It is clinically intuitive that physiological deterioration precedes critical illness, and there is growing research evidence to support this belief.1–6 The use of track and trigger early-warning systems to facilitate the early recognition of the acutely deteriorating patient outside the critical care area of a hospital is based on this premise. The first such scoring system, based on a single parameter trigger, was developed in Liverpool, Australia as part of the Medical Emergency Team (MET) concept.7 This was followed by the development of a range of similar systems that have entered clinical practice around the world.8,9 These incorporate a variety of physiological variables in scoring, and trigger algorithms that are often more complex than the original MET’s calling criteria. They use a wide range of weightings and cut-off points to develop a score designed to trigger activation of a response system. However, most parameters and cut-off points are still based on clinical intuition, rather than the use of rigorous derivation and validation methodologies.

Recent work casts doubt on the clinical utility of the systems currently in use. The MERIT study, a randomized, controlled trial by Hillman and colleagues,10 failed to demonstrate benefit partly because of the calling criteria’s sensitivity and specificity of <50%. Indeed, the majority of ‘deteriorating’ patients were not detected until <15 min before they suffered cardiac arrest, died or were admitted to an intensive care unit. The SOCCER study by Jacques and colleagues5 attempted to identify early and late predictors of adverse outcome and suggested that the failure of MERIT to demonstrate efficacy for METs may also be due to the current MET call criteria, which are actually late signs of deterioration.5 Two other studies have demonstrated the influence of relatively small changes in physiological cut-off points on improving the sensitivities and specificities of the MET criteria.11,12 More complex early-warning systems have not convincingly demonstrated that they are useful for the early detection of the deteriorating patient. When applied in an ideal mathematical way, these types of scores have reasonable sensitivity and specificity.13 However, in clinical practice, there are major concerns over their accuracy. A recent systematic review8 concluded that there was ‘little evidence of reliability, validity, and utility’, and that their ‘sensitivity was poor, which might be due in part to the nature of the physiology monitored or to the choice of trigger threshold’. Other work has also demonstrated inaccuracies in the calculation of scores by staff,14,15 and significant intra- and inter-rater reliability error.16 Unless scoring systems have appropriate sensitivity and specificity, and minimize errors associated with documentation and scoring, they will fail to identify patients who need additional care and will increase workload in circumstances where no intervention is required. There is a clear need to re-assess the use of early-warning scoring systems as they are currently used and to use a structured, scientific approach to their development and evaluation.
First, it is essential to identify which physiological variables are important as early identifiers of deterioration. This requires that study designs should ensure the collection of all potentially appropriate, routinely available items every time a measurement set is made. Restricting the data collection to objective variables is preferable, because using terms such as ‘worried’ as triggers has obvious problems with interpretation. Next, it is important to determine the correct cut-off values, and this requires studies to include the whole ‘at risk’ population (i.e. the denominator). The available literature appears to suggest that many existing scoring systems use the incorrect physiological variables (e.g. temperature),¹³⁻¹⁹ fail to use all of the appropriate ones (e.g. oxygen saturation),¹³⁻¹⁷⁻¹⁹ or use the wrong cut-off points (e.g. oxygen saturation).⁵¹⁻¹³⁻¹⁷⁻¹⁹

Having identified the important variables and cut off values, it is then important to derive and validate a scoring system de novo. This requires the use of derivation and validation cohort methodology and has been undertaken in only a few studies to date, usually in emergency departments or assessment units.¹⁷⁻¹⁹ Studies in these clinical settings have the obvious advantage that the identification of a uniform sampling point (e.g. on admission) for physiological variables is easy. Identifying appropriate sampling points for use in derivation and validation cohort methodology on general hospital wards is more challenging. Which value should one use? Which value should one use on admission? Which value is the most deranged? Which value is the most recent before the arrival of an outreach or MET?

In this edition of the British Journal of Anaesthesia, Duckitt and colleagues²⁰ describe a derivation and validation study of 4384 patients in a UK Medical Assessment Unit. They describe the values of physiological variables that were important as descriptors of in-hospital mortality (e.g. heart rate >101 beats min⁻¹, ventilatory frequency >19 bpm, oxygen saturation <96%, and systolic arterial pressure <100 mm Hg) and have applied weightings to them in a new scoring system, which they then tested in a separate validation cohort. The new scoring system has reasonable diagnostic accuracy and is more accurate than another commonly used scoring system. Higher scores are also associated with higher mortalities and longer hospital length of stay. However, before the Worthing system could be adopted into clinical practice in other areas of their hospital, or in other hospitals, further validation using other data sources is required. This poses a clinical dilemma as this study²⁰ used saturation values measured on air. It is questionable whether it is ethical to delay administering oxygen to, or perhaps worse still to remove it from, sick patients simply to perform an early-warning score. Patients on the steep curve of the oxygen dissociation curve, who are likely to be among those at greatest risk of an adverse outcome, may well suffer adversely if oxygen is withhold or removed. Ignoring the delivered oxygen concentration is equally problematic, because the treatment drives oxygen saturation towards a goal of 100%. Future studies will need to link oxygen saturation to delivered oxygen concentration.

Nevertheless, Duckitt and colleagues²⁰ are to be congratulated on taking a robust approach to their study, which adds considerably to the understanding in this field. Their work, and that of others,¹⁷⁻¹⁹ demonstrates that scoring systems of high accuracy can be developed de novo and may, after appropriate development and validation in varied patient groups, be able to replace the existing scores that have low or poor diagnostic accuracy.

Finally, to overcome the inaccuracies and miscalculations related to manual data collection, it may be necessary to adopt continuous vital signs monitoring,⁶ or electronic data management and scoring systems.²¹⁻²² Such systems should minimize intra- and interrater reliability error and facilitate the use of potentially more accurate discriminate functions that are less amenable to manual calculation error. However, to date, such systems have not demonstrated improved clinical or other outcomes, and further research is required.

In conclusion, the existing combinations of early-warning scoring systems and outreach/MET responses have failed to demonstrate improved outcomes in studies that attempt to eliminate bias.²³ This may be because of the lack of sensitivity and specificity of current systems to allow accurate detection of early critical illness. The use of scores that use parameters and cut-off points that are inappropriate may be unhelpful and should be replaced by ones that are derived using reliable methodologies and have high diagnostic accuracy. Further work is required to derive and validate such scores and to explore the role of electronic data capture and scoring systems. Until this work is completed, we should not accept that the existing early-warning scoring systems are necessarily any better at detecting the deteriorating patient than high-quality clinical assessment and judgement by appropriately skilled and experienced personnel.

Declaration of interest

G.S. is leading a team of researchers who have contributed to the development of an electronic vital signs data gathering system for use in general ward areas of hospital, and the team is currently investigating the function of early-warning scoring systems. Professor Smith’s wife is a creditor of the company that is developing the system.

B.H. Cuthbertson*  
Health Services Research Unit  
Health Sciences Building  
University of Aberdeen  
Aberdeen  
UK

G.B. Smith  
Department of Critical Care  
Portsmouth Hospitals NHS Trust and University of Bournemouth  
Portsmouth  
UK

*E-mail: b.h.cuthbertson@abdn.ac.uk
References

3. Goldhill DR, McNarry AF, Mandersloot G, McGinley A. A physiologically-based early warning score for ward patients: the association between score and outcome. Anaesthesia 2005; 60: 547–53