European Journal of Internal Medicine xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

European Journal of Internal Medicine

journal homepage: www.elsevier.com/locate/ejim

Original Article

The clinical usefulness of prognostic prediction models in critical illness

Tim Baker [a,b,c,*], Martin Gerdin [a,d]

[a] Global Health—Health Systems & Policy, Department of Public Health Sciences, Karolinska Institutet, Stockholm, Sweden
[b] Perioperative Medicine & Intensive Care, Karolinska University Hospital, Stockholm, Sweden
[c] Department of Anaesthesia & Intensive Care, Queen Elizabeth Central Hospital, Blantyre, Malawi
[d] Department of Clinical Science, Innovation, and Technology, Karolinska Institutet, 171 77 Stockholm, Sweden

ARTICLE INFO

Article history:
Received 12 July 2017
Accepted 11 September 2017
Available online xxxx

Keywords:
Emergency medical services
Critical illness
Critical care
Prognosis
Decision support techniques
Hospital rapid response team

ABSTRACT

Critical illness is any immediately life-threatening disease or trauma and results in several million deaths globally every year. Responsive hospital systems for managing critical illness include quick and accurate identification of the critically ill patients. Prognostic prediction models are widely used for this aim. To be clinically useful, a model should have good predictive performance, often measured using discrimination and calibration. This is not sufficient though: a model also needs to be tested in the setting where it will be used, it should be user-friendly and should guide decision making and actions. The clinical usefulness and impact on patient outcomes of prediction models has not been greatly studied. The focus of research should shift from attempts to optimise the precision of models to real-world intervention studies to compare the performance of models and their impacts on outcomes.

1. Introduction

Health services throughout the world manage patients with a great spectrum of illness severity. From stable patients with mild conditions to those with serious conditions and critical illness. Critical illness is any immediately life-threatening disease or trauma and can affect anyone, irrespective of age, gender or underlying diagnosis [1]. Globally, critical illness results in several million deaths every year [2,3].

Systems for managing this spectrum are required for optimal care. The patients with life-threatening conditions require a responsive system that can provide rapid care with immediately available medicines, medical equipment and human resources. Quick and accurate identification of these critically ill patients is thus of paramount importance. This article aims to describe the concept of prognostic prediction models that have been developed for the identification of critical illness, and to discuss their clinical usefulness.

2. Clinical prediction models are different from risk adjustment models

A prediction model is an algorithm for estimating the probability of a specific outcome in an individual. A *prognostic* prediction model is a prediction model that is used in a *patient* to estimate the probability that a given outcome, for example mortality within a defined time frame, will occur in the future. The prognostic prediction model can be contrasted with the *diagnostic* prediction model, which is used to estimate the probability that an individual has a specific condition, for example type 2 diabetes.

The output of a prognostic prediction model is an estimated probability, and may be regarded as an estimate of disease *severity* when the outcome is death or a similar unfavourable event. In critical care, the main use of such models has been for risk adjustment when comparing outcomes and care across settings, for example between hospitals or of one hospital over time. Examples of such risk-adjustment models are the Acute Physiology, Age, Chronic Health Evaluation (APACHE), Simplified Acute Physiology Score (SAPS), and the Mortality Prediction Model (MPM) families of models [4–13].

We distinguish between these risk-adjustment models and clinical prediction models—models that are aimed primarily at aiding clinical decision making in the care of individual patients. Although there is no strict boundary between risk-adjustment and clinical prediction models, the aims of the models differ. Risk-adjustment models prioritise predictive performance among populations, and generally include many parameters—the latest versions of APACHE, SAPS, and MPM include at least 14 parameters. Clinical prediction models focus instead on ease of use for health workers with individual patients, especially if they are intended to be calculated manually bedside. They usually involve only a few parameters which may be weighted to produce a simple or binary score rather than an estimated probability—in some models a single deranged parameter is enough for a critical score [14,15].

* Corresponding author at: Global Health—Health Systems & Policy, Department of Public Health Sciences, Karolinska Institutet, Stockholm, Sweden.
E-mail address: tim.baker@ki.se (T. Baker).

http://dx.doi.org/10.1016/j.ejim.2017.09.012
0953-6205/

There are strong associations between patients' vital sign abnormalities and poor outcomes—it is well established that vital signs should be used in healthcare settings to monitor illness severity [15–19]. Much recent work has focused on designing clinical prediction models using these vital signs [20]. A recent systematic review identified 56 unique clinical prediction models to help detect acute illness [21]. Similarly, a systematic review for paediatric patients identified 33 different models [22]. The National Early Warning Score (NEWS) [23], a much used model, is based on respiratory rate, oxygen saturation, temperature, systolic blood pressure, pulse rate, and level of consciousness. Similarly, the newly-developed quick Sequential Organ Failure Assessment (qSOFA) score [24] uses respiratory rate, systolic blood pressure and level of consciousness.

3. Assessing model performance

When assessing the predictive performance of these models most studies have focused on discrimination—the ability to differentiate between individuals with and without the outcome [25]. For a binary outcome this can be measured using the area under a curve plotted for sensitivity versus 1—specificity, also called the receiver operating characteristics curve (AUROCC) [26]. The AUROCC ranges between 0 and 1 and indicates how well a model separates high risk patients from low risk patients. An AUROCC of 0.5 indicates that a model discriminates no better than chance alone, whereas values between 0.5 and 1 suggest discrimination better than chance. Values between 0 and 0.5 also suggests better discrimination than chance, but that the association is inverse.

Authors have interpreted similar discriminations in different ways. Smith et al. in their NEWS study from 2013 conclude that their model's AUROCCs of between 0.722 and 0.894 are “able to discriminate patients at risk” [27]. Kruisselbrink and colleagues in Uganda, interpret that an odds ratio of a high modified early warning score (MEWS) of 5.2 “enables identification of patients in need” [28]. In contrast, Wheeler and colleagues in Malawi found an AUROCC of 0.78 for their TOTAL model and concluded it is “not a clinically useful tool” [29].

Another measure of model performance is calibration—here defined as the association between the proportion of observed and predicted outcomes. For example, where mortality is the outcome, in a sample of 100 patients of which 10 patients died, did the model predict 10% mortality? Calibration can be fine-tuned, and assessed visually using calibration plots [30]. In such plots, it is common to plot the observed prevalence of the outcome across ten deciles of predicted probability.

A model's discrimination and calibration need not be correlated. A model may discriminate well, and have an AUROCC of >0.8, but calibrate poorly. This may be the case if a model is applied in a setting in which the prevalence of the outcome differs from the setting in which the model was developed, but the association between model parameters and outcome is largely the same [31]. Hence, whereas discrimination is insensitive to outcome prevalence, calibration is not, but can be regarded as a property of the model together with the particular cohort to which it was applied [32].

4. Clinical usefulness

A clinically useful model is one that can be used by health workers and that leads to improved care of the patient. To be clinically useful, the first requirement of a model is to have a good predictive performance—i.e. both good discrimination and calibration. Despite the substantial work that has gone into developing clinical predictive models for critical illness, predictive performance has never been perfect. This reflects the heterogeneity of critical illness, the complexity of biological systems and the limitations of vital signs. All clinicians know of patients who have severely deranged physiology and yet have a low risk of death in the immediate-term (for example hypoxia in COPD), and of patients with normal parameters but a high risk of death (for example a large stroke).

An over-reliance on prognostic models could even be harmful. A predictive value, defined as the percentage of correctly classified patients out of all patients, of 80%, for example, will misclassify one-in-five patients, and could result in both under-triage and missing patients with a high-risk of death and over-triage and aggressive management in patients with a low-risk of death. It must be remembered, however, that clinicians will also make incorrect decisions at times—i.e. the predictive value of clinicians' decisions without the help of models, will also not be perfect.

A good predictive performance is not enough to guarantee clinical usefulness [32]. An additional requirement is that the model is quick, easy, user-friendly and acceptable to health workers especially in clinical areas where human resources are low and time is short—for example in general wards of busy hospitals or in hospitals in low-resource settings. Too many parameters, complex or time-consuming parameters or parameters that require unavailable resources reduce a model's clinical usefulness.

Thirdly, a model needs to assist a clinician's decision making, which implies that the decision is difficult and that there are options to choose from. When the clinical suspicion is already very high, i.e. the decision is clear-cut, then the model will not add information. To illustrate this, consider a dichotomous decision, such as whether to call a medical emergency team. A model could recommend an action based on a score in relation to a dichotomised cut-off. If the model was applied in a cohort of very sick patients, such as in an intensive care unit, then the score will be higher than the cut-off for all patients, and using the model will just add complexity without assisting decisions.

And lastly, a model is only clinically useful if it leads to improved care, ideally measured by improved outcomes for the patients. This requires that the decision being assisted by the model is clinically important, and that the choices available to the health worker have an impact on outcomes. An ideal model would give specific guidance to the clinician (Fig. 1). Many models only indicate a risk level for a patient, and do not indicate what should be done. Others, such as a single parameter system tested by one of the authors (TB), couple deranged signs to specific guidance to the health worker [15,33].

Measures have been proposed to assess the clinical usefulness of models [30,34]. These measures include decision analytic techniques, such as decision curve analysis [35] whereby the model is assessed together with data about the actual decision taken by the clinicians, to evaluate whether using the models provided added benefit. In the extensive literature on clinical prediction models for the detection of critical illness, it is rare that the reported measures of discrimination, and sometimes calibration, are complemented by decision analytic techniques. The continuing focus on producing and refining prediction models seems to be based on the premises that the improved detection of critically ill patients will lead to clinical interventions and that these interventions improve patient outcome. These premises have scarcely been researched.

Some related work is research on the introduction of medical emergency teams or rapid response teams. These teams are summoned, often from the Intensive Care Unit, to the bedside of patients who have been identified by clinical prediction models. A Cochrane systematic review from 2007 included two randomised controlled trials investigating the implementation of a clinical prediction model coupled with the activation of medical emergency teams [36]. The two studies reported conflicting results, one showed no significant difference in the primary outcome which was a composite score of mortality and unfavourable clinical events whereas the other study did show a significant difference in mortality [37,38].

Other systematic reviews have indicated that “robust evidence to support [medical emergency teams'] effectiveness in reducing hospital mortality is lacking” [39], that “[m]oderate strength evidence … showed that [medical emergency teams] are associated with reduced rates of

Please cite this article as: Baker T, Gerdin M, The clinical usefulness of prognostic prediction models in critical illness, Eur J Intern Med (2017), http://dx.doi.org/10.1016/j.ejim.2017.09.012