Alberdi, E., Becher, J.-C., Gilhooly, K. J., Hunter, J., Logie, R., Lyon, A., McIntosh, N. & Reiss, J. (2001). Expertise and the interpretation of computerized physiological data: implications for the design of computerized monitoring in neonatal intensive care. International Journal of Human-Computer Studies, 55(3), 191 - 216. doi: 10.1006/ijhc.2001.0477 <http://dx.doi.org/10.1006/ijhc.2001.0477>

Expertise and the interpretation of computerised physiological data: Implications for the design of computerised physiological monitoring in neonatal intensive care

Eugenio Alberdi[a], Julie-Clare Becher[b], Ken Gilhooly[a], Jim Hunter[a], Robert Logie[a],

Andy Lyon[b], Neil McIntosh[b], Jan Reiss[b]

[a]University of Aberdeen, Aberdeen, Scotland

[b] University of Edinburgh, Edinburgh, Scotland

Summary

This paper presents outcomes from a cognitive engineering project addressing the design problems of computerised monitoring in neonatal intensive care. Cognitive engineering is viewed, in this project, as a symbiosis between cognitive science and design practice. A range of methodologies has been used: interviews with neonatal staff, ward observations, and experimental techniques. The results of these investigations are reported, focusing specifically on the differences between junior and senior physicians in their interpretation of monitored physiological data. It was found that the senior doctors made better use than the junior doctors of the different knowledge sources available. The senior doctors were able to identify more relevant physiological patterns and generated more and better inferences than did their junior colleagues. Expertise differences are discussed in the context of previous psychological research in medical expertise. Finally, the paper discusses the potential utility of these outcomes to inform the design of computerised decision support in neonatal intensive care.

1. Introduction

Advances in medical informatics offer considerable potential for improving the quality of medical and nursing care in a variety of health care domains. However, there is extensive evidence to suggest that computerised aids in medicine are not always readily accepted or widely used by medical or nursing staff, and often fail to produce the sought-for clinical improvements (Green, Gilhooly, Logie, Ross, 1991; Cunningham, Deere, Simon, Elton, & McIntosh, 1998; Morgan, Takala, DeBacker, Sukuvaara, Kari, 1996). The most common reason given for these difficulties has been a failure in system design to incorporate an adequate knowledge of the cognitions and working practices of the eventual users (see e.g. Coiera, 1994).

One way of addressing the problem is to use cognitive engineering. This discipline has been traditionally characterised as the application of theories and models developed by cognitive psychologists to inform the design of human-computer applications (Norman, 1986). In the last two decades, a great deal of effort has gone into this enterprise. However there are strong suggestions that psychological knowledge has not had a significant impact on system design (see e.g. Barnard & Harrison, 1988; Carroll, 199; Landauer, 1987). This has led to new characterisations of cognitive engineering which essentially exclude psychological practices from human-computer studies (e.g. Long & Dowell, 1989, 1996). In contrast, we argue that cognitive psychology can play an important role in engineering design, and that system design and psychological theories and methods can support each other by maintaining a symbiotic relationship (Alberdi & Logie, 1998).

In our view, cognitive science can play (and, in fact, has played) important roles in the development of usable knowledge for human-computer interaction. If rightly applied, a great deal of what is known about human cognition can have important implications for design. The application of a cognitive model (e.g. Card, Moran, & Newell, 1983; Wickens, 1992), coupled with a sound analysis of the application domain and extensive empirical psychological investigations, has often resulted in successful contributions to the design process (e.g., Edworthy & Stanton, 1995; Egan, Remde, Gomez, Landauer, Eberhardt, & Lochbaum, 1990; Gray, John, & Atwood, 1993; Green, Logie, Gilhooly, Ross, & Ronald, 1996). Additionally the results of design-oriented task-specific psychological investigations can feed back into the cognitive theory from which they were generated. The drive to develop a particular piece of technology has often forced questions on the psychological theories which informed the development. And, as a result, these theories have been refined and enhanced. Hence our view of cognitive engineering as a symbiosis between cognitive science and design practice.

We have used this approach to deal with the problems of computerised monitoring in neonatal intensive care (Logie, Hunter, McIntosh, Gilhooly, Alberdi, Reiss, 1997; Alberdi, Becher, Gilhooly, Hunter, Logie, Lyon, McIntosh, & Reiss, 1999). Specifically, we have conducted a series of investigations in the neonatal intensive care unit (ICU) of the Simpson Maternity Hospital in Edinburgh (UK), where a PC based trend monitoring system (MARY^{TM})[1] has been in use for more than 10 years (McIntosh, Ducker, & Bass, 1989). The computerised system was generally welcomed by the clinical staff, who positively valued its utility (Deere, Cunningham, McIntosh, 1992). However, recent studies at the unit have shown that the presence of a computerised trend monitoring system does not in itself result in better outcomes in terms of morbidity and mortality (Cunningham et al., 1998).

The goal of these investigations was to study the users' (physicians and nurses) cognitions and working practices, with a view to evaluating the usability of the currently implemented system and contributing to the design of computerised decision support in intensive care. Our work has been partly guided by psychological theories of medical expertise (e.g., Gilhooly, 1990). Expertise differences are undoubtedly relevant to the design problem we are dealing with. Typically, computerised systems are designed by medical experts but the main users are nursing and junior medical staff in training. It is likely that these staff have different personal knowledge bases and so interpret data in different ways from the experts. Therefore, we need to know how the front-line users of the system differ from experienced clinicians in the way they interpret data and use their knowledge.

[1] MARY^{TM} is a trademark of Meadowbank Medical Systems.

In consonance with current research in complex naturalistic decision making environments (including intensive care; see e.g. Patel, Kaufman, & Magder, 1996), our approach has been to use a range of methodologies, namely, interviews with and observations of clinicians (physicians and nurses) working in the neonatal unit, as well as experimental work ("off-ward" simulations) to study the cognitions of the clinicians.

The rest of this paper is organised as follows. The next section presents a brief introduction to the problems of computerisation in neonatal intensive care, focusing on the monitoring system under study. This is followed by a review of the psychological literature on medical expertise. The following section presents a brief report of results from the interviews and observations. Next the "off-ward" simulations are discussed in detail. A discussion follows which highlights the implications of our results for cognitive psychology and for the design of human computer interaction in neonatal intensive care.

2. Computerisation in the neonatal ICU

The clinical monitoring of patients in the neonatal ICU has three objectives: (1) to confirm that the baby is stable and responding appropriately to therapy; (2) to allow early detection of abnormal physiological events, with a view to rectifying problems before they become too established; (3) to detect situations in which the baby is *not* responding appropriately to treatment thereby requiring alternative action.

Information technology is intended to assist in the achievement of these objectives, and intensive care wards for both adults and infants have seen a rapid increase in the data available to the clinical staff. Current monitoring systems can display information on a variety of physiological parameters: heart rate, blood pressure, blood gases, respiratory rate, body temperature, etc. Often each physiological parameter is displayed on a separate monitor and in a different format. However, physiological conditions can be indicated by changes in several of these parameters; such an arrangement can therefore result in significant complications for scanning and assimilating the data displayed.

An important development in the last decade has been the use of computers to collect data from different monitors and to display them in a more uniform format (Green *et al.,* 1996). Computer systems offer a means to avoid some of the information overload arising from multiple monitors through use of integrated and flexible displays. The computerised system used in our studies (Cunningham, Deere, Elton, McIntosh, 1992) is a good example of this type of software. One of the most distinctive features of this system is its presentation of monitored physiological data as trend graphs. The system shows physiological trends over long periods of time, in contrast with most conventional monitors which only present the value at a particular moment in time. Data presentation in the form of trends is deemed to facilitate clinicians' assessment of the data and propitiate rapid and effective decision making in emergency situations (Cunningham *et al.*, 1992). The system allows continuous collection of physiological information which is automatically recorded and displayed on a PC at the cotside. It allows the display of real time and previously recorded trend data: when monitoring in real time, data from any period of the infants' monitored stay in the ICU can be recalled. Important features of the system are the flexibility of its display and the ease with which this can be manipulated. Furthermore, the user can enter information or comments in real time by a cursor in the recorded trend data; nursing staff are encouraged to enter comments about procedures and tests performed on a baby, as well as about relevant clinical events occurring to an infant. The whole system is based on menus, which the users can access using a standard keyboard.

A further potential contribution of computer technologies is the development of decision support systems to assist in the interpretation of monitored data (Coiera, 1993). Such systems can support medical decision making by optimising the display content and format for the physiological condition of the patient, by detecting patterns of change or stability in several different parameters, and by recording parameter values for cumulative displays. The major difference between a computerised monitoring system and a decision support system is in the level of interpretation, organisation and selection of available data. The medical decision support system in the ICU has to meet all of the objectives of the monitoring system, but also has to make data available in a form which facilitates decision making. Advances in artificial intelligence (e.g., Salatian & Hunter, 1996; Taboada, Arcay, Arias, 1997) and in the World Wide Web (e.g., Nenov & Klopp, 1996; Norris, Dawant, Geissbuhler, 1997) are contributing to the development of medical decision support in intensive care. However, as suggested above, a common concern is that much of this work is still technology-driven rather than user-driven (Coiera, 1994; Gremy & Bonnin, 1995).

3. Literature on expertise and medical reasoning

A great deal of the psychological research conducted in medical reasoning has been aimed at determining the nature of expertise in diagnostic thinking. Research expertise in a range of non-medical domains suggests the following (Ericsson and Charness, 1997; see also, Feltovich, Ford and Hoffman, 1997; Ericsson and Smith, 1991; Chi, Glaser and Farr, 1988): (1) experts perform better than novices because they possess superior domain knowledge accumulated after many years of extensive practice, and not because of superior basic capacities (Ericsson and Lehman, 1996); (2) because they have a richer repertoire of relevant schemata, experts can remember more new information in their field than novices (Chase and Simon, 1973), and (3) have better problem representations in terms of the deep structure of the problem whereas novices are led by the surface features of the problem (Larkin, 1983); (4) experts tend to work forwards (i.e. from the starting state to the goal state), whereas novices work backwards from the unknown to the givens (Larkin, McDermott, Simon and Simon, 1980).

Two significant discrepancies between research on medical problem solving and general research in expertise have been noted (Gilhooly,1990; Patel and Groen, 1986, 1991; Patel, Arocha and Kaufman, 1994; Patel and Ramoni, 1997). One discrepancy is that medical experts do not seem to show a strong tendency to work forward to a goal state. In pioneering studies of diagnostic thinking (Elstein, Shulman, & Sprafka, 1978), it was found that expert physicians generated hypotheses very early in the process, after seeing just a few signs or symptoms; these hypotheses were then tested, checking for the presence or absence of symptoms deduced from the hypotheses. This approach, which can be characterised as one of hypothetico-deductive reasoning, involves reasoning backwards from the goal (the hypothesis) to the given (the symptoms). A reasonable explanation for this type of processing is that in diagnostic thinking not all the necessary information is presented initially; hence, the task requires information search, and this search is usefully guided by hypotheses (see discussions in: Gilhooly, 1990; Elstein, Shulman, & Sprafka, 1990; Patel and Ramoni, 1997).

A second discrepancy between research on medical diagnosis and general expertise research is that medical experts do not always remember more information about new cases than less experienced subjects. These findings have emerged from studies in which subjects with various degrees of expertise are presented with a short text containing details of a clinical case and, after a brief study period, are requested to recall the text and state the most likely diagnosis. Such studies show that subjects of intermediate levels of expertise recalled case information better than either more expert or less expert subjects (e.g., Claessen & Boshuizen, 1985; Patel & Groen, 1986). This pattern of results is generally known as the “Intermediate Effect” on memory (Schmidt, Boshuizen, & Hobus, 1988; Schmidt and Boshuizen, 1993; Patel and Groen, 1991). However, even if experts have poor memory for the specifics of a case, the diagnoses produced by them for that same case tend to be more accurate than those produced by less experienced subjects. This data pattern has been explained by the different forms of knowledge brought to bear on the task by expert and less expert subjects (Schmidt, Norman and Boshuizen, 1990). Experts seem to use what is usually referred to as “clinical knowledge”, that is, compiled knowledge in the form of “illness scripts” which contain prototypical information about diseases (Feltovich &

Barrows, 1984). This usually allows a reasonable diagnosis with little processing of the textual (case) information. On the other hand, subjects with intermediate levels of expertise seem to rely on “biomedical knowledge”, that is, knowledge of underlying pathophysiology and anatomy. Since such knowledge is not grounded on the personal experience of the subjects, they have to reason from first principles, resulting in a slower and often less accurate processing of the information.

Many studies on medical reasoning support the notion that less experienced subjects make extensive use of biomedical knowledge whereas expert subjects use it sparingly (e.g., Boshuizen and Schmidt, 1992; Elstein *et al.*, 1978; Lemieux and Bordage, 1986). However these findings are contradicted by research that looks at diagnostic tasks involving the interpretation of visually presented patient data, such as radiology and ECG interpretation. For example, in studies dealing with the interpretation of X-rays, Lesgold and colleagues (Lesgold, 1984; Lesgold, Glaser, Rubinson, Klopfer, Feltovich & Wang, 1988) found that expert subjects made more explicit use of biomedical knowledge. Their studies suggest that experts’ diagnostic reasoning is, in fact, opportunistic and will exploit whatever knowledge sources are available in the task. Similarly, Gilhooly and colleagues, (Gilhooly, McGeorge, Hunter, Rawles, Kirby, Green, and Wynn, 1997) in a study of the interpretation of ECG traces found that expert subjects used both clinical and biomedical knowledge more frequently than novices or intermediates. In contrast, less experienced subjects tended to generate a larger proportion of trace descriptions. Furthermore, novices and intermediate subjects were less likely to reach hypotheses than were the more experienced subjects, and when they reached hypotheses they were less likely to evaluate them by biomedical knowledge than the experts. Gilhooly and colleagues explained their findings, and those from Lesgold and colleagues, by noting that radiology and ECG interpretation tasks usually involve uncontextualised information: the experimental tasks were tackled without provision of background clinical information, in contrast with the other studies discussed, in which such information was available (e.g., Feltovitch & Barrows, 1984; Boshuizen & Schmidt, 1992). They argued that, when interpreting uncontextualised perceptual information, experts may need to use biomedical knowledge to discriminate amongst hypotheses that make similar predictions about the surface appearance of the data. Gilhooly and colleagues' conclusion is that experts do indeed use the "short-cuts" (e.g., application of "illness scripts") facilitated by their acquired clinical knowledge if relevant contextual information is available. However, when such information is missing, experts can effectively reason from underlying principles (i.e., application of biomedical knowledge).

It is important to note that not all the findings arising from studies of diagnostic thinking reflect significant expertise differences. Some results suggest important commonalities in the reasoning processes of experienced and less experienced clinicians. For example, early studies (Elstein *et al.*, 1978) found no skill-related differences in diagnostic process between experts and less qualified subjects. In particular, no quantitative differences were found in hypothesis processing or information use patterns between the two groups. Similarly, in a recent study of mammography interpretation, Azevedo (1998) found that experts and non-experts did not differ in the types of problem solving operators and diagnostic plans they used, or in the number and types of errors they committed. Differences were found, however, in terms of processing speed: experts scanned the radiological information significantly faster than the less experienced subjects. Similar results were also found by Joseph and Patel (1990).

To sum up, the literature on expertise differences in diagnostic thinking seems to suggest the following major patterns: (a) experts perform better than novices not because they use superior skills, but because they possess superior domain knowledge (a richer repertoire of schemata); (b) as a consequence, experts have a better representation of the domain than do novices; this allows them to focus on those aspects of the task which are more relevant, and thus process information faster and more accurately; (c) experts' problem solving is opportunistic: they make better use than novices of whatever sources of information are available and relevant to the task, and search effectively for relevant missing information.

4. Study 1: Interview and Ward Observation

All the studies reported in this paper were conducted in the neonatal ICU of the Simpson Maternity Hospital in Edinburgh (UK). The unit has 12 intensive care cots, 14 high dependency cots and 14 special care cots, with about 650 admissions per year. Four general paediatric surgeons and 6 consultant neonatologists, among other staff, work regularly at the unit.

As noted earlier, we will focus here on data relating to expertise differences between senior and junior physicians. Other findings from our interviews and observation sessions are reported elsewhere (Alberdi, Becher, Gilhooly, Hunter, Logie, Lyon, McIntosh, Reiss, in press).

4.1 INTERVIEWS

The purpose of the interviews was to obtain a subjective view of working practices, staff attitudes and perceived expertise, as well as information about their data interpretation procedures and their use of information sources.

Participants

Seven senior and eight junior physicians, working in the Neonanatal Unit of the Simpson Maternity Hospital, participated in our interviews. Five of the senior physicians were the consultants working at the time in the unit. They had an average of 12 years of experience (minimum five years, maximum 26 years) in neonatal care. The other two senior doctors were senior registrars who had had five and nine years of experience respectively in neonatal care. On the other hand, six of the junior physicians were senior house officers (SHOs), whose experience in neonatal care ranged from 4 months to 2 years (an average of less than one year). Two of the junior physicians occupied slightly more senior positions, namely, a registrar with less than a year of experience in neonatal intensive care, and a staff grade doctor who had worked in neonatal intensive care for less than five years.

Procedure

The questions asked during the interviews covered the following areas: (a) position and clinical experience of the interviewees as well as their responsibilities at the unit; (b) sources of information used to make clinical decisions on the ward; (c) the ways in which staff deal with monitoring artefacts (i.e., changes or disturbances of the monitored data which do not reflect the real state of the baby); (d) experience with computers, attitudes towards the computerised monitor, and the ways of interacting with the system.

Results

The most relevant findings from the interviews can be summarised as follows:

1. Most staff (94-95%) reported that the system (MARY) was useful and noted that trend monitoring (one of the most distinctive features of the system) was very helpful for their decision making.
2. Fewer junior doctors (75%) mentioned the system as a source of information they would consider when making decisions about the state of a baby, compared with 100% of the senior doctors.
3. In contrast, when asked specifically about how often they used the system, more junior doctors (75%) reported using the system "very frequently" or "constantly", compared with the senior doctors (57%).
4. Junior doctors were less likely (25%) to know how to alter various aspects of the data display on the computer monitor than were senior doctors (71%).
5. Only a small proportion of the junior doctors (37.5%) was able to suggest ways in which the computerised monitor could be improved; in contrast all senior doctors suggested improvements.
6. Whereas all the senior physicians reported being able to identify at least some of the most frequently occurring monitoring artefacts, only 33% of the junior doctors reported that they were able to do so.

7. Few of the interviewees reported receiving any training on the system (14% of the senior doctors, and 37.5% of the senior doctors); but the junior doctors were clearly more concerned about lack or shortage of training (100%) than were the senior doctors (57%).
8. More junior physicians (75%) reported having experience with various computer applications (other than MARY), compared with the seniors doctors (57%).
9. The following sources of information were suggested by interviewees (especially senior physicians) as data they would like to have online to take better advantage of trend monitoring: (a) information about test results (e.g., arterial samples, X-rays); (b) information about ventilator and incubator settings; and (c) in general, all information about the history of the baby and the mother, as well as nursing and medical notes, to get rid of paper notes altogether.

4.2 OBSERVATIONS

We conducted extensive observations of the neonatal ward to obtain a more objective picture of clinicians' working habits and performance. The observations provided an interesting contrast to some of the interview data.

Procedure

We conducted 8 observation sessions at the neonatal unit of the Simpson Maternity Hospital. Each session lasted from 1 to 2 hours, giving a total 13 ½ hours worth of observation data. Many of the members of staff who participated in the interviews were present at the ward during the observation sessions. In a preliminary session, the observer sat at the unit and noted all the different activities that staff conducted. This produced an encoding scheme which was used in the following sessions. In each session, a record was kept of the frequency with which each activity was conducted by different members of staff.

Results

Table 1

The following activities were identified: interacting with the computerised monitor, looking at the baby, handling the baby, handling equipment and substances, talking to colleagues, writing/reading paper notes, dealing with the alarms, and "other" rarely observed activities (i.e., supervising the ward, interacting with relatives, and looking at X-rays).

Table 1 summarises observation data, including information about nurses. The table suggests that the junior doctors were the staff group who interacted the least with the computerised monitor, compared with the senior doctors (and even the nurses). Furthermore, interaction with the computer was one of the least frequently observed activities amongst the junior doctors. The use of the system accounted for only 4.5% of the activities conducted by them. Furthermore, this small percentage accounted for all the interactions with the system that occurred in only two of the eight observation sessions. Many of the junior doctors were never seen using the system at all. In contrast, the senior doctors were the staff group who used the system most frequently. Interaction with the system amongst the senior doctors accounted for 13.50% of all the recorded activities and was the fourth most frequent activity undertaken by them.

The low frequency with which staff were seen interacting with the computerised monitor is particularly significant if we consider that most (or all) of the staff being observed knew that the observer was involved in a research project related to the computer system. There was no indication that staff tended to use the system more often while being observed than they normally would.

In summary, our interview and observation data suggest that junior doctors use the computerised monitor less frequently than do the senior doctors and are thus less likely to benefit from the monitoring information provided by the system.

5. Study 2: Off-ward simulations

The purpose of the off-ward simulations was to study, in an experimental setting, the cognitions of clinical staff while interpreting monitored data. Our simulations are a substantial simplification of the interactions that staff may have with the monitoring system in real life. During the experiments, staff had to rely mostly on the information provided by the monitored trends. Our aim was not to replicate in detail the complex decision making scenario of a neonatal unit. Rather we wanted to assess how much could be inferred about the condition of a baby by using only monitoring information. The participants were allowed to request extra information from the experimenter only after they had exhausted all the interpretations they could derive from the trends. One of our goals was to find out what other information they needed, in addition to the monitored trends, to make decisions efficiently about a patient.

5.1 METHOD

Participants

The data reported in this paper correspond to the off-ward simulations run with 5 senior doctors and 5 junior doctors. All the senior doctors who took part in the simulations had previously participated in the interviews reported above. The junior doctors were all Senior House Officers (SHOs) who had been recently appointed in the unit and had less than six months of experience in neonatal intensive care.

Material

Each staff member viewed on the computer screen 14 different physiological traces recorded from previous patients (babies) on the ward. The traces were selected by clinical experts (McIntosh & Reiss) from a database kept at the Simpson Maternity Hospital. Each trace comprised two hours of recorded data. In all traces, the same five physiological measurements were displayed in the following order (from top to bottom): heart rate, trans-cutaneous oxygen, trans-cutaneous carbon-dioxide, toe-core temperature differential, mean blood pressure. For those trace samples in which a given parameter was not monitored, the parameter would still be shown on the screen (although blank).

The vertical axes of the parameter graphs were scaled to show the appropriate physiological range for each baby. This range was determined by the clinicians involved in the selection of the stimuli (as well as running the experiments).

Each trace (with the exception of two control traces) contained a key clinical "event" that the participants were expected to identify. In particular, four types of "key events" (or "non-events") were used on the traces:

Baby's reaction to the administration of drugs

- Traces 1 & 2 contained the key event "administration of surfactant"
- Traces 3 & 4 contained the key event "administration of dopamine"

Spontaneously occurring pathological key events

- Traces 5 & 6 contained the key event "developing pneumothorax"
- Traces 7 & 8 contained the key event "blocking of the endo-tracheal tube"

Baby's reaction to regular procedures

- Traces 9 & 10 contained the key event "electrode change"
- Traces 11 & 12 contained the key event "all care" (i.e., a regular procedure which involves cleaning the baby, reapplying or fixing tubes, electrodes & probes, and various other activities)

The last two traces (Traces 13 & 14) were "control traces" that only contained artefacts.

In addition to the "key event", each trace contained several other clinically significant events or noteworthy artefacts which ought to be identified by qualified staff. These were recorded by a clinical advisor (McIntosh) prior to the running of the simulations. In this paper, we will refer to these secondary, but noteworthy, patterns as the "relevant patterns" or the "relevant events", to differentiate them from the "key events".

The traces were selected in such a way that, when possible, they all possessed the following characteristics: a) they all have elements of ambiguity, that is, the identification of the event is not obvious; b) they all contain roughly the same number of artefacts; c) the onset of the "key events" does *not* appear in the same place on all traces.

All participants saw the same 14 traces, but the presentation order was randomised for each participant: each saw a different sequence of traces. Figure 1 shows an example of the 7-minute blocks of monitored data used during the simulations.

Figure 1

Procedure

Each participant was told that the goal of the experiment was to study how computerised trend data influence the way s/he thinks about the neonatal ICU patients. The participant was then informed that s/he was going to view some trends of past babies on the computer screen, and that some of those trends were going to be uneventful, some were going to show normal events, and that some were going to show developing pathology. The participant was told that s/he was going to see in all traces the same five channels of data, scaled physiologically, and in two different time scales.

The participant was instructed to think aloud while looking at the traces, reporting everything that went through her/his mind. S/he was instructed to point at the abnormalities or artefacts that s/he saw on the traces and, if possible, to provide an interpretation.

Each trace, which contained 2 hours worth of data, was shown on a computer screen as a series of seven minute blocks of data; subsequently, the trace was shown again on a different time scale, namely as two 1.5 hour blocks (with ½ hour overlap) of compressed data.

The experimenter had full control over the manipulation of the computer display (i.e., scrolling between blocks of data, modifying the graph scales, etc.). The participant was told that s/he was allowed to ask the experimenter to scroll back and look at what had happened earlier, but was also informed that the experimenter would never scroll forward to the next block of data until the participant had said all s/he wished to say about the trace to that point. S/he was informed that at that point s/he could ask for more information which would be given to her/him if it would have been available at that time clinically. The participant was instructed to clearly state when s/he wanted to move on to the next block of data. Prior to the presentation of each trace, the participant was given a card with basic information about the baby from whom the trace was derived. This information consisted of: a) baby's weight; b) baby's gestation; c) baby's age; d) whether the baby was ventilated; e) percentage of ventilating air given to the baby at the start (if ventilated).

The only way in which the participant could interact with the computer system was by pointing at the display to clarify what specific physiological patterns s/he was referring to in her/his speech. All sessions were recorded on video to capture the computer display with the participant on the side speaking and pointing at the screen. The participant's speech was captured by a microphone attached to the video camera.

5.2 RESULTS

The simulation sessions generated a total of 140 video-recorded protocols (i.e., 14 protocols per participant). Two of those protocols (one from one senior doctor and one from one junior doctor) could not be used in the analyses because of technical problems. The analyses we report below were conducted on the remaining 138 protocols.

The study was a mixed between and within-participants design; with expertise as the between-participants factor and trace as the within-participants factor.

Protocol analyses

The resulting video-recorded think-aloud protocols were transcribed and analysed using standard protocol analysis procedures (Ericsson and Simon, 1984).

The transcription of the protocols involved: (a) *transcribing* verbatim the participants' verbal reports to reflect as accurately as possible their speech, as well as noting non-verbal aspects which may be meaningful (e.g., pauses, emphasis, etc.); (b) *dividing* the transcript into as many paragraphs as trace segments were seen by the participants in each trace – that is, into about 19-21 paragraphs, corresponding to the 17-19 "seven minute blocks" plus the two "compressed" blocks; each transcript paragraph was marked with a time interval which denoted its corresponding time segment on the trace; (c) *noting* on the transcript whether a participant points at the screen, and marking the parameter change s/he is pointing at, as well as the time that change is taking place on the trace.

Subsequently, each protocol was *segmented*. This involved dividing a participant's comments into statements, and listing them one per line. A statement represents a single idea, a basic unit of thought. Typically a statement contains a comment which refers to only one of the 5 physiological parameters displayed on the screen (example: "There is a significant drop of pO2 towards the end of the screen"). This process yielded 17,888 statements (11,278 for the senior doctors and 6,610 for the junior doctors).

A major component of the protocol analysis was the generation of an encoding scheme to characterise the cognitive processes used by the participants. The development of the encoding scheme involved two procedures. The first procedure was generating a label to describe the behaviour represented in each statement. The labels were meant to be mutually exclusive. However, often the same statement could involve two different behaviours, in which case two different labels were applied. Following with the example above, the statement comprises two behaviours: the description of a physiological change (i.e., "drop in pO2"), and an interpretation of that change (i.e., "significant"). The second procedure involved generating a list of the labels, including a description of the behaviour associated with each label.

Table 2

As a result, 10 labels/categories were generated. This set was produced after analysing the first protocol of the first participant (a consultant) and was partly guided by the authors' previous experience in a similar investigation (Gilhooly et al., 1997). Since the creation of an encoding scheme is always a dynamic process, the scheme suffered a few minor refinements as new protocols were analysed. Eventually, a set of criteria was established to determine the application of a code to a statement; these criteria are outlined in table 2.

Table 3

Using this scheme, a total 20,608 codes were generated for the whole set of protocols (12,850 codes for the senior doctors and 7,758 for the junior doctors). A highly significant correlation was found between the percentages of code frequencies for the senior doctors and the corresponding proportions for the junior doctors [$r(8) = 0.99$; $p < .001$]. This strongly suggests that both groups used essentially the same processes, and with equivalent relative frequencies, during the simulations. The most frequently identified processes for both groups were "Describe", "Interpret" and "Hypothesis". They account for 68% of all the coded behaviours. The category "Other" also accounts for an important proportion of the codes; but, as noted, this category comprises many other sub-processes which were not deemed to be relevant behaviours individually.

Table 2

The data in Table 3 suggest some differences between the two staff groups. For example, the usage of "Describe" and "Interpret" accounted for a larger proportion of junior doctors' behaviours than those of the senior doctors, although the differences were not statistically significant (see column 4). In contrast, the proportions with which the remaining types of behaviour were used were always higher for the senior doctors than for the junior physicians. The only behaviour for which there were statistical differences between senior and junior doctors was the frequency with which they noted artefacts (see column 4).

Requests for "extra information"

We noted earlier that, during the simulations, staff were allowed to ask for extra clinical information to complement the information provided by the monitored data. As reported earlier, about 6% of the statements produced by the senior doctors and 4% of the statements of the junior doctors contained requests for extra information. Those statements were analysed in detail to identify the most common types of information requested. We focus here on those aspects of these analyses which are relevant to expertise differences. More details can be found elsewhere (Alberdi *et al.*, in press).

The analyses showed that staff required information about: (a) the *baby*, more specifically, its state and appearance; (b) *procedures* conducted on the baby, for example, whether the baby has been handled in some way or some drug has been administered; (c) the *settings* of the machinery attached to the baby, more specifically, the ventilator settings and the incubator settings; (d) clinical *tests* and examinations conducted on the baby, for example, arterial blood samples and X-rays; (e) changes to the computerised *monitor display*, for example, requests to change the axis scale or requests to scroll back to previous data blocks; (f) *colleagues'* impressions or knowledge about the state of the baby; (g) the *calibration* of probes or leads, that is, whether the probes are correctly calibrated and whether they show real physiological values; and (h) finally, *"other"* statements where a person indicates that s/he would need further information but does not clearly specify what that required information is.

A tally was made of the number of times each type of extra information was requested in the protocols. Interesting differences were found between senior and junior doctors. On the one hand, the most frequent requests by senior doctors were to change the displays on the monitoring system (42% of the requests); they requested this information considerably more often than did junior doctors (11.50%). On the other hand, the most frequent requests from the junior doctors concerned information about procedures conducted on the baby (33% of their requests); in contrast with the senior doctors, who did not request this information quite so often (7%). Further, the second most frequently requested type of information by the junior doctors was information about the baby (19%), which suggests that the junior doctors are more likely to rely on the information obtained from a direct contact with the baby than on other sorts of information (e.g., the information provided by the computerised monitor). The senior doctors did not request information about the baby as often as did the junior doctors (14% of the requests). As regards the other types of information noted, the differences between senior and junior doctors were not remarkable.

It is interesting that much of the information requested during the simulations coincided with what many staff members reported in the interviews when asked about the additional information they would like to find online in a computerised monitoring system (see Section 4).

Identification of "key events"

As noted earlier, each of the traces (except the control traces) contained a key clinical "event" that the participants were expected to identify. A participant was said to identify one of these events if s/he generated a hypothesis containing the name of the event (or a synonym), and this hypothesis was generated to explain the physiological changes on the trace associated with the event. An expert clinical advisor (McIntosh) assisted in the analysis. The control traces (13 & 14) were excluded from the analyses. It was found that the senior doctors identified, as an average, 8.30 out of the remaining 12 "key events" (sd= 1.30; 69%), whereas the junior doctors recognised 7.80 (sd= 1.64; 65%). ANOVA showed no influence of seniority [F <1].

There is no conclusive evidence to suggest that the more experienced doctors have any advantage over the more junior staff when detecting "key" physiological events. In fact, the event recognition standards of both groups were fairly low. However, as shown below, more obvious expertise differences arise when more fine-grained data analyses are involved.

Detection of relevant patterns

As noted earlier a medical expert (McIntosh) identified, for each trace, a set of "relevant events" in addition to the "key event". The clinical expert generated a total of 314 events, with an average of 22 events per trace (minimum: 1 event for the control trace 14; maximum: 42 events for trace 10). In his records of "relevant events", the medical expert noted: (a) the time each event started; (b) the duration of the event; (c) the nature of the event (i.e., what physiological parameter changed and the type of change); (d) an inference about the possible causes for that parameter change or whether it is an artefact.

The records of events generated by the expert advisor were used as a "gold standard" with which to compare the participants' performance during the simulation. For each trace, a tally was made of the number of events reported by each participant that matched the events recorded by the expert. A participant is said to report an event that matches an event recorded by the expert if s/he either: (a) *describes* a change on a physiological parameter on the trace and this pattern coincides (both in its nature and the time of occurrence) with an event reported by the independent expert; or (b) does *not* describe a pattern but provides an interpretation or inference which refers to a particular monitoring pattern noted by the independent expert, hence it can be assumed that the participant has detected the pattern.

A larger proportion of the "relevant events" (N = 304 excluding the events in the control traces) was identified by the senior doctors (mean = 206.72; sd = 41.22; 68%) than by the junior doctors (mean = 164.16; sd = 47.80; 54%). ANOVA (2 x 12) showed significant influence of seniority [$F(1, 8) = 8.21$; $p < 0.05$] and trace [$F(11,88) = 4.55$; $p < 0.001$]; but no trace x seniority interaction [F<1] was found. The control (13 & 14) traces were excluded from the analyses because each of them contained a very small number of events (9 & 1 respectively).

Inferences about relevant patterns

As noted earlier, the records of relevant events generated by the expert clinical advisor included, when possible, inferences to explain the underlying causes for the physiological changes characterising the events. The expert provided causal inferences for 179 (57%) of the events he recorded (a mean of 12.5 per trace; maximum 23, minimum 1; *sd* = 7.3). A tally was made of the number of events (out of this subset of events) identified by each participant. Subsequently, the protocol statements were analysed to determine whether the participant had provided an inference for each of the identified events and, if so, whether the participant's inference matched the inference recorded by the clinical advisor. The clinical expert was partly involved in these analyses, providing advice in those cases in which the participants' reports were unclear.

Again the control traces were excluded from the analysis. A larger proportion of inferences that agreed with the expert's inferences (N= 172 excluding the control traces) was provided by the senior doctors (mean = 96.60; sd = 11.03; 56%) than by the junior doctors (mean = 48.50; sd = 26.88; 28%). ANOVA (2x12) showed significant influence of seniority [$F(1,8) = 14.06$; $p < 0.001$] and trace [$F(11,88) = 2.15$; $p<0.05$]; but no seniority x trace interaction was found [$F(11, 88) = 1.02$, NS]. These differences can be partly explained by the fact that the senior doctors identified a larger proportion of relevant patterns and generated many more inferences than did the junior doctors.

6. Discussion

The results of our investigations can be summarised as follows:

1. Both senior and junior doctors used essentially the same cognitive processes, and with equivalent relative frequencies.

2. Senior doctors showed a superior ability to focus on relevant aspects of the monitored data; the senior doctors, for example, were able to identify more relevant events than the junior doctors and tended to note a larger number of relationships among physiological parameters (see “Correlate” code in Table 1) than did their junior counterparts.

3. The junior doctors seemed to focus on more superficial aspects of the physiological data: the proportion of merely descriptive statements (“Describe” code) was higher among the junior doctors than among the senior doctors.

4. In contrast, the senior physicians generated more inferences (“Hypothesis” code) and their hypotheses tended to be of better quality than those generated by the junior physicians.

5. The senior physicians tended to revise their hypotheses more often than did the junior doctors (“Test/Revise” code) and generated a larger number of statements in which they showed uncertainty (“Uncertainty” code).

6. The senior doctors generated many more requests for extra information than did the junior physicians. Further, the senior doctors were far more likely to request modifications to the monitoring display than were the junior doctors.

7. The senior doctors knew how to take advantage of the information provided by the monitoring system better than did the junior physicians. The interviews showed that they were more familiar with the functionality of the system than were the more junior staff; in our observations we saw that they interacted with the system more often than other members of staff; the simulations showed, for example, that the senior doctors were far likelier to recognise monitoring artefacts than were their junior counterparts.

8. Some of the above differences may be a consequence of lack or shortage of training with the computerised system on the ward; in the interviews, junior staff were more likely to highlight this as a limitation than were the senior staff.

In the rest of this section we discuss the implications of these results for cognitive psychology and for the design of human computer interaction in neonatal intensive care.

6.1 IMPLICATIONS FOR COGNITIVE PSYCHOLOGY

The results of our investigations are essentially consistent with the literature on expertise and diagnostic thinking, as reviewed in Section 3. They corroborate the conclusions from previous research that differences in expertise are not so much due to skill or processing differences as to differences in domain knowledge and knowledge representation. Our data support the generally accepted view that experts possess superior domain knowledge and, as a consequence, a superior representation of the domain. Additionally, our studies support the view that medical experts' reasoning is opportunistic (Gilhooly et al., 1997; Lesgold *et al.*, 1988). The senior doctors seemed to make more efficient use than the junior doctors of whatever knowledge source was available and relevant. This is especially apparent in the differences between senior and junior doctors in their requests for extra information (consistent with e.g., Faremo, 1997).

The fact that the senior doctors were far more likely to request modifications to the monitoring display than were the junior doctors highlights an aspect of the participants' expertise which is not normally considered in studies of medical reasoning, namely, their experience with the data presentation devices. Because the senior doctors were far more familiar with the features of the computerised monitor they knew, better than their junior colleagues, what changes to the monitoring display would be most helpful for data interpretation and were better prepared to recognise artefacts. This is particularly relevant if we consider that overall the junior doctors were, by self report, more computer literate than the senior doctors.

Our data are also consistent with the conclusions of studies of diagnostic thinking in visual domains, such as radiology (Lesgold *et al.*, 1988) and ECG interpretation (Gilhooly *et al.*, 1997), in that the senior doctors seemed to make more use of biomedical knowledge than did the junior doctors. Although the off-ward simulations were not specifically designed to investigate expertise differences in the use of biomedical versus clinical knowledge, our findings reveal interesting patterns. Arguably, the usage of biomedical knowledge is indicated in the participants' protocols by the statements in which they recognise the relationships between two or more parameters ("Correlate" code). When noting relationships among parameters (especially if those relationships are used as the basis for a hypothesis), a participant is invoking and processing knowledge about neonatal pathophysiology. The senior doctors generated, as an average, more than twice as many such statements as did the junior doctors (see Table 3). As noted in Section 3, this is at odds with the generally accepted conclusion from studies of non-visual diagnostic domains that novices make more use of biomedical knowledge than do experienced clinicians (see, e.g., Boshuizen and Schmidt, 1992; Elstein *et al.*, 1978; Lemieux and Bordage, 1986). Gilhooly *et al.* (1997) argued that, in visual diagnostic domains, experts may need to invoke biomedical knowledge to make sense of surface physiological patterns that can be plausibly explained by more than one hypothesis. This indeed applies to the task (and experimental stimuli) in our simulations. The senior doctors seemed to be more aware of the ambiguity associated with the traces than were the junior doctors. The fact that they generated more hypotheses than the junior doctors indicates that they were able to think of more alternative explanations for the monitored data. This is further supported by the senior physicians' greater tendency to revise their hypotheses and to generate a larger number of "uncertainty" statements than their junior counterparts.

Further, Gilhooly *et al.* (1997) noted that an important difference between studies of the interpretation of visual medical data (ECG and radiology) and other studies of diagnostic thinking was the lack of contextual clinical information in previous studies of the former. This was not the case, however, in our off-ward simulations, where the stimuli presented to the participants *were* contextualised . Before being presented with the monitored traces, the participants in our off-ward simulations were given clinical information about the babies from whom the traces were derived. Furthermore, during their interpretation of the traces, the participants were allowed to request extra clinical information. It is interesting that, in spite of having a considerable amount of clinical information available, the senior doctors still seemed to make use of biomedical knowledge more frequently than did the junior doctors. This data pattern suggests that there must be something specific about the interpretation of perceptual physiological data (regardless of the presence or absence of contextual information) that elicits processes in experts not elicited in other less perceptually based diagnostic tasks.

6.2 IMPLICATIONS FOR HUMAN COMPUTER INTERACTION

As noted in the introduction, earlier studies in the neonatal ICU where we conducted our investigations showed that the implementation of the trend monitoring system did not result in better clinical outcomes (Cunningham et al., 1998). This is not surprising if, as our observations revealed, the staff who spent most time in contact with the patients (i.e., the junior clinicians and nurses) interacted rarely with the monitoring system. Further, our off-ward simulations showed that the junior doctors often failed to take full advantage of the information provided by the system.

It was not within the scope of these investigations to provide detailed specifications for the design of a particular computer aid. However, the expertise differences highlighted by our studies can contribute to the definition of a set of general guidelines for the design and implementation of efficient and usable computerised monitoring in neonatal intensive care. These guidelines are discussed below.

We saw that an important difference between senior and junior physicians was the formers' superior *experience with the monitoring device*. It can be argued that their ability to modify more efficiently the computer display gave senior doctors an advantage over the junior physicians, partly explaining the seniors' superior performance during the simulations. Although more computer literate than the seniors, the junior doctors were unaware of many of the display management features of the monitoring system. This can be a consequence of the lack or shortage of training on the functionality of this specific computer aid. Formal ongoing training has already been highlighted by previous human factors research as an essential requirement for the successful implementation of a computerised system in an intensive care unit, as it may affect staff's acceptance and subsequent usage of a system (Green et al., 1991). It is obvious that new staff need to be familiarised with the system in a more systematic fashion that has been done to date in the unit. For example, staff should be made aware that they can alter the scales of the physiological parameters and shown how such action can help optimise the interpretation of trend data.

However it can be argued that training on the use of sophisticated devices is time consuming and can add to the already heavy workload of temporary junior physicians. Furthermore, poor training may not be the only reason why junior staff failed to use the system efficiently. We saw that a crucial difference between experts and non experts is the formers' *opportunistic use of knowledge*: the senior doctors seemed to make more efficient use than the junior doctors of whatever knowledge source was available and relevant, whether it was biomedical knowledge, clinical information, or experience with monitoring devices. An implication of the discussion thus far is that some physiological patterns are easier to identify when displays are in some way adapted to them (by e.g. altering the parameter scales). Junior staff may not have the time or the abilities to extract this sort of information. Therefore, efforts should be made to provide them with monitored data which are easy to interpret without requiring sophisticated manipulations. A practical approach would be to design displays that are reconfigured automatically for each type of event.

A related issue is the detection of *artefacts*. The senior doctors, on average, referred to artefacts during the simulations seven times as frequently as did the junior doctors (see Table 3). Junior staff need to be made more aware of this limitation of existing monitoring methods. Furthermore, the design of decision support should take this into account and introduce mechanisms which either eliminate or minimise artefact, if clinically relevant, to highlight its appearance.

Another characteristic of the seniors' more efficient problem solving was their use of *biomedical knowledge*. We saw that this was reflected in the statements in which they noted concurrent changes amongst parameters. This is also an indication of the seniors’ superior representation of the domain. Junior doctors were less likely to produce such statements. However, concurrent changes are often indicative of relevant physiological events and it can be argued that noting them facilitates data interpretation and hypothesis generation. Therefore, the juniors’ poorer domain representation must be compensated by making this information explicit. A desirable feature of decision support would be the presentation of data in such a way that relevant links amongst parameters are highlighted.

We showed that the senior doctors' hypothesis generation and testing was superior to that of the juniors. Our data suggest that the senior doctors were more likely to generate *alternate explanations* of the data and tended to revise their hypotheses more often than their junior counterparts as they were more aware of the ambiguities associated with monitored data patterns. Arguably, this awareness should be encouraged in the juniors. The role of decision support, therefore, would be to draw the attention of junior staff to alternate competing hypotheses.

Some of the requirements we have just highlighted are being addressed as part of our ongoing investigations in artificial intelligence. Specifically, our approach is to use temporal trend templates (Coiera, 1990; Haimowitz, Le, & Kohane, 1995; Salatian & Hunter, 1996). The goal is to develop a system that provides: (a) algorithms for automatically identifying and interpreting relevant monitored patterns and artefact; (b) “intelligent” alarming, that is, using the system's interpretations to warn the staff working at the cotside on the possible onset of life-threatening clinical events; (c) summarisation of monitored events over an extended period at a high level of abstraction.

A considerable amount of information can be extracted from our simulation data to aid in the development of such algorithms. For example:

(a) The participants' errors during the simulations provide information about the sorts of monitored patterns that clinicians (especially junior staff) find most difficult to interpret, and must therefore be dealt with by the computerised system.

(b) Further analyses can be conducted to identify the precise circumstances in which the experts requested changes to the monitoring display, in order to discover what features of the events they were trying to bring out.

(c) The senior doctors' protocols will be used as a baseline for the evaluation of the data interpreting algorithms; the idea is to develop algorithms that perform at least as well as the more experienced clinicians.

(d) Finally, our interviews and simulations provide insights about the types of information that staff would need to have online to interpret developing trend data more efficiently. A pending matter, however, is to assess whether incorporating exhaustive information in the computer is the best option to propitiate effective work. In a decision making environment such as the ICU, the interactions among members of staff are crucial. It is therefore arguable whether all the information that staff need should be available online, or whether at least some of the data should be retained in more conventional methods to encourage exchanges of information among staff members, and so guarantee human contact.

7. Final remarks

In this paper we have presented outcomes of a cognitive engineering study that looks at the problems of computerised monitoring in neonatal intensive care. We have focused on those aspects which are relevant to understanding expertise differences in the interpretation of physiological monitored data, and we have indicated how the usage of these data can inform the design of human computer interaction in neonatal intensive care.

Our outcomes provide support for an approach to cognitive engineering which views the discipline as a symbiotic interaction between cognitive science and human computer interaction design (Alberdi & Logie, 1998).

On the one hand, addressing a specific human computer interaction problem in a relatively realistic decision making scenario has provided interesting insights about expertise in the interpretation of physiological data. Specifically, our findings support the generalisability of many well-known conclusions about expertise. Our data corroborate that (a) expertise differences are not so much due to different processing skills but to differences in domain knowledge, (b) experts are able to focus on relevant domain features better than less experienced subjects, and (c) experts' problem solving is opportunistic. Additionally, our data provide further support to the view that the interpretation of perceptual clinical data is influenced by certain constraints that make it, in some ways, different from other diagnostic tasks - as indicated, for example, by the ways in which novices and experts make use of biomedical knowledge (Gilhooly *et al.*, 1997).

On the other hand, the use of psychological theories (models of medical expertise) and methodologies (interviews, observations, and psychological experimentation) has allowed us to identify some of the limitations of monitoring software currently in use and has contributed to the specification of a series of design guidelines for the development of computerised decision support in intensive care. For example, our data support the well reported need for continuous formal training of staff on the functionality of the computer aids implemented in clinical settings. Furthermore, our conclusions on expertise have suggested ways in which the presentation of monitoring data can be enhanced to facilitate junior staff's trend interpretation; for example, (a) the automatic reconfiguration of data displays to feature specific patterns more clearly, (b) the elimination or minimisation of non relevant artefacts, (c) the need to highlight relevant concurrent changes in several parameters, and (d) the need to emphasise the ambiguity associated with the monitored trends by supporting the generation and revision of competing alternate hypotheses.

In summary, our studies have highlighted knowledge limitations of less experienced practitioners which need to be considered when developing systems which are meant to facilitate their work. Although we have focused on a particular medical domain, we believe our results and, most importantly, our methodologies are applicable to other areas of human computer interaction in which expertise may also play a role.

8. Acknowledgements

The authors would like to acknowledge the support of the UK Economic and Social Research Council, grant number L127251019.

9. References

ALBERDI, E., BECHER, J-C., GILHOOLY, K., HUNTER, J., LOGIE, R, LYON, A., MCINTOSH, N., REISS, J., (1999). Decision support in the neonatal intensive care unit: Expertise differences in the interpretation of monitored physiological data. In D. HARRIS, Ed. *Engineering Psychology and Cognitive Ergonomics*, vol. 3, pp. 397-404. Ashgate, Aldershot, UK.

ALBERDI, E., GILHOOLY, K. J., HUNTER, J., LOGIE, R. H., LYON, A, MCINTOSH, N. AND REISS, J. (2000). Computerisation and decision *making* in neonatal intensive care: a cognitive engineering investigation. *Journal of Clinical Monitoring and Computing*, **16** (2), 85-94.

ALBERDI, E & LOGIE, R. (1998). Applying Cognitive Theories and Methods to the Design of Computerised Medical Decision Support. In *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*, pp. 30-35. Lawrence Erlbaum: Mahwah.

AZEVEDO, R. (1998). Expert problem solving in a visual medical domain. In *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*, pp. 72-77. Lawrence Erlbaum: Mahwah.

BARNARD, P. & HARRISON M. (1988). Integrating cognitive and system models in human-computer interaction. In A. SUTCLIFFE & L. MACAULAY, Eds. *People and Computers V.* Cambridge, UK: Cambridge University Press.

BOSHUIZEN, H. P. A., & SCHMIDT, H. G. (1992). On the role of biomedical knowledge in clinical reasoning by experts, intermediates and novices. *Cognitive Science,* **16***,* 153-184.

CARD, S. K., MORAN, T. P., & NEWELL, A. (1983). *The psychology of human-computer interaction.* Hillsdale NJ: Lawrence Erlbaum.

CARROLL, J. M. (1991). The Kittle House manifesto (Introduction). In J. M. CARROLL, Ed. *Designing Interaction*. Cambridge, MA: Cambridge University Press.

CHASE, W. G. AND SIMON, H. A. (1973). The mind's eye in chess. In W.G. Chase (Ed.), *Visual information processing*. New York: Academic Press.

CHI, M. T. H., GLASER, R. AND FARR, M. J. (Eds) (1988). *The nature of expertise*. Hillside, NJ: Erlbaum.

CLAESSEN, H. F., & BOSHUIZEN, H. P. A. (1985). Recall of medical information by students and doctors. *Medical Education,* **19***,* 61-67.

COIERA E. (1990) Monitoring diseases with empirical and model generated histories. *AI in Medicine*, **2**, 135-147.

COIERA E. (1993). Intelligent monitoring and control of dynamic physiological systems (Editorial). *AI in Medicine*, **5**, 1-8.

COIERA, E. (1994). Question the assumptions. In P. BARAHONA & J. P. CHRISTENSEN, Eds., *Knowledge and decisions in health telematics*, pp. 67-72. IOS Press: Amsterdam.

CUNNINGHAM S, DEERE S, ELTON RA, MCINTOSH N. (1992). Neonatal physiological trend monitoring by computer. *International Journal of Clinical Monitoring and Computing,* **9**, 221-227.

CUNNINGHAM, S., DEERE, S., SIMON, A., ELTON R. A., & MCINTOSH, N. (1998). A randomised control trial of computerised physiological trend monitoring in an intensive care unit. *Critical Care Medicine*, **26:12**, 2053-60.

DEERE, S., CUNNINGHAM, S., & MCINTOSH, N. (1992). Staff acceptance of computerised cot monitoring in a neonatal (NN) intensive care unit. *Biology of the Neonate,* **62,** 185.

EDWORTHY, J. & STANTON, N. (1995). A user-centred approach to the design and evaluation of auditory warning signals: 1. Methodology. *Ergonomics,* **38***,* 2262-2280.

EGAN, D. E., REMDE, J. R., GOMEZ, L. M., LANDAUER, T. K., EBERHARDT, J., & LOCHBAUM, C. D. (1990). Formative design-evaluation of SuperBook. *ACM transactions on Information Systems,* **7***,* 30-57.

ELSTEIN, A. S., SHULMAN, L. S., & SPRAFKA, S. A. (1978). *Medical Problem Solving: An analysis of Clinical Reasoning*. Cambridge, MA: Harvard University Press.

ELSTEIN, A. S., SHULMAN, L. S., & SPRAFKA, S. A. (1990). Medical problem solving: A ten-year retrospective. *Evaluation and the Health Professions,* **13***,* 5-36.

ERICSSON, K. A. AND CHARNESS, N. (1997) Cognitive and developmental factors in expert performance. In P. J. FELTOVICH, K. M. FORD AND R. R. HOFFMAN (Eds).*Expertise in context*. Cambridge, MA: MIT Press.

ERICSSON, K. A., SIMON, H. (1984). *Protocol Analysis: Verbal Reports as Data*. Cambridge, MA: MIT Press.

ERICSSON, K. A. AND LEHMAN, A. C. (1996). Expert and exceptional performance: evidence on maximal adaptations on task constraints. *Annual Review of Psychology,* **47**, 273-305.

ERICSSON, K.A. AND SMITH, J., (EDS.) (1991). *Towards a general theory of expertise: Prospects and limits.* Cambridge, England: Cambridge University Press.

FAREMO, S. (1997). *Novice diagnostic reasoning in a visual medical domain: Implications for the design of a computer-based instructional system for* *undergraduate medical education.* Unpublished Master's thesis, Concordia University, Montreal, Quebec, Canada.

FELTOVICH, P. J. & BARROWS, H. S. (1984). Issues of generality in medical problem solving. In H. G. Schmidt & M. L. De Volder, Eds. *Tutorials in problem-based learning: a new direction in teaching the health professions*, pp. 128-142. Assen, The Netherlands: Van Gorcum.

FELTOVICH, P. J. , FORD, K. M. AND HOFFMAN, R.R. (Eds.) (1997). *Expertise in context.* Cambridge,MA: MIT Press.

GILHOOLY, K. J. (1990). Cognitive psychology and medical diagnosis. *Applied cognitive psychology*, **4**, pp. 261-272.

GILHOOLY, K. J., MCGEORGE, P., HUNTER, J., RAWLES, J. M., KIRBY, I. K., GREEN, C., & WYNN, V. (1997). Biomedical knowledge in diagnostic thinking: The case of Electrocardiogram (ECG) interpretation. *European Journal of Cognitive Psychology,* **9***,* 199-223.

GRAY, W. D., JOHN, B. E., & ATWOOD, M. E. (1993). Project Ernestine: Validating a GOMS analysis for predicting and explaining real-world task performance. *Human-Computer Interaction,* **8,** 237-309.

GREEN, C. A., GILHOOLY, K. J., LOGIE R., ROSS D. G. (1991). Human factors and computerisation in Intensive Care Units: A review. *International Journal of Clinical Monitoring and Computing*, **8**, pp. 95-100.

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.

GREMY, F. & BONNIN, M. (1995). Evaluation of automatic health information systems. What and How? In E. M. S. J. VAN GENNIP & J. L. TALMON, Eds. *Assessment and evaluation of information technologies in medicine*, pp. 9-20. Amsterdam, The Netherlands: IOS Press.

HAIMOWITZ, I. J., LE, P. P., & KOHANE, I. S. (1995). Clinical monitoring using regression-based trend templates. *AI in Medicine,* **7,** 473-496.

JOSEPH, G.-M. & PATEL, V. L. (1990). Domain knowledge and hypothesis generation in diagnostic reasoning. *Medical decision Making,* **10**, 31-46.

LANDAUER, T. K. (1987). Relations between cognitive psychology and computer design. In J. M. Carroll, Ed. *Interfacing Thought: Cognitive Aspects of Human-Computer Interaction*. Cambridge, MA: MIT Press.

LARKIN, J. H. (1983). The role of problem representation in physics. In D. GENTNER AND A. L. STEVENS (Eds.), *Mental models.* Hillsdale, NJ: Lawrence Erlbaum Associates.

LARKIN, J. H., MCDERMOTT, J., SIMON, D. P. AND SIMON, H. A. (1980). Models of competence in solving physics problems. *Cognitive Science,* **4**, 317-345.

LEMIEUX, M. & BORDAGE, G. (1986). Structuralisme et pedagogie medicale: Etude comparative des strategies cognitives d'aprentis-medecins (Structuralism and medical education: A comparative study of the cognitive strategies of novice physicians). *Recherches Semiotiques,* **6,** 143-179.

LESGOLD, A. (1984). Acquiring expertise. In J. R. Anderson & S. M. Kosslyn, Eds. *Tutorials in learning and memory: Essays in honor of Gordon Bower*, pp. 31-60. San Francisco, CA: Freeman.

LESGOLD, A., GLASER, R., RUBINSON, H., KLOPFER, D., FELTOVICH, P., & WANG, Y. (1988). Expertise in a complex skill: Diagnosing X-ray pictures. In M. T. H. CHI, R. GLASER, & M. J. FARR, Eds. *The nature of expertise*, pp. 311-341. Lawrence Erlbaum: Hillsdale.

LOGIE, R. H., HUNTER, J., MCINTOSH, N., GILHOOLY, K., ALBERDI, E. & REISS, J. (1997). Medical cognition and computer support in the intensive care unit: A cognitive engineering approach. In D. Harris, Ed. Engineering Psychology and Cognitive Ergonomics. Aldershot, UK: Ashgate.

LONG, J. & DOWELL, J. (1989). Conceptions of the discipline of human-computer interaction: Craft, Applied Science, and Engineering. In A. SUTCLIFFE & L. MACAULAY, Eds. *People and Computers V.* Cambridge, UK: Cambridge University Press.

LONG, J. & DOWELL, J. (1996). Cognitive engineering or `Getting users interacting with computers to perform effective work'. *The Psychologist,* **9,** 313-317.

MCINTOSH, N., DUCKER, D. A., BASS, C.A. (1989). MARY - a computerised neonatal cot monitoring system. *ITCM*, November, pp. 272-282.

MORGAN, C. J., TAKALA, J., DEBACKER, D., SUKUVAARA, T., KARI, A. (1996). Definition and detection of alarms in critical care. *Comput Methods Programs Biomed,* **51**, 5-11.

NENOV, V. & KLOPP, J. (1996). Remote analysis of physiological data from neurosurgical ICU patients. *J Am Med Inform Assoc,* **3**, 318-327.

NORMAN, D.A. (1986). Cognitive engineering. In D. A. Norman & S. W. Draper, Eds. *User Centered System Design. New Perspectives on Human-Computer Interaction*, pp. 31-61. Lawrence Erlbaum: Hillsdale.

NORRIS, P. R., DAWANT, B. M., & GEISSBUHLER, A. (1997). Web-based data integration and annotation in the intensive care unit. In *Proceedings of the AMIA, Annual fall Symposium,* pp. 794-798.

PATEL, V. L., AROCHA, J. F., KAUFMAN, D. R. (1994). Diagnostic reasoning and expertise. *The Psychology of Learning and motivation: Advances in research and theory,* **31**, 137-252.

PATEL, V. L. AND GROEN, G. J.(1986). Knowledge-based solution strategies in medical reasoning. *Cognitive Science,* **10**, 91-116.

PATEL, V. L. & GROEN, G. J. (1991). Developmental accounts of the transition from student to physician: Some problems and suggestions. *Medical education,* **25**, 527-535.

PATEL, V. L., KAUFMAN, D. R., AND MAGDER, S. (1996). The acquisition of medical expertise in complex dynamic environments. In K. A. ERICSSON, Ed. *The road to excellence. The acquisition of expert performance in the arts and sciences, sports, and games*, pp. 127-163. Lawrence Erlbaum: Mahwah.

PATEL, V. L. AND RAMONI, M. (1997). Cognitive models of directional inference in expert medical reasoning. In K. FORD, P. FELTOVICH, & R. HOFFMAN, Eds. *Expertise in context: Human and machine.* Cambridge, MA: MIT Press.

SALATIAN, A. & HUNTER, J. (1996). ASSOCIATE: An approach to the interpretation of ICU data. In *Working Notes of IDAMAP-96, ECAI-96*, pp. 73-78, Budapest.

SCHMIDT, H. G. AND BOSHUIZEN, H. P. A (1993). On the origin of intermediate effects in clinical case recall. *Memory and Cognition,* **21***,* 338-351.

SCHMIDT, H. H., BOSHUIZEN, H. P. A., & HOBUS, P. P. M. (1988). Transitory stages in the development of medical expertise: The "intermediate effect" in clinical case representation studies. In *Proceedings of the Tenth Annual Conference of the Cognitive Science Society.* Lawrence Erlbaum: Hillsdale

SCHMIDT, H. G., NORMAN, G. R., & BOSHUIZEN, H. P. (1990). A cognitive perspective on medical expertise: Theory and implications. *Academic Medicine*, **65**, 611-621.

TABOADA, J. A., ARCAY, B., ARIAS, J. E. (1997). Real time monitoring and analysis via the medical information bus. *Med Biol Eng Comput,* **35,** 528-534.

WICKENS, C. D. (1992). *Engineering Psychology and Human Performance*. New York, NY: Harper Collins.

TABLES

TABLE 1. Relative frequencies (%s) of activities recorded during observations

Nurses	
Handle baby/equipment	26%
Talk to colleagues	21.50%
Deal with alarm	18%
Write/Read paper notes	16%
Look at baby	7%
Other	6.50%
Interact with computerised monitor	5%
Junior Doctors	
Handle baby/equipment	35%
Talk to colleagues	25%
Write/Read paper notes	17%
Look at baby	8.50%
Deal with alarm	6%
Interact with computerised monitor	4.50%
Other	3.50%
Senior Doctors	
Talk to colleagues	32%
Handle baby/equipment	21.50%
Look at baby	16%
Interact with computerised monitor	13.50%
Other	10.50%
Write/Read paper notes	5.50%
Deal with alarm	1.50%

NOTE: Each percentage indicates the proportion of times each type of activity was recorded throughout the observation sessions for each staff group. The proportions are shown in decreasing order of frequency for each group.

TABLE 2. Encoding scheme derived from the protocol analysis

Types of behaviours (codes)	Examples
Describe Pattern Describe a change (rise, drop) or variability/stability in a physiological parameter.	"A dip in heart rate"; "a drop of pO2"; "blood pressure is rising"; "CO2 is decreasing" "the heart rate is variable"; "the blood pressure stabilises out";
Interpret Pattern Note whether a change in a parameter is normal, abnormal, desirable, worrying, etc.	"It's OK"; "It's worrying"; "It's acceptable"; "A serious problem"; "It's normal"; "It's satisfactory"; "I wouldn't get too excited about it"; "there is some event happening";
Correlate Parameters Note the relationship between changes which occur on more than one parameter.	"BP peak coincides with temperature gap opening up" "HR dip is associated with rise in BP"; "first a drop in pO2 and then a rise in CO2".
Hypothesis Suggest the cause for a physiological change which appears on the screen.	"Here it seems that the baby is developing pneumothorax" "it may as well be worsening lung disease at this point"; "there might be the possibility of an intra-tracheal haemorrhage that causes that".
Artefact Explicitly attribute a change on the monitor to a mechanical disturbance, as opposed to a "genuine" clinical change.	"That looks like the probe's off the baby. So that's actually an artefact with the O2 and CO2".
Extra Information The volunteer explicitly requests extra information or states that some further information would be necessary to make the right interpretation of a physiological change.	"I don't know if there is a blood gas available at this time to confirm some of these changes"; "I think I would be thinking about an X-ray"; "can you re-scale the screen?"; "was the baby's inspired oxygen increased at this point?".
Suggest Action State the sort of clinical action which needs be conducted to deal with a given clinical condition suggested by monitored data.	"then the baby needs re-intubating"; "how would you change ventilation to improve the pO2"
Uncertainty Uncertainty or insecurity on a given interpretation or hypothesis.	"I'm not sure"; "I wonder"; "I don't know".
Test/revise (an hypothesis or an interpretation) This rarely used category refers to volunteers' statements in which they are explicitly revising (or confirming) a previously stated hypothesis or interpretation	"Against that [hypothesis] is the fact that the temperature differential doesn't open"; "these [various symptoms] have all been in keeping with that [hypothesis]".
Other Statements which cannot be categorised by any of the labels above.	(a) indications to the experimenter to scroll to the next segment; (b) appeal to background knowledge, theoretical expectations, or information about the baby provided by experimenter; (c) statements not related to the task; (d) implicit behaviours; (e) repetition of a previous statement; (f) statements whose meaning cannot be understood by the coder.

TABLE 3. Frequencies and relative frequencies (%s) of protocol codes

	Senior Doctors Mean (sd)[a]	**Junior Doctors** Mean (sd)[a]	**Senior Doctors** (N= 12,734) Mean % [b]	**Junior Doctors** (N=7,758) Mean % [b]	**ANOVA**
Describe	886.40 (403.23)	657.80 (212.62)	34.80%	**42.39%**	F(1,8)=1.65; NS
Interpret	561.00 (118.21)	381.80 (187.28)	22.03%	**24.61%**	F<1
Correlate	70.20 (31.09)	33.60 (20.95)	**2.76%**	2.17%	F<1
Hypothesis	274.40 (141.66)	139.60 (52.28)	**10.77%**	9.00%	F<1
Artefact	107.40 (32.82)	14.60 (17.05)	**4.22%**	0.94%	**F(1,9)=66.91; p<0.001**
Request Info	163.80 (89.58)	64.40 (43.78)	**6.43%**	4.15%	F(1,9)=1.50; NS
Suggest Action	16.20 (15.18)	8.60 (9.13)	**0.64%**	0.55%	F<1
Uncertainty	60.00 (35.95)	22.80 (11.45)	**2.36%**	1.47%	F(1,9)=2.43; NS
Test Hyp.	24.40 (14.04)	7.80 (5.63)	**0.96%**	0.50%	F<1
Other	383.00 (114.55)	220.60 (99.37)	**15.04%**	14.22%	F<1

[a] Average use of each type of code by participants in each group; the standard deviation (sd) is shown in brackets.

[b] Average proportion of code use by participants in each group.

ANOVA was calculated on the percentages. The purpose was to assess whether the frequency with which each code was used in relation to the other codes was significantly higher in one staff group than in the other.

FIGURES

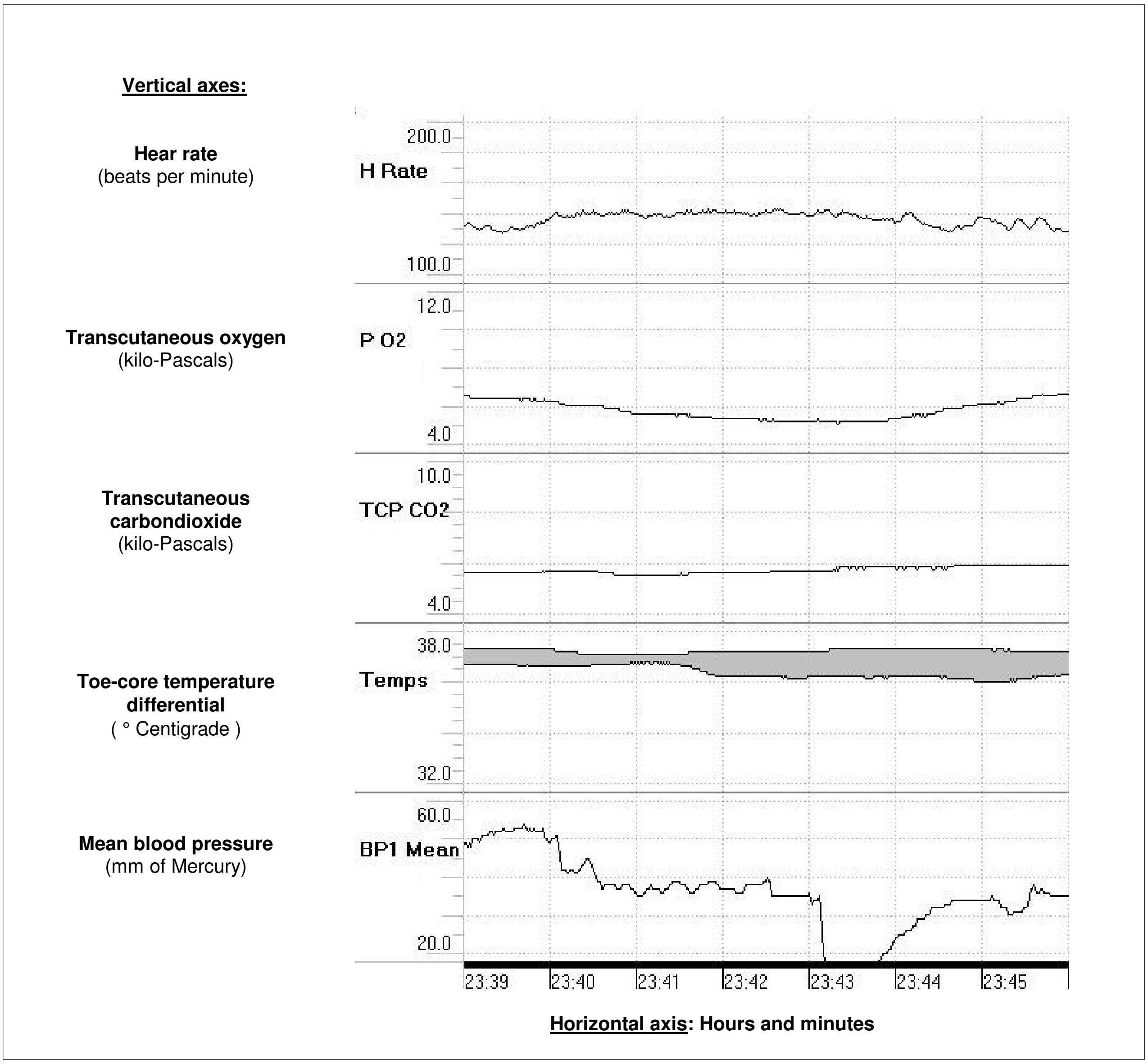

Figure 1. Trend monitoring sample