Comorbidities and medical history essential for mortality prediction in critically ill patients

See Articles page e78

Critically ill patients are a highly heterogenous population who tend to have many comorbidities. Often, patients admitted to intensive-care units (ICUs) with the same diagnosis and similar risk profiles according to available risk prediction scores have completely different clinical trajectories and outcomes. Even with increasingly large amounts of electronic health record data available, including clinical notes, vital sign measurements, laboratory data, and imaging data, the goal of unravelling complex disease mechanisms to better forecast patient outcomes remains largely unattained in critical care.[1]

Motivated by this problem, in *The Lancet Digital Health*, Annelaura Nielsen and colleagues[2] present the results of an innovative, exploratory analysis predicting in-hospital, 30-day, and 90-day mortality on the basis of a large and uniquely detailed cohort of patients in ICUs. In addition to laboratory data and other clinical parameters obtained during the first 24 h of an ICU stay for more than 10 000 patients, this dataset also included detailed, 10-year medical histories before ICU admission for more than 230 000 individuals. Factors present before ICU admission, such as comorbidities and medical history, have long been known to affect the risk of future complications or chance of survival.[3] However, even previous machine learning efforts that included broad health record data paid insufficient attention to these factors,[1,4] and Nielsen and colleagues' study is the first to link detailed medical history data from a highly heterogeneous patient population to clinical parameters measured during ICU stays. Remarkably, the authors concluded that a simple feed-forward neural network model including only age, sex, and patients' previous 10-year disease history performed similarly (in terms of prediction of mortality risk) to the two most commonly used ICU risk scores (the Simplified Acute Physiology Score II and the Acute Physiologic Assessment and Chronic Health Evaluation II), and that the combination of medical history and comorbidities with high-frequency ICU data outperformed both scores (Matthews correlation coefficient 0·391 for in-hospital mortality *vs* 0·347 with the Simplified Acute Physiology Score II and 0·300 with the Acute Physiologic Assessment and Chronic Health Evaluation II).

Medical history and comorbidity data are very important for predictions of survival in patients in critical care—and especially for efforts to increase the applicability of these models in clinical and research settings. Risk prediction can inform decisions, but an ideal decision support system would need to be dynamic and informative. The Artificial Intelligence Clinician, an algorithm that generates actual treatment decisions or suggestions, is an example of what decision support systems in the ICU could be.[5] When the algorithm successfully decreased sepsis-related mortality in an independent cohort in silico,[5] debate was sparked about the steps that should be taken to enable similar reinforcement learning models to be applied clinically.[6,7] Reinforcement learning models are developed on the basis of historical data for previous decisions made by clinicians.[5–7] Therefore, to generate good treatment decisions, all data used in clinicians' decision-making processes should be included to prevent confounding.[7] Additionally, after beneficial decisions are generated, any clinical application should be preceded by a clear mapping of the causal links that help clinicians to interpret the reasoning behind the decision. Both complete data collection and the identification of these causal links are notoriously difficult when observational data are used, because these data are often initially collected for a different purpose (ie, research or clinical). Therefore, some data used in clinicians' decision making might be missing, either because they were hard to identify or even unmeasurable, or because they were simply not included in the analysis despite being obtainable, making the dataset inappropriate for adjustment.[7–9]

Extensive patient characterisation is essential to maximise data quality, and, subsequently, the methodological correctness and clinical utility of machine learning models. However, even if all possibly relevant data were identified to minimise confounding, definition of outcomes of interest and strategies to gather these data prospectively can be equally challenging. As Gottesman and colleagues emphasised, focusing on short-term outcomes remains a challenge even for topics that are already broadly studied in critical care.[7] Studies focused on long-term outcomes,

however, face a different type of problem. The difficulty with defining short-term targets for critical illnesses stems from the intricacy of the pathophysiology of these illnesses, which is undoubtedly a major issue in critical care, but one that will probably be solved with further research.[10] Long-term outcomes are different in that they relate to the less obvious core goal of critical care: healthy recovery after an acute ICU admission. Gathering the data necessary for research focused on long-term outcomes will require changes to data collection strategies and infrastructure, including closer collaboration between clinicians, researchers, and data scientists, and national medical data registries (appendix).

See Online for appendix

Overall, Nielsen and colleagues provide captivating evidence for the inclusion of comorbidities and medical history in mortality prediction models for ICU patients.[3] However, their findings also contribute to a broader debate that extends beyond predictive modelling, which was prompted by advances in machine learning research in critical care, and increasing awareness of heterogeneity in treatment response and issues with long-term patient-centered outcomes. It is clear that interpretability and trustworthiness need to be achieved before decision support systems for prediction and decision policy recommendations can be applied in clinical contexts. The causal links between predictions, outcomes, and automated policy recommendations will have to be studied further, starting with the exploration of detailed comorbidity and medical history data.

*José Castela Forte, *Iwan C C van der Horst*
Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence, University of Groningen, Groningen, Netherlands (JCF); and Department of Critical Care, University Medical Center Groningen, Groningen, Netherlands (JCF, ICCvdH)
i.c.c.van.der.horst@umcg.nl

We declare no competing interests.

1 Marafino BJ, Park M, Davies JM, et al. Validation of prediction models for critical care outcomes using natural language processing of electronic health record data. *JAMA Netw Open* 2018; **1:** e185097.
2 Nielsen AB, Thorsen-Meyer H-C, Belling K, et al. Survival prediction in intensive-care units based on aggregation of long-term disease history and acute physiology: a retrospective study of the Danish National Patient Registry and electronic patient records. *Lancet Digital Health* 2019; **1:** e78–89.
3 Beck MK, Jensen AB, Nielsen AB, Perner A, Moseley PL, Brunak S. Diagnosis trajectories of prior multi-morbidity predict sepsis mortality. *Sci Rep* 2016; **6:** 36624.
4 Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. *NPJ Digit Med* 2018; **1:** 1609.
5 Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. *Nat Med* 2018; **24:** 1716–20.
6 Jeter R, Josef C, Shashikumar S, Nemati S. Does the "Artificial Intelligence Clinician" learn optimal treatment strategies for sepsis in intensive care? *arXiv* 2019; published online Feb 8. DOI:1902.03271 (preprint).
7 Gottesman O, Johannson F, Meier J, et al. Evaluating reinforcement learning algorithms in observational health settings. *arXiv* 2018; published online May 31. DOI:1805.12298 (preprint).
8 Bareinboim E, Pearl J. Causal inference and the data-fusion problem. *PNAS* 2016; **113:** 7345–52.
9 Pearl J. The seven tools of causal inference, with reflections on machine learning. *Commun ACM* 2019; **62:** 54–60.
10 Hernández G, Ospina-Tascón GA, Damiani LP, et al. Effect of a resuscitation strategy targeting peripheral perfusion status *vs* serum lactate levels on 28-day mortality among patients with septic shock: the ANDROMEDA-SHOCK randomized clinical trial. *JAMA* 2019; **321:** 654–64.

Survival prediction in intensive-care units based on aggregation of long-term disease history and acute physiology: a retrospective study of the Danish National Patient Registry and electronic patient records

Annelaura B Nielsen, Hans-Christian Thorsen-Meyer, Kirstine Belling, Anna P Nielsen, Cecilia E Thomas, Piotr J Chmura, Mette Lademann, Pope L Moseley, Marc Heimann, Lars Dybdahl, Lasse Spangsege, Patrick Hulsen, Anders Perner, Søren Brunak

Summary

Background Intensive-care units (ICUs) treat the most critically ill patients, which is complicated by the heterogeneity of the diseases that they encounter. Severity scores based mainly on acute physiology measures collected at ICU admission are used to predict mortality, but are non-specific, and predictions for individual patients can be inaccurate. We investigated whether inclusion of long-term disease history before ICU admission improves mortality predictions.

Methods Registry data for long-term disease histories for more than 230 000 Danish ICU patients were used in a neural network to develop an ICU mortality prediction model. Long-term disease histories and acute physiology measures were aggregated to predict mortality risk for patients for whom both registry and ICU electronic patient record data were available. We compared mortality predictions with admission scores on the Simplified Acute Physiology Score (SAPS) II, the Acute Physiologic Assessment and Chronic Health Evaluation (APACHE) II, and the best available multimorbidity score, the Multimorbidity Index. An external validation set from an additional hospital was acquired after model construction to confirm the validity of our model. During initial model development data were split into a training set (85%) and an independent test set (15%), and a five-fold cross-validation was done during training to avoid overfitting. Neural networks were trained for datasets with disease history of 1 month, 3 months, 6 months, 1 year, 2·5 years, 5 years, 7·5 years, 10 years, and 23 years before ICU admission.

Findings Mortality predictions with a model based solely on disease history outperformed the Multimorbidity Index (Matthews correlation coefficient 0·265 *vs* 0·065), and performed similarly to SAPS II and APACHE II (Matthews correlation coefficient with disease history, age, and sex 0·326 *vs* 0·347 and 0·300 for SAPS II and APACHE II, respectively). Diagnoses up to 10 years before ICU admission affected current mortality prediction. Aggregation of previous disease history and acute physiology measures in a neural network yielded the most precise predictions of in-hospital mortality (Matthews correlation coefficient 0·391 for in-hospital mortality compared with 0·347 with SAPS II and 0·300 with APACHE II). These results for the aggregated model were validated in an external independent dataset of 1528 patients (Matthews correlation coefficient for prediction of in-hospital mortality 0·341).

Interpretation Longitudinal disease-spectrum-wide data available before ICU admission are useful for mortality prediction. Disease history can be used to differentiate mortality risk between patients with similar vital signs with more precision than SAPS II and APACHE II scores. Machine learning models can be deconvoluted to generate novel understandings of how ICU patient features from long-term and short-term events interact with each other. Explainable machine learning models are key in clinical settings, and our results emphasise how to progress towards the transformation of advanced models into actionable, transparent, and trustworthy clinical tools.

Funding Novo Nordisk Foundation and Innovation Fund Denmark.

Lancet Digital Health 2019; 1: e78–89

See Comment page e48

Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark (A B Nielsen PhD, H-C Thorsen-Meyer MD, K Belling PhD, A P Nielsen MD, C E Thomas PhD, P J Chmura MSc, M Lademann PhD, Prof P L Moseley MD, Prof S Brunak PhD); **Department of Intensive Care, Rigshospitalet, Copenhagen University Hospital, Copenhagen, Denmark** (H-C Thorsen-Meyer, Prof A Perner PhD); **Centre for IT, Medical Technology and Telephony Services, Capital Region of Denmark, Copenhagen, Denmark** (M Heimann MSc); **and Daintel, Lyngby, Denmark** (L Dybdahl MSc, L Spangsege MSc, P Hulsen BSc)

Correspondence to:
Prof Søren Brunak, Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, DK-2200 Copenhagen, Denmark
soren.brunak@cpr.ku.dk

Introduction

Intensive-care units (ICUs) handle patients from all medical and surgical specialties. Therefore, their populations are highly heterogeneous, and consist of mainly elderly patients who often have a long history of disease. Prediction of prognosis to inform decision making in the ICU is difficult because of the severity of patients' current illness and their disease history.[1]

Mortality risk estimates based on acute physiology scores—such as the Simplified Acute Physiology Score (SAPS) and the Acute Physiologic Assessment and Chronic Health Evaluation (APACHE)—are sometimes used in clinical practice to assess disease severity.[2,3] They are based on logistic regression of specific markers of patient physiology that are recorded during the first hours after ICU admission. In the past 10 years, advanced

Research in context

Evidence before this study

Intensive-care units (ICUs) treat highly heterogeneous patients at high risk of mortality. The heterogeneity of the patient population complicates treatment. Linear models that predict patient outcome on the basis of acute physiology measurements are used in clinical practice to support decision making. We based our search for ICU mortality prediction models on a comprehensive 2016 review by Johnson and colleagues. We searched Google Scholar for studies citing the Medical Information Mart for Intensive Care database, the largest online ICU database, that were published in English between Dec 1, 2011 (the date of the first Medical Information Mart for Intensive Care publication), and Aug 1, 2018. In many studies published in the past 10 years, advanced modelling and machine learning techniques, some of which included text-mined features from electronic patient records, have shown promise for improving prediction of prognosis in the ICU based on acute physiology. However, we identified no studies or available severity scoring systems (eg, the Simplified Acute Physiology Score, the Acute Physiologic Assessment and Chronic Health Evaluation) in which long-term disease history before ICU admission was used to predict ICU mortality, even though ICU patients often have long disease histories. In the past 5 years, we have done several disease-spectrum-wide studies showing that factors such as total disease burden, time between diagnoses, and the sequential order of diagnoses affect the risk of future complications. Therefore, the aggregation of long-term disease history and acute physiology measures in a machine learning model could potentially provide a more accurate outcome prediction model to support decision making in the ICU.

Added value of this study

Long-term disease history is seldom systematically exploited in clinical settings, and is rarely used to inform clinical decision making. To our knowledge, our study is the first in which machine learning was used to predict mortality for ICU patients on the basis of long-term disease history. We used disease histories from more than 230 000 ICU patients with up to 23 years of available data before ICU admission stored in a national disease registry. These data were aggregated with acute physiology measurements obtained from electronic patient records. Models based on previous disease history could predict mortality as well as the clinical severity scores in use. Aggregation of long-term disease history and acute physiology measures in a neural network showed that health-related events from different timepoints in a patient's life interact in a non-linear manner, and that taking account of these interactions between long-term and short-term disease history gave more precise prognostic estimates than either long-term disease history or short-term disease history individually.

Implications of all the available evidence

Our study shows the importance of previous disease history in predictions of mortality in ICU patients and thus, the importance of these data in clinical decision making. The predictive value of long-term disease history was stable over time compared with that of physiology measures, is independent of ICU care, and can be made available at admission to the ICU.

modelling and machine learning techniques have shown promising results in improving prediction of ICU prognosis based on acute physiology.[4]

The extent to which disease events before ICU admission affect prognosis has been debated.[1,5] An increasing number of chronic comorbidity categories have been included in SAPS and APACHE as they have been updated over the years (from three in SAPS II to seven in SAPS III, and from five in APACHE II to seven in APACHE IV), yet still only a small number of comorbidities are included. The predictive value of comorbidity categories as defined by Charlson and colleagues[6] (ten diagnoses) and Elixhauser and colleagues[7] (30 diagnoses) has been studied repeatedly in relation to prediction of mortality in the ICU. In several studies,[8–10] Charlson comorbidity categories had better predictive power than the APACHE II comorbidity categories and similar predictive power to the SAPS II, SAPS III, and APACHE IV scores.[8–10] In another study,[11] comorbidity categories did not contribute significantly to the discrimination of the APACHE II score, and replacement of APACHE II comorbidities with Charlson or Elixhauser comorbidities did not improve discriminatory ability. Two studies[12,13] done in the same homogeneous population of predominantly male, American, military veteran, medical ICU patients showed that the Elixhauser comorbidity categories outperformed APACHE II categories and had better ability to predict ICU mortality because of the inclusion of more than 5000 unique diagnosis codes. Several disease-spectrum-wide studies[14–16] have shown that factors such as total burden of disease, time between diagnoses, and order of diagnoses affect the risk of future complications. However, because long-term disease histories are often not systematically recorded everywhere, integration of such information into clinical decision support is still largely unexplored.

In this study, we used neural networks to combine long-term disease histories before ICU admission for more than 230 000 patients from a population-wide disease registry with acute physiology measures obtained from electronic patient records (EPRs) during the first 24 h of the ICU stay to predict in-hospital, 30-day, and 90-day mortality.

Methods

Data sources

In this study, we used registry data for long-term disease history and ICU admission, and acute ICU clinical data

from EPRs, to create a machine learning model to predict 30-day and 90-day mortality. Long-term disease histories for all medical and surgical patients admitted to an ICU in Denmark between Jan 1, 2004 (when systematic registration of ICU admissions began), and July 1, 2016, were extracted from the Danish National Patient Registry.[17] The Danish National Patient Registry contains disease histories coded according to the 10th revision of the International Classification of Diseases (ICD-10) dating back to the introduction of ICD-10 terminology in Denmark in 1994. ICD-10 terminology is organised into levels: 21 chapters, 227 diagnosis blocks, and 1698 level-3 diagnoses with more detailed descriptions. The numbers of diagnosis blocks and level-3 diagnoses can vary slightly between countries.

Registry data were excluded if they related to patients younger than 18 years or if outcome information was unavailable because of migration, change of personal identification number, or other similar reasons.

Additionally, raw data for acute ICU measures were extracted from EPRs covering individual medical and surgical patients at three ICUs in hospitals in the Capital Region of Denmark, from which harmonised, high-frequency data were available. Data were extracted from the Daintel Critical Information System, a specialised, commercial data collection and EPR system for ICUs.

After we developed our model, we obtained an external validation dataset from a fourth hospital in the Capital Region of Denmark between June 7, 2012, and May 20, 2016. These data were processed in the same way as the EPR data included in our model. The study was approved by the Danish Patient Safety Authority (3-3013-1723), the Danish Data Protection Agency (DT SUND 2016-48 and 2017-57), and the Danish Health Data Authority (FSEID 00003724).

Procedures

Dates of admission to, and discharge from, hospital before and after ICU admission were also extracted from the registry and used to calculate time from diagnosis to admission, length of hospital stay before ICU admission, and time to outcome (ie, survivor or non-survivor). Patients with multiple ICU admissions were included in the study (all admissions were used in the model). We retrieved information about date of birth, date of death, and sex from the Danish Central Person Registry. On the basis of this information, we calculated in-hospital mortality (which was defined as death on any ward during hospital admission), 30-day mortality (death within 30 days of ICU admission) and 90-day mortality (death within 90 days of ICU admission). The in-hospital mortality was calculated based on how many patients die while still at the hospital, which could be any number of days after admission. The 30-day mortality was measured 30 days after ICU admission, irrespective of whether patients were inpatients or outpatients. The acute ICU measures from EPR data included in this study were those from the original SAPS II and APACHE II scores, which were in clinical use during the study period.[2,3] SAPS II measures include age, type of admission, three chronic disease variables (metastatic cancer, haematological malignancy, and AIDS), and 12 physiological variables all recorded within the first 24 h of ICU admission (ie, heart rate, temperature, systolic blood pressure, partial pressure of arterial oxygen in inspired air, fractional concentration of oxygen in inspired air, urine output, white blood cell count, bilirubin concentration, and serum urea, bicarbonate, sodium, and potassium concentrations). SAPS II scoring is based on physiological extremes, and therefore both minimum and maximum values for physiological variables were included in our model as variables. These categorical physiological variables were represented as individual binary features, resulting in 27 variables in our model. APACHE II also includes physiological, chronic disease, and admission data. We created maximum and minimum features for the additional physiological variables of APACHE II that are not included in SAPS II, and represented additional comorbidity categories in binary form. 44 variables from the two scores were included in our model. We applied the same exclusion criteria used for the registry data to EPR data. ICU stays shorter than 24 h were also excluded, as were patients with missing outcome data at 90 days. We set a cutoff of one missing value per admission.

For all EPR variables, possible physiological ranges were defined from clinical experience (appendix) and values outside these ranges were defined as missing. Missing values were imputed with the median and mode from the feature distributions of continuous and binary features, respectively. We tried other methods, including multiple imputation, but use of the median and mode was simpler and was sufficient for this work in which the amount of missing data was very low.

See Online for appendix

We used a combination of Tukey's and Winsor's methods[18,19] to normalise data for the continuous variables. Values outside the upper and lower bounds of 1·5 times the IQR were set to the upper and lower limits of the range (appendix) to reduce the effects of outliers and produce a normal distribution while conserving the topology of the data. Finally, continuous variables were scaled to a mean of 0 with an SD of 1. The external validation dataset was imputed and normalised with mean and SD as described for the dataset used to develop the model. For both imputation and normalisation, the mean and SD values found for the development dataset were used. All patients were assigned a specific letter (A, B, or C), which represented the hospital that they had been admitted to. This information was presented to the model by means of one-hot encoding with three input units. For the data from the fourth hospital (the external test set), this information was omitted. Our external predictive performance estimate was thus fully independent of information about the group of hospitals represented in the training data.

Model development

Neural network models were trained on data that met inclusion criteria from both sources (ie, registry data and EPR data) both separately (registry data) and in combination with a cross-validation scheme (for data present in both registry and EPRs that met all inclusion criteria) in addition to an independent test set (figure 1). Registry data for ICU admissions were randomly split into a training set (85%) and an independent test set (15%). To avoid overfitting, five-fold cross-validation was done during training. Training data were thus divided into training (80%) and validation (20%) sets for each cross-validation fold. With these percentages (by contrast with a ten-fold cross-validation, for example) the independent test and validation sets were reasonably large, which reduced the risk that they might not be representative of the underlying population—a scenario that could have resulted in overestimation of the predictive performance.[20] A feed-forward neural network with one hidden layer was trained by backpropagation on the training set for 3000 epochs. In each epoch, the performance was assessed against the validation set, and the optimal model for each cross-validation fold was selected from the epoch with the highest performance. We used balanced training to ensure optimal prediction of both classes. Balancing was done in each epoch by randomly picking an equal number of patients from the majority (survivors) and minority (non-survivors) classes.

To establish how length of previous disease history affects ICU mortality, neural networks were trained for datasets with disease history of 1 month, 3 months, 6 months, 1 year, 2·5 years, 5 years, 7·5 years, 10 years, and 23 years before ICU admission. To establish how detailed the data needed to be represented, additional neural networks were trained: 5 years and 10 years accumulated or year-wise. For example, in the accumulated representation, 5 years of chapter-level diagnoses resulted in 21 variables (one for each chapter), whereas the year-wise representation of five years of ICD-10 chapters resulted in 105 variables. To establish the effect of cohort size, neural networks were trained for cohorts of 250, 500, 1000, 2500, 5000, 10 000, 20 000, 40 000, 80 000, 160 000, and 230 000. All models were trained with chapter-level (n=21) and block-level (n=227) ICD-10 diagnosis codes and with 90-day mortality as the outcome. The optimisation function (RMSprop), activation function (sigmoid), number of batches (one), and learning rate (0·001) were kept constant across models. The number of hidden neurons was optimised (0, 25, 100, 250, 500, 750, and 1000 units). We also ran grid searches on optimisation function, activation function, learning rates, number of batches, and different drop-out models. However, we concluded that the number of hidden neurons was most important in relation to predictive performance. We did not include balanced bootstrapping because of the additional computational time that this process would necessitate. A single-layer network of the type that we used can approximate any continuous function from a compact interval of the real numbers, and is therefore not limited in terms of mapping input vectors to mortality labels.[21] The choice of a simple network structure also makes it easier to explain how the model works to users.

The subset of patients for whom EPR data were available was also randomly split into a training set (85%) and an independent test set (15%). Training was done by balanced training with five-fold cross-validation. A one-hidden-layer neural network was trained on the training set for 5000 epochs. For each dataset, the mortality modelling was based on five configurations: "history before admission" was based on 10 years of disease history before ICU admission, age, and sex; "history at admission" was based on data available at ICU admission, including 10 years of disease history, length of hospitalisation before ICU admission, transfer category (ie, medical, scheduled surgery, or unscheduled surgery), and hospital code; "SAPS II" was based on the original SAPS II score; "APACHE II" was based on the original APACHE II score; and "aggregated history" was based on the features from SAPS II, APACHE II, and "history at admission" (appendix). All models were trained for in-hospital, 30-day, and 90-day mortality.

Optimisation of hyperparameters was done for 0 (logistic regression), 3, 5, 10, 25, 50, 100, 200, and 300 hidden neurons. Optimisation function (RMSprop), activation function (sigmoid), number of batches (one) and learning rate (0·001) were kept constant across EPR models. The optimal models identified in each cross-validation fold and for each dataset and outcome were used to predict outcome for admissions in the independent test set. The five predictions, one for each cross-validation fold, were gathered in an ensemble (by mean) into the final prediction. The Python code for training the neural network for "aggregated history" is available in the appendix.

Model interpretation

For the best-performing model (trained on aggregated data), we analysed feature importance and feature interactions. We interpreted overall feature importance with the test set. Importance was assessed by estimating the effect of feature absence—ie, in an iterative process, each feature was set to 0 (mean or mode) and new predictions were computed. For each patient, the effect of feature absence was calculated by obtaining the distance to the original prediction. A negative distance was obtained if the new predicted value was higher (ie, towards non-survival) than the original prediction, suggesting that the patient in question had a value that lowered mortality risk compared with the mean risk for the population. A mean of absolute distances across patients was used to generate a ranked list of variables according to effect on outcome predictions across all patients. Estimation of variables' importance with a more sophisticated approach (ie, local interpretable

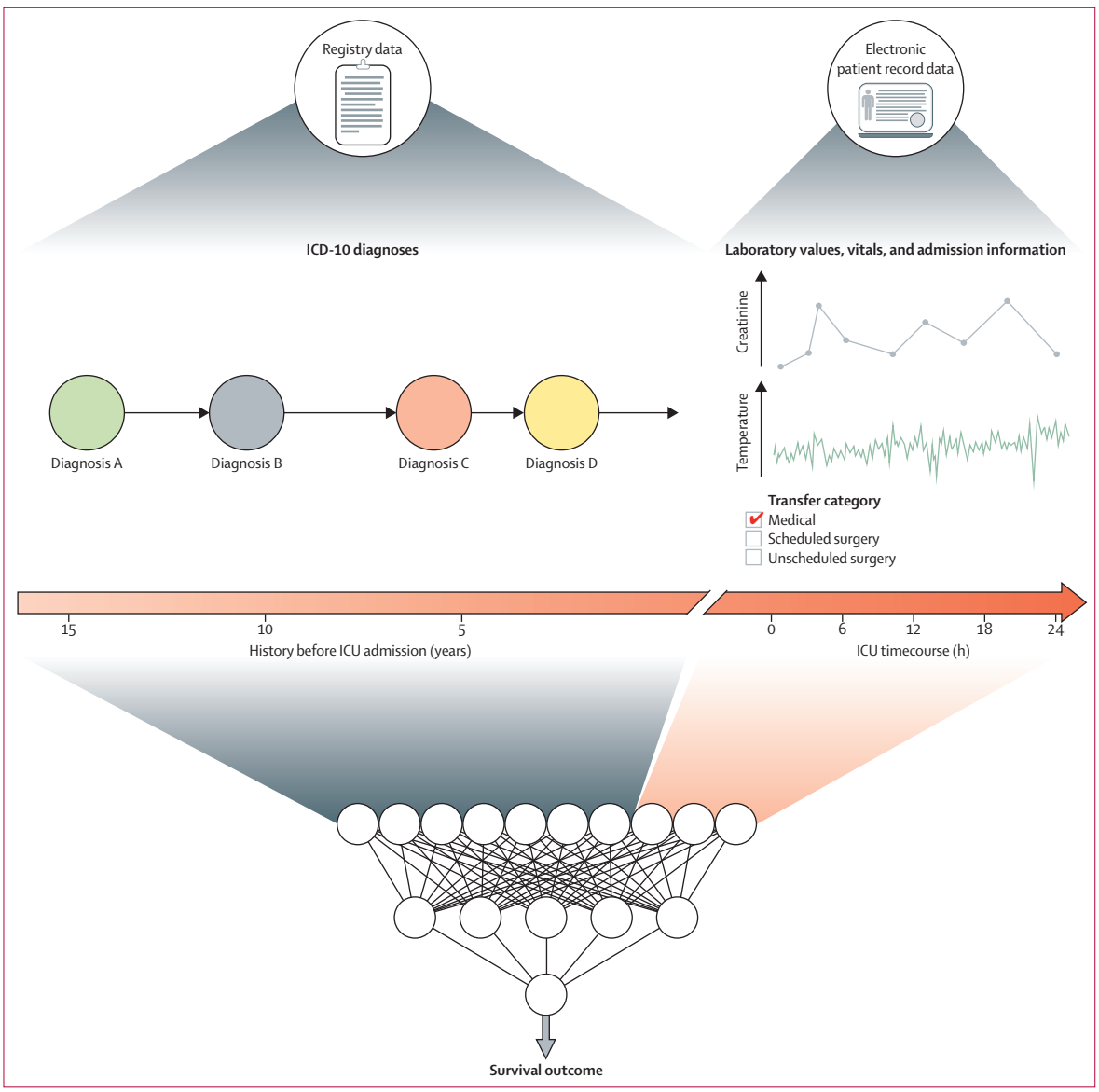

Figure 1: **Study overview of data and time-course aggregation**
Previous disease history and acute physiology data from the first 24 h of ICU admission were used separately or in combination as inputs to a neural network algorithm to predict in-hospital, 30-day, and 90-day mortality. Disease history data were retrieved from a population-wide registry that included up to 23 years of disease history before ICU admission. Laboratory values, vital signs, and admission information for the first 24 h of ICU admission were retrieved from electronic patient records from hospitals in the Danish Capital Region. ICU=intensive-care unit. ICD-10=10th revision of the International Classification of Diseases.

model-agnostic explanations) produced very similar results.[22]

Individual variable interactions were calculated for the top-ranking variables. Again, the variable in question was set to 0 for the entire dataset, and each of the remaining features were iteratively set to 0 to estimate the effect of both features being absent at the same time. The effect of the interaction was compared with the additive effect of the variables individually. The difference between the sum of the effects when removing the two features individually and the effect of removing the two features simultaneously was calculated for each patient and the mean of the absolute distances was used to find the top interacting features. Because many of the binary variables were present in only some patients, importance was re-estimated by correcting for the number of patients in whom the variable was present. Some variables were present in less than 1% of patients (n=15) in the independent test set and were not included in the analysis. Analyses were repeated for each model in the ensemble (each cross-validation fold) and the effect was averaged per patient before calculation of the overall importance and strength of the feature interaction.

Model performance was assessed with the Matthews correlation coefficient, the area under the receiver operating characteristics (AUROC) curve, and positive predictive value. The Matthews correlation coefficient—which is a quality measure for multiclass classifiers that takes the entire confusion matrix into account and, thus, fairly summarises the prediction on unbalanced datasets—was calculated as

$$\textit{Matthews correlation coefficient} = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

in which TP=true positives, TN=true negatives, FP=false positives, and FN=false negatives. AUROC was obtained by plotting the rate of correctly classified positives among all positive predictions (ie, the true positive rate) as a function of incorrect positives among all negatives (ie, the false positive rate), at varying thresholds. Positive predictive value, which is also known as precision and is used in clinical practice to assess the performance of alarm systems (for which a low ratio of false alarms is essential), was calculated as

$$\textit{Positive predictive value} = \frac{TP}{TP + FP}$$

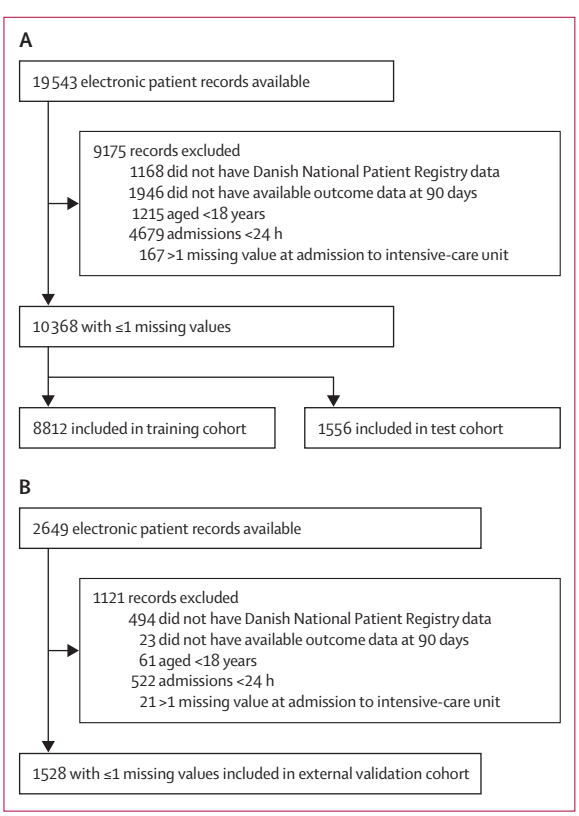

Figure 2: **Selection of electronic patient records for admissions to intensive-care units in the training and test cohort (A) and in the external validation cohort (B)**

The positive predictive value decreases with increasing number of false positive results. In this study, "negatives" refers to survivors, and "positives" to non-survivors. All analyses were done in Python (version 2.7), in which neural networks were trained with Keras. Plots were generated in R (version 3.4.0).

Role of the funding source

The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Results

10 years of disease history before ICU admission was available for 231 633 unique patients with 275 143 ICU admissions in Denmark. At a cutoff allowing only one missing value per ICU admission, only 421 (0·1%) of all values were missing. The final dataset included 8817 individual patients with 10 368 ICU admissions, which were subsequently split into training (n=8812) and test sets (n=1556; figure 2), and had both long-term disease history and ICU data from EPRs. 10 223 (98·6%) of the 10 368 admissions did not have any missing data. 3462 (33·4%) of this cohort died in hospital, 2729 (26·3%) died within 30 days of ICU admission, and 3671 (35·4%) died within 90 days of ICU admission (table 1).

Model performance, as indicated by the Matthews correlation coefficient, increased as additional years were included in the model, with saturation at 5 years and 10 years of disease history for the chapter and block level diagnoses, respectively (appendix). At both the chapter level and block level of the ICD-10, representation of the data in a year-wise manner resulted in improved mortality prediction compared with accumulated data representation (appendix). Furthermore, increasingly large cohorts were needed to obtain optimal performance when data were represented with increasing granularity—ie, block-level diagnoses and year-wise representation (appendix). The addition of further years to disease history or inclusion of level-3 diagnoses (ie, diagnosis codes) did not improve the model's performance (data not shown).

The Multimorbidity Index was reported to significantly outperform (ie, to have a greater AUROC than) other multimorbidity scores in the prediction of ICU mortality compared with other multimorbidity scores.[13] We also applied the Multimorbidity Index method to our long-term disease history data to predict mortality risk. 10 years of previous history was used with year-wise representation. Our neural-network approach performed better than the Multimorbidity Index in the independent test set for all performance measures (Matthews correlation coefficient 0·265 *vs* 0·065; AUROC 0·713 *vs* 0·554; positive predictive value 0·388 *vs* 0·257; appendix).

After first training models solely on previous disease history, we studied the predictive effect of previous

disease history and acute physiology measures separately and in combination (appendix). Remarkably, the network that included only age, sex and 10-year disease history before ICU admission as model inputs (ie, the history before admission configuration) performed similarly (Matthews correlation coefficient 0·326; AUROC 0·731) to SAPS II (0·325; 0·735) and APACHE II (0·285; 0·715; figure 3, table 2). The aggregated history model outperformed all other models as measured by the Matthews correlation coefficent (0·410; figure 3). AUROC and positive predictive value were also improved in the aggregated history model (table 2).

Data from the external validation set were obtained between June 7, 2012, and May 20, 2016, and comprised 1384 medical and surgical patients with 1528 ICU admissions (figure 2B). The best-performing model, the aggregated history model, was validated in the external dataset. For 90-day mortality, the Matthews correlation coefficient was 0·382, the AUROC was 0·746, and the positive predictive value was 0·576 (table 2).

To assess if our initial predictions were reliable as a reference at later stages during ICU stays, we investigated how accurate our initial predictions were for patients remaining in the ICU. A substantial proportion of patients was discharged within the first 10 days of ICU admission (appendix). Importantly, the performance of the models including previous disease history and admission information did not deteriorate during the entire admission irrespective of length of stay in the ICU, whereas predictions based on data collected within the first 24 h of ICU admission become less reliable as time passed (appendix). For the model based on history at admission, the Matthews correlation coefficient was 0·357 after 24 h and 0·337 after 14 days, whereas the Matthews correlation coefficient for the aggregated history model was 0·410 after 24 h and 0·242 at 14 days.

We next interpreted the machine learning model.[22,23] The ten most important variables contributing to the performance of the aggregated history model with aggregated time-course are listed in figure 4A. Low age decreases mortality risk whereas high age increases mortality risk (figure 4A). The only binary variables among the top ten were hospital A and mechanical ventilation (figure 4A).

For example, a previous history of diagnoses related to reproduction (ICD-10 diagnoses Z30–39) generally decreased mortality risk, and this effect was stronger for older patients than for younger patients (figure 4B). We also investigated these diagnoses individually: low-risk women had previous diagnoses related to pregnancy and deliveries (data not shown). Figure 4C shows the interaction of length of stay before ICU admission and the history of haematological malignancy. Mortality risk increases with increasing length of hospital stay before ICU admission, and this effect is enhanced for patients with a history of haematological malignancies (figure 4C). Yet, in the

	Model construction data (training cohort; n=8812)	Internal validation data (test cohort; n=1556)	External validation data (n=1528)
Sex			
Female	3413 (38·7%)	576 (37·0%)	668 (43·7%)
Male	5399 (61·3%)	980 (63·0%)	860 (56·3%)
Hospital stay before admission to ICU, days	2 (0–10)	2 (0–9)	2 (0–7)
Age, years	64 (52–73)	64 (53–73)	68 (60–76)
Arterial pH			
Minimum	7·31 (7·22–7·38)	7·31 (7·22–7·38)	7·29 (7·20–7·37)
Maximum	7·43 (7·38–7·47)	7·43 (7·38–7·47)	7·42 (7·37–7·47)
Bilirubin, µmol/L	10 (6–20)	10 (6–19)	13 (9–21)
FiO_2, %	50 (40–70)	50 (40–70)	45 (30–69)
Glasgow Coma Scale score	15 (12–15)	15 (11–15)	13 (7–15)
Heart rate, beats per min			
Minimum	75 (64–88)	75 (63–86)	76 (65–88)
Maximum	115 (100–131)	114 (100–130)	120 (104–135)
Haematocrit, % of blood volume			
Minimum	28·0 (25·0–32·2)	28·0 (25·0–32·0)	28·4 (25·0–33·6)
Maximum	34·0 (30·0–38·0)	33·0 (30·0–38·0)	32·2 (28·0–37·4)
Mean arterial pressure, mm Hg			
Minimum	57 (50–65)	57 (50–65)	54 (47–61)
Maximum	107 (90–135)	110 (90–141)	125 (101–188)
$PaCO_2$, kPa	5·6 (4·9–6·6)	5·6 (4·9–6·5)	5·2 (4·4–6·5)
PaO_2, kPa	10·5 (9·0–13·1)	10·3 (9·0–13·1)	10·0 (8·7–12·8)
Respiratory rate, breaths per min			
Minimum	13 (10–16)	13 (10–16)	12 (9–15)
Maximum	27 (21–35)	26 (20–33)	32 (27–38)
Serum bicarbonate, mmol/L	22·0 (18·6–25·0)	22·0 (18·5–25·0)	20·4 (16·8–24·0)
Serum creatinine, mmol/L			
Minimum	85 (60–140)	88 (62–144)	99 (65–178)
Maximum	100 (67–174)	103 (69–180)	116 (72–216)
Serum potassium, mmol/L			
Minimum	3·5 (3·2–3·8)	3·5 (3·3–3·8)	3·6 (3·3–3·9)
Maximum	4·4 (4·1–4·8)	4·4 (4·1–4·8)	4·4 (4·1–4·9)
Serum sodium, mmol/L			
Minimum	136 (133–140)	136 (133–140)	136 (132–139)
Maximum	140 (137–144)	140 (137–144)	140 (137–144)
Serum urea, mmol/L	9·0 (5·6–15·1)	9·0 (5·8–15·0)	10·0 (6·0–16·0)
Systolic blood pressure, mm Hg			
Minimum	85 (71–98)	84 (71–98)	79 (67–93)
Maximum	150 (130–170)	150 (130–170)	153 (135–173)
Temperature, °C			
Minimum	36·5 (35·6–37·0)	36·5 (35·6–37·0)	36·5 (35·7–37·0)
Maximum	37·7 (37·0–38·4)	37·7 (37·0–38·4)	37·4 (36·8–38·1)
Urine output, mL	2000 (1119–2800)	2000 (1015–2800)	1500 (800–2270)
White blood cell count, 10^9 per L			
Minimum	10·0 (6·8–14·0)	9·9 (6·5–14·0)	10·6 (7·0–15·0)
Maximum	13·0 (9·0–18·4)	13·0 (9·1–18·0)	14·0 (9·5–20·0)
Acute renal failure	1227 (13·9%)	227 (14·6%)	210 (13·7%)
AIDS	38 (0·4%)	5 (0·3%)	1 (0·1%)
Cardiac failure	752 (8·5%)	155 (10·0%)	132 (8·6%)

(Table 1 continues on next page)

	Model construction data (training set; n=8812)	Internal validation data (test set; n=1556)	External validation data (n=1528)
(Continued from previous page)			
Immunocompromised	1138 (12·9%)	200 (12·9%)	166 (10·9%)
Liver failure	575 (6·5%)	91 (5·8%)	67 (4·4%)
Malignant haematology	477 (5·4%)	86 (5·5%)	78 (5·1%)
Mechanical ventilation	5450 (61·8%)	963 (61·9%)	774 (50·7%)
Metastatic cancer	368 (4·2%)	65 (4·2%)	52 (3·4%)
Renal failure	550 (6·2%)	123 (7·9%)	98 (6·4%)
Respiratory failure	1277 (14·5%)	233 (15·0%)	228 (14·9%)
Transfer type			
Medical	4615 (52·4%)	817 (52·5%)	1076 (70·4%)
Scheduled surgery	415 (4·7%)	85 (5·5%)	46 (3·0%)
Unscheduled surgery	3782 (42·9%)	654 (42·0%)	406 (26·6%)
Mortality			
In-hospital mortality	2944 (33·4%)	518 (33·3%)	569 (37·2%)
30-day mortality	2315 (26·3%)	414 (26·6%)	525 (34·4%)
90-day mortality	3121 (35·4%)	550 (35·3%)	638 (41·8%)

Data are n (%) or median (IQR). ICU=intensive-care unit. FiO_2=fractional concentration of oxygen in inspired air. $PaCO_2$=partial pressure of carbon dioxide in inspired air. PaO_2=partial pressure of arterial oxygen in inspired air.

Table 1: **Characteristics of ICU admissions used to develop the model and in the external validation data set**

case of increased mortality risk caused by both a long length of stay and haematological malignancy, the effects of the two variables are not additive; instead a plateau is reached for patients with long hospital stays (>18 days) before ICU admission.

The two largest subgroups of ICU patients were patients with respiratory insufficiency, who accounted for 618 (39·7%) of 1556 test set patients, and patients with sepsis, who accounted for 224 (14·4%) of test set patients. Models based mainly on acute measures (eg, SAPS II, APACHE II) generally performed worse in patients with sepsis than in the overall cohort, whereas the aggregated model and the models based on long-term disease history performed well in patients with sepsis (figure 5A). Furthermore, long-term disease history seemed to contribute more than acute measures to models for estimation of prognosis in patients with sepsis (figure 5A). Performance of all models in the respiratory insufficiency subgroup was similar to that in the full ICU population (figure 5B).

Discussion

In this retrospective study of registry data for all Danish ICU admissions between 2004 and 2016 and EPR data for more than 10 000 patients, prediction models for mortality risk in ICU patients based solely on previous disease history performed as well as mortality scores in clinical use (ie, SAPS II and APACHE II). Furthermore, models based on aggregation of long-term disease history and acute physiology measures in a neural network showed that health-related events from different timepoints in a patient's life interact, and that exploitation of these interactions between long-term and short-term disease history gives more precise prognostic estimates than either short-term or long-term history individually.

On the basis of a review of mortality prediction models that included data from the critical care database Medical Information Mart for Intensive Care, we identified studies that applied various machine learning methods and that had outcome measures and inclusion criteria similar to those used in our study.[23–30] Sufficient information to calculate the Matthews correlation coefficient, AUROC, and positive predictive value was available in only three studies.[25,27,28] Comparison of prediction performance showed that other models had higher AUROC but worse Matthews correlation coefficients and positive predictive values than our model (appendix). The Matthews correlation coefficient and positive predictive value are useful performance measures for imbalanced datasets, because they show how well models perform in the two balanced classes and the minority class, respectively. Thus, our model is associated with fewer false predictions of death than those in previous studies.

The value of longitudinal historical health data in predictions of future health has been questioned.[31] In our study, the accuracy of mortality risk prediction was improved by the inclusion of increasingly long disease histories despite changes in clinical practice over the years. Furthermore, our neural network predicted ICU prognosis more accurately than the best-performing multimorbidity score reported so far, the Multimorbidity Index.[13] In the study of the Multimorbidity Index,[13] Min and colleagues argued that diagnoses should not be assumed to independently affect the odds of mortality. Our findings support this position. We did not have long-term data for medications prescribed, adherence to medication, or polypharmacy. Such data could be relevant to our model.

We could not study model interactions of long-term disease history and 24 h ICU measures in a year-wise manner because of the small size of the ICU cohort with both registry and EPR data available (n=10 368). Yet in the larger, population-wide ICU cohort (n=275 143 ICU admissions), in the model based on previous disease history alone, time-wise separation of an individual's diagnoses improves mortality prediction. Furthermore, mortality risk is affected not only by the diagnoses but also the temporal order of these diagnoses (appendix). A larger cohort with fine-grained EPR data would probably improve model performance. Recent disease history (ie, disease history immediately before the hospital admission) affects a patient's risk of ICU mortality more than diseases diagnosed 5–10 years ago. It seems intuitive that diseases that are diagnosed close to ICU admission affect mortality risk more than diseases diagnosed several years previously, although the inclusion of these older diagnoses still further improves mortality prediction models.

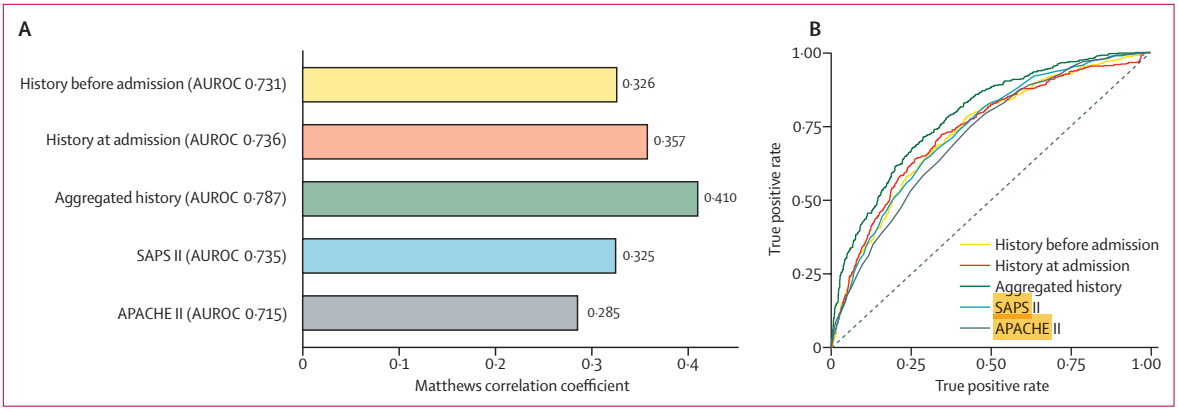

***Figure 3*: Matthews correlation coefficients (A) and AUROC (B) for models with input data from different timepoints**
90-day mortality was the outcome measure for this figure. The history before admission model was trained on age, sex, and 10-year disease history before ICU admission. The history at admission model also included transfer category (ie, medical, scheduled surgery, or unscheduled surgery), length of stay before ICU admission, and hospital code. The aggregated history model included an aggregated dataset with both long-term disease history (ie, the data used in the history at admission model) and data for the first 24 h of ICU admission (ie, features included in SAPS II and APACHE II combined). In these three models, disease history was presented as an accumulated representation because the cohort was too small for year-wise representation (appendix). In figure 3B, the dotted line represents the AUROC of a model with a random guess. AUROC=area under the receiver operating characteristic curve. ICU=intensive-care unit. SAPS II=Simplified Acute Physiology Score II. APACHE II=Acute Physiologic Assessment and Chronic Health Evaluation II.

A general limitation of our study was that the reliability of the predictive value of the aggregated history models deteriorates over time. Thus, the physiology measures included in the model should be updated dynamically, and prognosis reassessed on a day-by-day basis. However, one of the key clinical findings of our study was that the mortality predictions with models based on previous disease history alone remained stable and reliable throughout an ICU admission, whereas the accuracy of predictions with models based on ICU data deteriorated more rapidly over time (appendix). Furthermore, mortality prediction models based on previous disease history were independent of the care provided during the ICU admission and thus are an unbiased severity measure. The model can be used in the same way as severity scores like SAPS II. The addition of real-time data from ICU care could make the model dynamic and could, for example, help to provide hourly predictions. Additionally, this type of prediction does not require new data collection because it is based on data from an existing high-quality, population-wide registry. Thus, this model overcomes the burden of data harmonisation and quality control to some extent.[17] Practically, mortality predictions could be made before potential ICU admissions and available immediately, by contrast with the scores used in ICUs at present.

Clinical intervention trials do not use the methods that we used in this study, which allowed for initial stratification based on a more accurate understanding of disease history. Patterns in disease history data could probably be used effectively in trial design to detect a difference between groups and to avoid use of non-comparable subgroups.[32] Furthermore, the effect of previous disease history—both diagnosis and the temporal order of diagnoses—on patient outcomes could reflect a transition of physiological systems that predisposes individuals not only to certain diagnoses but also to certain outcomes. This could explain why some survivors of sepsis return to their baseline performance status whereas others progressively decline (a decline not predicted by hospital course).[33–35]

	MCC	AUROC	Positive predictive value
In-hospital mortality			
Disease history before admission	0·326	0·732	0·562
Disease history at admission	0·330	0·724	0·563
Aggregated history*	0·391 (0·341)	0·792 (0·733)	0·575 (0·534)
SAPS II	0·347	0·742	0·556
APACHE II	0·300	0·720	0·493
30-day mortality			
Disease history before admission	0·243	0·688	0·446
Disease history at admission	0·279	0·700	0·422
Aggregated history*	0·368 (0·341)	0·787 (0·737)	0·481 (0·492)
SAPS II	0·349	0·752	0·477
APACHE II	0·287	0·730	0·409
90-day mortality			
Disease history before admission	0·326	0·731	0·550
Disease history at admission	0·357	0·736	0·569
Aggregated history*	0·410 (0·382)	0·787 (0·746)	0·589 (0·576)
SAPS II	0·325	0·735	0·565
APACHE II	0·285	0·715	0·507

Values in parentheses relate to the additional, external validation set. MCC=Matthews correlation coefficient. AUROC=area under the receiver operating characteristics curve. SAPS II=Simplified Acute Physiology Score II. APACHE II=Acute Physiologic Assessment and Chronic Health Evaluation II. *Best model for prediction of in-hospital, 30-day, and 90-day mortality.

***Table 2*: Performance comparison of models predicting risk of mortality in the intensive-care unit with mortality outcome data**

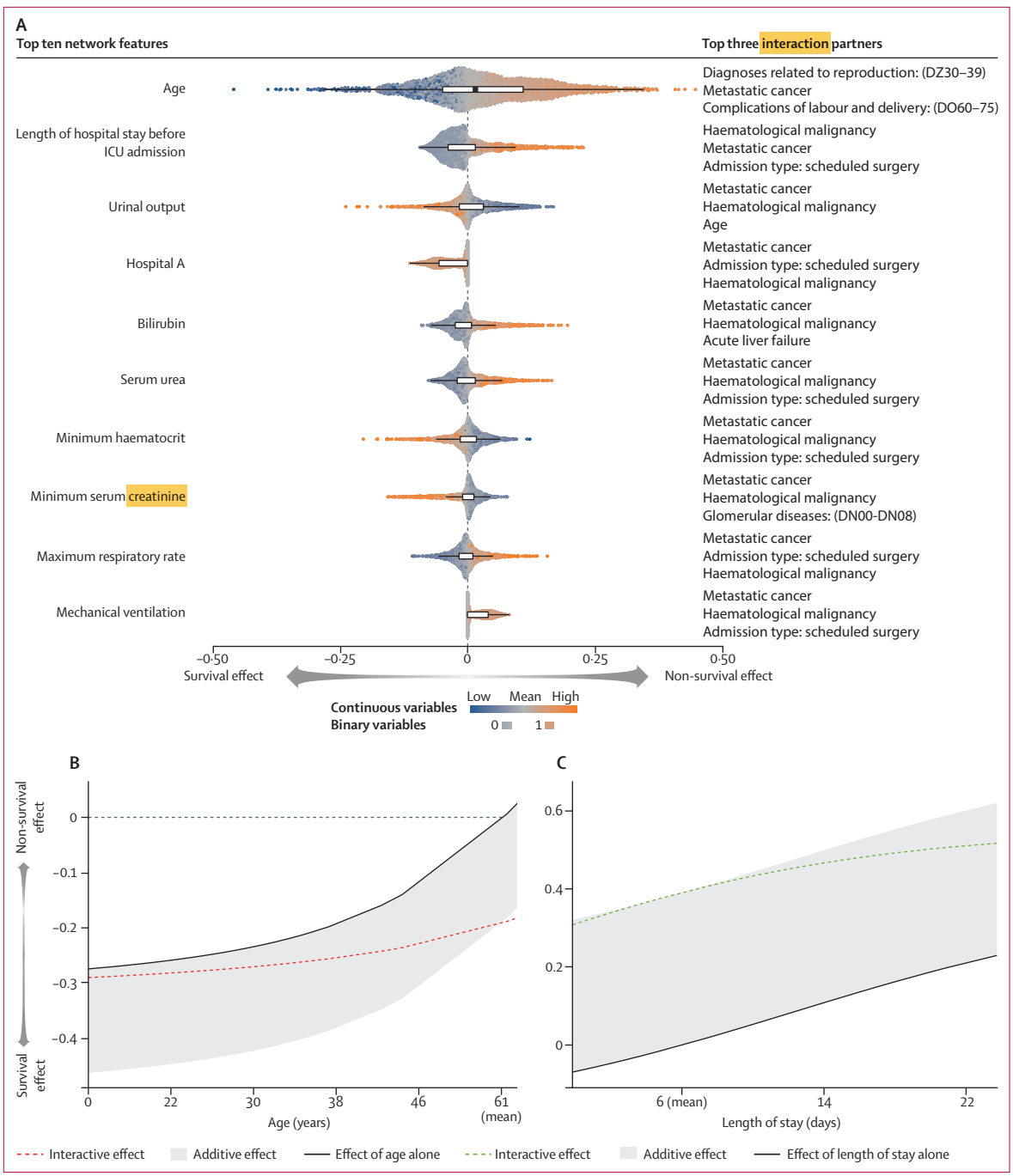

Figure 4: **The ten most important variables in the aggregated history model (A), and effect on the aggregated history model of interaction between age and history of diagnoses related to reproduction (B), and interaction between length of stay before ICU admission and history of haematological malignancy (C)**
In (A), each patient is represented by a dot in the distribution for each variable. For binary variables, the interaction importance is corrected according to the number of patients in whom the variable is present. In (B) the diagnoses related to reproduction are those covered by block Z30–39 of the 10th revision of the International Classification of Diseases. In (C), length of stay refers to the duration of the hospital stay before ICU admission. ICU=intensive-care unit.

Numerous studies have shown the rapid changes that occur in clinical practice when physicians are offered tools that enable data-guided decision making.[36] However, a clinical trial is required to fully uncover the value of mortality prediction as a decision-support tool in ICUs. We specifically used neural networks because we wanted to study the interaction of health-related events from different timescales. We compared the performance of our aggregated history model with other ICU mortality prediction models. Yet, as highlighted by Johnson and

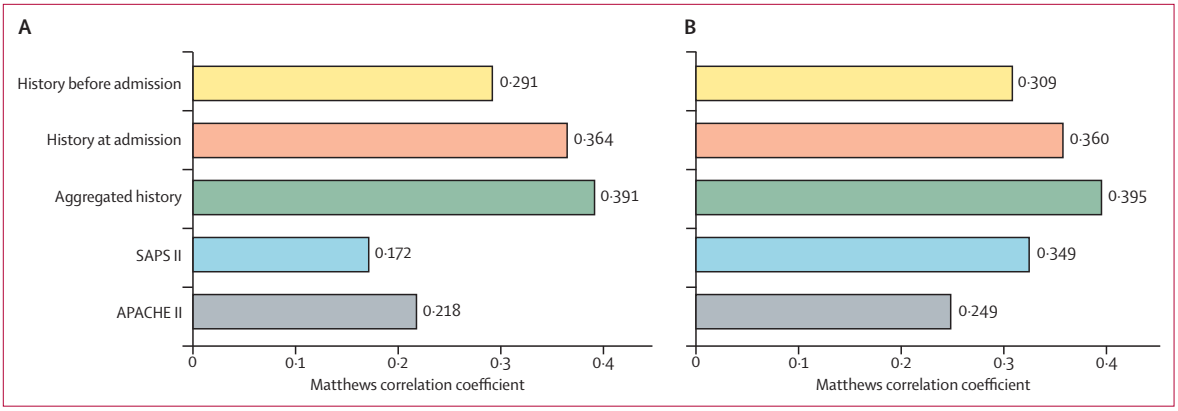

Figure 5: **Mortality prediction for patients in intensive-care units with sepsis (A) and respiratory insufficiency (B) as primary diagnoses**
In patients with sepsis, the models built on long-term disease history were better than those built primarily on acute measures. In patients with respiratory insufficiency, performance of all models was similar to that in the full population. SAPS II=Simplified Acute Physiology Score II. APACHE II=Acute Physiologic Assessment and Chronic Health Evaluation II.

colleagues,[23] comparison of studies is difficult because of variation in outcome measures, inclusion criteria, feature availability, and time from data collection to mortality prediction. Additionally, AUROC was the only performance metric reported consistently across studies, and this metric does not always adequately reflect model performance for data with unbalanced classes.[37] We found that studies with higher class imbalance (ie, few non-survivors compared with survivors) had better AUROC results but worse Matthews correlation coefficients and positive predictive values than studies with lower class imbalance. That is, high AUROC results can be at the expense of more false positive results, which are penalised more heavily in the calculation of the Matthews correlation coefficient and positive predictive value. Class imbalance was accounted for during training of our model, because there is clinical demand for accurate prediction, especially among non-survivors (ie, the minority class). Minimisation of the number of false positive predictions—ie, survivors predicted not to survive—was imperative because such predictions can have fatal consequences if, for example, treatment is stopped on the basis of prediction of a high risk of mortality. Even though our model was validated against external, retrospective data, validation in a prospective study is also important.

Neural networks can predict mortality in patients in the ICU as well as or better than other machine learning techniques.[29,38,39] However, interpretation of the feature interaction effects modelled here has not been previously reported. The advantage of the simple approach that we used was higher transparency and much faster computational time. Our study shows that more realistic interpretations of the risk of death can be achieved with neural network models than by studying the additive effects of disease history and ICU measures. Binary variables make a constant contribution to predictions in linear models (eg, SAPS II, APACHE II), but in the neural network their effect varies because of interactions with other features. Because binary features were mostly absent from patients in the EPR cohort, continuous features dominate the list of features whose effect is strongest on mortality across all patients. Explainable neural networks are also gaining popularity in the context of deep learning.[22,40] We believe that explainable models are a key to the transformation of advanced models into actionable, transparent, and trustworthy clinical tools.

Contributors

AP and SB conceived the study, which was designed by ABN. ABN and H-CT-M. did the data analysis, which was interpreted by ABN, H-CT-M, KB, APN, CET, ML, PLM, AP and SB. PJC, MH, LD, LS, and PH extracted and handled data. All authors contributed to the preparation of the Article and approved the final version.

Declaration of interests

LD, LS, and PH are employed by Daintel (which is taking part in the BigTempHealth project funded by the Danish Innovation Fund). AP reports grants from Ferring and the Novo Nordisk Foundation. SB reports personal fees from Intomics and Proscion. All other authors declare no competing interests.

Data sharing

Data are available for use in secure, dedicated environments via application to the Danish Patient Safety Authority and the Danish Health Data Authority.

Acknowledgments

This work was supported by the Novo Nordisk Foundation (grants NNF17OC0027594 and NNF14CC0001) and the Danish Innovation Fund (grant 5153-00002B). We thank the ICUs at Rigshospitalet (Copenhagen, Denmark), Hvidovre Hospital (Hvidovre, Denmark), Bispebjerg Hospital (Bispebjerg, Denmark), and Herlev Hospital (Herlev, Denmark) for their role in data acquisition, and particularly the staff at Rigshospitalet for their crucial clinical input.

References

1 Christiansen CF, Christensen S, Johansen MB, Larsen KM, Tønnesen E, Sørensen HT. The impact of pre-admission morbidity level on 3-year mortality after intensive care: a Danish cohort study. *Acta Anaesthesiol Scand* 2011; **55:** 962–70.

2 Le Gall JR, Lemeshow S, Saulnier F. A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. *JAMA* 1993; **270:** 2957–63.

3 Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: a severity of disease classification system. *Crit Care Med* 1985; **13:** 818–29.

4 Johnson AE, Ghassemi MM, Nemati S, Niehaus KE, Clifton DA, Clifford GD. Machine learning and decision support in critical care. *Proc IEEE Inst Electr Electron Eng* 2016; **104:** 444–66.

5 Esper AM, Martin GS. The impact of comorbid [corrected] conditions on critical illness. *Crit Care Med* 2011 **39:** 2728–35.

6 Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J Chronic Dis* 1987; **40;** 373–83.

7 Elixhauser A, Steiner C, Harris DR, Coffey RM. Comorbidity measures for use with administrative data. *Med Care* 1998; **36:** 8–27.

8 Poses RM, McClish DK, Smith WR, Bekes C, Scott WE. Prediction of survival of critically ill patients by admission comorbidity. *J Clin Epidemiol* 1996; **49:** 743–47.

9 Stavem K, Hoel H, Skjaker SA, Haagensen R. Charlson comorbidity index derived from chart review or administrative data: agreement and prediction of mortality in intensive care patients. *Clin Epidemiol* 2017; **9:** 311–20.

10 Christensen S, Johansen MB, Christiansen CF, Jensen R, Lemeshow S. Comparison of Charlson comorbidity index with SAPS and APACHE scores for prediction of mortality following intensive care. *Clin Epidemiol* 2011; **3:** 203–11.

11 Ho KM, Finn J, Knuiman M, Webb SAR. Combining multiple comorbidities with Acute Physiology Score to predict hospital mortality of critically ill patients: a linked data cohort study. *Anaesthesia* 2007; **62:** 1095–100.

12 Johnston JA, Wagner DP, Timmons S, Welsh D, Tsevat J, Render ML. Impact of different measures of comorbid disease on predicted mortality of intensive care unit patients. *Med Care* 2002; **40:** 929–40.

13 Min H, Avramovic S, Wojtusiak J, et al. A Comprehensive multimorbidity index for predicting mortality in intensive care unit patients. *J Palliat Med* 2017; **20:** 35–41.

14 Beck MK, Westergaard D, Jensen AB, Groop L, Brunak S. Temporal order of disease pairs affects subsequent disease trajectories: the case of diabetes and sleep apnea. *Biocompute* 2017; **22:** 380–89.

15 Jensen AB, Moseley PL, Oprea TI, et al. Temporal disease trajectories condensed from population-wide registry data covering 6·2 million patients. *Nat Commun* 2014; **5:** 4022.

16 Beck MK, Jensen AB, Nielsen AB, Perner A, Moseley PL, Brunak S. Diagnosis trajectories of prior multi-morbidity predict sepsis mortality. *Sci Rep* 2016; **6:** 36624.

17 Schmidt M, Schmidt SA, Sandegaard JL, Ehrenstein V, Pedersen L, Sørensen HT. The Danish National Patient Registry: a review of content, data quality, and research potential. *Clin Epidemiol* 2015; **7:** 449–90.

18 Tukey JW. Exploratory data analysis (Reading, MA: Addison-Wesley, 1977).

19 Hastings C, Mosteller F, Tukey JW, Winsor CP. Low moments for small samples: a comparative study of order statistics. *Ann Math Statist* 1947; **18:** 413–26.

20 Almagro Armenteros JJ, Tsirigos KD, Sønderby CK, et al. SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat Biotechnol* 2019; **37:** 420–23.

21 Hornik K, Stinchcombe M, White H. Multilayer feedforward networks are universal approximators. *Neural Networks* 1989; **2:** 359–66.

22 Ribeiro MT, Singh S, Guestrin C. 'Why should I trust you?' Explaining the predictions of any classifier. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; San Francisco, CA, USA; Aug 13–17, 2016.

23 Johnson AEW, Pollard TJ, Mark RG. Reproducibility in critical care: a mortality prediction case study. *Proceed Mach Learn Healthcare* 2017; **68:** 1–16.

24 Johnson AEW, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. *Sci Data* 2016; **3:** 160035.

25 Caballero Barajas KL, Akella R. Dynamically modeling patient's health state from electronic medical records. 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; Sydney, NSW, Australia; Aug 10–13, 2015.

26 Ghassemi M, Pimentel MAF, Naumann T, et al. A multivariate timeseries modeling approach to severity of illness assessment and forecasting in ICU with sparse, heterogeneous clinical data. *Proc Conf AAAI Artif Intell* 2016; **2015:** 446–453.

27 Ghassemi M, Naumann T, Doshi-Velez F, et al. Unfolding physiological state: mortality modelling in intensive care units. *KDD* 2014; **2014:** 75–84.

28 Lehman L-W, Saeed M, Long W, Lee J, Mark R. Risk stratification of ICU patients using topic models inferred from unstructured progress notes. *AMIA Annu Symp Proc* 2012; **2012:** 505–11.

29 Zahid MAH, Lee J. Mortality prediction with self normalizing neural networks in intensive care unit patients*. IEEE EMBS International Conference on Biomedical & Health Informatics; Las Vegas, NV, USA; March 4–7, 2017.

30 Hoogendoorn M, El Hassouni A, Mok K, Ghassemi M, Szolovits P. Prediction using patient comparison *vs* modeling: a case study for mortality prediction. *Conf Proc IEEE Eng Med Biol Soc* 2016; **2016:** 2464–67.

31 Chen JH, Alagappan M, Goldstein MK, Asch SM, Altman RB. Decaying relevance of clinical data towards future decisions in data-driven inpatient clinical order sets. *Int J Med Inform* 2017; **102:** 71–79.

32 Wong JLC, Mason AJ, Gordon AC, Brett SJ. Are large randomised controlled trials in severe sepsis and septic shock statistically disadvantaged by repeated inadvertent underestimates of required sample size? *BMJ Open* 2018; **8:** e020068.

33 Cuthbertson BH, Wunsch H. Long-term outcomes after critical illness. The best predictor of the future is the past. *Am J Respir Crit Care Med* 2016; **194:** 132–34.

34 Shankar-Hari M, Ambler M, Mahalingasivam V, Jones A, Rowan K, Rubenfeld GD. Evidence for a causal link between sepsis and long-term mortality: a systematic review of epidemiologic studies. *Crit Care* 2016; **20:** 101.

35 Davis JS, He V, Anstey NM, Condon JR. Long term outcomes following hospital admission for sepsis using relative survival analysis: a prospective cohort study of 1092 patients with 5 year follow up. *PLoS One* 2014; **9:** e112224.

36 Ferguson JS, Van Wert R, Choi Y, et al. Impact of a bronchial genomic classifier on clinical decision making in patients undergoing diagnostic evaluation for lung cancer. *BMC Pulm Med* 2016; **16:** 66.

37 Lever J, Krzywinsky M, Altman NS. Points of significance: classification evaluation. *Nat Publ Gr* 2016; **13:** 603–05.

38 Kim S, Kim W, Park RW. A comparison of intensive care unit mortality prediction models through the use of data mining techniques. *Healthc Inform Res* 2011; **17:** 232–43.

39 Clermont G, Angus DC, DiRusso SM, Griffin M, Linde-Zwirble WT. Predicting hospital mortality for patients in the intensive care unit: a comparison of artificial neural networks with logistic regression models. *Crit Care Med* 2001; **29:** 291–96.

40 Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Neural Information Processing Systems; Los Angeles CA, USA; Dec 4–9, 2017.