, whilst OR for the high-risk group vs. remaining patients was 7.6 (e-Figure 4). One advantage of SeF-ML is the continuous probability output that allows custom thresholds to be set depending on the care environment versus the five potential operating points for CURB-65, for example. For comparison, three potential scenarios are shown below: 1) *PSI=5* (N=98) 30-day mortality = 21.9% (PSI), 24% (SeF-ML); OR for high-risk vs. low risk = 11.5 (PSI), 14.3 (SeF-ML); 2) PSI ≥4 (N=428) 30-day mortality = 8.6% (PSI), 8.4% (SeF-ML); OR for high-risk vs. low risk = 9.3 (PSI), 7.7 (SeF-ML); 3) CURB-65 ≥3 (N=153) 30-day mortality = 14.4% (CURB-65), 18.3% (SeF-ML); OR for high-risk vs. low risk = 6.8 (CURB-65), 12.8 (SeF-ML). A fuller picture of the effect of choosing different operating points is shown in Supplementary e-Figure 4 and e-Figure 5. It shows the 30-day mortality as functions of the size of the high-risk group along with operating points for PSI (state) and CURB-65. The smaller the high-risk group, the higher the 30-day mortality.

Discussion

In our study, we applied a ML approach to develop and validate a 30-day mortality prediction clinical model in patients with CAP. Although not pneumonia-specific, the SeF-ML model shows potential for improving mortality prediction amongst patients with CAP. Remarkably, SeF-ML's performance in the validation set matched that of the training data in the derivation cohort, as did those of PSI and CURB-65, whereas the performance of SOFA and qSOFA scores did not match. SeF-ML not only enhances the mortality prediction ability of currently available tools but optimizes the use and quality of available electronic health records (EHR) data. Hence, although the advantages conferred by SeF-ML need further refining and interpretation, this ML model might be applied in clinical practice in the near future, i.e., patients can be stratified by their probability of death and such stratification might then be used to determine a patient's care trajectory and diagnostic workup.

This study is one of the first to use a large cohort of hospitalised patients with CAP to generate a ML model validated through an external cohort of patients with CAP. A barrier for physicians in using ML is its potential "black box" opacity. However, studies like ours show that results obtained with ML predictions are consistent with other severity scores that we have used so far[26,27]. In the near future, ML techniques will allow us to analyse a large volume of data that current techniques cannot do, facilitating the possibility amongst investigators to directly collect data from EHR [7,26–28].

The potential of SeF-ML to improve the current ability of available clinical scores for CAP primarily relies on the findings suggesting that SeF-ML better predicts 30-day mortality than qSOFA and CURB-65 according to our data, which nonetheless require further clinical validation. Compounding this is also the fact that SeF-ML had a higher

AUC than SOFA and PSI, albeit non-significant. Furthermore, the AUC value of SeF-ML was consistent between the predictive and validation model. There is strong evidence supporting the importance of accurately assessing the severity of CAP and stratifying patients based on their mortality risk to improve clinical outcomes[1,3,29]. Despite that, PSI and CURB-65 remain the most widely used CAP severity scores and recommended by international guidelines[3,29,30]. The ability of these scores to predict mortality has some limitations[31]. Moreover, although qSOFA score ≥2 is strongly associated with mortality in patients with pneumonia, its use in early identification of patients with CAP and mortality risk is hindered[32,33]. SeF-ML seems to provide better discriminative capacity to discern between high- and low-risk patients, which is key to adapting the intensity of care and resources per foreseen prognosis. However, this still needs further validation to prove its actual clinical validity. In addition, SeF-ML increases efficiency in exploiting available data. The lower requirements set for minimum data and the ability to handle missing data mean that 99% of patients, on average, would have sufficient data for predictions to be established. This would facilitate optimization of EHR use, with fewer investments on data collection and curation when compared to other scores. Also, the continuous output achieved with SeF-ML allows for adaptive fine-tuning of patient classification. Cut-offs for defining risk can be smoothly and accurately adjusted. In particular, the ability of SeF-ML to identify high-risk patients with low CURB65 and PSI scores, as well as low-risk patients with high CURB65 and PSI scores paramount as these CAP-specific clinical scores are the two most widely used. It is worth noting, however, that enhanced mortality prediction through continuous appraisals is not unique to ML but is also found in other models with continuous

output, e.g. the eCURB model, which achieved greater predictive accuracy using CURB-65 variables using regression splines[34].

The ability of SeF-ML to reclassify risk of patients with CAP might be more beneficial for certain patient subgroups, e.g., those with low CURB-65/PSI yet high SeF and mortality, or those with high CURB-65/PSI yet low SeF and mortality.

A foremost strength of our study is the use of a large set of clinical data that are representative of the population in a defined area of influence. This study feature enabled the integration of surveillance data into direct clinical care of individual patients and can be helpful in making decisions by applying ML models like SeF-ML. When developing and validating the algorithm, we used two large patient slices to ensure that our model can be implemented using real-time patient data.

However, some limitations need to be highlighted. This score validation against mortality is only the first step toward clinical utility. Remarkably, the ability of SeF-ML to predict ICU admission remains unproven. In particular, our study lacked information on post-admission disposition, and therefore we were not able to test SeF-ML against potentially different clinical pathways other than direct admission to the ICU. Furthermore, the use of closed databases does not incorporate new information and therefore does not allow "learning" of ML models. Besides, despite the large sample sizes of both cohorts, some baseline variables and clinical features, including CAP severity and outcomes, differed. However, this does not hamper the training and validation of the CPN model. There are components of the current SeF that were not available in the dataset due to them not generally being measured in the ED for CAP. However, SeF performs well despite not including these variables, which helps to demonstrate its robustness to missing values.

Interpretation

SeF-ML performance at predicting 30-day mortality appears to be overall superior than that of existing CAP-specific scores, with the exception of PSI in the validation model. SeF-ML offers some advantages over current scoring methods, eg., calculations easily made on routinely collected data and based upon; structured EHR data (vs. subjective criteria and arterial blood gas sampling needed for PSI for instance); and tunable performance so as to allow risk cut-offs to be tailored to workflow requirements and capacities of the individual institution (compared with fewer states in CURB-65). In addition, SeF-ML performance seems to not be dependent upon data availability, therefore allowing for more effective calculation of risk scores for CAP based upon data sources with limited access to or completeness of certain variables. Our findings need further validation in other cohorts from different settings to assess the actual clinical utility of SeF-ML in predicting CAP prognosis.

Acknowledgements

Guarantor:

Dr. Antoni Torres is the guarantor of the article

Author contributions:

Conception and design of the study: CC, LW, ML, MF, JP, RM, RM, AT. Data acquisition: all authors. Data analysis and interpretation: LW, ML, CC, AG. Manuscript drafting or critical revision for important intellectual content: all authors. Final approval of the submitted version: all authors.

Financial/nonfinancial disclosures:

The authors have reported to Chest the following: This study was supported by *CIBER de Enfermedades Respiratorias* (CIBERES CB06/06/0028), and by 2009 Support to Research Groups of Catalonia 911, IDIBAPS. The founders of the study had no role in the study design, data collection, analysis, or interpretation, writing of the report, or decision to submit for publication. SEPAR integrated respiratory infections program.

Role of sponsors:

The funders did not have a role in the conduct of the study; in the collection, management, analysis, or interpretation of data; or in the preparation of the manuscript.

References

1. Torres A, Cilloniz C, Niederman MS, et al. Pneumonia. *Nature Reviews Disease Primers* 2021;7(1):1–28.

2. GBD 2019 Diseases and Injuries Collaborators. Global burden of 369 diseases and injuries in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet* 2020;396(10258):1204–1222.

3. Metlay JP, Waterer GW, Long AC, et al. Diagnosis and Treatment of Adults with Community-acquired Pneumonia. An Official Clinical Practice Guideline of the American Thoracic Society and Infectious Diseases Society of America. *Am J Respir Crit Care Med* 2019;200(7):e45–e67.

4. Fine MJ, Auble TE, Yealy DM, et al. A prediction rule to identify low-risk patients with community-acquired pneumonia. *N Engl J Med* 1997;336(4):243–250.

5. Lim WS, Lewis S, Macfarlane JT. Severity prediction rules in community acquired pneumonia: a validation study. *Thorax* 2000;55(3):219–223.

6. Peiffer-Smadja N, Rawson TM, Ahmad R, et al. Machine learning for clinical decision support in infectious diseases: a narrative review of current applications. *Clin Microbiol Infect* 2020;26(5):584–595.

7. Garcia-Vidal C, Sanjuan G, Puerta-Alcalde P, Moreno-García E, Soriano A. Artificial intelligence to support clinical decision-making processes. *EBioMedicine* 2019;46:27–29.

8. Gonem S, Janssens W, Das N, Topalovic M. Applications of artificial intelligence and machine learning in respiratory medicine. *Thorax* 2020;75(8):695–701.

9. Kang SY, Cha WC, Yoo J, et al. Predicting 30-day mortality of patients with pneumonia in an emergency department setting using machine-learning models. *Clin Exp Emerg Med* 2020;7(3):197–205.

10. Chumbita M, Cillóniz C, Puerta-Alcalde P, et al. Can Artificial Intelligence Improve the Management of Pneumonia. *J Clin Med* 2020;9(1):248.

11. Ward L, Andreassen S, Astrup JJ, Rahmani Z, Fantini M, Sambri V. Clinical- vs. model-based selection of patients suspected of sepsis for direct-from-blood rapid diagnostics in the emergency department: a retrospective study. *Eur J Clin Microbiol Infect Dis* 2019;38(8):1515–1522.

12. Ward L, Møller JK, Eliakim-Raz N, Andreassen S. Prediction of Bacteraemia and of 30-day Mortality Among Patients with Suspected Infection using a CPN Model of Systemic Inflammation. *IFAC-PapersOnLine* 2018;116–121.

13. Andreassen S, Møller JK, Eliakim-Raz N, Lisby G, Ward L. A comparison of predictors for mortality and bacteraemia in patients suspected of infection. *BMC Infect Dis* 2021;21(1):864.

14. Leisman DE, Harhay MO, Lederer DJ, et al. Development and Reporting of Prediction Models: Guidance for Authors From Editors of Respiratory, Sleep, and Critical Care Journals. *Crit Care Med* 2020;48(5):623–633.

15. Torres A, Barberán J, Falguera M, et al. [Multidisciplinary guidelines for the management of community-acquired pneumonia]. *Med Clin (Barc)* 2013;140(5):223.e1-223.e19.

16. Singer M, Deutschman CS, Seymour CW, et al. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA* 2016;315(8):801–810.

17. Logan Morgan Ward. Gradation of the Severity of Sepsis: Learning in a Causal Probabilistic Network. 2016;55.

18. Vincent JL, Moreno R, Takala J, et al. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. On behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine. *Intensive Care Med* 1996;22(7):707–710.

19. ARDS Definition Task Force, Ranieri VM, Rubenfeld GD, et al. Acute respiratory distress syndrome: the Berlin Definition. *JAMA* 2012;307(23):2526–2533.

20. Rhodes A, Evans LE, Alhazzani W, et al. Surviving Sepsis Campaign: International Guidelines for Management of Sepsis and Septic Shock: 2016. *Intensive Care Med* 2017;43(3):304–377.

21. Bellomo R, Ronco C, Kellum JA, Mehta RL, Palevsky P, Acute Dialysis Quality Initiative workgroup. Acute renal failure - definition, outcome measures, animal models, fluid therapy and information technology needs: the Second International Consensus Conference of the Acute Dialysis Quality Initiative (ADQI) Group. *Crit Care* 2004;8(4):R204-212.

22. Steen Andreassen, Leonard Leibovici, Mical Paul, Anders D. Nielsen, Alina Zalounina, Leif E. Kristensen, Karsten Falborg, Brian Kristensen, Uwe Frank, Henrik C. Schønheyder. A Probabilistic Network for Fusion of Data and Knowledge in Clinical Microbiology [Internet]. springerprofessional.de. 2005 [cited 2022 Mar 22];Available from: https://www.springerprofessional.de/en/a-probabilistic-network-for-fusion-of-data-and-knowledge-in-clin/1020738

23. Lauritzen SL, Spiegelhalter DJ. Local Computations with Probabilities on Graphical Structures and Their Application to Expert Systems. *Journal of the Royal Statistical Society: Series B (Methodological)* 1988;50(2):157–194.

24. Cole TJ. Applied logistic regression. D. W. Hosmer and S. Lemeshow, Wiley, New York, 1989. *Statistics in Medicine* 1989;10(7):1162–1163.

25. Altman DG, Bland JM. Diagnostic tests 3: receiver operating characteristic plots. *BMJ* 1994;309(6948):188.

26. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. *Lancet* 2019;393(10181):1577–1579.

27. Mekov E, Miravitlles M, Petkov R. Artificial intelligence and machine learning in respiratory medicine. *Expert Rev Respir Med* 2020;14(6):559–564.

28. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. *Nat Med* 2019;25(1):30–36.

29. Menéndez R, Cilloniz C, España PP, et al. Community-Acquired Pneumonia. Spanish Society of Pulmonology and Thoracic Surgery (SEPAR) Guidelines. 2020 Update. *Arch Bronconeumol* 2020;56 Suppl 1:1–10.

30. Lim WS, Baudouin SV, George RC, et al. BTS guidelines for the management of community acquired pneumonia in adults: update 2009. *Thorax* 2009;64 Suppl 3:iii1-55.

31. Loke YK, Kwok CS, Niruban A, Myint PK. Value of severity scales in predicting mortality from community-acquired pneumonia: systematic review and meta-analysis. *Thorax* 2010;65(10):884–890.

32. Jiang J, Yang J, Jin Y, Cao J, Lu Y. Role of qSOFA in predicting mortality of pneumonia: A systematic review and meta-analysis. *Medicine (Baltimore)* 2018;97(40):e12634.

33. Ranzani OT, Prina E, Menéndez R, et al. New Sepsis Definition (Sepsis-3) and Community-acquired Pneumonia Mortality. A Validation and Clinical Decision-Making Study. *Am J Respir Crit Care Med* 2017;196(10):1287–1297.

34. Jones BE, Jones J, Bewick T, et al. CURB-65 pneumonia severity assessment adapted for electronic decision support. *Chest* 2011;140(1):156–163.

528 **Table 1. Variables used in scoring systems**

Variable	SeF	SeF-ML	PSI	SOFA	qSOFA	CURB-65
Demography						
Age	Yes	Yes	Yes			Yes
Sex			Yes			
Nursing home resident			Yes			
Comorbidities						
Neoplastic disease			Yes			
Liver disease			Yes			
Congestive heart failure			Yes			
Cerebrovascular disease			Yes			
Renal disease			Yes			
Vitals						
Temperature	Yes	Yes	Yes			
Chills	Yes	Yes				
Systolic Blood Pressure			Yes		Yes	Yes
Diastolic Blood Pressure						Yes
Mean Arterial Pressure	Yes	Yes		Yes		
Heart Rate	Yes	Yes	Yes			
Respiratory Rate		Yes	Yes		Yes	Yes
Mental Status	Yes	Yes	Yes	Yes, GCS	Yes	Yes
SpO2		Yes				
Labs						
CRP	Yes	Yes				
Creatinine	Yes	Yes		Yes		
Albumin	Yes*	Yes*				
Lactate	Yes*	Yes*				
Platelets	Yes	Yes		Yes		
Neutrophils%	Yes	Yes				
Bilirubin		Yes		Yes		
BUN or Urea		Yes	Yes			Yes
Sodium			Yes			
Glucose			Yes			
pH		Yes	Yes			

Variable	SeF	SeF-ML	PSI	SOFA	qSOFA	CURB-65
Hematocrit		Yes	Yes			
PaO2		Yes	Yes	Yes		
FiO2		Yes		Yes		
Bicarbonate		Yes				
Leukocytes		Yes				
Radiology						
Chest x-ray			Yes			
Treatment/devices						
Mechanical ventilation				Yes		
Pressors				Yes		

Abbreviations: Glasgow coma scale (GCS); * not part of standard testing for CAP patients at the emergency department (ED)

534 **Table 2. Clinical characteristics of the studied cohorts**

Variable	Derivation cohort (N = 4,531)	Validation cohort (N = 1,034)	P-value
Age, median (Q1; Q3), years	73 (55; 82)	72 (59; 81)	0.650
Male sex, n (%)	2,708 (60)	642 (62)	0.176
Previous antibiotic, n (%)	1,057 (25)	338 (33)	<0.001
Influenza vaccine, n (%)	1,770 (44)	438 (42)	0.350
Pneumococcal vaccine, n (%)	695 (17)	79 (8)	<0.001
Previous inhaled corticosteroids, n (%)	818 (18)	-	-
Previous systemic corticosteroids, n (%)	180 (4)	-	-
Previous episode of pneumonia (last year), n (%)	630 (14)	-	-
Comorbidities, n (%) [a]	2,982 (66)	731 (71)	0.005
Chronic respiratory disease	1,729 (39)	325 (31)	<0.001
Chronic cardiovascular disease	635 (14)	347 (34)	<0.001
Diabetes mellitus	914 (20)	288 (28)	<0.001
Neurologic disease	836 (19)	176 (17)	0.137
Chronic renal disease	318 (7)	139 (13)	<0.001
Chronic liver disease	212 (5)	38 (4)	0.146
Nursing home, n (%)	352 (8)	43 (4)	<0.001
Confusion, n (%)	820 (18)	77 (7)	<0.001
Respiratory rate, median (Q1; Q3), breaths/min	24 (20; 30)	18 (16; 24)	<0.001
Heart rate, median (Q1; Q3), beats/min	97 (84; 110)	97 (85; 110)	0.685
Systolic blood pressure, median (Q1; Q3), mmHg	129 (112; 148)	134 (118; 152)	<0.001
Diastolic blood pressure, median (Q1; Q3), mmHg	72 (64; 80)	70 (61; 80)	0.007
Creatinine, median (Q1; Q3), mg/dL	1.0 (0.8; 1.4)	1.0 (0.8; 1.4)	0.003
Glucose level, median (Q1; Q3), mg/dL	124 (105; 157)	-	-

Variable	Derivation cohort (N = 4,531)	Validation cohort (N = 1,034)	P-value
PaO_2/FiO_2, median (Q1; Q3)	281 (238; 327)	271 (238; 311)	0.012
PSI score, median (Q1; Q3)	98 (74; 123)	86 (66; 105)	<0.001
Severe CAP, n (%)	868 (26)	144 (14)	<0.001
Bacteraemia, n (%) [b]	390 (12)	54 (8)	0.001
Appropriate empiric treatment, n (%)	2,844 (96)	-	-
Length of hospital stay, median (Q1; Q3), days	7 (4; 11)	6 (5; 9)	0.261
Mechanical ventilation, n (%)	375 (10)	26 (3)	<0.001
Non-invasive	169 (4)	-	-
Invasive	206 (5) [c]	-	-
In-hospital mortality, n (%)	272 (6)	35 (3)	<0.001
30-day mortality, n (%)	293 (7)	43 (4)	0.004

Abbreviations: CAP indicates community-acquired pneumonia; PSI, pneumonia severity index; Q1, first quartile; Q3, third quartile. Percentages calculated on non-missing data. [a] May have >1 comorbid condition. [b] Calculated only for patients with blood samples (3,206 in the derivation cohort and 696 in the validation cohort). [c] Patients who initially received non-invasive ventilation yet needed intubation subsequently were included in the invasive mechanical ventilation group.

Table 3. Discriminatory ability of SeF-ML vs. other scoring systems for 30-day mortality in the derivation cohort

Model/Score	Availability	AUC (95% CI)	P-value (vs. SeF-ML)
SeF-ML	99.3%	0.801 (0.777-0.825)	-
(limited to those where PSI was available)		0.808 (0.776-0.839)	-
(limited to those where SOFA was available)		0.804 (0.780-0.829)	-
(limited to those where CURB-65 was available)		0.814 (0.788-0.839)	-
PSI (raw score)	57.7%	0.799 (0.768-0.830)	0.58
SOFA	91.4%	0.671 (0.638-0.704)	<0.001
qSOFA [a]	100%	0.642 (0.611-0.673)	<0.001
CURB-65	82.5%	0.759 (0.732-0.786)	<0.001

Abbreviations: AUC indicates area under the receiver operating characteristic curve; CI, confidence interval; PSI, pneumonia severity index. [a] Calculated assuming missing values were normal.

Table 4. Discriminatory ability of SeF-ML vs. other scoring systems for 30-day mortality in the validation cohort

Model/Score	Availability	AUC (95% CI)	P-value (vs. SeF-ML)
SeF-ML	100%	0.826 (0.753-0.899)	-
PSI (raw score)	100%	0.830 (0.753-0.90)	0.92
CURB-65	100%	0.764 (0.694-0.834)	0.03
SOFA	23.1%	0.728 (0.588-0.869)	0.85 [a]
SOFA-imputed [b]	100%	0.771 (0.706-0.836)	0.14
qSOFA	98.3%	0.729 (0.653-0.804)	0.005 [a]

Abbreviations: AUC indicates area under the receiver operating characteristic curve; CI, confidence interval; PSI, pneumonia severity index. [a] Compared only for patients with complete SOFA/qSOFA score. [b] Calculated assuming missing values were normal.

Figure Legends

Figure 1. Model performance for 30-day mortality prediction in the derivation cohort. A: ROC curves B: Calibration curves for SeF models

Figure 2. Survival curves for patients stratified according to Pmort in the derivation cohort

Figure 3. Survival curves stratified by SF risk group, set together according to CURB-65/PSI score in the derivation cohort

Figure 4. Model performance for 30-day mortality prediction in the validation cohort. A: ROC curves B: Calibration curves for SeF models

Take-home Points:

Study Questions: Did a CPN model could predict mortality in patients with CAP better than the frequent severity scores?

Results: SeF-ML performance at predicting 30-day mortality was overall significantly better than that of existing CAP-specific scores.

Interpretation: Our results showed that SeF-ML shows potential for improving mortality prediction amongst patients with CAP using structured health data.