Mortality prediction in the intensive care: Role of mathematical models in benchmarking and decision-making
Minne, L.

Citation for published version (APA):
Minne, L. (2013). Mortality prediction in the intensive care: Role of mathematical models in benchmarking and decision-making
Summary
In the Intensive Care Unit (ICU), one faces the difficult task of predicting whether a patient will survive intensive care treatment or not. Two important uses of mortality prediction in the intensive care are *benchmarking* and informing *decisions* about individual patients. Several mathematical models have been devised to predict mortality in the ICU, but not much is known about their comparison to clinicians and their role in clinical practice. This PhD dissertation consists of three parts aiming to better understand 1) characteristics of current prognostic models in the ICU, 2) the implications for their use over time in benchmarking, and 3) the role they can play in supporting decision-making for individual patients.

**Part 1: Characteristics of prognostic models for elderly (ICU) patients**

Elderly patients form a particularly interesting target group in which concerns regarding withholding therapy and end-of-life decisions are common, and comfort-focused end-of-life care may be preferred when chances of survival are low. Mathematical prognostic models of mortality may have the potential to inform clinical decisions of individual patients. No previous literature review addressed however (the validity and use of) prognostic models of mortality and/or survival in this population. We aimed to identify and characterize prognostic models of mortality or survival for elderly (ICU) patients, their validity and their use in clinical practice.

Chapter 2 describes the results of a systematic review of literature on prognostic models of mortality and/or survival, their (internal and external) validity and whether they can be used in clinical practice. We included 103 studies reporting on 192 models. The intended use of these models was often not reported, but the general potential use most frequently included support of individual clinical decisions. Variables that could have specific clinical relevance in an older subgroup include data on functional and cognitive status [19] but these were included in only 32% of the models. The majority of the studies were concerned with the development of a new prognostic model (68%). The degree of model validation was moderate and others rarely used them. Even more common models developed some 20 years ago (e.g. the Charlson Index, Deyo Score, SAPS-II and APACHE-II) have not been frequently validated in an older subgroup.

Chapter 3 contains a more detailed description of the 17 ICU models reported in 7 of the included studies in chapter 2. We assessed the methodological quality, clinical credibility and external validity of these models. Although methodological quality was generally good, we found caveats in the following elements of clinical credibility: inclusion of clinically relevant variables, avoiding use of arbitrary thresholds for categorizing continuous variables, avoiding subjective data collection and reporting the range of probabilities provided by models. The models were rarely cited nor validated by others and no studies were found on their impact in clinical practice.

**Part 2: Role of prognostic models in benchmarking**

In benchmarking, one compares the quality of care among different centres (ICUs) by using a predictive model as reference. Specifically, one compares the observed to the predicted mortality in each centre by calculating Standardized Mortality Ratios (SMR; i.e. observed mortality divided by predicted mortality). This results in a ranking of the centres based on their SMR, which is considered as an indication of their delivered quality of care.
The validity of these rankings is dependent on the assumption that the applied prediction model is accurate. We hypothesized that model performance decreases over time due to changes over time in the patient population (e.g. different prevalence of diseases) and the arrival of new technologies in healthcare. This will outdate the model which will no longer fit the data well. Therefore we need to understand model behaviour over time, its effect on ICU rankings, and whether continuous updating of models over time can alleviate this problem. We aimed to gain a better understanding of the behaviour of prognostic model performance over time and the effect of this behaviour on quality of care assessment.

We monitored model performance of a logistic regression model and a binary classification tree in a prospective dataset of 12,143 consecutive admissions of 21 Dutch ICUs of patients aged 80 years and older. The logistic regression model is a parametric model in which the form of the relationship between predictors and outcome is pre-specified and the tree-model is a non-parametric model with little assumptions about this relationship. We used a technique called Statistical Process Control, which combine rigorous time series analysis and graphical data presentation to identify structural changes in a process. In our case, the “process” indicates model performance.

In chapter 4, we monitored the performance of a recalibrated SAPS-II model. Model performance was stable in terms of discriminative ability (AUC; Area Under the Curve) and accuracy (Brier score), but we observed a gradual deterioration of model calibration as measured by the SMR. Repeated recalibration of the model rendered the SMR stable again. After repeated recalibration the proportion of hospitals with SMR>1 (indicating worse quality of care than expected) and SMR<1 (indicating better quality of care than expected) changed from 15% versus 85% to 35% versus 65%. This means that quality of care was overestimated in part of the hospitals by using an outdated model.

In chapter 5, we monitored the performance of a classification tree. The tree-model was stable over time in terms of discriminative ability (AUC), but unstable in terms of accuracy (Brier score). Although the observed mortality was lower than predicted, the tree-model was still able to identify subgroups of patients with high risk of mortality. The proportions of patients assigned to the tree’s leaves (which contain subgroups of patients with similar characteristics) were not stable, however. In chapter 6, we monitored the SMR of the tree-model over time and compared it to that of the SAPS-II model. In contrary to the SMR of the SAPS-II model, the SMR of the tree-model was stable over time. Although this illustrates that the tree-model is less sensitive to changes in the patient population, it may be preferable to use a model that does pick up these changes, such that the model can be recalibrated when necessary.

Part 3: Role of prognostic models in individual decision-making

Prognostic models may have the potential to improve individual decision-making, but there is still considerable resistance against their use in clinical practice [23,24]. Trust is built by evidence about their clinical credibility, external validity (especially for their use in individual patients) [25] and clinical effectiveness. Clinical credibility may be enhanced by the incorporation of clinically relevant variables, such as daily degree of organ failure (rather than admission data only). Not much is known about models that provide a prediction on each day of stay and how they compare to clinicians. We aimed to gain a better understanding of clinicians’ subjective predictions, individual mathematical predictions, and how they compare to each other.
In chapter 7, we systematically reviewed literature on SOFA-based prognostic models of mortality and/or survival. We found that the SOFA score is competitive with admission models such as SAPS-II and APACHE-II while based on less physiologic parameters. Models combining information of severity of illness at admission and information on the course of illness during treatment achieved better performance. Two strategies were used to incorporate SOFA in the logistic regression model: 1) SOFA derivatives (e.g. maximum SOFA score) and 2) frequent sequential patterns of SOFA (e.g. a low score on day 2 followed by a high score on day 3).

Chapter 8 and chapter 9 describe the results of a prospective analysis of (patterns in) clinicians’ daily subjective predictions of survival probabilities in the ICU (i.e. expert opinion) and their predictive value. Although we found that nurses and physicians have good predictive ability, which is superior to the SAPS-II model, they do not always agree and are not always consistent over time. In chapter 10, we compared the validity of physicians’ predictions (i.e. expert opinion) to three types of temporal models based on the two strategies found in chapter 7: SAPS-II + SOFA patterns; SAPS-II + SOFA patterns + SOFA derivatives; SAPS-II + SOFA patterns + SOFA derivatives + expert opinion. The mathematical models without subjective information (i.e. expert opinion) had fair but worse discrimination than physicians, but better calibration. The models significantly improved by including subjective information and the combination of subjective and objective information was superior to either one alone.

Conclusion

The use of mathematical models for individual clinical decision-making is still very premature as their clinical credibility is low and evidence for their external validity is scarce. Both require much more consideration to enhance model acceptance both by physicians as by patients and carers in the nearby future.

Model calibration decreases over time due to changes in healthcare and patient populations. The use of outdated models for benchmarking generates inaccurate views of the delivered quality of care. Our findings signify the importance of continuous monitoring of risk models and subsequent recalibration at least when overall calibration is observed to be worsening. Statistical process control is a useful technique to enable the visualization of performance behaviour over time. Future work should address the required frequency for recalibrating models, various approaches for recalibration, and ways to determine the extent of influence that older observations should exert when recalibrating models in a dynamically changing environment.

Temporal models providing a prediction of mortality at each day of ICU stay may potentially be useful for support of individual clinical decisions. SOFA-based models have significant predictive value and are competitive with physicians, especially when combined with subjective information. Physicians have an unfair advantage over the models, however, as their views influence the final decisions they make. Nurses and physicians have good, comparable predictive ability, but their predictions are not consistent and they do not always agree about a patient’s prognosis. This is not necessarily bad, but stresses the importance of inter-collegial communication. More research is needed on how to improve the end-of-life decision-making process and the possible role of objective mathematical models.