Observation chart design features affect the detection of patient deterioration: a systematic experimental evaluation

Melany J. Christofidis, Andrew Hill, Mark S. Horswill & Marcus O. Watson

Accepted for publication 24 August 2015

Abstract

Aim. To systematically evaluate the impact of several design features on chart-users’ detection of patient deterioration on observation charts with early-warning scoring-systems. Background. Research has shown that observation chart design affects the speed and accuracy with which abnormal observations are detected. However, little is known about the contribution of individual design features to these effects. Design. A $2 \times 2 \times 2 \times 2$ mixed factorial design, with data-recording format (drawn dots vs. written numbers), scoring-system integration (integrated colour-based system vs. non-integrated tabular system) and scoring-row placement (grouped vs. separate) varied within-participants and scores (present vs. absent) varied between-participants by random assignment. Methods. 205 novice chart-users, tested between March 2011–March 2014, completed 64 trials where they saw real patient data presented on an observation chart. Each participant saw eight cases (four containing abnormal observations) on each of eight designs (which represented a factorial combination of the within-participants variables). On each trial, they assessed whether any of the observations were physiologically abnormal, or whether all observations were normal. Response times and error rates were recorded for each design. Results. Participants responded faster (scores present and absent) and made fewer errors (scores absent) using drawn-dot (vs. written-number) observations and an integrated colour-based (vs. non-integrated tabular) scoring-system. Participants responded faster using grouped (vs. separate) scoring-rows when scores were absent, but separate scoring-rows when scores were present. Conclusion. Our findings suggest that several individual design features can affect novice chart-users’ ability to detect patient deterioration. More broadly, the study further demonstrates the need to evaluate chart designs empirically.

Keywords: design, deterioration, human factors, nursing, observation chart, subjective judgements, systematic
Why is this research needed?

- Recently, researchers and clinicians have developed new paper-based observation charts that are specifically designed to make abnormal observations easier for chart-users to detect.
- However, these chart designers have not directly assessed the effects of individual chart design features on the detection of patient deterioration.
- In the absence of objective evidence, there is no way to ascertain which design options represent best practice.

What are the key findings?

- Chart-users detected patient deterioration faster and more accurately using designs with a drawn-dot data-recording format (as opposed to written numbers) and an integrated colour-based scoring-system (rather than a non-integrated tabular one).
- In recent years, chart designs with these features have been shown to consistently out-perform other Australian observation charts in similar user-performance experiments.

How should the findings be used to influence policy/practice/research/education?

- As the first study to evaluate several features systematically, the findings suggest that future observation charts should adopt drawn-dot observations and integrated colour-based scoring-systems.
- The results also illustrate that chart designs should be evaluated objectively, through behavioural experimentation or alternative techniques that yield unbiased evidence.

Introduction

Inevitably, some patients will experience physiological deterioration while in hospital. Early recognition of the deteriorating patient is essential because delayed or missed recognition can result in adverse events including respiratory or cardiac arrest, unplanned admission to intensive care and even unexpected death (Franklin & Mathew 1994, Goldhill et al. 1999, Hillman et al. 2001). Given that deranged vital signs can signal deterioration as early as 48 hours before an adverse event (Franklin & Mathew 1994, Goldhill et al. 1999, Hillman et al. 2001, Endacott et al. 2007), one promising avenue for improving early recognition is to develop patient charts specifically designed to make abnormal observations easier for chart-users (including the least experienced nurses and doctors) to detect.

In recent years, clinicians and researchers in Australia and the UK have created new charts with this precise objective in mind and have employed several techniques to examine the effects of chart design on the detection of patient deterioration, including: prospective before-and-after controlled intervention trials (Mitchell et al. 2010); comparative clinical evaluations (Chatterjee et al. 2005, Elliott et al. 2014); and behavioural experiments (Preece et al. 2012a, Christofidis et al. 2013, Fung et al. 2014). In each of these studies, the performance of chart-users (including nurses) was compared across two or more charts and, in almost all cases, designs that included early-warning scoring-systems yielded the best results. On charts of this type, each value in a set of observations can be scored according to its degree of deviation from the normal range and these scores totalled to obtain an ‘early-warning score’ that summarizes the patient’s overall physical condition and can be used to trigger appropriate clinical actions (Prytherch et al. 2005, Lawson & Peate 2009).

In experimental studies that compared multiple charts with early-warning scoring-systems, two designs consistently yielded the fastest and most accurate identification of abnormal observations (Preece et al. 2012a, Christofidis et al. 2013) and both were versions of Horswill et al.’s (2010) Adult Deterioration Detection System (or ADDS) chart. This chart was designed by a multi-disciplinary team of human factors specialists and clinicians who took into account a wide range of usability considerations (Horswill et al. 2010, Preece et al. 2013). However, the precise reasons for its superior performance (and potential avenues for further improvement) remain unclear, because several design features varied unsystematically between it and the charts with which it was compared. For example, unlike some other charts with early-warning scoring-systems, the ADDS incorporates separate (vs. overlapping) blood pressure and heart rate graphs, drawn-dot observations (vs. written numbers), an integrated colour-based scoring-system (vs. a non-integrated tabular system) and scoring-rows grouped together at the bottom of the page (vs. presented separately, immediately below the corresponding vital sign data). A more recent experimental study has demonstrated that abnormal blood pressure and heart rate observations can be detected more quickly and accurately when these two vital signs are plotted separately, especially on charts with an integrated colour-based early-warning scoring-system (Christofidis et al. 2014). However, no empirical study to date has directly assessed the effects of other indi-
vidual observation chart design features on the detection of patient deterioration.

Background

Before supplying a new medical device to the market, manufacturers must obtain empirical data to support their claims about its safety and performance (TGA 2011). However, when paper-based observation charts are designed (or re-designed), this level of evidence-based accountability is seldom demanded despite comparable potential risks to patient safety. Instead, the efficacy of patient charts is typically assessed only via subjective judgements made by the health professionals who designed them and their colleagues (Chatterjee et al. 2005, Preece et al. 2012a). Consequently, observation chart designs (Preece et al. 2013) and health professionals’ perceptions of good design (Preece et al. 2012b), can vary considerably between locations. In the absence of objective evidence, however, there is no way to ascertain which design options represent best practice (Preece et al. 2012b).

The traditional subjective approach to chart development is inherently risky, as mounting research evidence suggests that health professionals’ preferences for particular chart features are not always consistent with objective performance data (Preece et al. 2012b). For instance, in a recent survey study, most health professionals reported that they preferred, and found it easier to detect patient deterioration, when blood pressure and heart rate were plotted together on the same graph (Preece et al. 2010). However, these opinions are at odds with more recent objective data. In Christofidis et al.’s (2014) experiment, overlapping blood pressure, and heart rate plots actually impeded recognition of abnormal vital signs by experienced nurses and novice chart-users alike, slowing them down and increasing their error rates. It has been suggested that performance-preference dissociations like this arise due to the inordinate influence of extraneous factors, such as familiarity and aesthetics, on people’s judgements and preferences (Andre & Wickens 1995). Given that such dissociations occur, it is possible that charts designed and endorsed on the basis of subjective judgements have contributed to documented failures (Franklin & Mathew 1994, Goldhill et al. 1999, Endacott et al. 2007) by hospital staff to record observations correctly and to detect or anticipate deterioration.

Indeed, the results of another two experimental studies suggest that poor design decisions may have potentially catastrophic consequences (Preece et al. 2012b, Christofidis et al. 2013). In these studies, participants were asked to detect abnormalities among vital sign observations presented on six observation charts of varying design quality, including four charts used in Australian hospitals and two versions of the ADDS chart, which had been designed as a more ‘user friendly’ alternative (Horswill et al. 2010, Preece et al. 2013). Both novice chart-users (Preece et al. 2012a) and health professionals (Preece et al. 2012a, Christofidis et al. 2013) made the least errors and responded fastest when using ADDS charts. These effects even held for clinicians who had prior clinical experience with one of the other charts used in the experiment, or a similar design (Christofidis et al. 2013). In fact, compared with the ADDS charts, the worst-performing design yielded up to 5.4 times as many errors by nurses and doctors who were experienced with a similar chart (Christofidis et al. 2013). As well as illustrating the dangers of poor design, these findings suggest that clinical experience alone may not be enough to overcome design deficiencies.

Given that improved observation charts could potentially deliver substantial patient safety gains, it is crucial that we develop a clear and thorough understanding of how precisely their design can be optimized. However, in past studies comparing the detection of deterioration across two or more charts (Chatterjee et al. 2005, Mitchell et al. 2010, Preece et al. 2012a, Christofidis et al. 2013), the unique contributions of specific design features to the outcomes were unclear, because the charts varied unsystematically on more than one dimension. For instance, we cannot infer that every design feature included in the ADDS chart positively contributed to its superior performance (Preece et al. 2012a, Christofidis et al. 2013). Rather, there may be room for further improvement and, in some cases, health professionals’ subjective preferences might still lead to better detection of patient deterioration. After all, even human factors-based chart design involves opinion-based compromises between competing design considerations (Preece et al. 2013). Hence, without systematic and objective comparisons, the efficacy of individual design features cannot be determined.

The study

Aims

This study aimed to systematically evaluate three design features that vary across Australasian charts with early-warning scoring-systems (Preece et al. 2013). Specifically, we manipulated data-recording format (drawn dots vs. written numbers), scoring-system integration (integrated colour-based system vs. non-integrated tabular system) and scoring-row placement (grouped vs. separate). For each of
these design features, the first listed alternative had been incorporated into the ADDS chart (Horswill et al. 2010, Preece et al. 2013), which was designed as part of a national initiative to develop a standardized adult general observation form (ACSQHC 2009). Using a similar methodology to prior experimental studies (Preece et al. 2012a, Christofidis et al. 2013, 2014), we evaluated each feature by testing charts-users’ ability to recognize abnormal observations on eight chart designs representing a factorial combination of these alternatives. Consistent with recent indicative findings (Chatterjee et al. 2005, Mitchell et al. 2010, Preece et al. 2012a, Christofidis et al. 2013, Fung et al. 2014) and the human factors-based design choices made in the development of the ADDS chart (Horswill et al. 2010, Preece et al. 2013), we predicted that chart-users would be faster and more accurate when using chart designs with: drawn-dot observations (Hypothesis 1); an integrated colour-based scoring-system (Hypothesis 2); and grouped scoring-rows (Hypothesis 3).

Design

The study employed a $2 \times 2 \times 2 \times 2$ mixed factorial design with data-recording format, scoring-system integration and scoring-row placement varied within-participants. In addition, the presence vs. absence of scores (i.e. overall early-warning scores and the scores for individual vital signs from which they are derived) was varied between-participants (see Scores for details and rationale). The dependent measures were response time and error rate.

Patient data

To ensure content validity, 64 cases of genuine de-identified patient data, each spanning 13 consecutive time-points, were used in the study. The cases, which were collected from several Australian hospitals, included data for the 10 parameters included in the ADDS chart (Horswill et al. 2010): respiratory rate, oxygen delivery, oxygen saturation, systolic and diastolic blood pressure, heart rate, temperature, four hour urine output, consciousness and pain. Half of the cases contained only normal observations (Table 1 for normal ranges: ACT Health 2011) and the others each included one abnormal observation: a derangement in oxygen saturation (eight hypoxic cases), systolic blood pressure (four hypotensive and four hypertensive cases), heart rate (four bradycardic and four tachycardic cases) or temperature (four hypothermic and four febrile cases).

The original data were only modified if either: (a) a vital sign remained abnormal for more than one time-point (excess abnormal data-points were shifted into the normal range); or (b) a data-point was missing (a plausible value was extrapolated or interpolated). Most of the cases (75%) had been used in previous studies employing a similar experimental paradigm (Preece et al. 2012a, Christofidis et al. 2013).

### Table 1 Vital sign normal ranges used in the experiment (table adapted from Preece et al. 2012a).

<table>
<thead>
<tr>
<th>Vital sign</th>
<th>Normal range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Respiratory rate</td>
<td>9-20 breaths per minute</td>
</tr>
<tr>
<td>Oxygen delivery</td>
<td>Patient is receiving oxygen at ≤1 l per minute</td>
</tr>
<tr>
<td>Oxygen saturation</td>
<td>93-100%</td>
</tr>
<tr>
<td>Systolic blood pressure</td>
<td>100-160 mmHg</td>
</tr>
<tr>
<td>Heart rate</td>
<td>50-100 beats per minute</td>
</tr>
<tr>
<td>Temperature</td>
<td>36.1-37.9°</td>
</tr>
<tr>
<td>Four hour urine output</td>
<td>120-799 ml</td>
</tr>
<tr>
<td>Consciousness</td>
<td>Patient is classified as being alert</td>
</tr>
<tr>
<td>Pain</td>
<td>Patient is in no pain</td>
</tr>
</tbody>
</table>

Observation chart designs

The eight observation charts created for this study, which were based on a version of the ADDS chart (Horswill et al. 2010), represented a factorial combination of two options for each of three design features, namely: (1) data-recording format (drawn dots vs. written numbers); (2) scoring-system integration (integrated colour-based system vs. non-integrated tabular system); and (3) scoring-row placement (grouped vs. separate) (Figure 1). Apart from these manipulations, the charts were identical. The designs were created and each set of patient data plotted onto each design, using Adobe InDesign CS5.5 (Adobe Systems Incorporated, 2011).

Scores

In real-world clinical situations, chart-users interpret observation charts that are in different states of completion. Sometimes, all vital sign data, individual vital sign scores and early-warning scores to date will already be present before a particular clinician picks up the chart. In other cases, some or all of the scores will be missing, either because compliance with the scoring-system is less than 100% (Odell et al. 2009), or because the nurse is in the process of recording the vital signs and has yet to complete the scoring. It is not necessarily the case that the same design options would be beneficial in all circumstances. Therefore, to obtain results generalizable to a broader range of real-world clinical situations, we manipulated whether or not scores were provided to participants.
Prior to testing, participants were assigned to one of two conditions using a random sequence generated by Microsoft Excel 2011: (1) ‘scores present’, where all charts had real scores recorded on them \((n = 102)\); or (2) ‘scores absent’, where all charts contained uninformative fillers (the letter ‘U’) in place of the real scores \((n = 103)\) (Figure 2). These fillers prevented the presence vs. absence of scores from being confounded with the absence vs. presence of blank scoring-rows. To account for this manipulation, the task instructions (see below) informed participants in the scores absent condition that ‘U’ was an abbreviation for ‘unrecorded’.

### Participants

Given that initial decisions about deteriorating patients are often made by relatively inexperienced nurses and doctors (Endacott et al. 2010), the present study focussed on novice performance. Power analysis using G*Power 3.1.9.2 (Faul et al. 2007, with ‘ANOVA: Repeated measures, between

---

**Figure 1** Four examples of chart designs used in the study, with: (a) an integrated colour-based scoring-system and grouped scoring-rows; (b) an integrated colour-based scoring-system and separate scoring-rows; (c) a non-integrated tabular scoring-system and grouped scoring-rows and (d) a non-integrated tabular scoring-system and separate scoring-rows. Each example includes either drawn-dot (a and d) or written-number (b and c) observations. The remaining four designs were identical, except that each used the alternative data-recording format option.
factors’ as the statistical test) indicated that a minimum sample of 180 participants was necessary to detect medium-sized effects (partial $\eta^2 = 0.06$) with 95% power and alpha set at 0.05, assuming a .85 correlation between repeated measures. A convenience sample of 205 novice chart-users, recruited from a Brisbane university (QLD, Australia), received psychology course credit for participating. Only individuals with no prior hospital chart experience were eligible, to ensure that no particular design option could be advantaged by participants’ previous chart-related preferences or experiences. In our prior experiments addressing observation chart design (Preece et al. 2012a, Christofidis et al. 2013, 2014), samples of naïve participants (recruited through the psychology research participation scheme) and health professionals consistently yielded very similar patterns of results across charts. Therefore, we reasoned that there would be no additional value in including a group of non-naïve novices, such as medical or nursing students.

After participating in the experiment, participants were excluded if they answered one or more items incorrectly in the postexperiment multiple-choice examination (see Data collection) or their overall error rate exceeded 50% (Figure 2). This was to ensure that, in the final sample, failure to understand the training instructions or retain the key information could not provide an alternative explanation for the results. Nevertheless, the overall patterns of results reported below remained unchanged when statistical analyses were re-run with these participants included.
Data collection

Participants were recruited and tested between March 2011–March 2014. Each participant was trained and tested individually in a quiet room. After completing a demographic questionnaire, participants watched a series of training videos that explained: (a) the ten vital signs included in the chart and their normal ranges; (b) track-and-trigger systems; and (c) how to use each chart design (presented in a different random order for each participant).

Next, the key concepts and vital sign normal ranges were tested with a 10-item multiple-choice examination. Participants scoring below 100% studied a summary and retook the examination until they answered everything correctly. A final video explained the experiment and indicated that responses and response times would be recorded.

Using a similar methodology to previous studies (Preece et al. 2012a, Christofidis et al. 2013), participants completed 64 experimental trials where they were presented with a patient chart containing a different case of patient data. For each participant, cases were randomly assigned to charts with the constraint that each design was assigned four normal and four abnormal cases (comprising derangements in oxygen saturation, systolic blood pressure, heart rate and temperature). To prevent order effects, trials were presented in a different random order for each participant.

In each trial, a chart appeared on a computer monitor and the participant responded by clicking on a green 'normal' button at the bottom of the screen (to indicate that all observations were normal) or on the appropriate vital sign graphing area (to indicate an abnormality). SuperLab experimental software (Cedrus Corporation, 2007) was used to present the images and to record the responses and response times (in milliseconds) for each trial. After completing all 64 trials, participants re-sat the multiple-choice examination.

---

**Figure 1** Continued.
Ethical considerations

This study was granted ethical approval in accordance with the review processes of the university ethics committees.

Data analysis

For each trial, the response was coded as ‘correct’ if the participant clicked on the appropriate area of the screen, identifying an abnormal vital sign or classifying a normal case as normal. Each participant’s average response time and error rate (percentage of incorrect responses) were calculated for each design. Statistical analyses were performed using IBM SPSS 21.0 (IBM Corp., Armonk, NY, USA) with statistical significance set at $\alpha = 0.05$. Separate mixed-design (data-recording format $\times$ scoring-system integration $\times$ scoring-row placement $\times$ scores) analyses of variance (ANOVAS) were conducted on response times and error rates, with $\eta^2$ as the measure of effect size (Howell 1997). T-tests were used to follow-up significant interactions, with Cohen’s $d$ as the effect size measure (Cohen 1992).

Results

Participant characteristics

Table 2 presents participant characteristics for the final sample of 188.

Response time

Analysis of the response time data revealed a significant main effect of data-recording format, $F(1, 186) = 82.05$, $P < 0.001$, $\eta^2 = 0.27$, qualified by a significant data-recording format $\times$ scores interaction, $F(1, 186) = 38.56$, $P < 0.001$,  

---

**Figure 1** Continued.
Participants for whom scores were absent responded 2.24 seconds faster (CI 1.76-2.72) using drawn-dot (vs. written-number) observations, \( t(1, 94) = -9.21, P < 0.001, \) Cohen’s \( d = -0.55 \) and participants with access to scores responded 0.42 seconds faster (CI 0.10-0.74), \( t(1, 92) = -2.58, P < 0.05, \) Cohen’s \( d = -0.13. \)

We also found a significant main effect of scoring-system integration, \( F(1, 186) = 195.83, P < 0.001, \) \( \eta^2 = 0.41 \), qualified by a significant interaction with scores, \( F(1, 186) = 96.90, P < 0.001, \) \( \eta^2 = 0.20 \) (Figure 3b). Participants for whom scores were absent responded 3.94 seconds faster (CI 3.40-4.48) using an integrated colour-based (vs. tabular) system, \( t(1, 94) = -14.52, P < 0.001, \) Cohen’s \( d = -0.95 \) and participants with access to scores responded 0.69 seconds faster (CI 0.32-1.06), \( t(1, 92) = -3.68, P < 0.001, \) Cohen’s \( d = -0.22. \)

Although there was no significant main effect of scoring-row placement, \( F(1, 186) = 0.01, P = 0.941, \) there was a significant interaction with scores, \( F(1, 186) = 13.60, P < 0.001, \) \( \eta^2 = 0.07 \) (Figure 3c). Participants for whom scores were absent responded 0.62 seconds faster (CI 0.14-1.09) using grouped (vs. separate) scoring-rows, \( t(1, 92) = -2.58, P < 0.05, \) Cohen’s \( d = -0.15. \) However, participants with access to scores responded 0.59 seconds faster (CI 0.15-1.04) using separate (vs. grouped) scoring-rows, \( t(1, 92) = 2.64, P < 0.05, \) Cohen’s \( d = 0.18. \)

Further, there was a significant scoring-system integration \( \times \) scoring-row placement interaction, \( F(1, 186) = 16.82, P < 0.001, \) \( \eta^2 = 0.08 \) (Figure 3d). Participants responded
2.89 seconds faster (CI 2.38-3.39) using an integrated colour-based (vs. tabular) system when scoring-rows were grouped \( t(1, 187) = -11.23, P < 0.001 \), Cohen’s \( d = -0.55 \) and 1.78 seconds faster (CI 1.32-2.23) when scoring-rows were separate, \( t(1, 187) = -7.70, P < 0.001 \), Cohen’s \( d = -0.32 \).

In addition, there was a main effect of scores, indicating that participants for whom scores were present (vs. absent) responded faster overall, \( F(1, 186) = 194.80, P < 0.001 \), \( \eta^2 = 0.52 \). However, this effect was also qualified by the interactions with data-recording format, scoring-system integration and scoring-row placement outlined above.

### Error rate

The ANOVA on error rate data revealed a significant main effect of data-recording format, \( F(1, 186) = 14.88, P < 0.001 \), \( \eta^2 = 0.07 \), again qualified by a significant data-recording format \( \times \) scores interaction, \( F(1, 186) = 6.36, P < 0.05 \), \( \eta^2 = 0.03 \) (Figure 4a). Participants for whom scores were absent made 2.57% fewer errors (CI 1.19-3.94) using drawn dots (vs. written numbers), \( t(1, 94) = -3.70, P < 0.001 \), Cohen’s \( d = -0.27 \). However, for participants with access to scores, there was no effect of data-recording format (\( P > 0.05 \)).

Once again, there was a significant main effect of scoring-system integration, \( F(1, 186) = 7.66, P < 0.05 \), \( \eta^2 = 0.04 \), qualified by a significant interaction with scores, \( F(1, 186) = 6.02, P < 0.05 \), \( \eta^2 = 0.03 \) (Figure 4b). Participants for whom scores were absent made 2.24% fewer errors (CI 0.75-3.73) using an integrated colour-based (vs. tabular) system, \( t(1, 94) = -2.98, P < 0.05 \), Cohen’s \( d = -0.23 \). However, this effect was not significant for participants with access to scores (\( P > 0.05 \)).

For error rate, scoring-row placement yielded no significant main effect or interaction with scores (\( P > 0.05 \); Figure 4c). However, as with response time, there was a significant scoring-system integration \( \times \) scoring-row placement interaction, \( F(1, 186) = 5.29, P < 0.05 \), \( \eta^2 = 0.03 \) (Figure 4d). Participants made 2.13% fewer errors (CI 1.01-3.25) using an integrated colour-based (vs. tabular) system when scoring-rows were grouped, \( t(1, 187) = -3.75, P < 0.001 \), Cohen’s \( d = -0.22 \). However, this effect was not significant when scoring-rows were separate (\( P > 0.05 \)).

### Error rate

![Figure 3](image-url)  
Figure 3: Response times for detecting abnormal observations, arranged by: (a) data-recording format and scores; (b) scoring-system integration and scores; (c) scoring-row placement and scores; and (d) scoring-system integration and scoring-row placement. Error bars indicate standard errors. Significant differences between adjacent bars are marked with an asterisk.
Again, there was a main effect of scores: participants for whom scores were present (vs. absent) made fewer errors overall, $F(1, 186) = 51.99, P < 0.001, \eta^2 = 0.22$. However, this effect was also qualified by the interactions with data-recording format and scoring-system integration reported above.

Discussion

This is the first study to evaluate systematically several observation chart design features to assess their contributions to the detection of patient deterioration. In support of Hypothesis 1, participants responded significantly faster (whether scores were present or absent on their charts) and made significantly fewer errors (if scores were absent) using drawn-dot (vs. written-number) observations. This suggests that, in a range of real-world contexts, drawn-dot observations may yield faster detection of abnormal vital signs and they may also prevent errors in some circumstances. In contrast, we found no evidence of any advantage for written-number observations, supporting existing indicative findings (Chatterjee et al. 2005, Fung et al. 2014). Considering this pattern of results, we argue that paper-based observation charts should use drawn-dot observations.

Our findings are consistent with the proposal that drawn-dot observations eliminated a potential source of unwanted workload, thus freeing chart-users’ cognitive resources for higher level tasks (Gerhardt-Powals 1996) by preventing the mental processing that numerical observations might have triggered (e.g. automatically reading the numbers and/or comparing them with clinical criteria stored in memory). In so doing, the use of drawn-dots also ensured that the task of searching for abnormal observations was not unnecessarily data-driven, potentially reducing the time that chart-users spent assimilating raw vital sign data (Gerhardt-Powals 1996). This interpretation is even more compelling when one considers that, just like the drawn-dot observations, the written numbers used in the experiment were presented graphically (as ‘quasi-graphs’; Preece et al. 2009). Hence, participants did not need to read the numbers to determine whether or not any particular observation was normal or abnormal, or to observe trends in the data. This contrasts with the many charts in clinical use that present observations as numbers written in a single row or column for each vital sign (Preece et al. 2012a, Christofidis et al. 2013). Indeed, charts featuring tabulated observations have yielded markedly slower response times and higher error rates (by both experienced clinicians and novice chart-users) in similar experimental studies (Preece et al. 2012a, Christofidis et al. 2013). Furthermore, written-number observations are arguably even more redundant considering that measurement error and transient variability (due to perturbations, natural steady-state variability, or clinicians’ technique; Reisner et al. 2012) can cause vital signs to fluctuate substantially over time.
Consistent with Hypothesis 2, participants were significantly faster (whether scores were present or absent) and significantly more accurate (absent scores only) when using an integrated colour-based (vs. non-integrated tabular) scoring-system. This finding has practical implications for chart-users’ efficiency: regardless of whether scores are recorded or not, an integrated colour-based system should lead to faster recognition of patient deterioration and, in some circumstances, fewer errors. Further, the study yielded no evidence of any circumstance where a non-integrated system would be advantageous. Given these results, we suggest that charts should also utilise colour-based, rather than tabular, scoring-systems. We propose that the presence of an integrated colour-based scoring-system automated chart-users’ unwanted workload by reducing the need for mental comparisons and unnecessary thinking (Gerhardt-Powals 1996). That is, participants did not need to consider normal ranges listed in a look-up table or held in memory. Instead, they could use the colour cues embedded in the graphs to identify criterion breaches rapidly; hence, the system also eliminated any need for the detection of abnormal observations to be a time-consuming, data-driven task (Gerhardt-Powals 1996).

The results relating to Hypothesis 3 were more mixed. The effect of scoring-row placement was confined to the response time data and differed in direction depending on whether scores were present or absent. Participants without access to scores were significantly faster when using charts that had scoring-rows grouped together at the bottom of the page (vs. separate), as predicted. However, when scores were present, charts with separate scoring-rows outperformed those with grouped rows. These findings should be read in conjunction with the results of a recent experimental study which found that participants were faster at determining and recording early-warning scores when the scoring-rows were separate, rather than grouped (Christofidis et al. 2015). If chart-related protocols are adhered to and all observations are scored, then the results of the present study also suggest, contrary to Hypothesis 3, that separate scoring-rows may be preferable. Hence, the optimal arrangement of scoring-rows may depend on the clinical context and compliance culture and we can make no overarching recommendation.

Interestingly, there was a significant interaction between scoring-system integration and scoring-row placement for both response time and error rate. Deconstruction of these interactions indicated that, irrespective of whether the chart design featured grouped or separate scoring-rows, participants performed better (in terms of response time, or both accuracy and response time) when the chart incorporated an integrated colour-based scoring-system. This suggests that the benefits of integrated colour-based scoring-systems are relatively robust to alternative scoring-row placements.

Unsurprisingly, participants were also faster and more accurate overall when early-warning scores were present (vs. absent), suggesting that they do assist chart-users to recognize deterioration. However, it should be noted that all of the scores recorded on charts in this study were accurate, which is not always the case in real clinical contexts (Christofidis et al. 2015).

The superior performance of the drawn-dot observations and integrated colour-based scoring-system validates several recommendations, based on cognitive engineering principles (Gerhardt-Powals 1996, Horswill et al. 2010), made in a systematic evaluation of Australasian observation charts (Preece et al. 2013). These recommendations also guided the design of the ADDS chart (Horswill et al. 2010), which has consistently out-performed other Australian observation charts in user-performance experiments similar to the present study (Preece et al. 2012a, Christofidis et al. 2014).

The results of the present experiment also have implications for the interpretation of previous research comparing the efficacy of observation chart designs. For example, Mitchell et al. (2010) re-designed an observation chart to include several potentially user-friendly features (e.g., quasi-graphs and a colour-coded aggregate weighted scoring-system) and conducted a prospective before-and-after intervention trial where the revised chart out-performed its predecessor on several clinical outcome measures (e.g. fewer unexpected ICU admissions and deaths). However, the contribution of specific design elements cannot be assumed because the study compared charts that varied on multiple dimensions, and implementation of the re-designed chart was accompanied by changes in vital sign monitoring policy and substantial education (Mitchell et al. 2010). Indeed, in subsequent empirical studies, Mitchell et al.’s design yielded more errors and slower response times compared with the ADDS chart, among both novice chart-users (Preece et al. 2012a) and health professionals (Preece et al. 2012a, Christofidis et al. 2013), including those trained and experienced in its use (Christofidis et al. 2013). The present findings suggest that this may be partially attributable to Mitchell et al.’s use of written-number observations (rather than drawn dots) for most vital signs, while the results of another recent experimental study suggest that plotting blood pressure and heart rate together on the same axes may also have compromised usability (Christofidis et al. 2014).
Limitations

As with our previous behavioural experiments (Preece et al. 2012a, Christofidis et al. 2013, 2014), we have not directly demonstrated that the results generalize to real-world settings. However, given that participants were not subject to the external pressures and distractions experienced by doctors and nurses in practice, it is plausible that the between-charts differences in response times and error rates would be larger in genuine clinical environments, where the impact of poor design on cognitive load would be more crucial (Preece et al. 2012a).

To maximize experimental control, we only recruited naïve participants. Consequently, we cannot, strictly speaking, generalize our results to experienced chart-users. However, these findings will still almost certainly apply to health professionals because: (a) samples of novices, nurses and doctors have consistently produced similar patterns of results across charts in our previous experimental studies (Horswill et al. 2010, Preece et al. 2012a,b, Christofidis et al. 2014); and (b) the effects of improved chart design on the detection of abnormal observations have been shown to outweigh health professionals’ prior chart experience (Christofidis et al. 2013). Given that initial decisions about deteriorating patients are often made by inexperienced doctors and nurses (Endacott et al. 2010), the inclusion of novices was important from a pragmatic perspective: observation charts must provide effective support for health professionals of all levels, especially the least experienced. Furthermore, in the future, all clinicians will effectively be novices in relation to paper-based charts once they are used exclusively as the backup for electronic systems (Christofidis et al. 2014).

Conclusion

Our findings suggest that chart design features have a substantial impact on chart-users’ ability to recognize patient deterioration. More importantly, they further illustrate the need to objectively evaluate the efficacy of observation chart designs. In sum, we suggest that, rather than relying on chart designers’ subjective judgements, or clinical trials with limited experimental control, new designs should also be evaluated objectively, through behavioural experimentation or alternative techniques that yield unbiased evidence (Preece et al. 2012a,b, Christofidis et al. 2013, 2014). Subsequent clinical studies can then focus on broader issues, such as chart utility post-implementation (e.g. Chatterjee et al. 2005, Mitchell et al. 2010, Elliott et al. 2011, 2014, Bunkenborg et al. 2014, Kyriacos et al. 2015) and subjective user experiences (e.g. Elliott et al. 2015). Like manufacturers of medical devices (Therapeutic Goods Administration 2011), chart designers should also be required to provide objective data to support their claims.

Acknowledgements

We thank Lynette Adams and David Collard for providing an Adobe InDesign CS5.5 copy of the Q-ADDS chart. We also thank staff from the participating university for accommodating and supporting this research project, and the students who volunteered to participate.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. The study was conducted as part of the first author’s PhD project, which is supported by an Australian Postgraduate Award and a Queensland Government Smart Futures Scholarship.

Conflict of interest

No conflict of interest has been declared by the authors.

Author contributions

All authors have agreed on the final version and meet at least one of the following criteria [recommended by the ICMJE (http://www.icmje.org/recommendations/)]:

- substantial contributions to conception and design, acquisition of data, or analysis and interpretation of data;
- drafting the article or revising it critically for important intellectual content.

References


The Journal of Advanced Nursing (JAN) is an international, peer-reviewed, scientific journal. JAN contributes to the advancement of evidence-based nursing, midwifery and health care by disseminating high quality research and scholarship of contemporary relevance and with potential to advance knowledge for practice, education, management or policy. JAN publishes research reviews, original research reports and methodological and theoretical papers.

For further information, please visit JAN on the Wiley Online Library website: www.wileyonlinelibrary.com/journal/jan

Reasons to publish your work in JAN:

- High-impact forum: the world’s most cited nursing journal, with an Impact Factor of 1·527 – ranked 14/101 in the 2012 ISI Journal Citation Reports © (Nursing (Social Science)).
- Most read nursing journal in the world: over 3 million articles downloaded online per year and accessible in over 10,000 libraries worldwide (including over 3,500 in developing countries with free or low cost access).
- Fast and easy online submission: online submission at http://mc.manuscriptcentral.com/jan.
- Positive publishing experience: rapid double-blind peer review with constructive feedback.
- Rapid online publication in five weeks: average time from final manuscript arriving in production to online publication.
- Online Open: the option to pay to make your article freely and openly accessible to non-subscribers upon publication on Wiley Online Library, as well as the option to deposit the article in your own or your funding agency’s preferred archive (e.g. PubMed).