APPLIED COGNITIVE PSYCHOLOGY
Appl. Cognit. Psychol. **24**: 77–89 (2010)
Published online 19 December 2008 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/acp.1545

When a Graph is Poorer Than 100 Words: A Comparison of Computerised Natural Language Generation, Human Generated Descriptions and Graphical Displays in Neonatal Intensive Care

MARIAN VAN DER MEULEN[1†,‡], ROBERT H. LOGIE[1*], YVONNE FREER[2], CINDY SYKES[2], NEIL MCINTOSH[2] and JIM HUNTER[3]

[1]*Human Cognitive Neuroscience-Psychology, University of Edinburgh, UK*
[2]*Simpson Centre for Reproductive Health, Edinburgh Royal Infirmary, UK*
[3]*Department of Computing Science, University of Aberdeen, UK*

SUMMARY

Volunteer staff from a Neonatal Intensive Care Unit (NICU) were presented with sets of anonymised physiological data recorded over approximately 45 minute periods from former patients. Staff were asked to select medical/nursing actions appropriate for each of the patients whose data were displayed. Data were shown in one of three conditions (a) as multiple line graphs similar to those commonly shown on the ward, or as textual descriptions generated by (b) expert medical/nursing staff or (c) computerised natural language generation (NLG). An overall advantage was found for the human generated text, but NLG resulted in decisions that were at least as good as those for the graphical displays with which staff were familiar. It is suggested that NLG might offer a viable automated approach to removing noise and artefacts in real, complex and dynamic data sets, thereby reducing visual complexity and mental workload, and enhancing decision-making particularly for inexperienced staff. Copyright © 2008 John Wiley & Sons, Ltd.

It is widely accepted that presenting data in the form of pictures, graphs or diagrams can enhance data comprehension, decision-making and communication of information about the data: 'A picture is worth a thousand words' (e.g. Carney & Levin, 2002; Roth & Bowen, 2003; Tory & Möller, 2004)). The mechanism behind this effect is thought to be through inducing cognitive processes such as visual chunking, mental imagery and parallel processing (Winn, 1994), and as an external aid to reduce demands on human memory (Card, Mackinlay, & Schneiderman, 1999), or to assist mental integration of complex data (e.g. Ratwani, Trafton, & Boehm-Davis, 2008). However, although the instructional and educational potential of graphs is widely acknowledged, in some cases graphs are not always more effective than other methods of representation.

*Correspondence to: Robert H. Logie, Human Cognitive Neuroscience-Psychology, University of Edinburgh, 7 George Square, Edinburgh EH8 9JZ, Scotland, UK. E-mail: rlogie@staffmail.ed.ac.uk
† Present address: Department of Clinical Neurosciences, University Hospital Geneva.
‡Post-doctoral researcher.

Law, Freer, Hunter, Logie, McIntosh, and Quinn (2005) demonstrated that when dealing with large volumes of complex medical data, presentation in the form of textual summaries leads to better decision-making than presentation in the form of graphs. They compared the effectiveness of decision-making with graphical or textual summaries of data recorded from patients in a Neonatal Intensive Care Unit (NICU). These data generally consist of continuously monitored physiological variables (e.g. heart rate, central and peripheral temperatures, transcutaneous oxygen and carbon dioxide, blood pressure, etc.) as well as discrete events (e.g. equipment settings, actions taken by staff, etc.). Computerised patient monitoring systems display these data on a single monitor as a set of trend (i.e. time series) graphs and annotations (e.g. Alberdi, Gilhooly, Hunter, Logie, Lyon, McIntosh, & Reiss, 2000; Ambroso, Bowes, Chambrin, Gilhooly, Green, Kari, Logie, Marraro, Mereu, Rembold, & Reynolds, 1992). However, previous research has shown that the introduction of systems that display data trends does not necessarily lead to clinical improvements (Cunningham, Deere, Symon, Elton, & McIntosh, 1998; McIntosh, Lyon, & Badger, 1996).

Law et al. (2005) investigated whether textual summaries of patient information might better support interpretation of the state of the patient and decisions regarding actions to be taken. In their experiment 40 neonatal ICU doctors and nurses with different levels of experience were presented with anonymised scenarios consisting of physiological data and discrete events previously recorded from real patients on a neonatal ICU. Participants were asked to select the most appropriate actions to be taken at the end of the time period for each scenario from a list of 18 possible actions (including 'no action'). Half of the scenarios were presented as trend graphs on a computer screen (e.g. Figure 1(a)), while the other half were presented on the screen as textual summaries of the data (e.g. Figure 1(b)). The textual summaries had been generated by two human experts and were intended to describe the changing values of the physiological parameters of the patient and any relevant medical interventions, but not to include any medical interpretation. Results showed that participants selected more of the appropriate actions when the information was presented as text than when it was presented as graphs.

An advantage of textual over graphical presentation of information has also recently been demonstrated in the context of mobile phone manuals (Langan-Fox, Platania-Phung, & Waycott, 2006). Other studies have demonstrated that students' understanding of graphs is rather limited (Mayer, 1993; Mevarech & Kramarsky, 1997; Shah & Carpenter, 1995; Shah & Hoeffner, 2002), and that people make numerous errors and require complex cognitive processes when interpreting even simple graphs (Carpenter & Shah, 1998; Guthrie, Weber, & Kimmerly, 1993; Romberg, Fennema, & Carpenter, 1993; Schnotz, 2002). Huang, Hong, and Eades (2006) argued that in the case of large and dense graphs, the human perception and cognitive system can get overburdened causing errors in graph interpretation, especially when high-level complex decision-making is required, or where data sets are large and complex.

The difficulty of interpreting trend graphs on the ICU monitors is compounded by a considerable number of artefacts, for example when recording probes are changed, moved, or fall off the patient. Doctors and nurses have to be able to discriminate between artefacts and changes in the trend lines that reflect the physiological state of the patient. This could be challenging especially for junior doctors and junior nurses, because they do not have the knowledge and experience required to quickly recognise artefacts. As such, junior staff are more likely to rely on their limited capacity working memory (Baddeley, 2007) and particularly on their visuo-spatial working memory (e.g. Logie,

DOI: 10.1002/acp

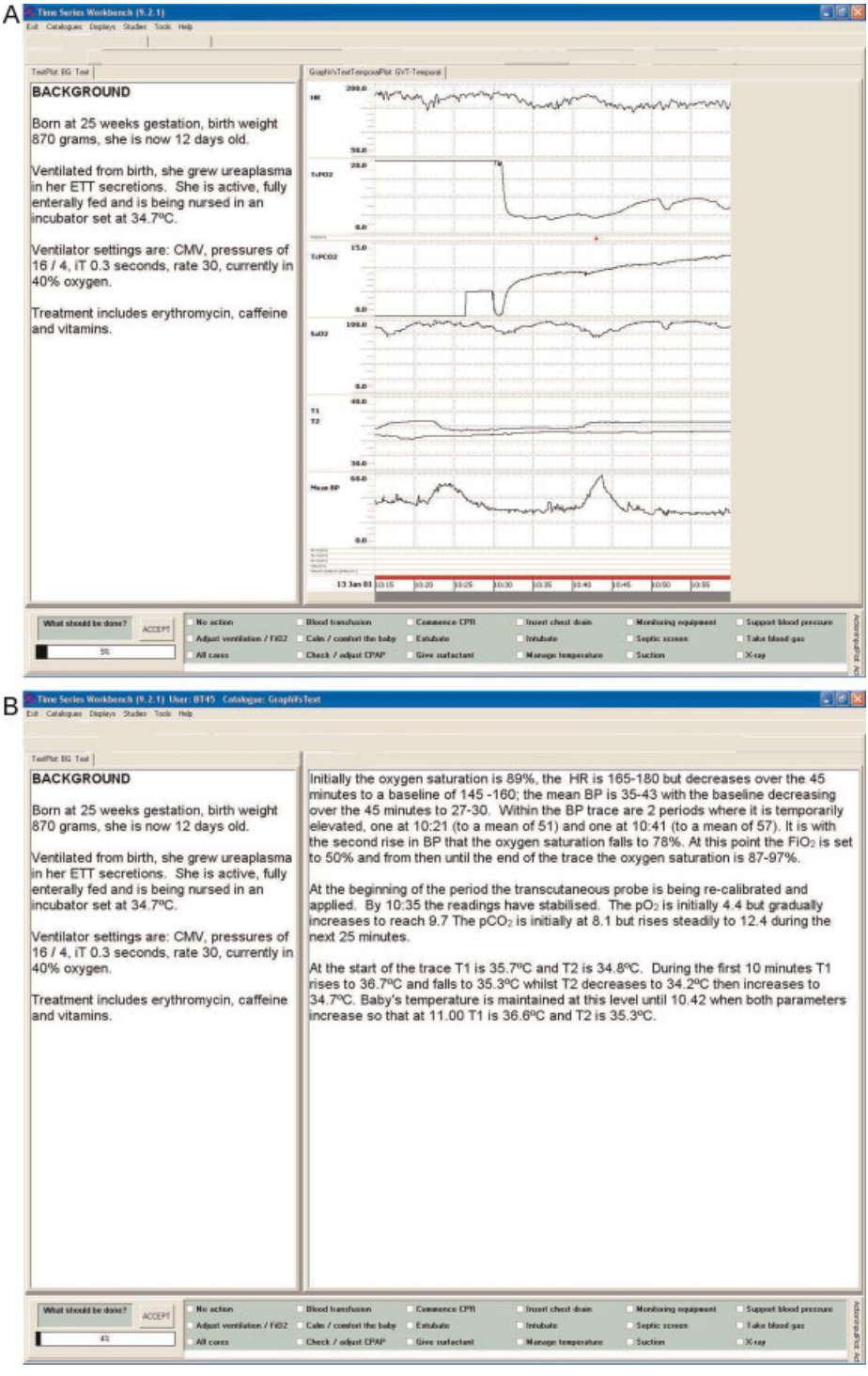

Figure 1. Screen shots of example displays showing, for the same scenario, (a) a graphical display, and textual descriptions generated by (b) human experts and (c) the BT-45 computer programme

DOI: 10.1002/acp

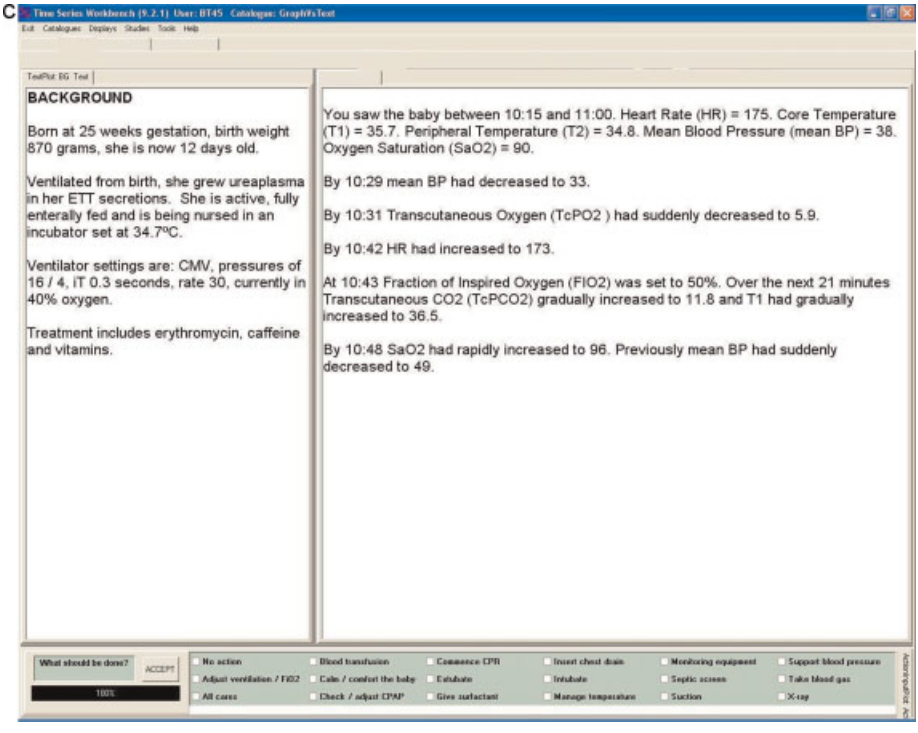

Figure 1. *Continued*

1995; Logie and van der Meulen, 2009) than would more experienced staff (e.g. Ericsson & Delaney, 1999). A textual summary would filter out the 'noise' in the data, thereby reducing the amount of information to be held and interpreted by staff, with a subsequent reduction in demands on working memory, thereby easing the task of interpreting data trends.

Although the presentation of data on a single monitor with trend lines is becoming standard practice in ICUs, these results emphasise the need to question whether trend graphs of physiological functions are the most appropriate, or should be the only format for use on the ward. Previous studies have indicated that different forms of graphical display might be helpful, for example with each physiological parameter shown on the same scale within an integrated 'polygon' display (e.g. Green, Logie, Gilhooly, Ross, & Ronald, 1996). The Law et al. (2005) findings suggest that the graphical displays could usefully be replaced, or at least supplemented, by textual summaries. However, the textual summaries in their study were generated by medical experts with considerable effort and involving a considerable amount of time. This would make it impracticable to exploit the benefits of textual presentation on an ICU in supporting the decisions made by less experienced staff. Nevertheless, recent developments have allowed computer generation of high-quality natural language textual descriptions and summaries from a range of data in health care settings (e.g. Cawsey, Webber, & Jones, 1995; Hallett & Scott, 2005; Hüske-Kraus, 2003b; Reiter, 2007; Reiter, Sripada, Hunter, Yu, & Davy, 2005). Another use of natural language generation (NLG) for health care is for personalised patient information (e.g. Cawsey, Jones, & Pearson, 2000; Reiter, Robertson, & Osman, 2003; Williams, Piwek, & Power,

DOI: 10.1002/acp

2007). Perhaps the most successful applications have been tools that (partially) automate the process of writing routine documents, such as Hüske-Kraus's (2003a) Suregen system, which is regularly used by physicians to create surgical reports.

The complete summarisation of ICU data has not been attempted before (see Portet, Reiter, Hunter, & Sripada, 2007). It is more complex, involving the processing of time series, discrete events and short free texts. This study reports the evaluation of automatically generated summaries of these types of data from the neonatal ICU, in an off-ward experiment following the procedure of Law et al. (2005). In the present experiment, graphical displays were compared with human generated textual summaries and with automatically generated textual summaries. One aim was to investigate how robust was the previously reported advantage for human generated texts in a paper geared towards applied cognitive psychology. A second novel aim was to assess whether staff could effectively use computer based NLG summaries for their decisions. This study was part of a larger project 'BABYTALK' focussed on developing computer generated natural language descriptions of physiological data. The study also capitalises on a level of access to practicing clinical staff that is quite rare in experimental studies. Equally rare is a systematic study of medical decision-making that is based on genuine patient data in a close simulation of a real medical setting.

METHOD

Participants

Participants were 35 staff working in the neonatal ICU at the Royal Infirmary of Edinburgh. They were allocated to one of four groups, depending on role and experience in neonatal care: Senior Doctors ($n = 9$), Junior Doctors ($n = 9$), Senior Nurses ($n = 9$), or Junior Nurses ($n = 8$). Those with 1 year or less of experience in their speciality were classified as junior; those with 8 years or more were classified as senior. Two other participants (both junior nurses) did not complete all three sessions of the experiment. A replacement could be recruited for only one of these. All the nurses were female except for one of the senior group. Six of the nine junior doctors were female, as were two of the senior doctors. The junior and senior doctors comprised all of the staff in these categories on the unit. The junior and senior nurses comprised all of the staff in those categories who were available to take part during the period of the data collection. None of the junior participants had taken part in the Law et al. (2005) study, but 6 senior nurses and 6 senior doctors took part in this previous experiment. There was a gap of approximately 3 years between sessions for data collection in the previous and in the current study.

Material

Scenarios

There were 24 experimental scenarios and 2 practice scenarios, comprising anonymised data sets of physiological measures recorded from babies that had previously been cared for in the unit. These data sets included a record of the actions that had originally been taken on the ward during the time covered by the scenario, as well as observations made by the staff and test results for those patients. None of these data sets had been used by Law et al. (2005), and they were selected to represent episodes of approximately 45 minutes that

DOI: 10.1002/acp

had led up to each of seven different 'target' actions taken on the ward, with three examples selected for each action. The seven actions were: adjust oxygen ventilation, check/adjust continuous positive air pressure, extubate, manage temperature, check monitoring equipment, carry out suction, support blood pressure. These target actions were identified simply to ensure a spread of different types of scenario that tend to appear for babies on the ward, and in each case, a range of additional actions would have been appropriate for the care of the baby at that point. A further three data sets were selected in which no action had been taken, and no specific intervention would have been necessary on the basis of the data available from the monitoring system. Our analyses were based on all appropriate actions selected by participants. The target actions were not considered to be different from other appropriate actions, and no reference will be made to them in the results and discussion sections.

The data sets for all 24 scenarios were already available for display as trend graphs showing heart rate, transcutaneous oxygen and carbon dioxide, oxygen saturation, central and peripheral temperatures and mean blood pressure, as well as the actions and observations. For the human generated texts, a consultant neonatologist, and two experienced neonatal nurses, individually produced a descriptive summary of the data for each scenario. These clinical staff then drew on their independent summaries to reach a consensus on a single summary for each scenario. The summaries were written to be descriptive only. Some degree of interpretation was inevitable, such as (i) excluding artefacts (e.g. a probe falling off would be characterised by one of the parameters rapidly dropping to zero) and (ii) focussing only on obvious changes in the data patterns such as sustained increases or decreases in the parameter values or the slopes of the trend lines. Any 'clinical' interpretation was avoided. The agreed textual descriptions were then read by a senior computer scientist whose research focuses on Artificial Intelligence in Medicine and who recommended changes to text that appeared to include a clinical interpretation. These changes were then agreed with the clinical experts.

For the computer-generated texts, an automatically generated summary for each data set was produced by prototype software system ('BT-45') developed within the BABYTALK project for summarising approximately 45 minutes of recorded physiological data. Both the continuous multi-channel time series data from the physiological monitors, and the discrete event data recorded for each scenario were fed into the program. BT-45 then generated summaries of the clinical data in four main stages, during which artefacts, patterns and trends were identified, and relations between events were inferred. These were translated into text according to an ontology of neonatal ICU concepts (for details of the program, see Portet et al., 2007).

To maintain the veridical character of the experiment, there were no restrictions on length of the human or computer textual descriptions. The brief was to provide an adequate summary of the physiological patterns for the patients over a 45 minute period. The information should be sufficient to allow staff on the ward to take the actions that were most appropriate and in the best interests of the patient.

Presentation of displays. All participants were familiar with the use of a trend monitoring facility known as the BADGER^{TM} (Clevermed Ltd.) system, which is in use in several neonatal ICUs around the UK. In the experiment, displays were constructed and presented in a research version of this system, known as the Time Series Workbench (TSW, Hunter, 2004). TSW was run on a Dell laptop computer with Windows XP Professional, and presented on an external (17 inch) monitor at a resolution of 1280 $\times$ 1024 pixels. Figure 1

DOI: 10.1002/acp

shows example screenshots for the same scenario in each of the three conditions. The scenarios (either in the form of graphs or text) were always presented in the main display area on the right hand side of the screen. With each scenario, some background information was given for that patient, consisting of a textual summary of the basic demographics (age, gestation, weight at birth) and the observations and actions recorded on the ward before the start of the scenario. This information occupied the panel on the left hand side of the screen and was identical in all three presentation conditions. The bottom panel of the screen contained 18 check boxes (corresponding to the 18 possible actions that the participants could select), and an *'Accept'* button. The TSW system recorded which actions participants selected, and the exact times at which these were clicked. A printed list with descriptions of all 18 actions was available to participants throughout the test sessions.

All physiological measures in the graph condition were shown on one screen with scales appropriate to each measure. Participants could click on one of the physiological traces causing a pop-up box to appear with the exact value. Beneath the graphs were coloured markers indicating events that occurred on the ward when the babies were originally observed; the events so represented were only those which were referred to in the human-generated texts. The user could click on these markers causing a box to pop up giving more information. These were the only types of interaction that the participant could have with the display, other than selecting the actions.

Design

We used a mixed design, with three levels of a within-participant factor (presentation of the scenarios in graphs (**G**), human generated texts (**H**) or computer generated texts (**C**)), and four levels of a between-participant factor (the four staff groups). The 24 scenarios (representing three examples of each of the 7 target actions plus the three 'no action' scenarios), were divided into three sets of 8, with one example of each target action in each set. For each participant, each of these three scenario sets was presented in a different format (graphs, human texts, computer texts) with allocation of scenario set to presentation format, and presentation order of the three formats counterbalanced across participants within each participant group. The order of presentation of the individual scenarios within each set was randomised and different for each participant.

Procedure

The experiment took place in a quiet room near the neonatal unit. Participants were tested individually during three separate test sessions. In each session they received one of the three scenario sets in a different presentation format. The three sessions were always given within a period of 10 days or less, with at least 8 hours between two consecutive sessions. Participants were not informed whether the textual summaries of the scenarios were generated by human experts or automatically. All data were recorded anonymously.

Each participant was shown the list of descriptions of actions and received the opportunity to read and ask questions about these. In an initial training phase, it was ensured that participants were familiar with these actions, the location on screen of the check box for each action, and the use of the mouse to indicate their choice.

Although real data were used, it was acknowledged to participants that the experiment was not wholly realistic in that they could not examine the babies whose data were being displayed. It was also emphasised that this was an assessment of the effectiveness of

DOI: 10.1002/acp

different modes of presentation and was not a test of their knowledge or abilities. Participants were then presented with the scenarios, and they were told that the baby could be improving, deteriorating or stable. Their task was to imagine that the time period covered by the scenario led up to the present, and they were to select the action(s) they thought were appropriate to be taken either by themselves or someone else on the unit. They were told that any number of actions could be appropriate but that some actions could be inappropriate. They were also told that they could select 'No Action' if they felt this was most appropriate.

Each scenario had to be completed within 3 minutes, and a dynamic 'time bar' was shown on the screen (see Figure 1). This time limit was introduced not to impose time pressure, but in order to guarantee the maximum length of an experimental session. A similar time limit was used for the Law et al. (2005) study and was found to be adequate for most participants and most scenarios. It was important to reassure nurses and doctors as to the maximum length of the experimental session, given that they were taken off their duties on the unit in order to take part in the study. Adequate back up cover for the unit was made available when necessary. Participants were told that they could complete their decisions in less than the maximum time allowed, and they were encouraged to make their decisions as fast as possible, but without rushing and making sure that they had considered all of the information available to them.

Participants were given two practice scenarios in each of the three sessions. These scenarios involved the presentation format to be used for the particular session. Participants were not asked formally for feedback on the sessions but any spontaneous feedback was recorded with consent of the participant.

RESULTS

Time outs

The TSW software recorded the timing of all responses to the nearest 10 milliseconds. If a participant had not pressed the '*Accept*' button within 3 minutes (to indicate that their response was complete), the scenario 'timed out'. This happened only in 10 trials out of a total of 840. There were too few time outs to analyse statistically, but they appeared with roughly equal frequency across participant groups (three for the junior nurses; three for the senior nurses; three for the junior doctors; one for the senior doctors) and presentation conditions (two for the graphical presentation; four for the human generated texts; four for the computer generated texts). The timed out trials were excluded from the analyses.

Response time

Response time was taken as the time to selection of the first action. To avoid the influence of a speed-accuracy trade-off, this analyses included only those trials on which the first action clicked was an appropriate action. The mean response time for the Graphs condition was 73.16 second (SD = 26.68), for the Human text condition it was 77.23 second (SD = 21.45) and for the Computer text condition it was 78.81 second (SD = 19.16). The response time data split by staff group are shown in Table 1.

A 3 × 4 mixed ANOVA (presentation format as within-subjects factor and staff group as between-subjects factor) revealed no main effect of group, $F(3, 31) = .579$, $p > .10$, no

DOI: 10.1002/acp

Table 1. Mean (SD) response time in seconds for each staff group in the three presentation conditions

	G	H	C
JN	**66.86** (31.06)	**78.33** (15.04)	**72.63** (15.30)
SN	**73.81** (27.09)	**84.14** (24.43)	**78.88** (21.08)
JD	**66.59** (23.86)	**70.81** (23.90)	**78.64** (21.97)
SD	**84.69** (25.29)	**75.77** (22.02)	**84.41** (18.84)

G, graphs; H, human generated texts; C, computer generated texts; JN, junior nurses; SN, senior nurses; JD, junior doctors; SD, senior doctors.

main effect of condition, $F(2, 31) = 1.085$, $p > .10$ and no interaction, $F(3, 31) = .924$, $p > .10$. There was no tendency for either the presentation format or the staff group to influence the time until the first appropriate action was selected. This replicated the results of Law et al. (2005), who found no effects of presentation condition (Graphs *vs*. Human generated texts) or staff group on response time.

Scores

For each scenario it was determined in advance by the three clinical staff who had generated the human text descriptions, which actions were appropriate (i.e. beneficial), which were inappropriate (i.e. could be harmful) and which actions were neutral (i.e. unnecessary but harmless). A participants' score for each scenario was derived by subtracting the proportion of inappropriate actions selected from the proportion of appropriate actions selected. If, for example, the appropriate actions for Scenario 1 were 'adjust ventilation' and 'order chest X-ray', and a participant selected one of these, plus 2 of 13 actions that were identified as inappropriate, then the score for this scenario would be 1/2 (the proportion of appropriate actions selected) minus 2/13 (the proportion of inappropriate actions selected) = 0.35. The maximum possible score for any scenario was 1.00, if all appropriate actions were selected, and none of the inappropriate actions. The scores for all 8 scenarios in each presentation condition were averaged for each participant to give one score for each of the three presentation formats.

The overall mean for the Graphs condition was 0.33 (SD = 0.14), for the Human text condition 0.39 (SD = 0.11) and for the Computer text condition 0.34 (SD = 0.14). The scores for each staff group are shown in Table 2.

A 3×4 mixed ANOVA showed a main effect of condition approaching significance, $F(2, 31) = 2.939$, $p = 0.060$, no main effect of group, $F(3, 31) = 1.238$, $p > .10$ and no interaction, $F(3, 31) = 1.095$, $p > .10$. Although the overall effect of condition was marginal, the Law et al. (2005) study had compared only the Graph and Human conditions. Therefore, we carried out separate analysis on just those two conditions to assess whether or not the earlier result had been replicated. In a 2×4 ANOVA, there was a significant main effect of condition, $F(1, 31) = 4.975$, $p < .05$, no main effect of group, $F(3, 31) = 1.328$, $p > .10$ and no interaction, $F(3, 31) = .687$, $p > .10$.[1] There was also a significant main effect of condition when the Computer and Human conditions were compared separately,

[1]Law et al. (2005) scored the performance of participants on the basis of the appropriate actions selected only, and did not take account of the inappropriate actions selected. Those previous data were re-analysed using the current scoring method. This resulted in a similar pattern, with performance significantly higher for text than for graphs, and with no difference between staff groups and no interaction.

DOI: 10.1002/acp

Table 2. Mean (SD) scores for each staff group in the three presentation conditions

	G	H	C
JN	**0.31** (0.16)	**0.43** (0.10)	**0.28** (0.09)
SN	**0.38** (0.09)	**0.42** (0.12)	**0.41** (0.15)
JD	**0.29** (0.14)	**0.33** (0.11)	**0.35** (0.14)
SD	**0.36** (0.17)	**0.39** (0.11)	**0.33** (0.13)

G, graphs; H, human generated texts; C, computer generated texts; JN, junior nurses; SN, senior nurses; JD, junior doctors; SD, senior doctors.

$F(1,31) = 5.266$, $p < .05$, with no main effect of group, $F(3,31) = .914$, $p > .10$ and a marginal interaction, $F(3,31) = 2.728$, $p = 0.061$. When the Graph and Computer conditions were compared separately, there was no main effect of group, $F(3,31) = 1.377$, $p > .10$, no main effect of condition, $F(1,31) = .056$, $p > .10$ and no interaction, $F(3,31) = .490$, $p > .10$.[2] From Table 2, it appeared that the Junior nurses showed the largest advantage for the human generated text compared with the other two conditions. It also appeared that the junior and senior doctors showed some advantage for the computer generated text over the graphs. However, given that the group by condition interaction was not near significant in the global ANOVA, no further analyses of participant groups were conducted.

The variability within participant groups appeared to be quite large compared with any possible between group differences, and this variability between participants would also have made the assessment less sensitive to differences between conditions. In this realistic setting with human experts, increasing participant numbers in each category to increase sensitivity is not an option. Therefore, we carried out a 'by items' analysis, across the different presentation conditions, treating the 24 scenarios as participants, and collapsing across participant groups. The one way ANOVA revealed a significant effect of presentation condition $F(2,188) = 6.2$, $p < 0.005$ and Newman-Keuls *post-hoc* tests showed that the human generated texts generated significantly higher scores than the other two presentation modes, which did not differ.

In sum, the advantage for the Human text condition compared with the Graphs condition was replicated in this study. The data also suggest that the Computer text condition was at least as good as the Graph condition when these were directly compared.

DISCUSSION

The present experiment replicated the main findings of Law et al. (2005). In both studies an advantage was found for human generated text over graphical presentation of data, with participants selecting a greater proportion of the appropriate actions in the former condition. In both studies, there was no tendency for either staff group or presentation format to influence the response time of participants. Overall, these results confirm that in a

[2]Performance of participants was also analysed using the scoring methods that were used by Law et al. (2005), namely the 'proportion of appropriate actions selected' (i.e. the number of appropriate actions selected divided by the total number of actions that are appropriate for that scenario) and 'the proportion of actions selected that was appropriate'. The former scoring method led to broadly similar results, with performance in the human generated text condition higher than performance in the graphs and computer generated text conditions. There were no significant differences using the latter scoring method.

DOI: 10.1002/acp

neonatal ICU, human generated descriptions of time series physiological measures are better able to support medical decision-making than graphs with trend lines.

Although the automatically generated textual summaries were less effective than the human generated texts, they were not statistically poorer than the graphical presentations with which staff are familiar and which are very similar to those used in the real clinical setting. There was a suggestion from the means in Table 2 that the junior and senior doctors could use the computer generated text more effectively than could the junior and senior nurses, but the variability within the groups undermined any statistical differences between different staff categories. This suggests that the effect sizes were rather small and that much larger numbers of participants might have made for a more sensitive statistical test. However, unlike a traditional laboratory experiment in which participants are plentiful, participants with this kind of expertise are rarely available for participation in a formally designed experiment using genuine physiological data. Indeed, the experiment included all of the junior and senior doctors on a large neonatal ICU, and the vast majority of senior and junior nurses. Moreover, small effects that appear only in very tightly controlled experimental conditions might be a rather poor basis for application in the real world setting (see e.g. Logie, Baddeley, & Woodhead, 1987). Effects that are found to be robust in a realistic simulation may have a greater chance of having a genuine impact if implemented in the applied setting.

The above arguments give us confidence that the advantage for textual descriptions is real. However, as mentioned earlier, the human generated texts require considerable human expertise and time to generate. The fact that the unfamiliar format of computer generated texts can support performance at least as effectively as the graphical displays with which the staff are very familiar in their clinical practice, suggests that there may be considerable potential in developing NLG for supporting human decision making in this setting.

The fact that the computer texts did not show the same advantage as the human generated texts raises the question as to what differed between these types of text. A detailed and formal analysis of the contents of the human and computer-generated texts is beyond the scope of this paper, but has been completed and will be reported in full elsewhere (McKinlay, McVittie, Reiter, Freer, Sykes, & Logie, 2008). However, those analyses indicated that the human generated texts tend to have a more coherent grammatical structure and 'narrative', show a greater tendency to group physiological measures together, and tend to be longer than the computer generated texts. Participants were not told explicitly that some of the textual descriptions they received in the experiment were automatically generated by a computer. None of the participants indicated that they were aware of the nature of the texts or how they had been generated. Consistent with the more detailed analyses by McKinlay et al. (2008), one participant commented in the human generated text session that the texts were longer than in the other session, and several participants noted that sometimes descriptions in the computer generated text condition seemed 'awkward' or 'inconsistent'.

The outcomes of this study with human domain experts suggest that despite its limitations, the BT-45 software can effectively support real clinical decision-making, and that further development of this technology is likely to be extremely fruitful in supporting complex real-world cognition. The study also points to ways in which the BT-45 program may be further refined (see Portet et al., 2007; Reiter, Gatt, Portet, & van der Meulen, 2008), perhaps to include additional knowledge about grouping of physiological functions. A possible implementation of a refined system would be to complement the graphical trend line display, although the added value of having both rather than just one display format

DOI: 10.1002/acp

would merit further investigation. A further use for NLG would be to summarise the condition of a patient for a senior clinician who might be at a remote location rather than at the bedside.

ACKNOWLEDGEMENTS

The authors are grateful to all of the staff who took part in this study, and to other members of the BABYTALK team—Albert Gatt, Francois Portet, Ehud Reiter and Somayajulu Sripada who, together with Jim Hunter, were responsible for generating the NLG software. This work was funded by research grant awards EP/D049520/1 and EP/D05057X/1 from the UK Engineering and Physical Sciences Council.

REFERENCES

Alberdi, E., Gilhooly, K., Hunter, J., Logie, R. H., Lyon, A., McIntosh, N., & Reiss, J. (2000). Computerisation and decision making in neonatal intensive care: A cognitive engineering investigation. *Journal of Clinical Monitoring and Computing*, *16*, 85–94.

Ambroso, C., Bowes, C., Chambrin, M. C., Gilhooly, K., Green, C., Kari, A., Logie, R. H., Marraro, G., Mereu, M., Rembold, P., & Reynolds, M. (1992). INFORM: European survey of computers in Intensive Care Units. *International Journal of Clinical Monitoring and Computing*, *9*, 53–61.

Baddeley, A. D. (2007). *Working memory, thought and action*. Oxford, UK: Oxford University Press.

Card, S. K., Mackinlay, J. D., & Schneiderman, B. (1999). *Readings in information visualization: Using vision to think*. San Fransisco: Morgan Kaufmann Publishers Inc.

Carney, R. N., & Levin, J. R. (2002). Pictorial illustrations still improve students' learning from text. *Educational Psychology Review*, *14*, 5–26.

Carpenter, P. A., & Shah, P. (1998). A model of the perceptual and conceptual processes in graph comprehension. *Journal of Experimental Psychology: Applied*, *4*, 75–100.

Cawsey, A., Jones, R., & Pearson, J. (2000). The evaluation of a personalised information system for patients with cancer. *User Modelling and User-Adapted Interaction*, *10*, 47–72.

Cawsey, A., Webber, B., & Jones, R. (1995). Natural language generation in health care. *Journal of the American Medical Informatics Association*, *4*, 473–482.

Cunningham, S., Deere, S., Symon, A., Elton, R. A., & McIntosh, N. (1998). A randomized, controlled trial of computerized physiologic trend monitoring in an intensive care unit. *Critical Care Medicin*, *26*, 2053–2059.

Ericsson, K. A., & Delaney, P. F. (1999). Long-term working memory as an alternative to capacity models of working memory in everyday skilled performance. In A. Miyake, & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 257–297). Cambridge, UK: Cambridge University Press.

Green, C. A., Logie, R. H., Gilhooly, K. J., Ross, D. G., & Ronald, A. (1996). Aberdeen polygons: Computer displays of physiological profiles for intensive care. *Ergonomics*, *39*, 412–428.

Guthrie, J. T., Weber, S., & Kimmerly, N. (1993). Searching documents: Cognitive processes and deficits in understanding graphs, tables, and illustrations. *Contemporary Educational Psychology*, *18*, 186–221.

Hallett, C., & Scott, D. (2005). Structural variation in generated health reports. *Proceedings of the 3rd International Workshop on Paraphrasing*, Jeju Island, Korea.

Huang, W., Hong, S. H., & Eades, P. (2006). Predicting graph reading performance: A cognitive approach. *Proceedings of the Asia Pacific symposium on Information Visualisation, 60*, 207–216.

Hunter, J. (2004). The Time Series Workbench: User Manual, University of Aberdeen. *Computing Science Technical Report*.

Hüske-Kraus, D. (2003a). Suregen-2: A shell system for the generation of clinical documents. *Proceedings of EACL*.

Hüske-Kraus, D. (2003b). Text generation in clinical medicine—a Review. *Methods of Information in Medicine*, *42*, 51–60.

DOI: 10.1002/acp

Langan-Fox, J., Platania-Phung, C., & Waycott, J. (2006). Effects of advance organizers, mental models and abilities on task and recall performance using a mobile phone network. *Applied Cognitive Psychology*, *20*, 1143–1165.

Law, A. S., Freer, Y., Hunter, J., Logie, R. H., McIntosh, N., & Quinn, J. (2005). A comparison of graphical and textual presentations of time series data to support medical decision making in the Neonatal Intensive Care Unit. *Journal of Clinical Monitoring and Computing*, *19*, 183–194.

Logie, R. H. (1995). *Visuo-spatial working memory.* Hove, UK: Lawrence Erlbaum Associates Ltd.

Logie, R. H., Baddeley, A. D., & Woodhead, M. M. (1987). Face recognition, pose and ecological validity. *Applied Cognitive Psychology*, *1*, 53–69.

Logie, R. H., & van der Meulen, M. A. (2009) Fragmenting and integrating visuospatial working memory. In J. R. Brockmole (Ed.), *Representing the visual world in memory* (pp. 1–32). Hove, UK: Psychology Press.

Mayer, R. E. (1993). Comprehension of graphics in texts: An overview. *Learning and Instruction*, *3*, 239–245.

McIntosh, N., Lyon, A., & Badger, P. (1996). Time trend monitoring in the Neonatal Intensive Care Unit: Why doesn't it make a difference? *Pediatrics*, *98*, 540.

McKinlay, A., McVittie, C., Reiter, E., Freer, Y., Sykes, C., & Logie, R. (2008). Design Issues for Socially Intelligent User-Interfaces: A Data-to-Text System for Summarizing Clinical Data. Submitted for publication.

Mevarech, Z. R., & Kramarsky, B. (1997). From verbal descriptions to graphic representations: Stability and change in students' alternative conceptions. *Educational Studies in Mathematics*, *32*, 229–263.

Portet, F., Reiter, E., Hunter, J., & Sripada, S. (2007). Automatic generation of textual summaries from Neonatal Intensive Care data. *Proceedings of the 11th Conference on Artificial Intelligence*.

Ratwani, R. M., Trafton, J. G., & Boehm-Davis, D. A. (2008). Thinking graphically: Connecting vision and cognition during graph comprehension. *Journal of Experimental Psychology: Applied*, *14*, 36–49.

Reiter, E. (2007). An Architecture for data-to-text systems. *Proceedings of the European Workshop of Natural Language Generation 2007*, 97–104.

Reiter, E., Gatt, A., Portet, F., & Van der Meulen, M. (2008). The importance of narrative and other lessons from an evaluation of an NLG system that summarises clinical data. In *Proceedings of the International Conference on Natural Language Generation*, 2008.

Reiter, E., Robertson, R., & Osman, L. (2003). Lessons from a failure: Generating tailored smoking cessation letters. *Artificial Intelligence*, *144*, 41–58.

Reiter, E., Sripada, S., Hunter, J., Yu, J., & Davy, I. (2005). Choosing words in computer-generated weather forecasts. *Artificial Intelligence*, *167*, 137–169.

Romberg, T., Fennema, E., & Carpenter, T. (1993). *Integrating research on the graphical representation of function*. Hillsdale, NJ: Erlbaum.

Roth, W. M., & Bowen, G. M. (2003). When are graphs worth ten thousand words? An expert-expert study. *Cognition and Instruction*, *21*, 429–473.

Schnotz, W. (2002). Towards an integrated view of learning from text and visual displays. *Educational Psychology Review*, *14*, 101–120.

Shah, P., & Carpenter, P. A. (1995). Conceptual limitations in comprehending line graphs. *Journal of Experimental Psychology: General*, *124*, 43–61.

Shah, P., & Hoeffner, J. (2002). Review of graph comprehension research: Implications for instruction. *Educational Psychology Review*, *14*, 47–69.

Tory, M., & Möller, T. (2004). Human factors in visualisation research. *IEEE Transactions on Visualization and Computer Graphics*, *10*, 72–84.

Williams, S., Piwek, P., & Power, R. (2007). Generating monologue and dialogue to present personalised medical information to patients. *Proceedings of the 11th European Workshop on Natural Language*.

Winn, W. D. (1994). Contributions of perceptual and cognitive processes to the comprehension of graphics. In Schnotz, W. & Kulhavy R. (Eds.), *Comprehension of graphics*. Amsterdam: Elsevier. 3–27.

DOI: 10.1002/acp