Are observation selection methods important when comparing early warning score performance?

Dr. Stuart W Jarvis, PhD, Senior Research Associate, Centre for Healthcare Modelling and Informatics, University of Portsmouth, Portsmouth, UK and [Department of Health Sciences, University of York, York, UK](#)

Mrs. Caroline Kovacs, BSc, Research Assistant, Centre for Healthcare Modelling and Informatics, University of Portsmouth, Portsmouth, UK

Dr. Jim Briggs, BA, DPhil, Principal Lecturer & Director, Centre for Healthcare Modelling and Informatics, University of Portsmouth, Portsmouth, UK

Dr. Paul Meredith, PhD, Data Analyst, TEAMS centre, Portsmouth Hospitals NHS Trust, Portsmouth, UK

Dr. Paul E Schmidt, MRCP, B.Med.Sc, MBA, Consultant In Acute Medicine, Portsmouth Hospitals NHS Trust, Portsmouth UK

Dr. Peter I Featherstone, FRCP, Consultant In Acute Medicine, Portsmouth Hospitals NHS Trust, Portsmouth UK

Professor David R Prytherch, PhD, MIPEM, CSci, Clinical Scientist, TEAMS centre, Portsmouth Hospitals NHS Trust & Visiting Professor, Centre for Healthcare Modelling and Informatics, University of Portsmouth, Portsmouth, UK

Professor Gary B Smith, FRCA, FRCP, Visiting Professor, School of Health & Social Care, University of Bournemouth, Bournemouth, UK

Correspondence to:

Professor G B Smith, FRCA, FRCP,
Centre of Postgraduate Medical Research & Education (CoPMRE),
The School of Health & Social Care, Bournemouth University,
Royal London House, Christchurch Road, Bournemouth, Dorset BH1 3LT, United Kingdom

Tel: +44 (0) 1202 962782

Fax: +44 (0) 1202 962218

Email: gbsresearch@virginmedia.com

Keywords: Early warning score; Illness severity score; Vital signs; Monitoring; Risk

ABSTRACT

Introduction:

Sicker patients generally have more vital sign assessments, particularly immediately before an adverse outcome, and especially if the vital sign monitoring schedule is driven by an early warning score (EWS) value. This lack of independence could influence the measured discriminatory performance of an EWS.

Methods:

We used a population of 1,564,143 consecutive vital signs observation sets collected as a routine part of patients' care. We compared 35 published EWSs for their discrimination of the risk of death within 24 hours of an observation set using (1) all observations in our dataset, (2) one observation per patient care episode, chosen at random and (3) one observation per patient care episode, chosen as the closest to a randomly selected point in time in each episode. We compared the area under the ROC curve (AUROC) as a measure of discrimination for each of the 35 EWSs under each observation selection method and looked for changes in their rank order.

Results:

There were no significant changes in rank order of the EWSs based on AUROC between the different observation selection methods, except for one EWS that included age among its components. Whichever method of observation selection was used, the National Early Warning Score (NEWS) showed the highest discrimination of risk of death within 24 hours. AUROCs were higher when only one observation set was used per episode of care (significantly higher for many EWSs, including NEWS).

Conclusions:

Vital sign measurements can be treated as if they are independent – multiple observations can be used from each episode of care – when comparing the performance and ranking of EWSs, provided no EWS includes age.

INTRODUCTION

Several prior publications by our group have assessed the performance of the early warning scores (EWS) used to identify patients' severity of illness.¹⁻³ EWS systems allocate points in a weighted manner, based on the derangement of a predetermined set of patient vital signs variables (e.g., blood pressure, heart rate, breathing rate, temperature) from an arbitrarily agreed "normal" range. The points for each variable are summed and the total is used to inform a change in the patient's vital sign monitoring schedule and/or trigger a call for expert help at the bedside.

Our performance evaluations of EWS systems have often used all the observations sets from a sample of patient episodes and, therefore, contain multiple vital sign observation sets from the same patient episode in the analysis.^{2,3} Multiple observations may be within 24 hours of death (or another adverse outcome). We have considered an EWS to be better than another if it has a significantly ($p < 0.05$) higher area under the ROC curve (AUROC⁴, a measure of discrimination). Sicker patients generally have more vital sign assessments, particularly immediately before an adverse outcome, and especially if the vital sign monitoring schedule is driven by an EWS value. A previous review of our manuscripts have suggested that this lack of independence of the data points in the sample data sets may influence the measured discriminatory performance of an EWS. By extension, it is possible that an EWS that appears significantly better than another when all observations are used may appear significantly worse if only one observation was used from each episode.

EWS systems are implemented clinically as if vital sign measurements and derived EWS values are independent. EWS escalation decisions are generally binary. For example, an EWS value of 4 might result in no clinical intervention, whereas a value of 5 might require both a change in vital signs frequency and an assessment by a doctor (irrespective of the fact that the previous EWS was 0 or 4). Consequently, it is the extent of derangement of physiology at any given time, and not the degree of abnormality of any previous measurements, that determines actions taken based on the EWS score.

One study by our group⁵ has suggested that treating vital signs and derived EWS values as independent may be reasonable, as an alternative technique of using one randomly chosen observation set per episode did not significantly affect discrimination of the combined outcome of cardiac arrest, unanticipated ICU admission or death within 24 hours. In this study,⁵ as with others,¹⁻³ the ability of the EWSs to discriminate the risk of a range of adverse outcome has been compared using the AUROC.⁴ The use of multiple observation sets per episode has the potential to bias the AUROC as episodes with more observations may disproportionately influence the AUROC compared to those with fewer observations.

The aim of this study was to determine whether a lack of independence between data points when sampling patient observations might significantly change the ranking of EWS systems by their AUROC (i.e. lead to one EWS having significantly higher AUROC than another under one method of choosing observations, but significantly lower AUROC than the other under another method). We compared the performance of EWSs using three methods of observation selection: (1) all observations, (2) one randomly chosen observation set per episode, and (3) one observation set per episode based on choosing a random point in time within each episode.

METHOD

This research falls within local research ethics committee approval (08/02/1394) from the Isle of Wight, Portsmouth and South East Hampshire Research Ethics Committee.

Vital signs test results database and its development

We constructed a database of vital signs collected from all adult (≥ 16 years old) patients admitted to Portsmouth Hospitals NHS Trust on or after 25/05/2011 and discharged on or before 31/12/2012. We excluded data from patients discharged alive on the day of admission. We also excluded data from episodes in which no observations were recorded during the final 24 hours (this was to reduce the numbers of patients in the dataset that might be on a recognised end of life pathway, where routine observations are halted). Vital signs data were recorded in real-time at the bedside using handheld electronic equipment running the VitalPAC software.⁶ Each vital signs measurement set contained: pulse rate, breathing rate, systolic and diastolic blood pressure, temperature, S_pO_2 , the inspired gas (e.g., oxygen or air) at the time of S_pO_2 measurement, and the patient's conscious level. Conscious level was recorded as alert (A), responds to voice (V), responds to pain (P) or unresponsive (U).

Within-episode dependence of vital signs observations

We tested for the level of within-episode dependence of the eight recorded vital signs and age at admission by calculating the intraclass correlation (ICC) for each using the episode identifier as the grouping variable.

Outcome

The outcome studied was death within hospital within 24 hours of a vital signs dataset. Where it occurred, patient death was identified from the patient administration system (PAS). Therefore, multiple observations within a single episode of care might be followed by death within 24 hours.

Observation selection methods

We used three different methods to select observations. Firstly, we used all the observations in our dataset (i.e., for each patient care episode, all observations taken during that episode were used).

Secondly, we made 10,000 samples of observation sets, each sample being constructed by selecting one observation set at random from every patient care episode (i.e., so each observation set in an episode had an equal chance of being selected in each sample). Thirdly, we made 10,000 samples containing one observation set per episode. The observations were chosen by first randomly selecting a time during every episode in every sample, and then selecting the observation set closest to it.

In the case of the time-based selection method, we encountered problems due to inaccurate recording of patient admission and discharge times (some episodes had observations recorded either before admission or after discharge). Observations are time-stamped automatically as they are entered at the patient's bedside, whereas admissions and discharges are recorded independently on PAS and are less likely to be accurate.² Therefore we used the first and last observation dates and times, as the basis of determining the time period from which to choose observations. To avoid biasing against selection of the first and last observation sets, we added to the beginning and end of the selection period a length of time equal to half of the mean time between observation sets for that patient episode.

Selection of early warning scores and assessment of performance

We compared 35 published EWSs – 33 previously compared by Smith et al.,^{1,2} plus the cardiac arrest risk triage (CART) model⁷ and the Centiles EWS.⁸ The full list of EWSs evaluated is listed in Table S1 of the supplementary information. EWS performance - the ability to discriminate risk of death within 24 hours of an observation set - was assessed by calculating the area under the ROC curve for each of the 35 EWSs under each of the three observation selection methods. When using all observations, we calculated a 95% confidence interval for the AUROCs using the methods set out by DeLong et al.⁹ When using 10,000 sample sets, we calculated an AUROC for each sample set and reported the mean AUROC and the 2.5 and 97.5 centiles of the AUROCs as the 95% confidence interval.

Effects of age

For those EWSs that include age, our analyses used age at admission in the EWS calculation. This is therefore constant throughout an episode and clearly not independent between observations. To study whether the inclusion of age in an EWS changed its ranking under different observation selection methods, we repeated the above analysis, removing the age component from EWSs that included it.

Assessment of significance of changes in EWS ranking

We compared EWS performance under each of the selection methods by ranking them by their AUROC. We used DeLong's method for paired AUROCs⁹ to determine significance ($p < 0.05$) of differences between AUROCs calculated on all observations. To determine whether two EWSs had significantly different AUROCs when using randomly chosen observations, the difference in their AUROCs was calculated for each of the 10,000 sample sets and the 95% confidence interval of the difference was estimated using the 2.5 and 97.5 centiles. The difference was considered significant if the 95% confidence interval for the difference between AUROCs did not include zero.

When ranking EWSs by AUROC, an EWS may change position either due to a change in its own performance or a change in the other EWSs (or both). When an EWS changed its rank, we made a list of EWSs that lay between its positions under the two observation methods. When an EWS rank increased (or decreased), the increase (or decrease) was considered significant if one or more of the listed EWSs that had significantly greater (or smaller) AUROC under the first selection method had significantly smaller (or greater) AUROC under the second method. For example, if EWS 'X' changed its rank from 10th when using all observations to 2nd when using one randomly selected observation set per episode, we made a list of all the EWSs that had rank between 2nd and 10th under both methods of observation selection. For the change in rank of EWS 'X' to be considered significant, we required that at least one of the listed EWSs had a significantly higher AUROC than that of EWS 'X' when all observations were used and a significantly lower AUROC than EWS 'X' when selecting only one observation set per episode.

Data analysis tools

All data manipulation was performed using Microsoft® Visual FoxPro 9.0. All analyses were undertaken in R version 3.02.¹⁰

RESULTS

In the study period, there were 64,285 episodes of care with admission on or after 25/05/2011 and discharge on or before 31/12/2012, where the patient was aged ≥ 16 , the patient was not discharged alive on the day of admission and one or more observations were taken during the last 24 hours of the stay. Associated with these episodes of care were 1,395,941 observation sets (mean 21.7 observation sets per episode). Of these episodes, 30,723 (48%) were for male patients and the mean age at admission was 61.8 years (standard deviation 20.4 years). 1697 (2.6%) of the episodes ended in death. For the two methods using one randomly chosen observation set per episode, there were 64,285 observations sets available for analysis. Table 1 shows the summary data regarding incidence of death within 24 hours and the mean NEWS values for observations chosen by each of the selection methods.

There is dependence between vital signs observations taken from the same episode, as demonstrated by the ICCs for each of the vital signs (Table 2, ICCs are significantly greater than 0). For age at admission, which cannot change during an episode and is therefore fully dependent, the ICC is 1.

The AUROC values, measuring the discrimination of the 35 EWSs, show large differences between the best and worst performing (Figure 1). There are also changes in rank depending on the observation selection method, but only one of these (for the EWS described by Bakir¹¹) is significant. Most variation in rank occurs towards the middle of the rankings, where differences in AUROC between similarly ranked EWSs are very low. When the age component is removed from those EWSs that include it, there are fewer changes in rank and none are significant (Figure 2).

The AUROC for any given EWS is lowest when all observations in the dataset are used, highest when one random observation is selected per episode and intermediate (though towards the top of the range) when one random observation is selected for each episode based on choosing a random time point in the patient's stay (Figures 1 and 2). The differences between AUROCs when using all observations sets and using one random observation per episode are significant (95% confidence intervals do not overlap – Table S2, supplementary information) for all EWSs except

the Centiles EWS and the EWS published by Allen. A lower number (16 of the 35 EWSs, including NEWS) have significantly higher AUROC when using one observation set per episode based on choosing a random point in the patient's stay compared to using all observation sets. There are no significant differences between AUROCs for the two methods of selecting one observation set from each episode of care.

DISCUSSION

For the three observation selection methods studies, there were no significant changes in the rank of EWSs by their AUROCs except for EWSs that included age. Overall, the findings of this research suggest that vital signs and derived EWS values for EWSs that do not include age can be treated as if they were independent (even though the ICCs demonstrate that there is within-episode dependence). Therefore use of multiple observation sets from a single episode in assessing the performance of EWS systems does not appear to bias the ranking of EWSs, as long as no EWS includes age. However, in reality only a minority of EWS systems does. In common with the findings of an earlier study,² we also found that the UK National Early Warning Score (NEWS)¹³ offers the best discrimination among the EWSs studied, irrespective of the method of observation selection.

The method of selecting observations only has a significant effect on the ranking of Bakir's EWS¹¹, which includes age. Non-significant changes were also observed for CART⁷ and Subbe MEWS¹⁴, both of which also include age. These variations can be explained by considering the distribution of ages in observations in our data. We know that older people are more likely to be in hospital – we have a progressively higher number of episodes of care for older people within our dataset than we do for younger people. Elderly patients are also likely to have longer stays (and therefore more observations). Age is a good predictor of risk of death, but only when the score assigned for age in an EWS varies between observations. When all observations are used, the combined effect of older patients having both more episodes and more observations per episode skews the dataset to contain many observations for older patients. This means that many observation sets – although representing differing risks – will score the same for age. The usefulness of age as a discriminator of risk is therefore diminished. When only one observation is used from each episode of care, fewer observations come from older patients as one source of bias – more observations per episode for older patients – is removed. There are then fewer observations that score the same for age and age becomes more useful as a discriminator of risk of death. The effect is greatest for Bakir's EWS as it has the greatest weightings for age. This finding is relevant to the use of EWSs in a hospital – age at admission is not useful for discriminating changes in risk during an episode

of care, but it is useful for discriminating risk between two episodes of care in which the patients have different ages.

Even for EWSs that do not include age, choosing only one observation per episode results in a higher AUROC (significantly higher for many EWSs – Table S2, supplementary information) than when multiple observation sets from a single episode are used. The AUROC values⁴ determined in this study represent the probability that any randomly chosen observation followed within 24 hours by death will have a higher EWS score than any randomly chosen observation not followed by death within 24 hours. When all observations are used, some pairs of observations (one followed by death within 24 hours and one not followed by death within 24 hours) will be for the same episode. A patient may have a high EWS score, receive an intervention that reduces the EWS score (though it may still be high) and then die.¹² If this process takes 24 hours or more then scores closer to death may be lower than those earlier in the episode, resulting in a lower AUROC when multiple observations are used from each episode of care and potentially reducing the mean difference between scores for observations that are and are not followed within 24 hours by death. This is reflected in the mean differences in NEWS (taken as an example – similar comments apply to other EWSs) scores between observations followed by death within 24 hours and those not followed by death within 24 hours under each selection method (Table 1). The difference is greatest when observations are selected at random, slightly lower when they are selected based on a random point in time and lowest when all observations are used.

There are differences in the incidence of death within 24 hours of selected observation sets, depending on the method of selection (Table 1). The differences between all and random observations may be explained by the variation in the number of observations per episode. When using all observations, 7637 observations from a total of 1,395,941 (0.55%) are followed by death within 24 hours. Some 328 episodes have all of their observations within 24 hours of death. Therefore, when using one randomly chosen observation per episode, a minimum of 328 selected observations from the 64,285 episodes (0.51%) are followed by death within 24 hours. A further 1369 episodes have some, but not all, observations within 24 hours of death. Random selection will choose, on average, an additional 268 observations from those 1369 episodes that have death

within 24 hours (based on the probability for each of those episodes of selecting an observation within 24 hours of death). This gives a total (on average) of 596 observations from 64,285 (0.93%) having death within 24 hours – a notable increase compared to the value of 0.55% when all observations are used. When choosing observations based on a random time within a patient's stay, 257 observations (on average) are followed by death within 24 hours from the 1369 episodes – a slightly lower number than when choosing observations at random.

When choosing observations based on the generation of random numbers, there is a possibility that a particular sample could be randomly atypical and give a different ranking of EWSs compared to another. To minimise such effects, we calculated the mean AUROC for each EWS from the AUROCs for 10,000 samples of the dataset. To satisfy ourselves that this was a sufficiently large number of sample sets, we repeated the process using another 10,000 sample sets starting from a different random seed in R's random number generator. These produced EWS rankings that were identical to those presented in this paper and with only small changes in reported AUROCs (well within the 95% confidence intervals).

This large study has several strengths. Of particular note is that all vital signs variables in the 1,564,143 sets were collected simultaneously in a standardised manner as part of the routine clinical process, and were stored electronically. Another is the use of many random sample sets (10,000) when studying the effects of random observation selection. Weaknesses include the fact that we excluded admissions that were admitted directly to critical care areas of the hospital, and included patients with 'Do Not Attempt Cardiopulmonary Resuscitation' decisions, although we took steps to exclude patients on a recognised end of life pathway. Also, for those EWSs that include age, our analyses used age at admission in the EWS calculation and we made no adjustment if the patient had a birthday during the admission. In practice, few patients experience a birthday during their admission and, even if they do, even fewer will move into a new age band that incurs more points than that applied on admission. Consequently, our a priori decision to use only age at admission is unlikely to have underestimated the early warning score in many cases, and therefore unlikely to affect the EWS performance significantly. Finally, as this has been a single centre evaluation, our findings require validation.

CONCLUSIONS

Using multiple observations from each episode of care does not significantly change the ranking of EWSs compared to using only one observation from each episode, as long as no EWS includes age. This is in spite of observed dependence between vital signs observations collected during the same episode of care.

The method of observation selection can affect the AUROCs recorded – higher AUROCs (significantly higher for many EWSs) are recorded when only one observation is used from each episode. For EWSs that include age, the observation selection method can be important in determining the rank order of EWSs.

NEWS performs better than the other EWSs in discriminating risk of death within 24 hours of an observation set, irrespective of the method used to select the observations. This supports the findings of earlier studies in which multiple observations were used from each episode of care.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the efforts of the medical, nursing and administrative staff at Portsmouth Hospitals NHS Trust who collected the data used in this study. Dr Stuart Jarvis had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

FUNDING

None

REFERENCES

1. Smith GB, Prytherch DR, Schmidt PE, Featherstone PI. Review and performance evaluation of aggregate weighted 'track and trigger' systems. *Resuscitation*. 2008;77:170-179.
2. Smith GB, Prytherch DR, Meredith P, Schmidt PE, Featherstone PI. The ability of the National Early Warning Score (NEWS) to discriminate patients at risk of early cardiac arrest, unanticipated intensive care unit admission, and death. *Resuscitation*. 2013;84:465-470.
3. Prytherch DR, Smith GB, Schmidt PE, Featherstone PI. ViEWS - towards a national Early Warning Score for detecting adult inpatient deterioration. *Resuscitation* 2010;81:932–937.
4. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982;143:29-36.
5. Badriyah T, Briggs JS, Meredith P, et al. Decision-tree early warning score (DTEWS) validates the design of the National Early Warning Score (NEWS). *Resuscitation* 2014;85:418-423.
6. Smith GB, Prytherch DR, Schmidt P, et al. Hospital-wide physiological surveillance - a new approach to the early identification and management of the sick patient. *Resuscitation* 2006;71:19-29.
7. Churpek MM, Yuen TC, Park SY, Meltzer DO, Hall JB, Edelson DP. Derivation of a cardiac arrest prediction model using ward vital signs*. *Crit Care Med* 2012;40:2102-2108.
8. Tarassenko L, Clifton D, Pinsky M, Hravnak M, Woods J, Watkinson P. Centile-based early warning scores derived from statistical distributions of vital signs. *Resuscitation* 2011; 82:1013-1018
9. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988; 44:837–845.
10. R Core Team 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org/> (accessed 16/08/2014)
11. Bakir A, Duckitt R, Buxton-Thomas R. A simple physiological scoring system for medical in-patients derived by modeling hospital mortality data. Poster presentation Intensive Care Society State of the Art Meeting. London: Intensive Care Society 2005.

12. Murray A, Kellet J, Huang W, Woodworth S, Wang F. Trajectories of the averaged abbreviated VitalPAC early warning score (AbEWS) and clinical course of 44,531 consecutive admissions hospitalized for acute medical illness. *Resuscitation* 2014;85:544-548.
13. National Early Warning Score (NEWS): Standardising the assessment of acute-illness severity in the NHS. Report of a working party. Royal College of Physicians, London, 2012.
14. Subbe CP, Kruger M, Rutherford P, Gemmel L. Validation of a modified Early Warning Score in medical admissions. *QJM* 2001;94:521-6.

LEGENDS FOR FIGURES:

Figure 1

Rank of early warning scores' discrimination (as measured by AUROC value, indicated on each bar) of risk of death within 24 hours of an observation set. Highest ranked EWSs are at the top. Un-shaded bars indicate an EWS that includes age. Lines through end of bars indicate 95% confidence interval for the AUROC. Lines linking bars indicate change in rank; black lines indicate a significant change.

Figure 2

Rank of early warning scores' discrimination (as measured by AUROC value, indicated on each bar) of risk of death within 24 hours of an observation set when age components of EWSs are excluded. Highest ranked EWSs are at the top. Un-shaded bars indicate an EWS that originally included age. Lines through end of bars indicate 95% confidence interval for the AUROC. Lines linking bars indicate change in rank; black lines indicate a significant change.

Table 1: Summary observations, incidence of death within 24 hours and mean NEWS values for observations chosen by each of the selection methods.

Dataset	Method of observation selection		
	All observations	One observation per episode of patient care, chosen at random (mean values from 10,000 sample sets)	One observation per episode of patient care, chosen as nearest to a randomly selected time in the patient's stay (mean values from 10,000 sample sets)
Number of observations used	1,395,941	64,285	64,285
Number of observations followed by death within 24 h	7637	596	585
% of observations followed by death within 24 h (95% confidence interval)	0.55 (0.53 – 0.56)	0.92 (0.89 – 0.97)	0.91 (0.87-0.95)
Mean NEWS value for observations not followed by death within 24 h	1.76	1.41	1.32
Mean NEWS value for observations followed by death within 24 h	6.74	6.85	6.64
Difference in mean NEWS values (observations followed by death - observations not followed by death within 24 h)	4.98	5.44	5.32

Table 2: Intraclass correlation, using episode identifier as the grouping variable, for the measurements included in the EWSs studied and the outcome variable (death within 24 hours of an observation set).

Variable	Intraclass correlation (95% CI)
Pulse rate	0.55 (0.53 - 0.56)
Breathing rate	0.30 (0.28 - 0.31)
Systolic blood pressure	0.48 (0.46 - 0.50)
Diastolic blood pressure	0.40 (0.39 - 0.42)
Temperature	0.21 (0.20 - 0.23)
SpO ₂	0.38 (0.37 - 0.40)
Inspired gas	0.44 (0.43 - 0.46)
Conscious level	0.17 (0.16 - 0.18)
Age at admission	1.00 (1.00 - 1.00)
Death within 24 hours	0.36 (0.35 – 0.38)