Clinical paper

Centile-based early warning scores derived from statistical distributions of vital signs∗

Lionel Tarassenkoa,∗, David A. Cliftona, Michael R. Pinskyb, Marilyn T. Hravnakb, John R. Woodsc, Peter J. Watkinsona,d

aUniversity of Oxford, UK
bUniversity of Pittsburgh Medical Centre, USA
cClarian Health Partners, Indianapolis, USA
dOxford Radcliffe Hospitals NHS Trust, UK

1. Introduction

In 1997, Morgan and colleagues described the first early warning score (EWS) system, designed to alert clinicians to deteriorating patients using aggregate weighted scoring of vital signs. Many variations of this scheme have since been published. Despite evidence that physiological instability precedes critical clinical deterioration, EWS systems have not been shown to improve patient outcomes.

Most EWS systems use weights and cut-off values for vital signs that are derived from expert opinion. We argue that clinical conjecture can be improved upon by statistical determination of the ranges of normality in the at-risk population, for each vital sign observation. In this paper we describe the development of a centile-based system derived from such information.

2. Methods

2.1. Data collection

A large dataset comprising 64,622 h of vital-sign data, acquired from 863 acutely ill in-hospital patients using bedside monitors, was used to investigate the statistical properties of the four main vital signs: heart rate (HR), respiratory rate (RR), peripheral arterial oxygen saturation (SpO2), and systolic blood pressure (SysBP). The data came from three clinical studies in the United Kingdom (UK) and United States (US) between 2004 and 2008.

The methodologies of these studies have previously been described. The CALMS1 study (Oxford, UK, OxREC No: C03.057) comprised a randomised controlled trial of the effects of mandated continuous physiological monitoring in both medical and surgical ward patients with a >5% risk of in-hospital death. Patients were
randomised to receive either mandated continuous 5-parameter physiological monitoring (for up to 72 h post-surgery or post acute medical admission), or to receive usual ward care. The Clarian Health study (Indianapolis, US, approved by the Clarian Methodist Hospital Institutional Review Board)\(^1\) was a prospective cohort study of the use of monitoring to detect cardio-respiratory instability in general medical/surgical patients receiving treatment on a progressive care unit in a teaching hospital. The University of Pittsburgh Medical Center (UPMC) study (Pittsburgh, US, approved by the UPMC Patient Safety Committee)\(^2\) was a quality improvement project for a 24-bed adult surgical trauma step-down unit in a level I trauma center hospital. In phase 1 (standard care), continuous vital-sign data were collected from bedside monitors over 8 weeks. Phase 3 was an 8-week intervention phase,\(^3\) during which nursing staff utilised a clinical algorithm to respond to alerts from a data fusion system (Visensia, OBS Medical, Abingdon, UK) connected to the bedside monitors. Phase 2 was a training phase for the nursing staff.

2.2. Statistics

Normalised histograms (unit area under the curve) and cumulative distribution functions were plotted for each physiological variable (HR, RR, \(\text{SpO}_2\) and SysBP), for each study and for the 3 datasets combined.

2.3. Data utilisation

A centile-based alerting system for the four variables was modelled using the aggregated database. The alerting system was constructed using the hypothesis that an EWS of 3 (which, in most systems, initiates a review of the patient) should be generated when a vital sign is below the 1st centile or above the 99th centile for that variable (for a double-sided distribution). We further considered that a score of 2 should correspond to the vital sign being between the 1st and 5th centiles or between the 95th and 99th centiles and that a score of 1 should correspond to the vital sign being between the 5th and 10th centiles or between the 90th and 99th centiles. (For \(\text{SpO}_2\), with a one-sided distribution starting at 100%, values above the 98th centile will give a score of 3, values between the 90th and 98th centiles a score of 2, and values between the 80th and 90th centiles a score of 1.)

2.4. Alert generation

The number of alerts which would be generated by our proposed EWS system was investigated, assuming that alerts occur whenever a score of 3 is assigned to a single variable and a score of \(\geq 4\) for the multivariate case.

To estimate the expected number of alerts generated for different frequencies of patient observation, the 4 vital signs in the database were evaluated using the proposed centile-based EWS at intervals of \(N\) hours, for \(N = 4, 8,\) and 12 h. The median of each vital sign was calculated over the 1-min interval centred on the observation time.

3. Results

The patient population characteristics in each study are shown in Table 1.

3.1. Vital sign characteristics

The distributions for each variable (see Table 2 for means and standard deviations) were very similar in all studies; acutely ill hospital patients in the UK and US have the same physiological characteristics. The vital sign data were aggregated together to produce an overall histogram for each variable. These histograms, computed from the 64,622 h of data, are shown in Fig. 1. The HR histogram shows small peaks at 60 and 70 beats/min, corresponding to the common settings for patients fitted with pacemakers.

3.2. Weightings and cut-offs estimated from vital-sign database

Histograms are estimates of the probability density function \(p(x)\) for the random variable \(x\). The cumulative distribution function (cdf), \(P(x)\), is the integral of \(p(x)\). The cdfs for each vital sign are shown in Fig. 2. The vertical lines on these plots allow the cut-off values to be determined for each vital sign. To take respiratory rate as an example, 1% of patients had a respiratory rate \(\leq 7\) breaths/min, 5% a rate \(\leq 10\) breaths/min, and 10% a rate \(\leq 13\) breaths/min. At the upper end, 90% of patients had a respiratory rate \(\leq 26\) breaths/min, 95% a rate \(\leq 29\) breaths/min and 99% a rate \(\leq 34\) breaths/min.

The information given in Fig. 2 can be converted to the tabular format used in track-and-trigger forms, as in Table 3.

3.3. Number of alerts

Table 4 shows the percentage of the patient population that would generate one or more alerts within \(N\) hours, when using our proposed centile-based EWS system with different frequencies of vital-sign observation. \(N\) can be thought of as the time interval during which a clinical team is on duty (e.g. \(N = 12\) for a 12-h shift) and \(M\) is the time interval between observations (e.g. \(M = 4\) for four-hourly observations). \(M\) is always less or equal to \(N\).

The percentage of patients that generate an alert decreases with decreasing frequency of observations, because each patient is being observed less often. When observations are made every 4 h during a 12-h shift, approximately 12% of the at-risk patients (1 in 8) would be expected to generate an alert during the shift.

This analysis allows us to estimate the resource required to attend to patients as a function of observation frequency. The program for performing this type of analysis with a user-defined EWS is available at http://www.robots.ox.ac.uk/~davidec/EWS. As the patients in our database were considered to be at high risk of clinical deterioration, the numbers in Table 4 may over-estimate the population alerting rates for the whole spectrum of hospitalised in-patients.

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Patient demographics for the CALMS1 (Oxford, UK), Clarian (Indianapolis, US) and UPMC (Pittsburgh, US) studies.</th>
</tr>
</thead>
<tbody>
<tr>
<td>CALMS1</td>
<td>Clarian</td>
</tr>
<tr>
<td>Number of patients</td>
<td>169</td>
</tr>
<tr>
<td>Age (mean ± SD)</td>
<td>71 ± 14</td>
</tr>
<tr>
<td>Age (median ± IQR)</td>
<td>74 ± 18</td>
</tr>
<tr>
<td>Sex (male)</td>
<td>97 (57.4%)</td>
</tr>
<tr>
<td>Medical</td>
<td>107 (53.3%)</td>
</tr>
<tr>
<td>Surgical</td>
<td>62 (36.4%)</td>
</tr>
<tr>
<td>Mortality</td>
<td>17 (10%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2</th>
<th>Vital sign means (standard deviations) from our database (CALMS1, Clarian and UPMC studies).</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heart rate</td>
<td>Breathing rate</td>
</tr>
<tr>
<td>JR (CALMS1)</td>
<td>85.0 (16.8)</td>
</tr>
<tr>
<td>Clarian</td>
<td>84.0 (18.6)</td>
</tr>
<tr>
<td>UPMC (P1)</td>
<td>83.1 (16.9)</td>
</tr>
<tr>
<td>UPMC (P3)</td>
<td>82.8 (17.0)</td>
</tr>
</tbody>
</table>

| Total | 84.2 (17.4) | 18.6 (5.3) | 96.0 (3.7) | 128.5 (21.4) |
4. Discussion

We have presented the ranges of values for the four main vital signs in at-risk hospitalised populations. We suggest that clinicians would want to be alerted when their patient strays outside the limits of normality defined by these ranges. Conversely, it seems reasonable that patients whose vital signs remain within the limits of normality should not score in an alerting system. Our EWS uses...
Table 3

Range of values for each weighting for the main vital signs (RR, HR, SysBP and SpO₂) in the proposed centile-based EWS derived from evidence base of 64,622 h of continuously recorded vital sign data from 863 patients.

<table>
<thead>
<tr>
<th>Score</th>
<th>3</th>
<th>2</th>
<th>1</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>BR</td>
<td>≤7</td>
<td>8–10</td>
<td>11–13</td>
<td>14–25</td>
<td>26–28</td>
<td>29–33</td>
<td>≥34</td>
</tr>
<tr>
<td>HR</td>
<td>≤50</td>
<td>51–58</td>
<td>59–63</td>
<td>64–104</td>
<td>105–112</td>
<td>113–127</td>
<td>≥128</td>
</tr>
<tr>
<td>SpO₂</td>
<td>≤84</td>
<td>85–90</td>
<td>91–93</td>
<td>≥94</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4

Percentage of patients that generate an alert (total score of 4 or more, or single variable with a score of 3) within a period of N hours with centile-based EWS, if vital signs are observed every M hours. The figure in brackets indicates the percentage of patients generating an alert as a result of a single vital sign scoring 3 (extreme value for a single parameter); the difference between the two figures indicates the percentage of patients generating an alert with a total score of 4 or more, but without a single vital sign scoring 3 (i.e. a combination of worsening vital signs without an extreme value in one or more parameters).

<table>
<thead>
<tr>
<th>Observation frequency (M)</th>
<th>Shift length (duration of observations – N)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4 h</td>
</tr>
<tr>
<td>Four-hourly</td>
<td>5.1%(4.1%)</td>
</tr>
<tr>
<td>Eight-hourly</td>
<td>N/A</td>
</tr>
<tr>
<td>Twelve-hourly</td>
<td>N/A</td>
</tr>
</tbody>
</table>

rather than relying solely on expert opinion, recent EWS systems10,14,15,18 (e.g. the viEWS system14) use a binary outcome (mortality/survival or admission/no admission to ICU). The cut-off values for each vital sign (measured at time T1) are set such that alerts are meant to predict the binary outcome at some later time T2 (often 24 h after T1). The prediction accuracy is assessed by comparing the areas-under-the-receiver-operator-characteristic (AUROC), improvements being achieved by adjusting the cut-offs for each vital sign.10,14 However, in any patient cohort used to create these EWS systems, some form of vital-sign monitoring will have been used (if only a traditional TPR chart). Many of the patients will have displayed vital sign abnormalities (identified by the chart), which led to an escalation of care at T1. Some of the patients in the “good outcome” group at T2 would have been in the “poor outcome” group but for the escalation of care triggered by the use of the chart at T1.

As shown in Fig. 3, there are 3 categories of patients at T1, but these are not known to the observer: patient will die regardless of intervention (“unsalvageable”); patient will survive, even without intervention (“stable”); and patient may be “salvageable” as a result of the intervention triggered by the observation of abnormal vital signs. Ideally, AUROC optimisation would be carried out using the 3 categories at T1 to set the EWS threshold for alerting at its optimal value, that which corresponds to the XX’ boundary in Fig. 3. If we could use XX’ to set the alerting threshold, an intervention would not be triggered for stable patients who would be fine in any case (category 1), but an intervention would be triggered for all others (categories 2 and 3), thereby maximising the number of salvageable patients who are salvaged. However, we cannot use XX’ as the threshold for alerting as the 3 categories on the left-hand side of Fig. 3 are not known.

For patients in the dataset retrospectively reviewed to create the outcome-based EWS system (the “training dataset”), abnormal vital-sign observations at T1 will have triggered an intervention that resulted in a proportion of the salvageable patients (category 2) being salvaged at T2 (category 2a), although not all of them were saved (category 2b). When the outcome-based EWS system is used prospectively on new patients, it will disadvantage category 2a patients (those who were previously “salvaged” following recognition by the vital sign monitoring system in use at the time the “training dataset” was acquired). This is because an EWS threshold based on the YY’ boundary will not generate an alert for these patients (since these patients in the “training dataset” belonged at T2 to the alive class) and the EWS system will therefore not prompt an intervention.

In summary, we argue that it is not valid to use AUROC optimisation to set the EWS threshold for making a decision at T1 (alert/no alert) based on classification (dead/alive) at T2, since this treats the two groups of patients below the YY’ boundary in the “training dataset” as the same (same classification at T2) and hence discriminates against category 2a patients. EWS systems which use outcome data at T2 to derive a predictive rule from vital-sign measurements at T1 are based on an approach originally developed for evaluating diagnostic tests.19 Such an approach is valid in the diagnostic setting as the disease (outcome) is either present or not at the time of testing.

the 10th and 90th centiles to define the no-score limits. A centile-based system provides a statistical basis for identifying vital signs outside the population limits of normality.

In comparison with existing EWS systems, our cut-off values differ most for RR and SysBP. Most acutely ill patients in hospital have elevated respiratory rates, as indicated by a mean value of 18.6 (SD 5.3) breaths/min (Table 2). In most current EWS, respiratory rates of 24 breaths/min, i.e. within one SD of the mean (broadly normal for the population), score highly.8,14–16 This may be because clinicians base “normal” values on those for young healthy populations, rather than older, hospitalised populations. SysBP shows the opposite behaviour. Several EWS systems do not score for hypertension <250 mmHg.7,14,15 Yet a patient with a SysBP of 190 mmHg is in the highest centile for blood pressure and so would be reviewed in our system. Hypertension, unlike hypotension, does not appear to be associated with imminent mortality,14,17 but this does not make it a poor marker of early clinical deterioration. If a patient is chronically hypertensive, the EWS system can be adjusted (at the first alert) to take account of this. If the hypertension is caused instead by an event leading to pain or distress, it may be an earlier marker of deterioration than when cardiovascular decompensation (with hypotension) occurs.

Fig. 3. Vital-sign measurements are made at time T1 when there are 3 categories of patients (stable, salvageable and unsalvageable) which are not known to the observer. There are only 2 outcomes at time T2 (dead or alive) separated by the classification boundary YY’.
There is a further problem associated with choosing mortality or ICU admission as the outcome. Recent UK guidelines recommend that EWS systems prompt immediate identification of those at risk of clinical deterioration. Focusing on vital sign values which are predictive of subsequent death/ICU admission is not necessarily the best way of finding the values which identify early deterioration, not least because delays in detecting clinical deterioration may increase morbidity without causing ICU admission or death. Unfortunately, there is no obvious binary outcome for early deterioration. Events such as escalation/no escalation of care are hard to define precisely, depend on clinical judgement and only take account of recognised deteriorations, making it impossible to set the EWS threshold for alerting on the basis of an “early deterioration outcome” at T1.

Our approach relies on different assumptions. Firstly, it assumes that the vital signs of normal patients (category 1) mainly lie close to the center of their probability distributions (i.e. they are the “most normal”), those of unsalvageable patients (category 3) at the extreme edges (“most abnormal”), and those of deteriorating but salvageable patients (category 2) somewhere between. This assumption ought not to be contentious, as it is the underlying principle of early warning scores based on vital-sign measurement. The vital-signal data from category 2a patients are included in the statistical model in the same way as the data from all other patients (avoiding the problem of mis-classification as in AUROC optimisation based on outcome at T2).

In setting the cut-off values for a score of 2 at the 5th and 95th centiles, we assume that the 10% most abnormal patients will score a 2 or 3 for at least one vital sign. To a first approximation, these 10% most abnormal patients consist of the 3.4% patients from categories 2b and 3 (given the 3.4% mortality rate in our dataset) and of the 6.6% of patients who survived but had abnormal vital signs. Most of the patients with a score of 2 for at least one vital sign must therefore belong to category 2a, since they are highly unlikely to be the stable category 1 patients.

Unless the vital signs of category 2a patients are under-represented in our dataset (and hence in the probability distributions), it seems likely that these patients will alert using the proposed centiles. Of course, the eventual recognition of abnormal physiology in these patients in our dataset will have led to intervention(s) and subsequent improvement in their vital signs. However, the most abnormal vital signs for these patients will have been recorded (every 5 s) in the hours preceding clinical intervention, and our dataset contains lengthy periods of vital signs gradually becoming more abnormal for patients who alerted and subsequently recovered.

Our system currently weighs low and high deviations equally. Whether this is optimal for detecting early deterioration requires further investigation. In the 24 h preceding death, cardiac arrest or unplanned ICU admission, low respiratory and heart rates are more prevalent than high values. This suggests that low rates should be given more weight. However, these low values may simply be markers of agonal events, rather than being of extra use in detecting early deterioration.

Our study is also limited in that it is based on retrospective analysis of high-risk hospitalised patients. However, the average values for each vital sign are remarkably similar to previously published values for hospitalised populations. The use of continuously recorded data from patient monitors has allowed us to estimate the distribution of physiological variables from thousands of hours of data. Although vital-signal data from patient monitors are unvalidated, the smoothness of our cumulative distribution functions suggests that artefactual measurements are relatively low in number (and randomly distributed).

Our databases did not include detailed information on urine output and conscious level, factors commonly included in EWS systems. As urine output is a continuous variable, its cut-off values could be set as we have described. Measures of conscious level, such as the Glasgow Coma Scale (GCS) or the AVPU scale, however, are not continuous, and the mapping between the 1/2/3 EWS scores and the GCS or AVPU scales may need to be non-linear. Clearly both urine output and conscious level could be added to our proposed system. Other EWS systems also include oxygen therapy in the SpO2 score. This factor was not recorded in our dataset and its use requires investigation.

Finally, an inherent weakness of all data-based EWS systems is the low numbers of data from young people used in model construction. If (as would be expected) the pattern of physiological deteriorations differs with age, younger patients may be disadvantaged as they are under-represented in study populations. Our initial work suggests that different cut-off values for low heart rate (bradycardia) should be used for younger adults. This topic requires further work, for all vital signs.

Despite these limitations, we believe that our methodology has produced an EWS system designed to alert the clinician to abnormal physiology in the at-risk population. As our cut-off values differ from those of previous systems, our proposed system will alert the clinician to either a different population or to a similar population at a different time.

5. Conclusion

Our centile-based statistical approach is in line with the original objectives of early warning scores, namely the identification of abnormal physiology at the time of the test. Observations are treated as being abnormal if they lie at the extremes of the distributions of vital signs acquired from representative sets of at-risk hospitalised patients. EWS systems based on such an approach have the potential to identify patients with abnormal vital signs and are more likely to trigger when presented with patients with redeemable morbidity or avoidable mortality (salvageable patients).

Whether the use of our system will result in improved patient outcomes needs to be tested. We are starting a stepped-wedge clinical trial gradually introducing an electronic version of our system (including temperature and AVPU) on the trauma wards in a medium-sized teaching hospital, which will enable us to measure its effect on patient outcomes.

Conflict of interest statement

No conflicts of interest to declare.

Acknowledgements

The work described in this paper was funded by the NIHR Biomedical Research Centre Programme. Dr David Clifton is supported by the Wellcome Trust and EPSRC under grant number WT 088877/Z/09/Z.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.resuscitation.2011.03.006.

References