Using Natural Language Generation Technology to Improve Information Flows in Intensive Care Units

Jim Hunter, Albert Gatt, François Portet, Ehud Reiter and Somayajulu Sripada[1]

Abstract. In the drive to improve patient safety, patients in modern intensive care units are closely monitored with the generation of very large volumes of data. Unless the data are further processed, it is difficult for medical and nursing staff to assimilate what is important. It has been demonstrated that data summarization in natural language has the potential to improve clinical decision making; we have implemented and evaluated a prototype system which generates such textual summaries automatically. Our evaluation of the computer generated summaries showed that the decisions made by medical and nursing staff after reading the summaries were as good as those made after viewing the currently available graphical presentations with the same information content. Since our automatically generated textual summaries can be improved by including additional content and expert knowledge, they promise to enhance information exchange between the medical and nursing staff, particularly when integrated with the currently available graphical presentations. The main feature of this technology is that it brings together a diverse set of techniques such as medical signal analysis, knowledge based reasoning, medical ontology and natural language generation. In this paper we discuss the main components of our approach with a critical analysis of their strengths and limitations and present options for improvement to address these limitations.

1 INTRODUCTION

The modern intensive care unit (ICU) is data rich. Most bedside computers are equipped with software that helps medical staff to record, retrieve and visualize patient-related information. Several megabytes of physiological data (e.g. heart rate), as well as laboratory results and the results of other investigations (such as X-ray results), are recorded per patient per day. Medical staff need to assimilate these data rapidly to provide real-time patient care. In the healthcare sector there is increased interest in deploying technology to deal with this information overload. Over several years, we have been working closely with medical staff in a neonatal ICU to understand how to achieve more effective use of patient data in order to improve the quality of patient care. Our studies led us to the following hypotheses:

1. Timely and accurate information flow between medical staff is the key to better patient care. Nurses and doctors require tools that accurately summarise the key facts about patients rather than expert systems that generate clinical recommendations.
2. The informational requirements of these summaries vary between the different groups of people involved in health care such as nurses, doctors and even family and friends. While nurses need information that helps them to make detailed patient care plans, doctors require information to support clinical decision making. Family and friends have more variable requirements depending on how close they are to the patient.
3. Staff roles in ICUs are well defined and any new technology has to suit the work flows associated with existing staff roles [16].

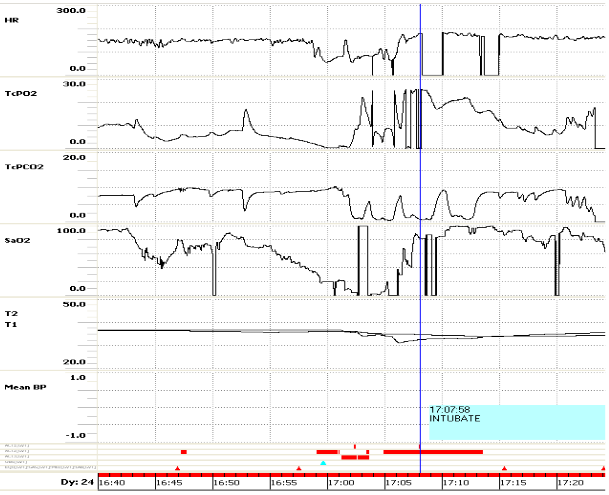

Figure 1. Graphical Display of NICU data

This paper presents an extensive programme of work whose ultimate goal is the automatic generation of summaries in natural language (text) of the history of a baby in a neonatal ICU over a period of several hours. In the rest of this introduction, we set out the background to this endeavour. In Section 2 we describe our initial prototype; the outputs from this system have been evaluated experimentally. This evaluation and its results are presented in Section 3; full details can be found in [17]. We conclude with a discussion of what has been achieved so far and the challenges for the future.

1.1 Data-intensive clinical environments

Apart from clinical environments in which imaging plays a central role, the largest volumes of clinical data are generated where several physiological variables are measured at high frequencies (say once per second) over considerable periods of time.

[1] Department of Computing Science, University of Aberdeen, UK. email: j.hunter@abdn.ac.uk

Examples of such environments are:

- intensive care units (both adult and neonatal);
- high dependency units (e.g. coronary care);
- operating theatres;
- renal dialysis;
- ambient monitoring in the home.

The data are collected for good reason, in that when correctly interpreted, they enable the tracking of the clinical state of the patient over time so that action can be taken to rectify any undesirable deviations.

For example, in an ICU, up to ten physiological variables (including heart-rate, blood pressure, O_2 and CO_2 saturations, and temperatures) may be measured continuously every second (i.e. almost a million measurements per day). In addition there will normally be a number of additional discrete data items which are entered sporadically: laboratory results, blood gases, daily organ failure scores, medication, equipment settings, patient observations, etc. The example from a neonatal ICU shown in Figure 1 illustrates many of the characteristics of data from such environments – they are multi-channel, voluminous, noisy and full of artifact.

Even in intensive care, where the ratio of patients to staff is low, medical errors are not uncommon. Errors often result from missed symptoms and signs, or a lack of appreciation of their importance which may be due to staff ignorance, to attentional or informational overload. Presenting these data in an effective way is central to informed medical decision making.

Normal practice is to present the physiological variables graphically as shown in Figure 1. However previous research [3] showed that displaying data in this way does not automatically lead to improved patient care. One of the reasons for this was established by the Cognate project [1], which showed that junior staff (who are responsible for most of the immediate care of the baby) spend a small fraction (about 5%) of their time looking at the displays. Cognate concluded that if the information present in these complex multi-channel time-series was to be fully utilised, some form of additional processing was essential.

In the Neonate project, a follow-up to Cognate, it was hypothesised that reporting information textually would be more effective than the graphical displays because textual descriptions, by virtue of their structure, can present a coherent picture of all the relevant information, while graphical displays require their user to filter out a lot of irrelevant detail in order to establish the relatively few important facts about the patient state. To test this hypothesis, an experiment was carried out to compare the effect on decision-making of two different modes of presentation [8]. One mode was graphical (as in Figure 1); the other was a summary of the data in natural language written by expert clinicians. The data periods presented had durations of about 45 minutes. This experiment demonstrated clearly that participants took better decisions when informed by the textual presentation. However it took considerable effort and involved a considerable amount of time for the medical experts to generate the text. If the benefits of textual summary are to be exploited, the texts need to be generated automatically. This can be achieved by mean of recent advances in data-to-text technology.

1.2 Data-to-text NLG

Natural Language Generation (NLG) is an area of research focused on developing techniques for automatically generating natural language (such as English) descriptions of non-linguistic information [11]. An implicit assumption of most NLG techniques is that the non-linguistic input information comes from knowledge bases with well-defined semantics. In practice, however, in most application domains where automatic textual descriptions are desperately required, such knowledge bases do not exist. The data-intensive clinical environments described in the previous section generate large amounts of clinical data that are not structured into logical forms in a knowledge base. Data-to-text NLG is a recent extension to traditional NLG to allow such naturally occurring data to be described linguistically.

Since the pioneering work of Kukich [9] in summarising stock market data, various applications of data-to-text NLG have been reported in the literature. Yu et al. [18] reported an NLG system that produces textual descriptions of time series data from an operational gas turbine. Sripada et al [13] reported an NLG system that produces textual weather forecasts for the offshore oil industry. This system has actually been deployed in the industry to produce drafts of marine weather forecasts for supporting oilrig operations in the North Sea. User evaluation studies found computer generated forecasts to be more consistent in their language and therefore of better value than the expert written forecasts. Sripada and Gao [15] reported a system that described data recorded by a scuba (Self Contained Underwater Breathing Apparatus) dive computer used by recreational scuba divers.

In the Medical domain, the recent CLEF project [4] aims at generating summaries of multiple text-based health reports. In [2], the authors describe a system that dynamically generates hypertext pages that explain treatments, diseases, etc related to the patient's condition, using information in the patient's medical record as the basis for the tailoring. Perhaps the most successful medical data-to-text applications have been tools that (partially) automate the process of writing routine documents, such as the Suregen-2 system [6], which is regularly used by physicians to create surgical reports; see [7] for a review of text generation in medicine. However, the complete summarisation of ICU data is more complex, involving the processing of time series, discrete events, and short free texts, which seems not to have been done before.

2 BABYTALK PROJECT

The studies discussed above led us to conclude (i) that textual summaries of ICU data might have a significant role in decision support, and (ii) that the automatic generation of textual summaries of complex time series data was feasible. A preliminary study [12] demonstrated the problems specific to the ICU domain. The main challenge in developing a computer application for the ICU context that delivers the required functionality is bringing together a diverse set of techniques that were originally developed quite independently - medical signal analysis for processing the raw physiological data, knowledge based reasoning to integrate and interpret the results of signal analysis along with other discrete data and finally the application of NLG to generate the required descriptions. These issues are being tackled in the BabyTalk project.

2.1 Example data

Input to our system consists of several channels of physiological data such as heart rate, blood pressure, and temperature. In addition to these continuous data, discrete data about equipment settings, laboratory results and actions taken by medical staff were also available to our system. Figure 1 shows a window of continuous channel data with annotations of discrete data.

"You saw the baby between 16:40 and 17:25.

Initially the HR baseline is 140-160; pO_2 is 8-10; oxygen saturation = 92%, T1 and T2 are 36.9° and 36.6ºC.
At around 16.45 ET suction is performed; there is a drop in oxygen saturation to 50% and pO_2 to 3.3 and a rise in pCO_2 to around 9. The FiO_2 is increased from 61 to 100%. By 16.51 the HR is at 155, the pO_2 is 6.7 and the oxygen saturation is 88% and the pCO_2 is 9.2. There is an upward spike in the pO_2 to16.9 and a corresponding downward one in pCO_2 to 3.1; the oxygen saturation has fallen to 78%. T1 is now reading 36.9ºC and T2 35.7ºC. At 16.57 the ventilator rate is increased to 30.

Baby is given Neopuff ventilation. The oxygenation continues to decrease: pO_2 = 0.2 and oxygen saturation = 20% at 17:00 and the HR falls to 60. The pCO_2 continues to rise to 10.1. The baby is pale and unresponsive. ET suction is given, baby is turned and at 17:02 the ETT is removed; the baby is again given Neopuff ventilation. Baby is re-positioned and the NGT aspirated. By 17:08 the baby is reintubated; the oxygen saturation has increased to the 80s and the HR has risen to 176 the pO_2 = 0.1 and pCO_2 = 0.2, T1 is 32.7ºC and T2 34.7ºC.

At 17:15 the FiO_2 is reduced to 33% and the rate put back to 15.

At 17:24 the oxygen saturation falls to 65 and the FiO_2 is increased to 56%.

At 17:25 the HR is 165, the oxygen saturation is 100%, T1 is 35.7° and T2 is 34.5ºC."

Figure 2. Human generated summary of the data shown in Figure 1.

Independently, our human experts produced their own summaries of the data; Figure 2 shows the text they generated for the data shown in Figure 1. An example of the automatically generated output of our system for the same data period is presented in Figure 3.

"You saw the baby between 16:40 and 17:25. Heart Rate (HR) = 155. Core Temperature (T1) = 36.9. Peripheral Temperature (T2) = 36.6. Transcutaneous Oxygen (TcPO2) = 9.0. Transcutaneous CO2 (TcPCO2) = 7.4. Oxygen Saturation (SaO2) = 94.

Over the next 24 minutes there were a number of successive desaturations down to 0. Fraction of Inspired Oxygen (FIO2) was raised to 100%. There were 3 successive bradycardias down to 69. Neopuff ventilation was given to the baby a number of times. The baby was re-intubated successfully. The baby was resuscitated. The baby had bruised skin.

Blood gas results received at 16:45 showed that PH = 7.3, PO2 = 5, PCO2 = 6.9 and BE = -0.7.

At 17:15 FIO2 was lowered to 33%. TcPO2 had rapidly decreased to 8.8. Previously T1 had rapidly increased to 35.0."

Figure 3. Computer generated summary of the data shown in Figure 1.

2.2 Architecture

The main architecture of the prototype is presented in Figure 4. All of the modules in the architecture are driven by information stored in the knowledge base of the NICU Ontology (1). While developing ontology-driven systems is nothing new in both medical and NLG applications, developing an ontology that captures the knowledge required by both of these components is one of the novel features of this architecture. Moreover, as shown in the figure, the ontology is programmatically integrated into the system.

The system creates a summary of the clinical data period in four main stages. All the terms used to describe the discrete events are related to our ontology which is still being extended to include about 1800 concepts. To enable future sharing of this valuable knowledge source we synchronised our ontology with UMLS, a meta-thesaurus that brings together several popular medical taxonomies (such as SNOMED-CT, WHO classification). The first stage of the processing is Signal Analysis (2) which extracts the main features of the physiological time series (artifacts, patterns, and trends) using modelling based on a baby's physiological values, auto-regressive filtering, and adaptive bottom-up segmentation techniques. Data Interpretation (3) performs some temporal and logical reasoning to infer higher medical abstractions and relations ('re-intubation', "A causes B", etc.) from the signal features and the clinical observations using an expert system linked to the ontology. From the large number of events generated, Document Planning (4) selects the most important based on an importance factor either determined by experts (e.g. a surgical operation is must always be reported) or computed from the signals (e.g. the importance of some patterns depends on their duration). The most important events are then structured into a tree. Finally, Microplanning and Realisation (5) translates this tree into coherent text. Details of all these stages are available in [10].

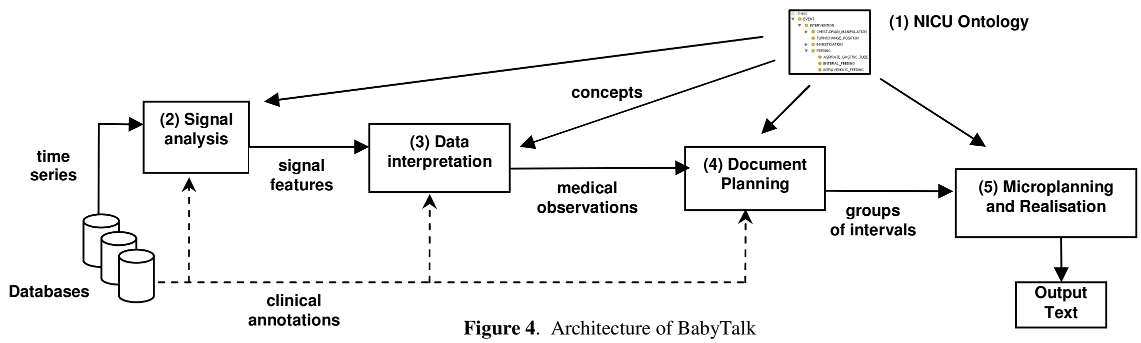

Figure 4. Architecture of BabyTalk

3 EXPERIMENTAL EVALUATION

There are several ways of evaluating the quality of computer generated texts, including post-editing [14] (where one measures the amount of editing required to render the text acceptable to a user) and detailed comparison with human generated texts based on the same input data. However, we are strong believers in task-based evaluation, where end-users are asked to perform a meaningful task under different conditions; a judgment is then made on the extent to which the task was carried out satisfactorily – any differences are attributed to the conditions under which the task was performed.

3.1 Experiment

The participants in our experiment were nurses and doctors from the neonatal ICU in Edinburgh with different levels of experience: 7 junior nurses and 8 senior nurses, 8 junior doctors and 8 senior doctors. They attended three separate sessions during each of which they had to look at the presentation of data from each of 8 scenarios and indicate which action or actions they would take at the end of the period; the actions were selected from a set of 18 possible actions. The experiments took place away from the ward and suitable practice was provided each time before the test began.

For each scenario a participant was given some information about the baby at the start of the period and the data for the period was presented graphically or as text (human-authored or computer-generated). Participants were under some time pressure (three minutes per scenario), both to simulate the pressures that arise on the ward and to ensure that the time that the participant was away from her duties was predictable.

Considerable care was taken to ensure that the human-authored text was purely descriptive i.e. that it did not interpret the data in a way which would have given a definite hint to the participant as to the correct action.

Further information about all aspects of the experiment (including more detailed results) is given elsewhere [17].

3.2 Results

For each scenario our experts judged some actions to be appropriate, some to be inappropriate and the remainder to be neutral. Each participant was scored on each scenario, the score consisting of the percentage of the appropriate actions that they selected minus the percentage of inappropriate actions that they selected. For each condition an average score over the eight scenarios was calculated.

Subjects from all the groups performed better with human-written textual summaries (mean score = 0.39), compared to their performance with graphical presentations (mean score = 0.33); note that graphical presentation is the currently available form of data presentation to medical staff on the ward. This result is in agreement with the earlier experiment reported in [8] and provides additional support for our hypothesis that textual summaries are valuable in presenting patient data to medical staff.

On the other hand, subjects did not perform as well with the computer-generated summaries (mean score = 0.34) as they did with human-written texts. Although this is disappointing, the observation that they found the computer-generated texts as useful as the existing graphical presentations is positive and encouraging.

As with the previous experiment, the textual descriptions (both human-authored and computer generated) deliberately avoided interpreting the data in ways that gave clues about the correct decisions to the subjects. However this is an artificial restriction imposed to allow meaningful comparison with the existing presentation and need not limit the future development of our approach.

Because human (expert) written summaries seem to be superior, we have started to analyze the differences between these texts and those generated automatically. Among other things, our initial analysis shows that the human written texts take a narrative approach to the presentation of the content; this is currently missing from the computer generated texts.

Moreover, in the experiment, textual and graphical presentations were treated as alternatives. But in the future there is no reason why they should not be used to complement each other with the textual summaries reinforcing important messages. One of the most important factors in favour of computer generated texts is the amount of time they save for already stressed medical staff, who may otherwise be required to write these textual summaries (e.g. nursing shift summaries).

4 CHALLENGES AND CONCLUSIONS

Besides the need to improve automatic generation to produce human-quality texts, there are several issues that need to be addressed before routine operational deployment.

The existing technology for producing textual descriptions of ICU data only processes data captured on the ward over 45 minutes whereas in reality we need to be able to handle several hours of data. Also the current textual descriptions are not targeted at any specific group of people in the ICU. In our experiment, the performance of different groups such as junior nurses and senior doctors was not the same, almost certainly because the informational requirements of these groups are different. So one objective of the ongoing project is to develop significant extensions to the current working prototype for three distinct purposes:

1. To assist nurses in writing summaries at the end of their 12 hour shifts; these shift summaries help the incoming nurse to plan her care of the baby; the benefits to the nurses of automatic generation would be time saving and consistency.
2. To generate summaries on demand for junior doctors; summaries will cover approximately 12 hours and be designed to support decision making.
3. To provide parents and relatives with a basic summary of the progress of the baby over the previous 24 hours; the Edinburgh neonatal ICU has recently made such reports (written by clinicians) available to parents and close friends; we note that immediate family members often require more detailed information than others.

These extensions will attempt much more complex tasks than the simple summarising of a 45 minute period. The times to be covered are an order of magnitude larger and the level of abstraction must be correspondingly greater. The systems are targeted at very different classes of user; the content, structure and language used must be appropriate. In addition to the above, the following issues require attention in future extensions:

1. Uncertainty in information: Many discrete events recorded by medical staff on the ward do not bear accurate time stamps. Some events may be recorded before they actually happen; most are recorded after their occurrence. This uncertainty in temporal information should either be resolved by the system or communicated to the user with appropriate linguistic cues. There are several domain independent techniques for handling uncertainty such as the possibility theory. We are currently

developing techniques for resolving uncertainty in temporal information which exploits domain knowledge. For example, if it is known that an event such as intubation (the task of inserting a tube for artificial breathing) has been performed, based on the domain knowledge about the expected duration of the event, it may be possible to estimate the start and end times of that event.

2. Missing information: Several events happen regularly on the ICU and it is humanly impossible to record all of them. When inspecting recorded data, medical staff can infer, from experience, the occurrence of some of the unrecorded events. Certain events also manifest their occurrence as observable patterns in the physiological data which can be computationally recovered. We are currently investigating techniques for automatically detecting 'missing' information from the existing recorded information. Certain types of information such as visual observations may always elude automatic detection. This may place an upper bound on the amount of information that can be communicated by the textual descriptions.
3. Unstructured (free) text: Medical staff often record details about specific patients in the form of unstructured (free) textual comments in the patient record. Summarising patient status without this information may be incomplete but it is not possible simply to insert the textual comments into the computer-generated descriptions. We are currently exploring Information Extraction (IE) techniques for automatically extracting important data from the free text comments and integrating them into our textual descriptions.
4. Integrating text and graphics: In our interactions with the medical staff, we found that the eventual acceptance of our new technology will only be possible if it integrates seamlessly into the existing informational infrastructure on the ward. Although our system presents patient information in textual form, there is much to be gained if the texts can be integrated into the graphical presentations. Textual presentations seem to be appropriate for certain contexts - graphical presentations for others.
5. Generating narratives: As observed earlier, human written texts follow a narrative style which helps readers' comprehension. We are currently working on developing techniques to produce texts that mimic expert narrative style.

ACKNOWLEDGEMENTS

The authors are grateful to the UK Engineering and Physical Sciences Research Council for funding of the BabyTalk project-project under award EP/D049520/1.

We would also like to thank the other members of the BabyTalk team: Yvonne Freer, Felix Gao, Robert Logie, Saad Mahamood, Neil McIntosh, Wendy Moncur, Marian van der Meulen and Cindy Sykes.

REFERENCES

[1] E. Alberdi, J-C Becher, K.J. Gilhooly, J. Hunter, R. Logie, A. Lyon, N. McIntosh and J. Reiss, 'Expertise and the Interpretation of Computerised Physiological Data: Implications for the Design of Computerised Physiological Monitoring in Neonatal Intensive Care', *International Journal of Human Computer Studies*,Vol 55, No 3, pp 191-216 (2001).

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

[3] S. Cunningham, S. Deere, A. Simon, R. A. Elton and N. McIntosh, 'A randomised control trial of computerised physiological trend monitoring in an intensive care unit', *Critical Care Medicine*, 26:12, 2053-60 (1998).

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

[5] J. Hunter, L. Ferguson, Y. Freer, G. Ewing, R. Logie, P. McCue and N. McIntosh, 'The NEONATE Database', *Workshop on Intelligent Data Analysis in Medicine and Pharmacology and Knowledge-Based Information Management in Anaesthesia and Intensive Care*, AIME-03, pp 21-24 (2003).

[6] D. Hüske-Kraus, 'Suregen-2: A Shell System for the Generation of Clinical Documents', *Proceedings of EACL-2003* (demo session) (2003).

[7] D. Hüske-Kraus, 'Text Generation in Clinical Medicine – a Review', *Methods of Information in Medicine*, 42, pp 51-60 (2003).

[8] A. Law, Y. Freer, J. Hunter, R. Logie, N. McIntosh and J. Quinn, '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, pp 183-194 (2005).

[9] K. Kukich, 'Design and Implementation of a Knowledge-Based Report Generator', *Proceedings of ACL-1983*, pp 145-150 (1983).

[10] F. Portet, E. Reiter, J. Hunter and S. Sripada, 'Automatic Generation of Textual Summaries from Neonatal Intensive Care Data', R. Bellazzi, A. Abu-Hanna and J. Hunter (eds.), *11th Conference on Artificial Intelligence in Medicine* (AIME 07), pp 227-236, (2007).

[11] E. Reiter and R. Dale, *Building Natural Language Generation Systems,* Cambridge University Press (2000).

[12] S. Sripada, E. Reiter, J. Hunter, J. Yu, 'Summarising Neonatal Time Series Data', *Proceedings of 2003 Conference of the European Chapter of the Association for Computation Linguistics, Companion Volume,* pp 167-170, Budapest, Hungary (2003).

[13] S. Sripada, E. Reiter, I. Davy and K. Nilssen, 'Lessons from Deploying NLG Technology for Marine Weather Forecast Text Generation', R Lopez de Mantaras, L Saitta (eds.), *European Conference on Artificial Intelligence (Valencia, Spain)*, pp 760-764, IOS Press (2004).

[14] S. Sripada, E. Reiter and L. Hawizy, 'Evaluation of an NLG System using Post-Edit Data: Lessons Learnt', *Proceedings of European Natural Language Generation Workshop*, pp 133-139 (2005).

[15] S. Sripada and F. Gao, 'Linguistic Interpretations of Scuba Dive Computer Data', *Proceedings of 11th International Conference on Information Visualization*, pp436-441 (2007).

[16] B. Strople and P. Ottani, 'Can technology improve intershift report? What the research reveals', *J Prof Nurs*, **22**(3), pp 197-204 (2006).

[17] M. van der Meulen, R. Logie, Y. Freer, C. Sykes, N. McIntosh and J. Hunter, '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', *Submitted* (2008).

[18] J. Yu, E. Reiter, J. Hunter and C. Mellish, 'Choosing the Content of Textual Summaries of Large Time-Series Data Sets', *Natural Language Engineering*, **13**, pp 25-49 (2007).