Computers in Intensive Care

If you visit a state-of-the-art intensive care unit (assuming that you are conscious!) you might be struck by the absence of the paper record which used to be clipped to the end of the bed. As in so many areas of life, the computer has taken over, and, sure enough, over there is the screen, keyboard and mouse for the nurses and doctors to look at the patient notes and keep them up-to-date.

It’s not my intention here to debate whether this is a ‘good thing’ or not, but rather to focus on certain consequences of the introduction of computers. The first is that it is now possible to acquire, store and save much more information than was possible in the past. Patients in intensive care are connected via sensors to monitors which can measure a number of things, including heart rate (HR), blood pressure (BP), temperatures (T) and the amount (saturation) of oxygen in the blood (SaO2). These measurements can be made frequently, often once every second. Before the introduction of computers, clinicians could get an instantaneous reading by looking at the monitor, but any idea of the way that the value had varied over the last few hours could only be obtained by entering the value on the paper chart, and this was usually only done once per hour. If you look at the traces on the following page, you will see that hourly sampling is liable to miss potentially important events. Computers can record the detailed information and display it on demand.

The second consequence is related to the first. Because data can be transferred automatically from one system to another, the time spent by the medical staff in entering information can be reduced. The results from laboratory analyses, and on-ward blood-gas analysers can be sent directly to the bedside computer, without the lengthy and error-prone business of someone having to phone up the lab and
laboriously transcribe the data by hand. Ventilators, which are used to support breathing, can be interfaced to the computer and their settings (e.g. respiration rate) acquired directly.

Having all of this information is all well and good, but how is the already overloaded nurse or doctor to make sense of it, in order to extract what they need and want. Let’s be clear, the introduction of computers hasn’t necessarily made this situation worse – ask any doctor who has had to try to find a particular bit of information in a file stuffed with pages of patient notes. The third consequence is that we can use the computer to do much more than just acquire, store and display the information. Research into Artificial Intelligence (or how to enable computers to undertake the sorts of mental tasks that humans can do) is starting to provide us with the tools to do more ‘intelligent’ things with the data.

**Summaries**

Well, what sorts of things? One approach that has received a lot of attention in the past is the so called ‘expert system’ – i.e. given the data on an individual patient and given a fair amount of medical knowledge, the expert system makes direct recommendations about a diagnosis and/or a treatment plan for the patient. A lot of research effort has gone into the development of clinical expert systems but very few are in routine clinical use. No doubt there are a number of reasons for this, but one reason must surely be that highly trained clinical staff are reluctant to be ‘told what to do’. Perhaps a more acceptable way of using the computer would be to extract the most important features of the data – in short to summarise them - and to
BABYTALK

make the summaries available to the clinicians and let them draw
their own conclusions.

The question now arises as to the best way to present such a
summary. You could imagine some sophisticated graphical plot, or
perhaps a set of bullet points, but we opted for what is arguably the
most natural way of imparting information – telling a story. The
following extract from the description of what happened to a baby
over a 45 minute period was written by an clinical expert:

"At 15:15 the incubator is open. The HR = 147–153; oxygen
saturation is 93%; mean BP = 40; T1 and T2 are 37.1º and 36.3º
respectively. Throughout the trace the HR and BP are stable at
around 150–160 and 39–47 respectively. By 15:26 T2 has crept up
to 36.6ºC. At 15:30 the arterial line is accessed to take a blood
sample; there is loss of HR reading during this. From this point the
T2 drifts down, ultimately to 36.1ºC. At 15:37 and 15:47 there are
brief desaturations to the mid 80s; recovery to baseline is
spontaneous."

The next obvious question is whether such textual summaries are
useful to clinicians. To test this we ran an experiment away from the
unit using historical data, in which we presented nurses and doctors
with a number of data sets each covering about 45 minutes. Half of
the time they were shown a graphical presentation (a bit like the one
shown earlier) and half of the time they were given a textual
summary. We then asked them what action(s) they would take at the
end of the 45 minutes. Surprisingly (perhaps) they made better
decisions when they were shown the text1.

This is all very interesting, but it’s not possible to have expert
clinicians standing round doing nothing but writing summaries.
Could computers undertake this task? And to try to answer this
question the BabyTalk project, funded by the Engineering and
Physical Sciences Research Council, was conceived and born in 2006.

Neonatal Intensive Care
As you may have gathered from the previous references to 'babies'
and 'incubators', BabyTalk was concerned with writing summaries of
data collected in a neonatal intensive care unit (NICU). Neil
McIntosh and his team in the Simpson Centre for Reproductive

1 Law A, Y Freer, J Hunter, R Logie, N McIntosh, J Quinn, “A Comparison of Graphical and
Textual Presentations of Time Series Data to Support Medical Decision Making in the Neonatal
Health at the Royal Infirmary of Edinburgh have been experimenting with the use of computers in the NICU for more than 20 years. By the end of 2006, their unit was on the verge of doing away with their paper charts and entering all of their data into a computerised system. This system, supplied by a small Edinburgh-based company called Clevermed, would collect and store all of the physiological data and provide a computer at each each cot-side to allow the clinical staff to enter detailed notes. These computers would also allow them to review what had been entered as well as the lab results and the detailed physiological time series. For the technically minded, all of the information was stored on a back-end database, and at the cot-side a web-browser was used to display web-pages prepared on demand by the server. Clevermed bent over backwards to help us by supplying details of their data formats and giving us access to the (anonymised) data.

We spent the first year building a system (called BT-45) to try to replicate the hand-written 45 minute summaries; we then repeated the previous experiment (with the added presentation of the computer-generated texts) using a new set of scenarios. Reassuringly, the performance using the human written texts was still better than that using the graphical presentations. Performance using the computer generated texts was not as good as using the human texts, but matched that using the graphics.

So far, so good. We’d shown that it was feasible to generate reasonable summaries automatically, but there was still some way to go before they were as good as summaries written by experts. However it was time for a reality check. Summaries covering 45 minutes are of no practical interest. Furthermore we had been using data specially collected for us by a research nurse. If we wanted to make a practical impact it was time to move into the real world. By this time the NICU was paperless - there was a rich source of data that we could use. We decided that we would only use the data that was routinely available - we wouldn’t ask for anything extra to be entered specially for us. This would mean that generating our summaries wouldn’t require any additional effort on behalf of the

---


medical staff. In deciding what period to summarise and for what purpose, we noted that nurses are obliged to produce summaries of their 12 hour shifts. The automatic generation of nursing shift summaries became our objective and we started work on BT-Nurse.

**BT-Nurse**

In order to see what we should be trying to achieve, we again got our nursing expert to generate a number of sample texts (based on real data). Here's an extract from one of them:

> “CMV rate 55, pressures 20/4, in 27% oxygen, giving tidal volumes of 1.5 ml (3.3 ml/kg). Very recent ABG good: pH 7.31, CO2 5.72. He received morphine prior to intubation at 00:30; no spontaneous respiratory effort noted since being re-ventilated. Desaturates during cares and suction but recovers afterwards; otherwise SaO2 has been fairly stable. Large ETT secretions, mucopurulent and blood stained.”

Analysing these texts led us to realise that BT-Nurse would have to have a number of capabilities.

1. The detailed physiological *signals* have to be *analysed* and *abstracted*. See how the text says that the baby "desaturates" - this means a temporary but significant decrease in the oxygen saturation in the blood (SaO2); BT-Nurse has to identify these events, as well as bradycardias (temporary decreases in the heart rate). The text also characterises the overall pattern of SaO2 as "stable" so we need to generate abstractions such as *stable, variable, increasing*, etc.

2. We need to look at the data the nurse has entered and work out when certain events happened - for example, "intubation [the insertion of a tube down the baby's windpipe to allow mechanical ventilation] at 00:30". The nurse will usually enter the fact that intubation has taken place, but she might not do so until some time after the event and we may need to infer a more exact time from other available information.

3. Events are often related to one another. The above text says that the baby "received morphine prior to intubation". These two events are presented together because morphine is often given before intubation in order to relax the baby's muscles and make it easier to insert the tube. If our summaries are going to link these events BT-Nurse has to
have access to that medical knowledge. In fact the acquisition of medical knowledge and its formalisation in software is one of the biggest bottlenecks in building an intelligent medical system.

4. After all of this, we end up with a large number (possibly hundreds) of facts - i.e. things we could say in the summary. But we don't want to say everything - if we did, it wouldn't be a summary! So somehow we have to pick the most important facts; to decide what is important and what isn't, we go back again to our medical experts. But just reporting important facts in any order doesn't make a good narrative. We have to organise them in a way which makes medical sense. The link between morphine and intubation which we established earlier is now used to group these two facts together. The result of all this is a sort of skeleton for the summary - a document plan.

5. The final step is to turn the document plan into good English. This is more difficult than it might seem. Look for example at the number of ways that time is referred to in the example: "He received morphine prior to intubation"; "no spontaneous respiratory effort noted since being re-ventilated". Achieving naturally flowing text is tackled incrementally. For example, "the baby is in air" → "the baby has been in air" → "the baby has been in air since this morning" → "he has been in air since this morning".

Here is an extract from the summary written automatically by BT-Nurse, for the same baby as we looked at earlier:

"The baby was intubated at 00:15 and was on CMV. Vent RR was 50 breaths per minute. Pressures were 20/4 cms H2O. FiO2 was 29%. Tidal volume was 1.5. He was given morphine and suxamethonium. MAP was raised from 6 cms H2O to 8 cms H2O. ...

Currently, the baby is on CMV in 27% O2. Vent RR is 55 breaths per minute. Pressures are 20/4 cms H2O. Tidal volume is 1.5. SaO2 is variable within the acceptable range and there have been some desaturations. The most recent blood gas was taken at around 07:45. Parameters are acceptable. pH is 7.3. CO2 is 5.72 kPa. BE is -4.6 mmol/L. The last ET suction was done at about 05:15."

No, it's not identical to the expert-written summary, but at first sight it seems quite good. The question is: how good?
Evaluation

This brings us to the difficult question of evaluation. In part we undertook this project as pure research, trying to gain insights into solving the problems of turning complex heterogeneous data into an informative and convincing narrative. But we also hoped that our systems would have the potential to bring benefits to real clinical practice. So we have to answer the (somewhat vague) question: how good is BT-Nurse. This question come in two parts: what are you going to evaluate and how?

We could have run the system on data already collected (historical data) and compared our summaries with texts written by experts. One of the problems with this is that you would need a large number of expert-written texts. We had already asked our experts to write 32 summaries so that we could understand what it was we were trying to generate - the excerpt you have seen was taken from one of these. But it is a fundamental law of research that you can’t test your system on data that you have already used to help build it. This is called the separation of ‘training’ and ‘test’ data sets. Writing a summary from historical data took our experts a considerable time - it just wasn’t feasible to ask them to write a lot more.

Another way of evaluating summaries based on historical data would have been to conduct another ‘off-unit’ experiment. But by this time our clinical collaborators were (understandably) a bit fed up with this artificial approach. They wanted to see how BT-Nurse would perform on real babies at the end of a real nursing shift. So we agreed that ‘what’ we would evaluate would be summaries generated on the unit in real time.

The ultimate evaluation of any clinical intervention is the randomised clinical trial (ideally double-blinded) where some patients receive the intervention and others don’t. You then look for any statistically significant difference in some previously agreed output measures. In our case these measures might have been mortality rate, length of stay on the unit, number of complications, etc. However although such trials can provide the desirable objective evaluation, they are very complicated and expensive to design and run. We weren’t ready for this. We decided for a more subjective approach by asking the nurses for their views on the summaries. Subjective evaluation is always open to criticism on the grounds that people don’t always know what is best for them! For example, in our first experiment, the participants said they preferred the graphical presentations whereas they actually performed better with the text
summaries. On the other hand, if the people who potentially are going to use the system don't like it, it's going to be a hard job to persuade them to accept it.

Our evaluation was therefore designed as follows. Near the end of the nursing shift a summary was generated and shown to the nurse whose shift was about to end. Again for the technically minded, Clevermed arranged for the cot-side system to send BT-Nurse a request for a summary; BT-Nurse interrogated the live database, and generated the summary as a web page which was returned to the nurse's screen for display.

We asked the nurses for their level of agreement with three statements:

- The BT-Nurse summary was *easy to understand*.
- The BT-Nurse summary is *accurate*.
- The BT-Nurse summary would *help me* write a shift summary.

The possible answers were: *agree*, *disagree*, *neither*. We also asked them to enter any comments that they had about the summary. In addition, we showed summaries to nurses who were coming on shift and asked them the same questions - except that the third statement about being helpful was worded:

- The BT-Nurse summary would *help me* in care planning*.

We ran live on the ward during December 2009 and January 2010, getting 165 separate evaluations involving 31 different babies and 54 individual nurses. Overall, there was 91% positive agreement that the summaries were understandable, 74% agreement that they were accurate, and 61% agreement that they were helpful. When we analysed the comments, we found that most of them were concerned with inaccurate or missing content; very few were critical of the way the summaries were written.

**Where Now?**

We believe that BabyTalk has demonstrated that it is possible automatically to write meaningful and useful clinical summaries from a complicated electronic patient record. The evaluation shows that not all of our summaries were considered to be useful but 61% isn't a bad starting point for what was, after all, an experimental system. It looks as if we have resolved most of the issues with generating text; what is needed is additional and more accurate medical knowledge.
BABYTALK

One thing that we are considering is whether clinicians are the most appropriate recipients of our summaries. There is an argument that they prefer to get their information in a more structured way, so that they can find a specific item more efficiently than they would if they had to read through a more informally written narrative. But what about non-medical recipients? Clearly in neonatal intensive care, our patients are unable to read, but their parents can, and within BabyTalk we are trying to generate the kinds of summary that they would find acceptable and useful. Let us be clear - we are not trying to supplant the role of doctors and nurses in imparting sensitive information about life-threatening conditions, but premature babies can spend months in the unit and it has already been shown that the parents welcome daily bulletins on how their baby is doing. Our objective here is to improve the quality of these reports.

However, we can look beyond neonates. If you had spent some time in intensive care (and were probably unconscious for part of it), wouldn't you welcome a summary of what had happened to you while you were there? In fact we needn't limit ourselves to intensive care. How often have you come away from a visit to your GP wishing that you had a permanent record of the various test results, what they meant, and how they had affected the decisions regarding the medication that you were taking.

The original BabyTalk project is now over. We believe that our original assumptions that it was possible to build a useful system to generate summaries from clinical data have been vindicated. The research has been done. If there is anyone out there who is involved in building commercial medical systems, our door is always open ... !

On a Personal Note
I am immensely grateful to all those involved in the BabyTalk project which has been one of the most enjoyable that I have been involved in: in Aberdeen - Ehud Reiter, Yaji Sripada, François Portet, Albert Gatt, Feng (Felix) Gao, Saad Mahamood, Wendy Moncur, Dave Westwater; in Edinburgh - Neil McIntosh, Yvonne Free, Bob Logie, Cindy Sykes, Marian van der Meulen, Rahul Sambaraju; and for Clevermed - Peter Badger, Andy Lyon, Tom Lyon.

When I started working in medical computing over thirty years ago, electronic clinical data were like hens' teeth. In fact one of the reasons I moved to Aberdeen was the chance to work with data from an adult ICU. Now, at least in the ICU, clinical databases are full to bursting and my only regret is that I'm not thirty years younger. So what else is new?!