FIRST WORKSHOP ON

Human Error and Clinical Systems

EDITOR: CHRIS JOHNSON

GAAG TECHNICAL REPORT G99-1.
Department of Computing Science, University of Glasgow, Scotland.

WORKSHOP TIMETABLE... **4**

RESTAURANTS IN THE LOCAL AREA **7**

THE "WHY" OF ERROR IN CLINICAL SYSTEMS **8**

MARILYN SUE BOGNER,

CRITICAL INCIDENT REPORTING IN ANAESTHESIOLOGY IN SWITZERLAND USING STANDARD INTERNET TECHNOLOGY **10**

S. STAENDER, M. KAUFMAN, D. SCHEIDEGGER

RISK ORIENTED DATA CAPTURE: STAGED MODELLING AND SAMPLING SOLUTIONS FOR PROBLEMS WITH DATA OVERLOAD IN CLINICAL SETTINGS; PARALLELS AND PARADIGMS **12**

JOHN ARTHUR AND HENRY WYNN

TOWARDS THE IDENTIFICATION OF PROTOTYPICAL RISK SITUATIONS IN ANAESTHESIA AS A COMPLEX SYSTEM... **22**

A.S. NYSSEN,

IMPROVING THE USER-INTERFACE TO INCREASE PATIENT THROUGHPUT...................... **29**

A.GUPTA, J. MASTHOFF, P. ZWART,

PROBLEMS CAUSED BY HUMAN ERROR IN THE OPERATION OF A CLINICAL DATA COLLECTION SYSTEM... **55**

BARBARA MCMANUS,

SAFETY AND COMPUTER AIDED DESIGN OF CHEMOTHERAPY PLANS **61**

P. HAMMOND, S. MODGIL, J.C. WYATT,

COMPUTER-AIDED MAMMOGRAPHY: A CASE STUDY OF ERROR MANAGEMENT IN A SKILLED DECISION-MAKING TASK **73**

MARK HARTSWOOD AND ROB PROCTER,

CRITICAL INCIDENT REPORTING IN AN INTENSIVE CARE UNIT: 10 YEARS' EXPERIENCE **99**

DAVID WRIGHT,

IDENTIFICATION AND ANALYSIS OF INCIDENTS IN COMPLEX, MEDICAL ENVIRONMENTS **101**

DANIELA K. BUSSE AND CHRIS W. JOHNSON

LEARNING FROM ERRORS AND RECOVERIES IN MEDICAL CHAINS **121**

TJERK W. VAN DER SCHAAF

APPLICATION OF A SYSTEM MODEL TO OBSTETRIC AND NEONATAL AUDIT **122**

J.M. DAVIES,

HUMAN FACTORS ENGINEERING IS THE BASIS FOR A PRACTICAL ERROR-IN-MEDICINE CURRICULUM **135**

J.W. GOSBEE,

FRAGMENTATION IN CARE AND THE POTENTIAL FOR HUMAN ERROR **146**

ANDREW SHEPHERD AND OLGA KOSTOPOULOU,

HUMAN ERROR IN THE CONTEXT OF WORK ACTIVITY SYSTEMS: A CASE STUDY OF A KNOWLEDGE-BASED COMPUTER SYSTEM IN ROUTINE USE IN OBSTETRICS **160**

M. HARRIS, A.P. JAGODZINSKI AND K.R. GREENE

WORKS IN THEORY BUT NOT IN PRACTICE? SOME NOTES ON THE PRECAUTIONARY PRINCIPLE. **171**

KENNETH CALMAN AND DENIS SMITH

WORKSHOP TIMETABLE

WEDNESDAY 14TH APRIL

17:00-19:30 Welcome reception, Department of Computing Science, Lilybank Gardens, University of Glasgow.

THURSDAY 15TH APRIL

09.30-10.00 Welcome and Introduction, Chris Johnson.

10.00-11.00 Keynote: The "Why" of Error in Clinical Systems
Marilyn Sue Bogner, Institute for the Study of Medical Error, Bethesda, USA.

11.00-11.30 Coffee

11.30-13.00 Session 1: Medical Incident Reporting. Chair: Tjerk van der Schaaf, Technical Univ. of Eindhoven.

Critical Incident Reporting in Anaesthesiology in Switzerland Using Standard Internet Technology
S. Staender, M. Kaufman and D. Scheidegger, Dept. of Anaesthesia, Univ. of Basel.

Staged Modelling and Sample Solutions with Data Overload in Clinical Settings

J. Arthur and H. Wynn, The Risk Initiative, Warwick University.

Towards the Identification of Prototypical Risk Situations in Anaesthesia
A.S. Nyssen, Psychologie du Travail, Universite de Liege.

13.00-14.30 Lunch

14:30-15:30 Session 2: Interface Design and Evaluation. Chair: Martin Gardner, University of Glasgow.

Improving the User Interface to Increase Patient Throughput
A. Gupta, J. Masthoff and P. Zwart, Philips Research Laboratories and Philips Medical Systems.

Problems Caused by Human Error in the Operation of a Clinical Data Collection System

Barbara McManus, Department of Health Informatics, Univ. of Central Lancashire.

15.30-16:00 Tea

16:00-17:00 Session 3: Diagnosis and Treatment Support Tools. Chair: John Gosbee, Michigan State University.

Safety and Computer Aided Design of Chemotherapy Plans
P. Hammond, S. Modgil and J.C. Wyatt, University College London.

Computer-Aided Mammography: Error Managment in a Skilled Decision Making Task

M. Hartswood and R. Procter, Communicating and Collaborating Systems, Edinburgh University.

FRIDAY, 16TH APRIL

09.30-10.30 Session 4: Longitudinal Studies. Chair: Tjerk van der Schaaf, Technical Univ. of Eindhoven

Ten Years of a Medical Incident Reporting System (provisional title)
David Wright, Intensive Therapy Unit, Western General Hospital, Edinburgh.

Identification and Analysis of Incidents in Complex, Medical Environments
Daniela Busse and Chris Johnson, Dept. of Computing Science, Univ. of Glasgow.

10.30-11.00 Coffee

11.00-12.30 Session 5: Learning from Medical Risk Analysis. Chair: Daniela Busse, University of Glasgow

Moving from Industry/Transportation to the Medical Domain
Tjerk van der Schaaf, Safety Management Group, Technical Univ. of Eindhoven.

Application of a System Model to Obstetric and Neonatal Audit
J. Davies, Foothills Medical Centre, University of Calgary.

Teaching about Error in Health Care
John Gosbee, Centre for Applied Medical Informatics, Michigan State University.

12.30-14.00 Lunch

14:30-15:30 Paper Session 6: Workplace Studies. Chair: Anne-Sophie Nyssen, Universite de Liege.

Fragmentation in Care and the Potential for Human Error
A. Shepherd and O. Kostopoulou, Cognitive and Ergonomics Research Group, Loughborough University.

Human Error in the Context of Work Activity Systems: A Case Study of Knowledge Based

Computer Systems in Routine Use
M. Harris, A.P. Jagodzinski, K.R. Green, Perinatal Research Group, Derriford Hospital.

Postgraduate Medical School, School of Computing, University of Plymouth,

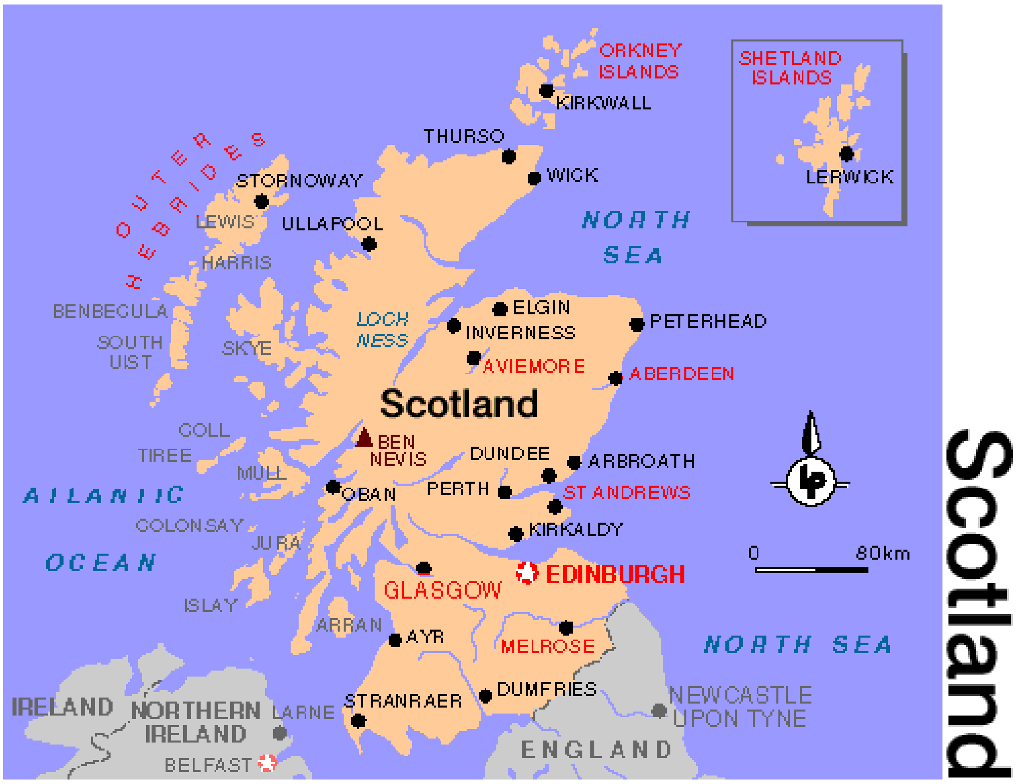
15.30-16.00 Tea

16:00-17:00 Keynote: Works in Theory but not in Practice? Some notes on the Precautionary Principle

Sir Kenneth Calman and Denis Smith, Centre for Risk and Crisis Management, Durham University.

17:00-17.15 Close and hand-over.

SATURDAY, 17TH APRIL



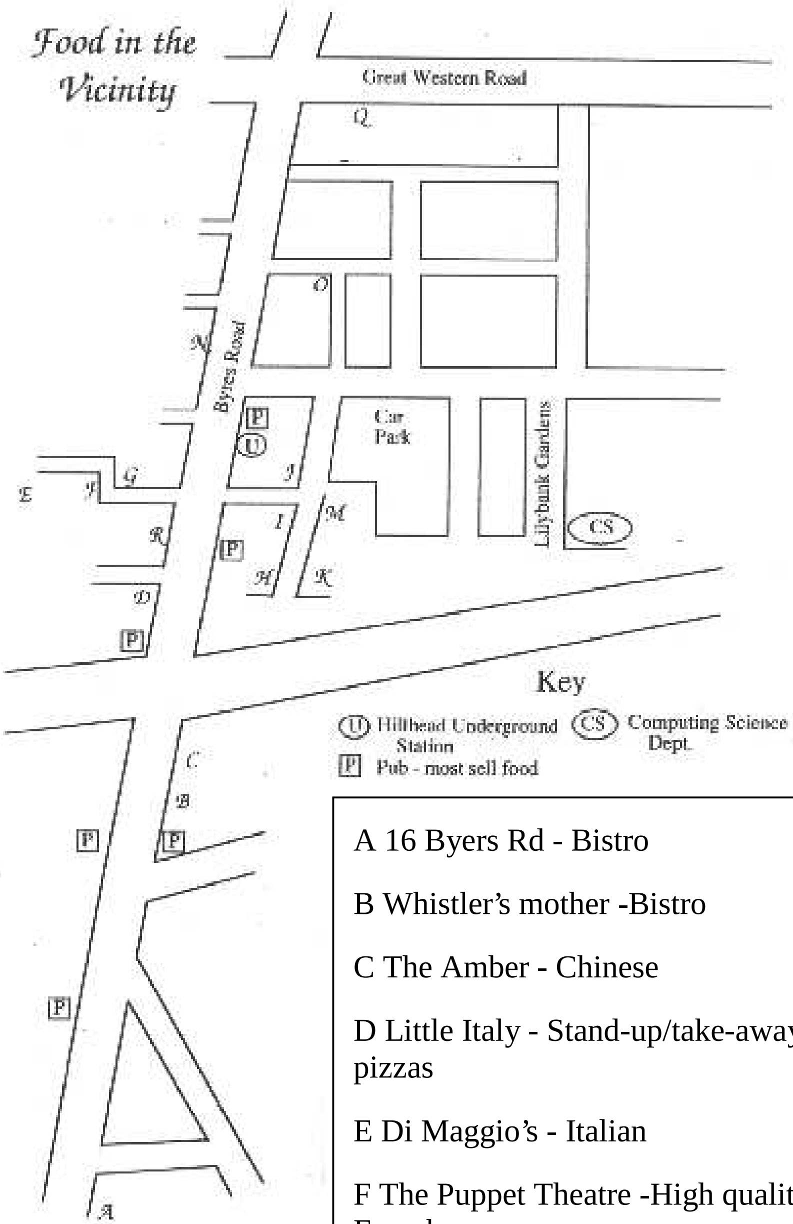
This will provide the opportunity for informal discussions about the issues raised during the workshop. The day will be spent on the Isle of Arran, off the west Coast of Scotland. The intention is to meet outside the Department of Computer Science at 07:30. We will be taking the train because this connects directly with the ferry onto the Island. Anyone who misses the first rendez-vous can meet us underneath the large clock at Central Station for 08:00 (Buchanan Street is the nearest Underground station).

Trains depart from Glasgow Central station at 08:33, arrives at Ardrossan harbour at 09:25. The ferry leaves for Arran at 09:45. Ferry arrives at Brodick on Arran at 10:40.

The ferry departs Brodick at 16:40, arrives Ardrossan 17:35. The train arrives at Glasgow Central 18:52. There is an additional service departing Brodick at 19:20, arriving at Ardrossan to connect with the 20:30 that arrives into Glasgow at 21:22. If anyone misses this departure then they will have to spend the night on the Island (there are lots of hotels and bed & breakfast places).

Arran Tourist Office can be contacted on 01770-302140 for hotel accommodation and other enquiries. The Ardrossan Tourist Board is on 01292 288688. The recently opened whiskey distillery is open for visits from 10.00-18.00 and can be contacted on 01290 553255.

RESTAURANTS IN THE LOCAL AREA



A 16 Byers Rd - Bistro

B Whistler's mother -Bistro

C The Amber - Chinese

D Little Italy - Stand-up/take-away pizzas

E Di Maggio's - Italian

F The Puppet Theatre -High quality French

G Back Alley - Burgers 'n Fries

H The Cul de Sac

The "Why" of Error in Clinical Systems

Marilyn Sue Bogner, Ph.D.,

Institute for the Study of Medical Error, Bethesda, USA.

Abstract

Adverse outcomes associated with presumed human error increasingly are receiving media attention with concomitant increased clinical concern, emphasis on reporting errors, and pressure to reduce the incidence of such events. The emphasis of error reporting systems is on the error or event which typically is identified through an associated adverse outcome of death or serious injury, and attributed solely to the care provider involved in the event. Although such attribution satisfies the need to explain the event and provides the clinician as a target for remedial action such as training, those actions seldom solve the problem. This presentation discusses an approach to identify factors other than the clinician that lead to error. Those are factors inherent in the systems that comprise the context in which care is provided.

The purpose of this presentation is to underscore the value of the systems approach in reducing the likelihood of human error in health care. First the systems approach and its application to error via heuristic incident reporting will be described with emphasis on how that approach changes the manner by which error is reported. Next, examples of heuristic incident reports are discussed. Finally, implications of the systems approach and heuristic incident reporting to error management and reduction are addressed. It is to be noted that this presentation proposes the systems approach to error management as central to quality health care.

Obtaining information for studying error is hindered by the specter of litigation. The presumption is that the care provider associated with an error was negligent; however, studies by several industries report that error is not the result of a single act, but is precipitated by the co- occurrence of a number of conditions or precursor mini-errors each of which is necessary, but not sufficient for the incident. Research in several domains has found these conditions and precursor events to be influenced by numerous interacting systems or categories of factors.

The immediate context in which health care is provided is impacted by a number of interacting systems ranging from the systems of laws, regulations, and cultural mores, through the organization system, the social system, the physical system, and the system of ambient conditions to the care providing system. The latter system contains three entities or subsystems: the care provider, the means for providing care which include medical devices and procedures, and the recipient of care, the patient. These systems can be conceptualized as concentric circles or layers of an onion from the most distal of the cultural mores to the most proximal to the nucleus of the care providing system.

Despite the findings that errors are the culmination of a cascade of mini-errors which may not have direct adverse outcomes, most error reporting systems focus on describing the specific incident, the error, often reported only when associated with an adverse outcome. Some incident reporting systems such as the Aviation Safety Reporting System (ASRS) provide the opportunity for individuals to describe a near-miss incident or an event in their own words. Such reporting can include distal system factors that contributed to the near-miss or incident, but no memory aides per se are provided. It is acknowledged that some ASRS data entry personnel may probe for additional information, but that is inconsistent. The ASRS raises 3 issues: the value of reporting near-miss incidents as well as events, the importance of those involved in the incident describing it in their own terms, and the utility of eliciting relevant systems factors. The heuristic incident reporting described in this presentation incorporates these three issues.

Health care providers are trained to be responsible which includes being responsible for adverse event. Because of this, providers tend not to consider system factors as contributing to the adverse event, a near miss incident, or an event waiting to happen. The heuristic reporting asks that each of those three types of incidents be described and aids the reporter in recalling factors that contributed to the incident by presenting a list of the systems comprising the context of care. As a quality assurance activity, the reporter is encouraged to note at least one potential factor for each system. Analysis of responses by type of incident as well as across incidents indicate system factors to be addressed to manage error and enhance quality of care. Examples heuristic reports and analysis of the reports will be presented.

The value the heuristic incident reporting is threefold. First, reporting near-miss incidents and events waiting to happen provides information to determine the "why" of error as well as adverse events. This could result in system change that avoids adverse outcomes. Secondly, repeated use of this heuristic incident reporting orients the focus and judgement to the system rather than the provider, and because the onus of presumed guilt is reduced through consideration of system factors, there is an impetus to report adverse events.

Thirdly, reporting information by system factors enhances quality of care by identifying issues to be addressed by the appropriate aspect of the system which provides a target for follow-up evaluation. Thus, rather than punishing the individual without concern for the context hence perpetuating the existence of factors that contribute to adverse events and compromise quality of care, heuristic incident reporting focuses the search for the "why"of the error or incident to system factors such as equipment design, excessive workload, and reimbursement policies. Addressing the identified system factors reduces the likelihood of subsequent events and adverse outcomes for all health care providers and enhances quality of care and patient safety.

Biography

Marilyn Sue Bogner, Ph.D., Institute for the Study of Medical Error, 9322 Friars Rd, Bethesda, MD 20817, Ph. +1 301.571.0078, Fax: +1 301.530.0679, e-mail: msbogner@erols.com

Critical Incident Reporting in Anaesthesiology in Switzerland using Standard Internet Technology

S. Staender MD, M. Kaufman MD, D. Scheidegger MD

Department of Anaesthesia, University of Basel, Kantonsspital, CH-4031 Basel, Switzerland

Keywords: complications, incident reporting, anaesthesics.

Abstract

Voluntary incident reporting can be one tool to detect safety deficits in anaesthesiology. Using standard Internet technology, we set up a registry for critical events in the anaesthetic domain in Switzerland. Human error was the leading course of error. On the other hand, human performance, like knowledge, skill, experience and situational awareness helped to prevent a critical incident to proceed to an accident.

Introduction: If we accept the fact that failures exist in everyday practice in hospitals we need to decide which method to choose to investigate these errors. Existing methods for error detection include: direct observation, chart reviews and/or computer screening, focus groups and incident reporting (ref. 1). All of them have advantages and disadvantages.

Voluntary reporting of critical events is another means to gain insight into dangerous situations.

The advantages of incident reporting are:

- Incidents happen far more often than severe accidents. Therefore, analysing incidents provides a far broader database in a certain amount of time.
- The analysis of incidents gives the opportunity to look for recoveries out of dangerous situations. Such analyses cannot be carried out as with accidents, simply because there is no recovery.
- Incident analysis can (under certain conditions) lead to root-causes of errors. These detected root-causes can then form the basis of very strong quality-improvement actions.
- Tracking incidents allows a monitoring of the quality of the structure and process of certain endeavours over time and over change in policies, if compliance is high.
- Exchanging experiences on critical events has a strong teaching potential even for old professionals, thus empowering people and teams.
- COSTS are low compared to other means as long as you are not having such a large, national organisation as the Aviation Safety Reporting System (ASRS) in the United States with a yearly budget of two million dollars. On a local level, incident reporting probably is good value for money when you keep in mind, that the patient care costs of a preventable adverse drug event were calculated at more than 4000 dollars (ref. 2).

At the Department of Anaesthesia at the University of Basel, Switzerland we have therefore set up an anonymous critical incident reporting system (CIRS©) for anaesthetists in 1996 based on standard Internet technology (ref. 3).

CIRS© is implemented on the local network using Internet technology. Against the Internet it is protected with a firewall that satisfies the strong legal requirements of the national data-security act. In the hospital, CIRS© can be reached from every networked computer. Computers are located in the offices, the theatre, the recovery room and the intensive care unit. Therefore, participating in CIRS© is possible from very different locations within the hospital. CIRS© is completely electronic. From a typical Internet page, several local links are possible: links to background information about the project, links to the questionnaire, links to already reported cases and links to extraordinary cases with a high teaching potential. The content of the detailed questionnaire (HTML-form) is automatically inserted into a central database (MS-ACCESS®). This database provides the users with the above mentioned details of every case and allows an easy compilation of the data. Furthermore, with the help of a kind of dynamic, database driven bulletin-board, an anonymous discussion of each case is possible.

CIRS© is now in its third year and we have already received some 200 cases. Analysing 132 critical incidents showed that in 64% of the cases, human factors such as tiredness, haste, wrong decision making and reduced situational awareness contributed to the incident. Adding the factor of problems in team performance, namely insufficient communication, increases the rate to 83%. On the other hand, human performance like experience, skills and situational awareness prevented in 75% of the cases an incident to procced to an accident.

Since the end of 1998 we have also used this system as a tool for a national incident survey in anaesthesiology in Switzerland (CIRS©-CH). This national system is under the auspices of the national society of anaesthesiology (SGAR) in order to facilitate reporting. Again, this system, also running on the Internet, is accessible only in a protected way after an individual login procedure. Therefore it cannot be visited from outside. A national board of experts from different regions of Switzerland and different classes of hospitals maintains this system, publishes the national results and acts as a facilitator for expert-opinion on selected cases. This feedback on individual cases was regarded as most important for national success. It is worth noting, that the analysis and comments on the events require at least as much expertise as is involved in their generation. Furthermore, an international version of CIRS© exists, that is very similar to the local and national system. The unprotected international version of CIRS© is free available on the Internet at: http://www.medana.unibas.ch/cirs/. This Internet-site offers insights into this system.

CIRS© is regularly scanned for certain extremely important cases that would require immediate action. Then, from time to time, summary statistics are performed and the findings published. An analysis of underlying error-causes can be done on the local level, where sufficient information about the system is available. On the national level, this root-cause analysis is not possible, simply because of the anonymity of the system and the limited amount of details in the questionnaire about the unit and environment the reporter works in. Nevertheless, we believe that the national exchange of information on critical events enhances patient safety in Switzerland.

Underreporting and the lack of an appropriate denominator is for sure a major handicap of voluntary reporting systems. There is evidence that we still face a huge underreporting of critical incidents. Another problem with the anonymous character of such an incident reporting systems lies in the impossibility to validate the submitted information and the possible bias of reporters using such systems. First, there can be some uncertainty about *who* is reporting and *what* gets reported. Knowing the outcome of a certain case can also influence both the reporter and the analyst. Keeping these limitations in mind, we nevertheless believe that certain trends can be detected from voluntary and anonymously reported critical incidents which can be used as the basis for action to improve safety.

References

1 Leape, L.L. A systems analysis approach to medical error. J Eval.Clin.Pract. 3, 213-222, 1997

2 Bates, D.W., Cullen, D.J., Laird, N., Petersen, L.A., Small, S.D., Servi, D., Laffel, G., Sweitzer, B.J., Shea, B.F., Hallisey, R. Incidence of adverse drug events and potential adverse drug events. Implications for prevention. ADE prevention study group. JAMA 274, 29-34, 1995

3 Staender, S., Davies, J., Helmreich, B., Sexton, B. and Kaufmann, M. The anaesthesia critical incident reporting system: an experience based database. Int.J Med Inf. 47 , 87-90, 1997

Biography

S. Staender MD, staff anaesthesiologist, Department of Anaesthesia, University of Basel, Kantonsspital, CH-4031 Basel, Switzerland. Telephone: ++41 (61) 265 2525, Facs: ++41 (61) 265 7320

Risk Oriented Data Capture: Staged Modelling and Sampling Solutions for Problems with Data Overload in Clinical Settings; Parallels and Paradigms

John Arthur and Henry Wynn,
Warwick Risk Initiative, University of Warwick , England

Abstract

In successful incident reporting systems there is a problem in dealing with data volume. This is complicated in that the high number of incidents is in a dynamic relationship with incident severity, (re)action feedback, ongoing data capture, data analysis and reporting level changes. Cause identification methodologies have sought to build core models of error. These are distinguished by interactive causalities, taking the exclusive emphasis away from human agents and sharing it with technological and organisational aspects. The problem of clinical incidents is further complicated by the dynamic nature of the clinical environment and its capacity to generate new incidents e.g. the introduction of a new technology. This paper proposes that risk oriented data capture is a technique which suits the design of an incident capture system and provides a flexible and dynamic alternative in the problem of data volume. Parallels are drawn between this and two applied case studies. The operational definition of risk and its application to data capture and/or analysis is discussed. This is seen as a useful sister process to root cause identification complimenting it by offering the possibility of dynamic modelling of error and the streamlining of the data analysis tasks.

Introduction

Incident reporting is an intense data-capture activity. Reporting should include near miss identification, dealt with in detail by Reason 1998 (ref. 1) but a universal idea. Also, if possible, some form of remedial strategy identification should be made. These three types of data provide the fullest possible picture. In any large setting such comprehensive activity, as well as being time consuming, creates a huge amount of data. This effort should be compensated by its results. It can be very effective results in terms of increased systems knowledge and identification of the causes, and root causes, of errors, again a widely endorsed idea. Results have to be applied to existing systems, e.g. in quality management. Integration with existing systems is preferable to complete innovation outwith it, e.g. see Battles et al 1998 (ref. 2).

Incident reporting systems, if they are a positive experience for the source employees, can become a victim of their own success. By their nature they provide instability. They generate an expectation for both their own perpetuation and oftentimes a rolling review of practices. This is the necessary ‘teeth’ to promote their validity to already busy people. The organisational “power-to-weight ratio” of the system is the key issue for their workload, maintenance and aforementioned integration. Efforts to reduce the weight (complexity) would be welcomed but the problem is that much of the validity of the methodology relies on maintaining, or even expanding, its rigour. This can be seen by the proliferation of systematic models and taxonomies from different disciplines: Failure Modes and Effects Analysis, Influence Diagrams, Fault Trees, Hierarchical Task Analysis, Hazard and Operability Studies, Tripod Delta, Medical Event Reporting System are a few examples. Thus reducing complexity without a loss of rigour leaves only a few avenues for development.

Why risk: The purpose of this paper is to outline some possibilities for these avenues. A central procedure is proposed, that of clarifying design issues of the data-capture, and then operating a continuous monitoring system. In putting forward such a scheme we will draw on related areas of modern management and statistically based quality improvement and, most importantly, risk-mediated design and data capture.

Risk must lie at the core of any approach because the long term aim is to minimise the risk in some safety-critical area. It is unfortunate but typical that the ‘risk(s)’ itself tends to remain at a high order conceptual level, rather than be distilled into operational definitions. The net result seems to be in part that risk management systems are viewed as an eventual goal or emergent property of the incident capture process. We would like to argue that it is at the operational level where risk of most practical use and the control of operational risk should be an explicit pre-cursor to the design of the incident capture. This is because too discrete a process e.g. continuous single error/near miss identification, does not always lend itself to integrated actions. Too compound a process e.g. strategic clinical risk management, may lack sufficient operational detail (although it draws its validity from another very important source). A simplified continuum can be proposed as shown in figure one:

Figure one

The focus of immature error capture systems is often cause identification from incident reports which initially are somewhat 'in a vacuum'. Many theorists concede that this is at least a mix of organisational factors in event reactivity and system proactivity, acceptability etc. As a system matures, its associated data usually becomes cumbersome, particularly if clarification processes are needed. It is at this point that the focus shifts to some form of risk based data handling model, however basic. We suggest that risk based modelling approaches should be used from the outset. Such models should identify the interaction *between* causes and operational risks using these interactions, and not causes alone, to direct future data capture and analysis process. The paradox is simple to sketch out. Does one begin with a simple expedient system to capture errors and deal with the complexity problems of the data later, or does one begin with a complex system and refine through controlled validation.

<u>Data-directed versus action-directed data-capture</u>: Volume is not the sole data-related problem incident systems have. At this stage we will consider a continuum between data-oriented and action-oriented incident collection.

Data-oriented: Incidents may simply lead to a data exercise. No directed action per se occurs, but strategic level planning is greatly informed. Ideally this approach should be a concrete decision i.e. potential users are clearly identified and designed into the information form. The strategic level data use can sometimes be to compare centres or units under study on raw contours. This kind of broad-brush, multipurpose, data-collection has an important advantage in inclusiveness. This potentially captures the full nature of the process, if there is time to make use of the data. It is essentially a data mining approach: collect large volumes of data for several reasons and use part of it to assess or detect incidents. Examples abound: such as credit analysis by banks, trying to detect telephone fraud from telephone use data, detecting computer hacking from computer-use data or tracing cause of failure in automobiles from warranty data.

Action-oriented: In this case the aim is only to collect data specifically related to the error reduction actions to be taken as part of the intelligent monitoring highlighted above. Feedback is an essential element in this type of system. Typically this model has discrete well-defined user groups.

<u>The data volume problem</u>: Whether a data-oriented technique or action-oriented approach is chosen, they share a common medium term problem of data volume. Either system, if accepted, typically creates a lot of data. Kaplan et al 1998 (ref. 3) note in the case of blood transfusion:

"The goal of error management should be to increase error detection and the reporting rate."

There are however some complexities from an operational risk point of view in such a data hungry goal. It also deviates from one of the cannons of quality management: decrease dependence on inspection but improve the process. Simplistically it would be more desirable for the number of reports to decrease due to error control. One needs to be able to detect if such a drop is an actual reduction in incidents and not a loss of confidence in the reporting system. Kaplan et al (op cit) suggest that the relationship between detection and reporting has a severity axis. The "event severity level" is seen as ideally dropping whilst the reporting rates remain fairly constant. This may be a stylised view which does not take into account the operational value of continuous reporting of low level events which has no visible effect on systems. Nor indeed does it overcome data volume problems in large scale situations.

'Common Sense' Solutions to Data Volume Problems

If reducing the workload in incident systems becomes a priority two common sense routes are popular: curtailing data; curtailing analysis. Curtailing the data is often seen as high risk. Later queries, unknown at present, are dependent on the availability of data. The received wisdom is that it is better to have too much data than not enough. This may lead to data stockpiling which is very common in large companies in many areas, and is one reason for the rapid growth in data-mining technologies. Do we have a good data mining reference

Data solution: This may be approached by reducing the data processing, we shall return briefly to the notion of **??** cost in the final section. The data level can be controlled, for example, maximising the use of numeric coding and minimising free text. The data entry may be accelerated through automatic options e.g. Optical Character Recognition (OCR). However these simplification approaches have a cost associated with accuracy and meaning which creates problems for the necessary rigour.

Analysis solution: These seek to streamline the analysis activity. Analysis may be stratified. Checklist approaches are often seen as a simple error-free way to quickly process data. The data may then be revisited on a need basis in the future. However, with very large numbers of incidents, even this approach can be onerous.

Mixed solution: These compromise using some salience drivers, e.g. incidents are given basic processing and some are chosen to be more in-depth. This choice may tend to be mediated by factors like novelty and repetition. Ultimately success is dependent on the intelligence of the database. That intelligence may initially be very neutral. Any approach to handling or modelling in incident systems with pre-existent low level of data structure is a complex trade off between time, costs and thoroughness.

Risk oriented solution: It is certain that data volume is a complex function of detection sensitivity, detection severity, feedback strength, actions, strategic purpose and a host of other organisational factors. It should be recognised therefore that incident capture has an already highly complex environment. It is crucial, once this is recognised, that the design of a system confronts complexity and does not roll out with an overly simplified aim of collecting and analysing incidents. Even with the more complex dynamics set aside, data volume creates an increase in the associated high level of "intelligent administration". This workload is directly correlated with system scope, e.g. regional pilot reporting system compared to a nation-wide version. Sheer volume of data not only pressurises efficiency and resources but it creates emergent system risk from computational factors and human factors in data handling.

In the final section we will sketch in more detail a number of statistical and decision-theoretic tools which can form the basis of this risk-oriented approach. It should be noted however that any new methodology should be parsimonious with respect to the factors above. A simplification must relate to the entire life cycle of the incident capture system. Any methodology must realise that system success, reputation and continuous acceptance require a well designed maintenance strategy.

Risk Oriented: System or Data?

Solutions are required not only to deal with the data volume problem but to maintain the best possible practice in incident capture within the constraints of what is practical. It is rare that significant new resources will be made available to do this and so the problem is one of directedness and efficiency of processes. This is to say that the data collection ultimately has to be about managing operational risk in some systematic form. Therefore the data capture process should have an a priori risk orientation, this will direct data sampling through species of modelling possibilities discussed later. It is not sufficient however just to get a different of data or even to model with it. Several human–systems paradigms are important to note for a source of fresh ideas. McCleod 1998 (refs. 4-5) gives a complex and fascinating account of an emerging distinction between systems (in his case machines) which give information and those which give advice. The error system one would ideally design for clinical situations would have some interactive notions i.e. it would advise practice. He notes:

These advances include new methods of data filtering, data fusion, and aids to assist the human operator to interpret and associate the fused data within their operating environment e.g. Knowledge Based Systems (KBS).

In particular he wants to emphasise that

Increasingly, with programmable electronic systems, the determination of system hazards must consider the work of the designer, the system physical build, and the tasks of the human operator. Complex systems design should assist the human operator to avoid serious errors and to achieve recovery from errors. The issue is not just a matter of engineered reliability or human reliability.

Furthermore if the system, *however technically sophisticated*, is to be used by humans then the risks inherent in that very relationship and how it is represented are a key factor. Speaking of the dulling of creativity which levels of automation can bring to health care settings Satchell 1998 (ref. 6) cautions:

While competencies and knowledge are as much issues in the medical profession as any other, health systems have generally shown little interest in their human resource, and have responded reactively to performance variation with overt behavioural change programmes. Task sharing technology struggles in this type of situation, and can amplify inappropriate behaviour as easily as appropriate behaviour.

The most obvious flaw in any data collection is a lack of orientation or to put it another way, uncertain or crude design. In the case of blood safety in the UK James 1997 (ref. 7) reports

Despite a massive scientific literature which speaks to several distinct aspects of this topic – including disease prevalence and incidence and laboratory test performance – a simple and comprehensive conceptual model has not been articulated.

The collection of undirected data is not confined to clinical errors. Crossland et al 1998 (ref. 8) note in their survey of current practice in managing design risk:

It seems unfortunate that of the many companies who take the trouble to collect data, only very few collect quantitative data which allow for meaningful rolling-up of complex sources of risk

We would argue that McCleod's advice-giving fused data, Satchell's concern for the human element, James' coherent model of safety and Crosslands meaningful rolling up are diverse examples of the need for risk oriented data capture approach. In short some pre-existent models at work a priori to data collection. Conventional cause identification paradigms use judgements about frequency or severity or novelty of an incident. Whilst these are essential building blocks in understanding causes something further is needed. There are two predominant issues, one is 'soft' the other 'hard'.

- If an incident capture system can be agreed as a method to control risk in a live setting, care has to be taken to define how operational the definitions of risk in use actually are.
- What risk orientation models can meaningfully and usefully be applied to incident data?

The next sections will deal with these two issues. Operational risk will be considered by briefly examining two real case studies of a 'risk issue' in keyhole surgery training. Guidelines for defining operational risk will be suggested. Risk oriented data capture models will then briefly be described and their applicability to incident capture discussed.

Risk

The methodology should consider the 'live' risks, coming from the incidents themselves, and 'inert' risks, coming from the process of recording, encoding, analysing and classifying the incidents. Most importantly the risk in reacting to the incidents has to be uncovered. For this purpose it is crucial to be absolutely clear about what constitutes a risk and how it functions. As the following case studies show vague definitions an early stage can deeply compromise coping system design and either fail to deal with identified risks or introduce new ones.

Case study one: some experience in the surgical area: One author was a researcher on The Minimal Access Therapy Training Initiative Evaluation carried out for the Dept. of Health 1996 – 1998. (ref. 9)

The use of training on simulations is a relatively recent development in laparoscopic surgery. It follows catastrophic mistakes in the late 1980's and early 1990's from training on live patients. Reviews around that time recommended error directed tuition in a safe laboratory environment i.e. a reduction in the overall risk to patients from surgeons in training. This was an upstream reaction to root causes (human only), to wider organisational issues, to political issues and to ethics. Such training systems needed to be designed to:

- deal with the problems
- not introduce new problems

Not introducing new problems was largely ignored as the main 'ethos' of training remained the same. Whilst a list of technical risks (akin here to simple causes) were identified industriously, through ad-hoc analysis of expert knowledge, operational risks were not identified or acted upon. This was for a number of reasons including:

- They were complex and the trainers lacked the knowledge and resources to research them
- They challenged the status (and low consensus) of current practitioners' technique
- They challenged a cherished status quo in terms of training delivery by mentor
- They could not easily address thorny issues like 'necessary risk' from a positivist stance
- The ramifications, e.g. certification of surgeons, were vast

A comprehensive incident capture system in general surgery in the UK has never been attempted. There are enormous restrictions on the possibility in the US, for a good discussion see Laing 1997 (ref. 10). Capturing a list of technical incidents which must be avoided, although an immensely powerful tool for increasing safety, was not sufficient for adequate training system design. This is because it was more akin to curriculum development than to safety critical training. It still left untouched a whole species of errors in the more powerful human dynamics at the interface between current professional judgement and training. These were to prove paradoxical.

Logically it was the population of trainers, i.e. the experienced surgeons at the time of the errors becoming widely noticed, who were "responsible" for the problem. It was however the trainees, i.e. those surgeons not yet trained, upon whom the remedial actions were taken. The actions were mediated by the current practice and views of the experienced surgeons in two important ways:

- They controlled the curriculum and, more importantly, its delivery. Courses highlighted extremely low consensus on technique.

- They controlled the ongoing experiential training after formal courses, which of course, in the structure of this profession, was commensurate with their own (original?) views on safe practice.

Attempting risk oriented training design was an important step. In this case the risks (error possibilities) were articulated at too low a level, i.e. mainly practical intra-operative considerations. In consequence the risk (generalised 'bad' outcome) to patients from surgeons in training was reduced but by a far smaller increment than could have been expected from the rhetoric of the reform. Without identification, agreement and definition of observable risk sources as live behaviours, the interface between training and practice is still more dangerous than it need be. In consequence the whole profession is still under a great deal of pressure in the UK following recent unfortunate deaths.

Case study two: Experiences from simulation design in Arthroscopic surgery: *The authors are working with Sheffield University medical physics Department in the development of their Virtual Reality Knee Arthroscopy Training System.*

The applied problem of a safety critical trainer for arthroscopy is the same as that in the previous case, namely, increasing surgeon safety and therefore increasing patient safety. Unlike the previous case this is an example of a technology-based solution to risk. Technology is particularly acceptable to the surgical professions because:

- They are very familiar with interacting with, if not necessarily understanding, high technology in work
- It is exciting to them
- It forms the second arm to solutions for pathology after surgical procedure development *(which it greatly supports)*
- It supports the self view of a high risk profession which makes de-facto use of high technology solutions to training, like airline pilots

The model needed to create the simulator is far more complex than that needed to create training. However, it should be based on the same articulation of what the risks are and how they operate. The prevalence of compromise and trade off in this kind of application is very marked, for a discussion see Arthur et al 1998 & McCarthy ?? (refs. 11-12). For example prototype testing has indicated a need for haptic feedback mechanisms which are enormously complex. So complex are they that the problem quickly and predictably shifts to a related technical problem of simulator capability. High levels of VR specification may be defensible for commercial, aesthetic, technical challenge and face validity reasons. However a design approach which refocuses on technology away from the original risks is a key risk driver for product failure. This is because, irrespective of the technological prowess of the simulator, the risks do in fact remain cultural, attitudinal, motivational and organisational issues. These define the imperative applied context of the final product. If the production of the error reducing system (trainer) is not cognisant of this it is in danger of introducing new latent risks into the organisation in a similar way to any other new or revised technology.

Points from the Case Studies

These examples illustrate that too small (error by error), and too large (quick fix?) definitions of risk muddy the waters when it comes to the design of a complex system. The efficacy of the risk reduction function of the initiatives was greatly decreased. Inevitable data overload problems associated with rigorous incident capture systems is another excellent example of such a problem. Re-casting the problem as a design issue creates an interesting paradigm shift for incident capture. It suggests one ought to treat it as any other design problem. In particular then it should pay attention to:

- The whole human context
- Statistically powerful modelling species

The Whole Human Context

Risk is a notoriously difficult term to define, the Royal Statistical Society abandoned lengthy attempts to do so. In practice however, all the information needed to define risk can be found within the behavioural language of an organisation. It is for this reason that we pursue the idea of operational risk definition. Since there is no parameter for the accuracy or purity of a risk definition what really matters is that it makes sense, can be accepted and is agreed. Incident capture systems which create root cause taxonomies are an excellent starting point for an operational definition of risk. Some might suggest that root causes are the correct form of risk but we would argue that there has to be a working model at the next level. This is based on synthesis between groups of causes and between those groups and the individual organisation members experiences and comprehension. Operational risks are a parsimonious world model of how risks behave which is particular to an organisation or even a part of it. To be operational risk definitions have to fulfil a set of base criteria.

- Always describe behaviours or groups of behaviours as a starting point - what people actually do described by the people who do it.
- Have an explicable organisational value to the operator in natural language. - account for why people do things in organisation centred terms.
- Have a direct link to executable remedial actions - low generality.
- Have the potential for simple monitoring and active feedback - is observable and the results of observation are usable immediately.
- Have instructional value – create learning opportunities for all users.
- Have explicit links to other risk centres – contribute to a wider communication and learning network.

Fulfilling these criteria creates better building blocks for decision making in the light of understanding the causes of errors. This avoids naïve notions like ‘risk free’ working and embraces the idea of situation-inherent risks, observable risk tolerance and risk sharing/contacting. For health care professionals in particular such a model creates a structural basis for a host of tactical decisions about dealing with the causes of risk. This has hitherto been impeded by multiple, diverse, vague and even fashionable definitions of risk. Incident capture coupled to operational risk definitions creates the possibility of auditable decisions which are data driven (evidence based?) and operationally grounded.

A key problem to overcome, to achieve usable risk concepts like those above, is the uncertainty in handling large amounts of data which is about risky events and risk in events. The following sections will deal with this notion of the risk orientation of data gathering.

Risk Orientation Models

On-line off-line: risk management as control:

Among many models for risk management one of the oldest is control, e.g. see Hotteling 1947 (ref. 13), sometimes called automatic control. In this model the long-term strategy is to bring a process under control. In the controlled state the process can be maintained in stability with short-term feedback form observation to control action. Before the control systems is set up there should be considerable off-line activity leading to the design and implementation of the control system. For, example the control systems may incorporate some empirically validated model of the real-world process being controlled. Alternatively certain parameters (set points) may need to be set taking the particular situation into account, such as the nature of the batch or raw material in a plant.

Simplistically then, one can talk about off-line that is 'design', versus the on-line automatic control that is 'performance'.

Of course, life is not that simple. In a continuous incident reporting system there may not be an automatic control action. However there ought to be some types, or classification, of action which are laid down. This weaker, more manual, version of control is analogous the statistical process control, MacGregor 1994 (ref. 14). The rule there is to look for special (local) cause or common (root) cause. In the more sophisticated, "intelligent", versions one may do a special analysis to detect patterns of root cause and link patterns via data-base searches etc.

Such is the technology of, for example, multivariate control charting. There a number of performance characteristics are measured through time. Clearly one can monitor each one individually and take action when one or a number exceed their control limits. More sophisticated versions build multivariate boundaries so that when the whole vector of characteristics crosses the boundary action is flagged. Over time the nature of the crossing can be linked to root cause. It is possible to do simultaneous statistical analysis in that, for example one can decide to monitor a few principle components or other multivariate statistics. These can include not just means but also variation itself, by analogy with monitoring volatility in share prices.

Having identified the pattern of causes one then proceeds to mediate/eliminate it. It is this "intelligent monitoring" model of mediation or elimination, weaker mathematically than automatic control, but stronger than simple control charting that is needed. It is then clear to see that it is risk metrics defining the link to risk reducing actions which will serve to strengthen the "intelligence" of the database.

Basic process for the design of an incident reporting system:

- Risk is the main source of metrics for both the monitoring and the assessment of effectiveness of actions.
- There should be a more static (off-line) phase which models the process and captures the taxonomy of the incidents.
- There should dynamic phase which is less data intense but concentrates on key items. This phase should detect individual incidents, model broad trends, recognise patterns, and point to causes.
- An action phase which takes the feedback from the monitoring and eliminates or reduces the cause: special or common.

<u>The static phase:</u> The static phase should provide all the necessary information for the design of the dynamic phase and it is here that risk methodologies are crucial. It is not enough to design the dynamic phase only on the basis of historical data alone. As explained, such data will usually not have been collected for this purpose. Rather, the static phase must be directed towards the design of the dynamic phase. It will usually be more data intensive, conducted over a short period

and on one or only a few sites. It should seek at least to

- test the administration and culture for acceptance of the future dynamic system.
- capture the basic system (process) using models of activity, task analysis, flows of information, medical informatics.
- be comprehensive in terms of severity, types of case, level of skill, organisation structure and, of course, type of incident.
- designed making use of social research techniques (survey, questionnaire, psychometric).
- submit to comprehensive statistical analysis.
- include risk-related methodologies of some kind: decision analysis, ranking, decision conference, Delphi etc.
- elicit special events which might occur in real operation and either note to remove them through the design or build contingencies into the design.

Most importantly this phase will deliver a list of key risk features (drivers, attributes, indicators). These are to be monitored at the heart of the dynamic phase. In clinical incidents capture that is the medium term data overload. Analysis will take advantage of established lead indicators some of which may be proxies for others and reflect special and root causes. Thus these lead indicators will have been derived from careful modelling *during the first phase.*

<u>The design phase:</u> The usual iterative steps amount to (i) capturing the specifications (ii) translating them into the high level functionality and thence to (iii) detailed architecture and prototyping.

Some of the above discussion can be seen as the beginnings of a specification. To be a little more comprehensive it is now possible to claim that the following are likely to be in the specification.

- High acceptability to the organisation, culture, environment
- Streamlined, low cost data capture
- Highly relevant, risk-directed data (lead indicators)
- Dynamic in nature

The actual functional design does not concern us here. There may be many.

<u>Ways of meeting these specifications or criteria</u>: Elsewhere Arthur et al (op cit) we have argued that it is dangerous to allow the technology to drive the design. The technology should be the slave of the requirements. A technically smart method of data capture may not of itself meet data capture requirements in terms of the risk objectives nor itself be risk free.

More interesting is the prototype or first version of the dynamic system. This is not the same thing as the static part of the methodology. To rush forward and test a prototype ahead of a initial phase assessment is dangerous, in the same way as doing clinic trials on a drug before the in vitro or early phase have been gone through. A prototype for example could be a paper version to be tested before a computerised version.

<u>Operational risk: running the dynamic part:</u> The main risk indicators to be used in expanding from the initial phase and pilot to “real” operation will have been caught at the initial stage. Other events may have been foreseen, such as failures of equipment. These should be considered as part of a wider risk control methodology. For example the incident capture could be for surgery, but power failures would have contingencies.

The dynamic capture should be linked to a feedback system for correction of smaller items and swift action on lager items. The flavour of this on-line control system is that these procedures should be automatic and accepted. In quality improvement one is taught that in-line measurement connected to process control is to be preferred to end of line inspection which requires cumbersome trace-back and is “after the horse has bolted”. Thus, the accent is on continuous improvement of the process being measured. There may be unforeseen risks associated with the dynamic capture, and it may be that these arise in some aspect of acceptance.

Information and Risk

We now want to set up in more detail a model for understanding risk-directed data capture. Fortunately there are a number of areas in which the idea features, in some form, of directing attention to what “matters” . Early work in cognitive science argued that attention was directed because of the limited computational capacity of the brain. For a time attention was seen as one of the keys to the study of perception Neisser, 1967 (ref. 15). One of these early models by Cherry 1953 (ref. 16) led to the growth of work in attention using an information-theoretic model. For a review see Allport 1989 (ref. 17).

An information approach assesses the quality of planned data collection, by the expected decrease in uncertainty, which is equal to the expected increase in information. This can be set up using the familiar Shannon information. Roughly, the hope is that the method of data collection will leave the scientist with most information in the sample and consequently least information in the un-sampled data. This is explained in recent work by the author (ref. 18). Important work in cognitive science is by Oaksford et al 1994 (ref. 19).

Understanding that uncertainty or entropy is negative information it is not too large a step to extend this idea to real risk. A simple example explains this. When crossing a road we look towards the oncoming traffic. A careful risk analysis explains why. If we look the wrong way then we remain by the roadside because that is the best option whatever we see. If we look the right way then if a vehicle is coming then we wait and are no worse off then the other case. But if a vehicle is not coming we “gain: by being able to cross (and save time etc).

This should make clear that the gain in directed attention, smart data collection, is not simply that we “see” this risk but that we have the better

Potentiality for action to mitigate it. In fact this is very close in spirit to the control paradigm we mentioned above. In control we take an action *now* taking into account the future action. This is the basis of dynamic programming. It is just that observation or data collection is a special kind of action.

With this we can build a sound and practical framework for thinking about data collection and action in a single system. But their remains a problem the solution of which is precisely covered by the static phase. The problem is that until we observe the system passively we are unable to model it sufficiently to understand the sources of risk (metrics) and thereby build our information/risk model. To summarise our risk-directed data capture should be *model-based.* To complete the analogy if we are landed in a country and we do not know whether people drive on the left or right we do not know which way to look. Only by a prior stage of observation can we determine this (or ask where we are and draw on our knowledge).

We have stuck in the article with a two stage method, but of course we can continuously refine the model. Drawing on control such a system might be called adaptive. Even the two stage procedure is really adaptive, just a simpler version.

A more complex example can be taken from medical screening. A data collection decision in this context is when to screen. If it is too early there are few cases and needless cost. If it is too late then the medical and financial costs are greater. Again the time of screening is predicated on actual later action and complex consequent cost structures or risk.

Notice in the screening case this case the cost of data collection (say for all women in an age-band) is pertinent. The analogy with incident capture is the cost, for example of storing and using large data bases. This is a good analogue of the computational models in cognitive science. Finally, one can turn the data based issue onto itself. There is even a cost with “mining” the data. Thus the same principles which apply to raw data collection apply to mining. In fact data-mining for model building is already a form of directed attention. I want to test this hypothesis so I will draw that data.

Conclusion

We have tried simply to draw on a range of ideas in and around the subject of risk to focus the activity of incident capture.

In particular we have suggested that one must neither over conceptualise risk, that is to say make it too strategic, or under conceptualise it that is to make it less than tactical. The operational definition of risk is a complex task requiring a balance of qualitative and ethnographic data collection. This should have as its basis live incidents of precisely the kind incident capture systems uncover.

We suggested that incident capture suffers from known problems of data directedness and data volume. There are a range of possibilities which could be adapted to organise and lift it to a more active mode. Or rather to have a suitable combination of passive and active modes.

Data collection is a special form of decision making directed towards future action to control risk. Active definitions of data, and of risk, are essential clarifications to the varied settings of the world of clinical incidents capture.

References

1. Reason, J. Managing the Risks of Organisational Accidents 1998. Ashgate.

2. Battles, J.B., Kaplan, H.S. Van der Schaaf, T.W., Shea, C.E. The Attributes of Medical Event-Reporting Systems, Experience With a Prototype Event-Reporting System for Transfusion Medicine. Arch Pathol Lab Med – Vol 122, March 1998, pp 231 –238.

3. Kaplan, H.S., Battles, J.B., Van der Schaaf, T.W., Shea, C.E., Mercer S. Q. Identification and Classification of the Causes of Events In Transfusion Medicine. Transfusion, Nov/Dec 1998.

4. McCleod, I.S. Information and Advice: Considerations on their Forms and Differences in Future Aircraft Systems. In Proceedings of Global Ergonomics Conference, Capetown, SA,.Elseivier Science, Amsterdam 1998

5. McCleod, I.S. System Cognitive Function Specification: The Next Step. In Engineering Psychology and Cognitive Ergonomics Vol. 3 In press.

6. Satchell, P. Innovation and Automation 1998. Ashgate.

7. James, R.C. Blood Safety: A Conceptual Model. In Proceedings of Risk and Decision Policy Conference, Oxford 1997.

8. Crossland, R., McMahon, C.A., Sims Willis, J.H. Survey of Current Practice in Managing Design Risk. EPSRC Grant No. GR/L38745, 1998.

9. Fletcher, J., Arthur, J.G. Sutton, F., Szczepura, A. Minimal Access Therapy Training Initiative Evaluation, report prepared for the Dept. of Health Feb 1998.

10. Liang B.A. Legal Issues in Medical Care Risk Reduction. In Proceedings of Risk and Decision Policy Conference, Oxford 1997.

11. Arthur, J.G., Wynn, H., McCarthy, A., Harley, P., (1998) “Beyond Haptic Feedback: Human Factors and Risk as Design Mediators in a Virtual Reality Knee Arthroscopy Training System (SKATS). Engineering Psychology and Cognitive Ergonomics Vol 3. In press.

12. McCarthy, A.D., Hollands, R.J., (1998) A commercially viable virtual reality knee athroscopy training system. Medicine Meets Virtual Reality: 6, San Diego, U.S.A. P 302-308

13. Hotelling, H., (1947), “Multivariate Quality Control”, in Eisenhart, Hastay and Wallis (eds) Techniques of Statistical Analysis, (New York)

14. Mac Gregor, J., (1994), “Statistical Process Control of Multivariate Processes”. Preprints of the IFAC ADCCHEM’94 Conference on Advanced Control of Chemical Processes, May, Koyoto Japan.

15. Neisser, U. (1967) Cognitive Psychology. New York: Appleton –Century-Crofts.

16. Cherry, E.C. (1953). Some experiments on the recognition of speech with one and two ears. Journal of the Acoustical Society of America, 25, 975-979.

17. Allport, A. (1989) Visual Attention in Foundation of Cognitive Science. Cambridge Mass: MIT press (chapt. 16)

18. Sabestini, P., Wynn, H.P., (1998) Risk based Optimal Designs in Atkinson, Pronzato and Wynn (eds) MODA 5 – Advances in Model Oriented Data Analysis and Experimental Design.

19.. Oaksford, M. Chater, N., (1994) A rational analysis of the selection task as optimal data selection. Psychology Review. 101, 608-631.

Towards the Identification of Prototypical Risk Situations in Anaesthesia as a Complex System

A.S. Nyssen,
Work Psychology Department, University of Liege, Belgium

Keywords: Human error, Systemic Analysis, Prototypical Risk Situations

Abstract

Accident investigations in medecine as in other complex domains tend to attribute most adverse outcomes to human error. Human error is defined in terms of act deviating from normal and accepted safe practice or working conditions. This is not the case in many clinical situations. In order to better undertsand the root causes of accidents in clinical systems, we analysed 30 problem situations in anesthesia. We complemented classical accident analyses, based on the external contributory factors, with a cognitive analysis, based on the decision functions involved in the problem situation. The results revealed that the critical phases in the decision making process is far from being always the diagnosis. Most often cited was failure to perceive information during surgery. The analysis of our data also revealed different cognitive difficulties with different degrees of expertise. This analytical perspective is interesting if we want to predict some cognitive failures connected with some particular work conditions (prototypical risk situations) and limit the risk through some means as training, technical or organizational improvements.

Introduction

Human error has been proposed as the most frequent cause of critical incidents in medicine as in other complex domains (refs. 1-3). Accident investigations often classify accidents into exclusive categories: human error, technical factor or other complication. Human error is defined by the criterion of "performance which deviates from the norm". This definition implies a well-structured task sequence performed in stable and predictable work conditions (ref. 4). This is not the case in many medical situations where there is high uncertainty. In anaesthesia for instance, it seems that at least 20 % of anaesthesia cases involve a problem event requiring intervention by the anaesthetist, while around 5% of cases involve a potentially catastrophic event (refs. 5, 6). Moreover, the patient's reactions are not always predictable and

his clinical state evolves constantly. From this, the medical procedures will necessarily be less specific than in industrial environments and the process of decision-making should be analyzed in relation to the dynamic context. Under such conditions, the usual convention to define human error in relation to the norm is difficult to apply. Furthermore, the classification of medical accidents into exclusive categories appears irrelevant in a clinical system in which humans, technology and organization are more and more interdependant. Understanding how this complex system poses demands on the physicians, and how the physicians meet those demands seems more important for the prevention and management of accidents than statistical and epidemiological data. This implies to explore the knowledge, attentional and contextual demands that operate in the field and which influence the behaviours and the decision-making process of the practitioners. Seen in this perspective, accident data provide valuable information about how the system in which practitioners are embedded functions and dysfunctions.

In general, accident reporting forms used by studies on risk in medicine include: data of the patient, type of surgery, detailed report of the incident and appreciation of the associated factors (refs. 7, 8). Although this information on what we term "extrinsic" factors is the most objective way of collecting information about critical events, it may not be sufficient when proposing an accident reduction strategy considering the evolution of the physician's task towards greater cognitive tasks. This evolution is due to the introduction of computerized technology in the hospital: for instance, in today's OR, the anaesthetist supervises, detects and regulates the case essentially from electronic displays. By providing more precise information on the patient, this technology facilitates the diagnosis and control of the patient but it is at the expense of a more elevated mental cost. This observation stresses the importance of analyzing critical events at the cognitive level in order to discover the difficulties encountered by the physicians in this man-machine system and to adapt training program and technology.

Several researchers emphasized the need for accident analyses to go beyond the superficial level and the particularness of an accident. (refs. 9, 10). Points of commonalty among accidents requires to use a mechanism for abstracting some kind of patterns among accident data. Holnagell (ref. 11) stimulated this search by making a distinction between error phenotype or domain description and error genotype or underlying cognitive mechanisms. Reason's latent failure model (ref. 12) has suggested that a variety of common latent factors present in most organizations can potentially degrade safety.

In this paper, we carried this construct further and explored the idea of prototypical risk situations that linked some kind of human performance issues that we have observed with some attributes of modern man-machine systems. We used the domain of anaesthesia as a laboratory for collecting, analyzing and learning about the prototypical risk situations. We will attempt to demonstrate how the concept of prototypical risk situation provide a mechanism for abstracting common patterns of breakdowns in complex systems. Although anaesthesia is usually referred to as a procedure for putting the patient into an unconscious state so that the surgery is possible, the action of the anaesthetist is highly contextual : it lies within an evolving situation marked by a transformation of the patient's state and by the surgeon's actions. Thus, it shares the common profile of complex systems as described by Woods (ref. 13): high dynamism and uncertainty, time pressure, ill-formed problems, complex human-machine interactions and risk.

<u>Material and Method</u>

The collection of accident cases was carried out at the University Hospital of Liège (Belgium). Between January, 1995 and April, 1996, 30 problem situations were collected by a senior anaesthetist. The selection also comprises cases reported by anaesthetists in training outside of the University Hospital in 16 other peripheral institutions. Each of these cases satisfied the four criteria : 1) the incident constituted a potential threat to the safety of the patient at one or several moments during its course; 2) the actions of one or several anaesthetists played a critical role in the evolution or the consequences of the incident; 3) the anaesthetist who experienced the situation was willing to present the situation to his peers; 4) an extensive reporting form was available for analysis. This reporting form was filled out as soon as possible after the incident by the anaesthetist involved. It included : data of the patient with appreciation of the risk (ASA Physical Status), type of surgery, detailed anaesthesia procedure from premedication to recovery and even post-operatively, detailed report of the incident and appreciation of the associated factors. In addition, the anaesthetist was asked to evaluate decision functions that could have played a role in the genesis of the problem situation. Labels for cognitive activities involved in decision making process were created using the well-known Rasmussen's model of decision making (ref. 4). In order to fit better the sequential structure of an anaesthesia case, three main phases of evaluation have been differentiated : preoperative anaesthetic risk evaluation, inadequate perioperative management and postoperative risk evaluation. Table 1 shows the proposed framework based on the Rasmussen's model being used in our reporting form.

Monthly "safety" conferences were organized in order to allow the anesthetist who experienced the situation to present the facts and to discuss them in the presence of the members of the Anesthesia-Resuscitation Department. These conferences played a determining role in the appreciation of the adequacy of reaction to the situation through the flow of communication between expert anesthetists and those in training.

Table 1 - Framework for Potential Cognitive Failures

Inappropriate preoperative anaesthetic risk evaluation
Failure to perceive information
Misdiagnosis
Inadequate perioperative risk management
Failure to follow established procedure
Inappropriate postoperative risk evaluation

Results

Tables 2 and 3 summarize the basic features of the problem situations and the critical decision making steps involved in the 30 problem situations collected.

Table 2 - Factors associated with Problem Situations

Associated Factors	n
inadequate experience	30
poor communication	26
lack of sleep, fatigue	24
inadequate supervision of juniors	16
productive pressure	16
unfamiliar with the operating room	13
anaesthetist/surgeon relations	14
presence of a distracter	10
inadequate monitoring	9
visual restriction	9
interruption	8
emergency	1

Table 3 - Distribution of the Critical Decision Functions to the Situation, according to degree of training of the anaesthetist involved. For one problem situation, there can be several critical stages; thus the number of responses is not equal to the number of problem situations analyzed.

	2nd	3rd	4th	5th	senior	mixed
pre. eva.	2	3	1	2	1	2
detection	3	4	3			3
misdiag.	2	4	1		1	
peri.man.	3	3	1			2
procedure	4	4	1	1		
post. eva.		2	1	1		1

From a systemic approach of the different contributory factors identified in our study, four prototypical risk situations emerge which bring into play the types of error, the complexity of the system, the contextual factors (including the role of the team), the cognitive processes brought into play in the decision making and the degree of training.

The first prototypical risk situation concerns failures to perceive information during surgery and failures of procedure. It concerns mainly anaesthetists of 2nd, 3rd and 4th years of training.

Failure of perceive information are related to situations where either the anaesthetist did not detect some valuable information about the problem or detected it too late (for instance, the bleeding, the ischemia and the airway obstruction). These type of failure are the most frequently cited at the origin of the problem situation with severe outcome for the patient. Failures of procedure or execution appear in an unfamiliar environment. Lack of experience, poor communication, fatigue and inappropriate supervision are often mentioned by the anaesthetist as contributory factors to the problem situation.

The second prototypical risk situations concern difficult diagnosis process and treatment. Results show that mainly anaesthetists of 3rd year of training experienced diagnostic difficulties. In-depth analysis of the problem situation showed that anaesthetist doesn't ask for help, as the procedure suggests. Fatigue, poor communication, distractor and inappropriate supervision are the most frequently associated factors mentioned by the anaesthetists.

The third prototypical risk situations concerns mainly novice anaesthetists of 2nd and 3rd year of training. It concerns inappropriate preoperative and postoperative anaesthetic evaluation. It is the dynamics of the world and the adaptation to its which holds our attention here. Fatigue, poor communication but mainly lack of experience are the most frequently cited associated factors.

The last prototypical risk situations involve more experienced anaesthetists and concern also inappropriate preoperative and postoperative anaesthetic evaluations. The most frequently cited factors associated to these problems are fatigue and poor communication.

We have attempted to validate these patterns by the means of the LISREL (Estimation of Linear Structural Equation Systems by Maximum Likehood Methods) but the limited number of our database and the binary form of the variables did not allowed to develop mathematically the relational structures.

However, the correspondence analysis procedure allowed to describe the proximities existing between the type of failure and the degree of expertise, taking into account the difference in numbers (figure 1). Given the restrictive number of our data, it has mainly a descriptive value. The proximity of the three points, 2nd, 3rd and 4th year of training signifies a similarity of the profiles of failure types. If we look at the distances between these 3 points, we see that the types of failure that most concerned this sub-category of anaesthetists are the failures of preoperatory risk evaluation, of detection, of diagnosis (particularly close to the 3rd-year subjects), of procedure and of execution, while the 5th-year anaesthetists and the seniors are more concerned by preoperatory and postoperatory risk evaluation failures.

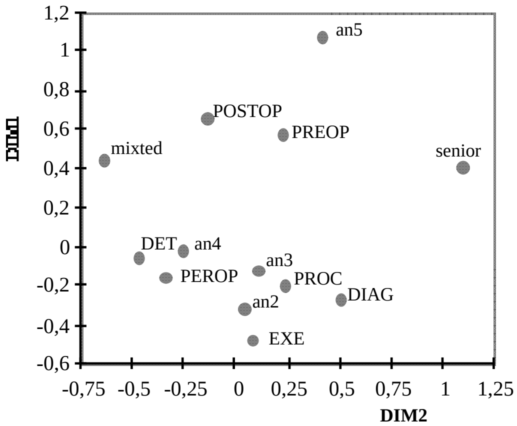


Figure 1 - The correspondence analysis -- Proximity between type of failure and degree of expertise

Conclusions

It is not our concern in this paper to discuss the results regarding the anaesthesia accident literature (see ref.14). We will focus on the systemic analytical perspective adopted in the study and discuss its interests in the view to escape from too superficial accident analyzes in medicine that emphasize the role played by one single factor, human error. In this discussion, we will also attempt to demonstrate that the prototypical risk situations that emerge from our accident data are not specific to one single domain but rather present characteristics that are shared by complex dynamic systems. Seen in this view, we believe that they can be cross-contextual.

We collected and analyzed 30 problem situations and described the contextual factors but also the critical decision functions that have played a role in the genesis of the problem situations.

We observed that the main critical decision function mentioned by the practitioners is : failure to perceive valuable information during surgery. The first configuration shows a link between this kind of difficulty and the fact that the process change not only as a direct consequence of the problem's solver's actions but also spontaneously, due to system's factors such as other agent's actions on the process (for instance surgeon's actions in our case). We can relate this configuration to problems of situation awareness that has been described as the origin of many accidents in other dynamic environments. In aeronautics, Sarter and Woods (ref. 15) have observed the difficulties experienced by pilots for constructing an uptodate representation of the state of the plane, notably concerning the status of the on-board automated systems. The information given out by these automated systems is often insufficient to allow the pilot a precise representation of the functioning (ref. 16). In the same way, the surgical draping limits the anaesthetist's access to visual information about the surgical procedure. If access to visual information is limited and if verbal exchange within the team is reduced, as we observed, the practitioner might not be informed of problems until critical signs appear on the displays. This can induce a delay in the diagnosis process and treatment. Collective work has become indispensable in modern working environments due to the development and complexity of techniques. It is the visibility of the behavior of all the agents (machine and/or human) involved in a task that is our concern here.

The second prototypical situation concerns problems of planification and anticipation. More and more studies describe planification and anticipation as a determining element for the recovery from problems in dynamic environment (refs. 17, 18). From the cognitive point of view, we can consider that the orientation of the attention before the appearance of a problem contributes to partially reducing the uncertainty of the worker regarding the place, the moment and the type of response he should make during the problem situation (ref. 19). It is easy to imagine the advantage of this benefit in dynamic system where time delays are crucial in avoiding permanent damage. In the domain of anaesthesia, this observation led several countries to put forth the preanaesthesia visit as an obligatory safety standard (ref.20). But making this visit mandatory is not a guarantee of safety. In most hospitals, preanaesthetic visits are often done late in the evening before the operation, under time pressure and even by the team on duty. In these conditions, the quantity of resources assigned to this visit, is perhaps not optimal. Bias as *reticent rationality* described by Reason (op.cit.) as the refusal of human cognition to engage in fastidious analytical reasoning under high workload pressure may shape the quality of the decision making process (see the 4th prototypical risk situation involving more experienced practitioners in our study). Multitasking is a common attribute of modern systems and human can easily be overwhelmed by work demands in such conditions.

Finally, the analysis of our data permit to link the type of cognitive difficulties encountered by the practitioner and his degree of expertise. This can improve our knowledge about the progressive development of skills and abilities required by the task in the context. We observed that mainly anaesthetists of 3rd year of training experienced diagnostic difficulties. At this intermediary period of learning, it has been shown from other studies (refs. 21, 22), that behaviour would be guided in a "descending" manner by more rigid structures of knowledge or rules, ignoring possible exceptions. This may leads to the development of "fixation errors" which have been widely recognized as a major source of human error in dynamic and complex environments (ref. 23).

To summarize, the table 4 presents the prototypical risk situations that emerge from the flow of accident data. These patterns emphasize the links between human performance, contextual factors, some attributes of complex work systems and the structuring of expertise. They emphasize the cognitive compromise firmly established in a complex context and the persistence of errors even in more experienced practitioners.

In conclusion, this study emphasized the need to adopt a systemic approach to analyze accident in complex systems such as clinical systems. To do that, we complemented classical accident analyses, based on the external contributory factors, with a cognitive analysis, based on the decision functions involved in the problem situation. This type of research is critical to better understand the demands and the constraints pose by the systems on the practitioners and predict some system breakdowns in modern human-machine working environments. The results should lead to the development of better training programs and technology developments that actually improve safety. We believe that the concept of prototypical risk situation provide a mechanism for abstracting common generic patterns of system breakdowns. We proposed this study as an encouragement to further investigations in order to test the hypothesis of prototypical risk situations on a larger sample of accident data but also in other complex domains.

Table 4 - Characteristics of the four prototypical risk situations

Critical Decisions Functions	Attributes of the system	Underlying cognitive mechanisms	Degree of training
failure of detection & per manag.	collective work dynamic	limited attention capacity	trainees
misdiagnosis	uncertainty	Heuristic fixation	3rd
failure of reevaluation anticipation	dynamic feedback	cope with time	novices
bad evaluation	Multi tasking	reticent rationality	more experts

References

1. Amalberti, R. Safety in flight operations. In B. Wilpert, & T. Qvale (Eds), Reliability and safety in hazardous work systems : approaches to analysis and design. Hillsdale : Erlbaum, 1993.

2. Chopra, V., Bovill, J.G., Spierdiyk, J., & Koorneef, F. Reported significant observations during anaesthesia : a prospective analysis over an 18-month period. British Journal of Anaesthesia, 1992; 68: 13-17.

3. Hollnagel, E. Models of cognition : procedural prototypicals and contextual control. Le travail humain, 1993; 56, 1/1993: 27-57.

4. Rasmussen J. Analysis of the tasks, activities and work in the field and in laboratories. Le travail humain 1993; 56, 133-147.

5. Gaba, D., De Anda, A. The response of anaesthesia trainees to simulated critical incidents. Anesthesia and Analgesia, 1989; 68: 444-451.

6. Sanborn, K.V., Castro, J., Kuroda, M., Thys, D.M. Detection of intraoperative incidents by electronic scanning of computerized anesthesia records. Anesthesiology, 1996; 85, 977-87.

7. Cooper, B., Newbower, R.S., Long, C.D. Preventable anesthesia mishaps : a study of human factors. Anesthesiology, 1978; 49: 399-406.

8. Vourc'h, G. Enquête épidémiologique sur les anesthésies. Ann. Fr. Anesth. Réanim., 1983; 2: 333-385.

9. Woods, D.D., Johannesen, L.J., Cook. R.I., Sarter, N. B. Behind human error: Cognitive systems,computers and hindsight. Sate-of-the-art report for CSERIAC, Dayton, OH, 1994.

10. Reason, J. The identification of latent organizational failures in complex systems. In J.A.Wise, V.D. Hopkin & P. Stager (Eds), Verification and Validation of Complex Systems: Human Factors Issues. Berlin: Springer-Verlag, 1993.

11. Hollnagel, E. The phenotype of erroneous actions. In G.R. Weir & J.L. Alty (Eds). Human-Computer Interactions and Complex Systems. London: Academic Press, 1991.

12. Reason, J. Human error. Cambridge: Cambridge University Press, 1990.

13. Woods, D. Coping with complexity : the psychology oh human behavior in complex systems. In L.P. Goodstein, H.B. Anderson, & S.E. OLsen (Eds), Tasks, errors and mental models. New-York : Taylor & Francis, 1988.

14. Nyssen A.S. Vers une nouvelle approche de l'erreur humaine dans les systèmes complexes : exploration des mécanismes de production de l'erreur en anesthésie. Thèse de doctorat (unpublished), Université de Liège, Liège, Belgium 1997.

15. Sarter NB, Woods DD.Strong, silent, and out-of-the-loop : properties of advanced (cockpit) automation and their impact on human-automation interaction. CSEL Report-95-TR-01. Columbus, Ohio State University, 1995.

16. Norman DA. The problem with "automation": inappropriate feedback and interaction, not "over-automation". Philosophical Transaction of the Royal Society of London, 1990; B327: 585-593.

17. Xiao Y, Milgram P, Doyle D. Off-loading, prevention, and preparation : planning behaviors in complex system management. 25th Annual Conference of the Human factors Association of Canada.. Mississauga, Ontario : Human Factors Association of Canada, 1992.

18. Hoc, J.M. . Effets de l'expertise des opérateurs et de la complexité de la situation dans la conduite d'un processus continu à long délais de réponse : le haut fourneau. Le Travail Humain, 1991; 54: 225-249.

19. Camus J.F.La psychologie cognitive de l'attention. Paris : Armand Colin, 1996.

20. Belgian Anaesthesia Patient Safety Steering-Committee: Belgian standard for patient safety in anaesthesia. Acta Aanesth. Belg. ,1989; 40: 231-238.

21. Lesgold A. Acquiring Expertise. In: J. Anderson & SM. Kosslyn (Eds). Tutorials in Learning and Memory. New-York: Freeman and Company, 1984; 31-58.

22. Anderson JR. Acquisition of cognitive skills. Psychological Review, 1982; 89: 369-406.

23. De Keyser V, Woods DD. Fixation errors: failure to revise situation assessment in dynamic and risky systems. In: Colombo AG, Saiz de Bustamente A, eds. Systems Reliability Assessment. Brussels, B: ECSC, EEC, EAEC, 1990; 231-251.

Improving the User-Interface to Increase Patient Throughput

A. Gupta, J. Masthoff,
Philips Research Laboratories, Redhill, UK

P. Zwart,
Philips Medical Systems Ned. B.V., Best, The Netherlands

Keywords: Medical, Error, Guidance, Customisation, Learning, Task Analysis, User Interface, Help.

Abstract

Philips Medical Systems is a leading supplier of equipment to the international health care industry. We have been exploring the potential for applying advances in user interface technologies (particularly customisation) in future medical X-ray systems. In this paper, we outline the rationale for this by looking at the trends in the health care industry, and the types and sources of error, which we have observed first hand in various hospital departments. We distinguish and discuss three ways to reduce errors: improved design of the user-interface, customised default parameter values, and active support for learning during system usage i.e. during examinations. We have based our work firmly on the principles of user-centred design: observations, task modelling, user involvement and prototyping have been undertaken.

Introduction

The research presented in this paper has been conducted in a collaboration between Philips Research Laboratories and Philips Medical Systems, with input from Philips Design. It is a three-year project that has been going on for a year.

As part of a team skilled in user-interface design, the authors of this paper from Philips Research were contracted by Philips Medical Systems to suggest improvements in the examination and control rooms. Our prior experience with the medical domain was rather limited. We needed to build up our knowledge of the clinical X-ray environment, so that we could identify real opportunities for improvement. Our customer did not see this as a burden – we were needed for our knowledge in the user-interface field and particularly in the methods underlying user-centred design, the customer provided the knowledge about the medical field. We therefore initiated a series of hospital observational studies, interviews, and training courses. We also needed to then conceptualise solutions that were technologically feasible. We organised a multi-disciplinary Design Workshop at which we held many brainstorms.

Structure of this paper

The structure of this paper is as follows. We begin by discussing what we mean by human error in the clinical domain. We then describe the observational studies undertaken. The process followed, methods used, and our findings, are described. The user-interface technologies that were judged to be applicable are then described. Summaries of all of the above were presented to participants at a Design Workshop as the raw material on which to base concept generation. The process we used is described. Additionally, the notation used to document the task model is also described. We then discuss a number of ways to reduce human error, namely improved design of the user-interface, default parameter values customised to the situation at hand, and active support for on-line learning. Customisation is the essential aspect and is described. We end with some conclusions.

Definition of human error in the clinical domain

Research on human error in safety-critical systems is generally concerned with the prevention and analysis of major accidents and near misses, concerning aircraft, trains, and oil and chemical plants. We are, however, not focusing on the extreme examples, but on more subtle, latent types of errors, where each occurrence in itself does not have catastrophic effects, but the cumulative impact on relatively less tangible aspects is considerable.

Leplat (as quoted in reference 1) states that "a human error is produced when a human behaviour or its effect on a system exceeds a limit of acceptability". Another definition of human error is "all the occasions in which a planned sequence of mental or physical activities fail to achieve its intended outcome, and when these failures cannot be attributed to the intervention of some chance agency" (ref. 2). Both of these definitions associate human error with not reaching a set goal ("exceed a limit of acceptability" and "fail to achieve its intended outcome"). So, in order to define human error in the clinical domain, we have to first assess what the goals in the clinical domain are.

Some important goals in the healthcare industry are to:

- maximise patient *safety and comfort*, for instance, by minimising dose and contrast fluid used, and by accurate diagnosis.
- maximise *patient throughput*, by reducing the time needed per examination,
- minimise the *resources* used per patient, such as film, contrast fluid, and staff time,

We take a broad view of error and define it as *human behaviour resulting in sub-optimal achievement of the goals listed above*. These errors are extremely likely to occur, but individually do mostly not have severe and obvious consequences. However, they pose a high risk even if the severity of their individual consequences tends to be low, as risk is defined as the probability or frequency of occurrence multiplied by the severity of the consequence.

According to our definition, the following are examples of errors:

- Applying a higher dose than strictly necessary (even when not violating safety limits!)
- Taking more time for an examination than necessary
- Using more contrast fluid than necessary
- Producing sub-optimal images and thereby endangering good diagnosis.

It should be noted that it is very difficult to assess whether one of these errors has occurred, because of the necessary balancing between different aspects. For instance, taking less time for an examination may lead to less patient comfort and less accurate diagnosis. Applying a lower dose may lead to a lower quality of images.

There are many reasons why an error may occur:

- Sub-optimal operation of the system, i.e., not using all functions of a system or incorrectly using them, or not operating the system in the most efficient way
- Sub-optimal system design, i.e., the system does not support users' tasks sufficiently, the system is difficult or slow to operate, or the system is not designed for the particular examinations for which it is being used,
- Sub-optimal workflow management.
- Non-optimal diagnosis, for instance, not optimally using results from medical research. This issue is addressed in evidence-based medicine (ref. 3), and is outside the scope of this paper.

Often a combination of these occurs: the clinical user does not use a certain function of the system because the design of the system does not make the purpose of the function clear.

The basis of good user-interface design lies in applying knowledge of users, tasks and the context of use. This is the essence of the user-centred design approach, which we have practised from the outset. The analytical activities undertaken are now described.

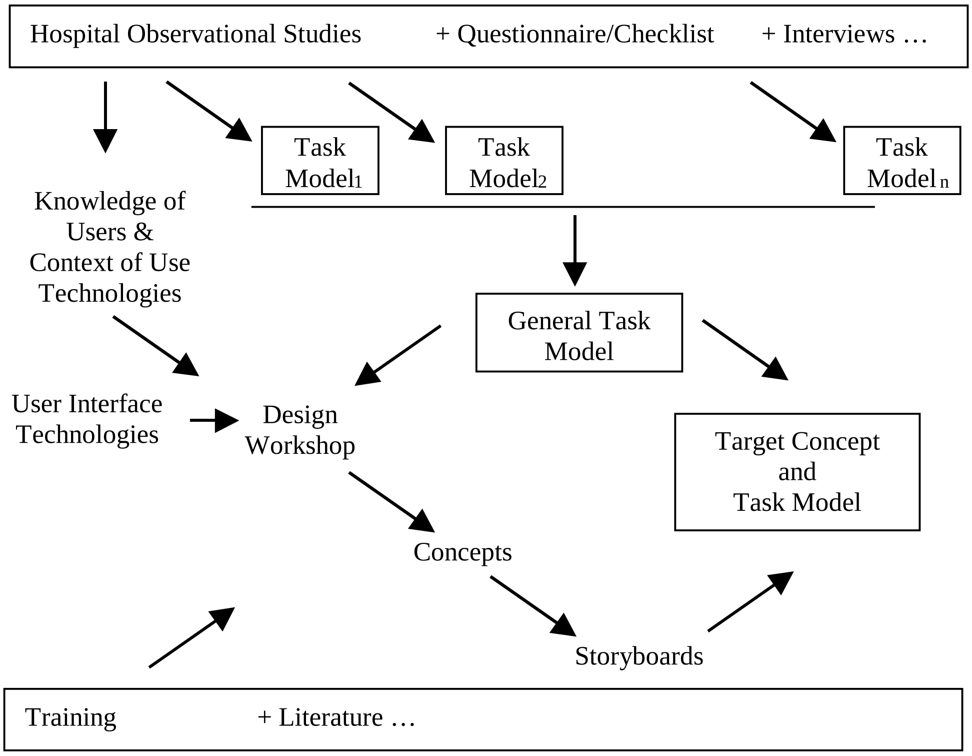


Figure 1 – User-centred Design Activities Conducted

Analysis

The range of activities we undertook is illustrated in the process map shown in Fig. 1. We have built up domain knowledge via training sessions at Philips Medical Systems (PMS), observational studies in hospitals, interviews with PMS' application specialists, and studies of manuals of some current Philips systems (the MD4 general radiology system, the BV300 surgery system, and the BH5000 cardiology system).

Organisation of observational studies: Visits were organised to five hospitals with which PMS has a special relationship. An application specialist of PMS who was very well acquainted with the hospital would accompany us. Application specialists are attached to hospitals and maintain frequent contact with clinical staff at all levels; they also deliver on-site training and support. The whole process was guided by a checklist (see Appendix).

Typically we would start and end a visit with a conversation with the application specialist. We were interested in the type of hospital, the specialisms of the department, the kinds of examinations conducted, the number and types of people working there, the types of systems used, the process used for introducing new systems into the hospital, trends in the hospital, and problems experienced and anticipated.

Next, we would be guided through the department and begin our observations; each observer typically doing about eight. During the examinations, we would stand in the control room, or the examination room wearing protective tunics, and observe what was going on. We would draw plans of the rooms, and record the actions of and communications between the clinical staff and patient, noting information present on walls, and ambient conditions, for example, lighting and noise. We would not only look at what happened during an examination, but also between examinations.

We questioned the staff present, namely radiographers, radiologists, cardiologists, and nurses. We were fortunate to be able to talk to very experienced staff and also to trainee radiographers and doctors. We would ask about the nature of their work, their roles, the problems they experienced, the changes they would like, the trends in their work, their training, and what they liked in their job.

<u>Observational studies conducted</u>: We have performed observations in hospitals in the Netherlands, Belgium, and the UK (ranging from film-based to a digital hospital). We have seen both diagnostic examinations and interventions, varying from general radiology, to MR, CT, and cardiology. We have also seen some other parts of the workflow, namely reporting and archiving.

The observational studies are an ongoing activity, starting off as a way to get to know the domain, the users, their tasks, and the environment, and now becoming more specific relating to the prototype system we are developing.

<u>Samples of information obtained:</u> We have recorded each hospital visit in a report (on average 25 pages, typical sections: users, systems, environment, tasks, information, training, UI issues, trends). Some fragments are included below (see also Figures 2 and 3):

"There is considerable ambient noise in the examination room: an intercom system is in use for broadcasting announcements, motors and jets may sound from the X-ray equipment, the people present in the working area talk amongst themselves, music may be played on a hi-fi system. The noise is intermittent.

A small bowel enema examination:

-The patient is selected from the RIS list.

-The examination-table is put in position. In the meantime, the patient is brought in.

-Local anaesthetic is applied (using a spray).

-A tube is prepared and inserted in the patient's nose.

-The patient is put on the table, on his back.

-The C-arm is moved.

-The monitors are moved.

-The lights are dimmed.

-The table is moved up, the C-arm down.

-An image is taken.

-The tube is inserted by a radiographer in the patient's nose and down to the stomach while looking at fluoroscopy images on the ceiling mounted monitor. The fluoroscopy is operated by the other radiographer at the table operator's control (TOC).

-The one radiographer tells that the tube is in position. The other radiographer releases the fluoroscopy

In the mean time, the patient is covered with a white blanket.

-Two images are made.

-They wait again. Fluoroscopy is used, images are made. The C-arm is moved, using the TOC. This is repeated a couple of times.

-The patient repeatedly tries to look at the monitor, but has difficulty doing this as the monitor is behind his head.

-The patient is put on his stomach. The tube in his nose is put in the right position, using a kind of scissors or tang.

- An alarm sounds (a collision?).

- The patient is put on his back again. The C-arm is moved. Fluoroscopy is used, an image is taken.

- The patient is instructed on the position. The table and C-arm are moved. The radiologist looks at the monitors, then walks towards the patient to change something: he is pressing the Barium along. This is repeated.

- A lot of changes in geometry take place, and images are being taken. At some point the radiologist says it is okay

Figure 2 – Sample of an Examination Observation

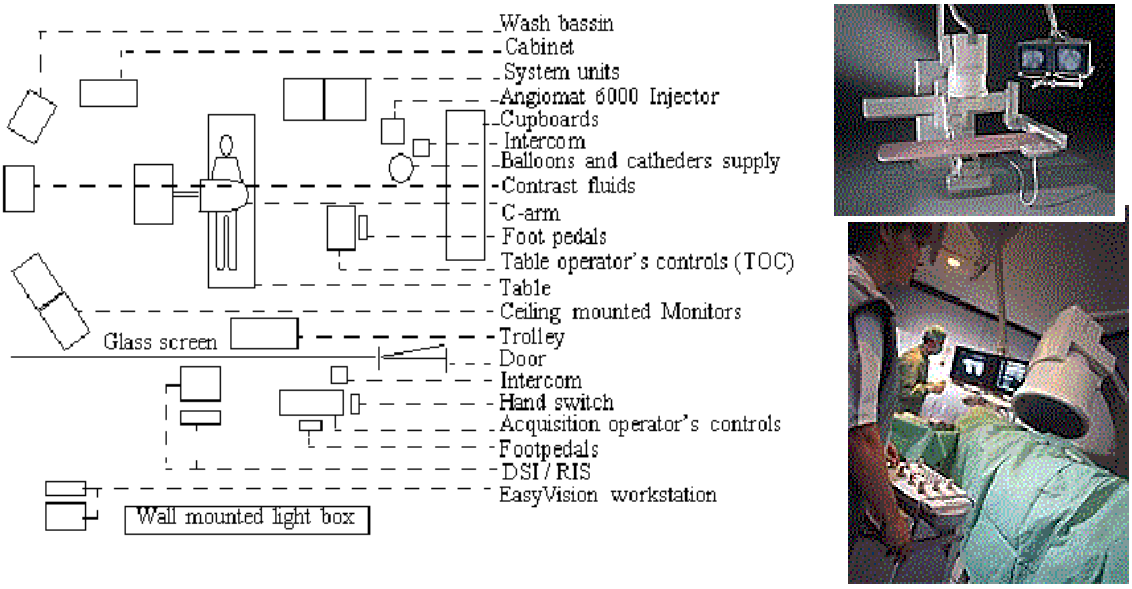


Figure 3 –Sketch of an Examination Environment and Pictures of an X-ray Setup.

There is no daylight, only artificial light - incandescent or fluorescent is available, usually lighting is soft, it may sometimes be dim, lighting levels may be changed during an examination."

Main observations: We have made the following observations during hospital visits:

- [Obs1] Radiologists and radiographers often only know part of the available system functionality, and not always how to accomplish their task efficiently.
- [Obs2] People with different backgrounds may have to operate the same system. For instance, surgery systems are often operated by cardiac radiologists as well as by nurses, for different types of examinations. The growing and complex functionality, for instance, for cardiac applications, further raises the threshold for the nurses to operate the systems efficiently.
- [Obs3] Hand-written reminders, tips, and other information, for instance on allergic reactions to contrast fluid, are stuck to the walls or the lead glass screen in the control room.
- [Obs4] *Training* takes place during a number of days on installation of a new system. A small number of clinical users are trained: initially by giving them a system overview, and then by assisting them during real examinations. These clinical users then go on to train their colleagues. So, a substantial part of training is conducted on the job. Colleagues exchange the odd tip about, for example, which command sequence gives a better image, and stick to this. In our hospital studies, user manuals were almost never seen in control rooms.
- [Obs5] *Device controls are spread around* – by the examination table, on a trolley-mounted console, a remote control for a ceiling mounted monitor, etc. We have observed physicians occasionally actually running from one to the other. Radiologists in the examination room have been observed shouting commands to the radiographer operating the system from the control room. Additionally, the latter may shout

instructions to the patient: “hold your breath!”, “don’t move!”, etc.

- [Obs6] Due to the range of systems from different manufacturers, and the lack of industry standards, controls vary from system to system causing confusion amongst clinical users who must repeatedly move from using one to another.

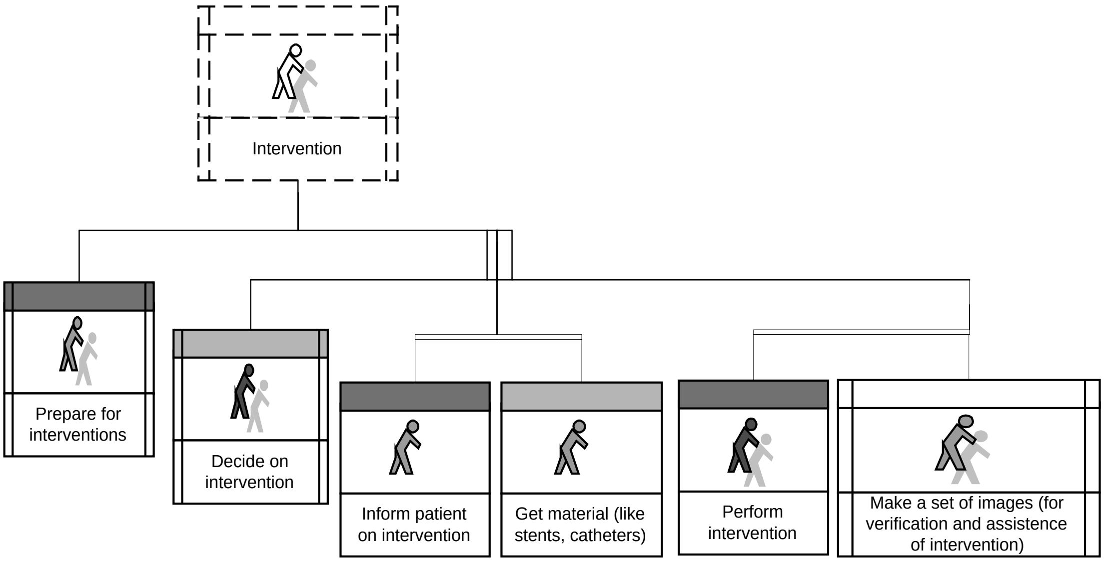


Figure 4 - Fragment of a generalised task model of existing systems

- [Obs7] When inserting a catheter, the physician must be sterile, which can make system control harder.

- [Obs8] We have observed instances where staff and patient do not have a common language making it difficult for correct instructions to be given to the patient. We will not address this issue in this paper.

Trends in the clinical domain: The following trends can be observed in the hospitals:

- [Trend1] An increasing amount of functionality is being offered in systems. For instance, the manual for Philips’ Cardiology X-ray system has almost doubled in size, increasing from 120 pages for the BH3000 to 245 pages for the following system, the BH5000.

- [Trend2] Boundaries between domains are becoming fuzzy. For instance, the boundaries between diagnostic and interventional radiology, and between cardiology and surgery are diminishing. Increasingly, for example, cardiologists prefer when performing a minimal invasive by-pass operation to also place an abdominal stent (i.e., a small intervention in a leg normally performed by a radiologist) when needed.

- [Trend3] There is a shift from stand-alone to networked systems where the medical equipment is effectively a peripheral. Increasingly, there is a requirement for equipment to be based on open standards, and for inter-operability with systems from other manufacturers.

- [Trend4] Information is increasingly stored digitally, giving rise to film-less departments.

The above observations and trends are not exhaustive, but are illustrative of some of the material on which we based subsequent work.

After each hospital study, we developed task models. This activity is now described.

Task Analysis and Modelling

A Task Model is a drawing that represents the tasks undertaken to achieve a goal. It is a hierarchical drawing as tasks may be decomposed into sub-tasks. It is read from left to right and captures the process used by the subject in going about his or her job. It is also independent of any technology, user-interface or other, as the level of abstraction is much higher than mouse and keyboard interactions. Task Models are of two types: those representing currently used systems, and those representing the target system to be designed. An example-fragment of the former kind is shown in Figure 4. We have based our task modelling work in part on the MUSE method described by Lim & Long (ref. 4). Note: The relation between a Task Model and a system is a subtle one: though a task model is rooted in one or more systems it is only a description of tasks, and is entirely independent of system-level details.

Benefits of task modelling:

1. Capturing important characteristics: In the observations we conducted it soon became clear that task characteristics such as location, communication, co-operation and duration would be useful to record. These could constrain design solutions (e.g. co-operation between users), or point to where improvements would be most effective (e.g. duration). These can easily be recorded as annotations on the Task Model.

2. Communication: Having a common frame of reference is useful both within the project team, and when communicating with partners. The task model proved to be very useful when collaborating with others during idea generation and interaction design.

3. As a basis for design: Comparing task models is a good starting point to find commonalities and differences which impact user-interface design. Once agreement on the desired task structure has been obtained, candidate design solutions can be brainstormed. Optimal functional allocation between system and user, can then be followed by interaction and graphic design.

4. Validation: Task Models can be shown to application experts to check that descriptions are correct. The notation (explained below) is simple enough to understand.

Notation:

Task Ordering: Two types of ordering are distinguished in the task model diagrams: Sequential Ordering (indicated by a single horizontal line, tasks ordered from left to right) and Free Ordering (indicated by a double horizontal line).

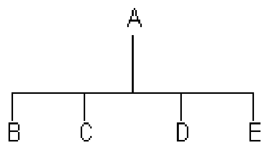


Figure 5 – Sequential Ordering

Suppose a task A consists of subtasks B, C, D, and E. Figure 5 indicates that the subtasks should be done in sequential order: B followed by C followed by D followed by E.

Figure 6 indicates that the ordering of the subtasks is free, i.e., unconstrained: B, C, D, and E can be done in any order.

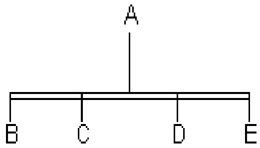


Figure 6 – Free Ordering

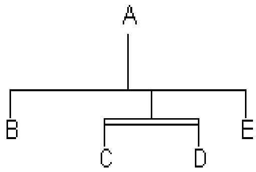


Figure 7 – Mixed Ordering

Figure 7 shows a mixture of sequential and free ordering: it indicates that B should be done before C and D, which should be done before E. The order in which C and D are done is not important.

Repeatable tasks: These are indicated by multiple vertical bars connecting a task to its parent task.

Optional Tasks: These are represented by boxes drawn with dashed lines.

Location of tasks: As we are only interested in two locations (examination or control room) we have used simple colour coding in the top of the box to represent this property. Brown (which prints as a dark shade in grey-scale) represents the examination room; blue (a light scale) represents the control room. An absence of colour implies no restriction on location.

Person: This indicates any demarcation on tasks amongst radiologists and radiographers. Red (which prints as a dark shade in grey-scale) represents tasks done by the former; green (a light shade) the latter.

UI-technology: Customisation

Reflecting on our observations, it occurred to us that there was a need for more flexible systems – unconstrained by requirements to be used at a fixed location, or by certain kinds of users, or by fixed interactions. Additionally, there was potential for a system which could be pro-active in enhancing the proficiency of the user. Customisation technology is the approach we followed. The Microsoft Office Assistant is an example of customised assistance (in this case of context-sensitive help). We can expect professional users to expect such features in the systems they use, after exposure to them in Word processing systems. We now discuss customisation in the user interface.

Customisation is the process of adapting a system's behaviour (by the designers i.e. system developers, by the system itself, or by the user) to suit the individual or group using it. In a sense, customisation takes place in each system design. The important variables are to *whom* the system is customised, *what* is customised, and *how* the customisation is done. These are discussed in turn.

To whom is it customised: Customisation is normally aimed at "the customer". In the medical domain, this may be the radiology department of a certain hospital. Each department may have different wishes with respect to the kind of functionality required, may want to use the system for different tasks, may have its own procedures, its own terminology, etc. Each customer can also have a different environment in which the system will be used, for instance, different lighting and noise conditions, different cultural background, etc. Usually a system is designed with certain prototypical customers in mind, having different versions of the system for different categories of customers, and providing options to configure a system to the wishes of the customer. In the medical domain, for instance, different versions may be made for the different countries (or regions), or for low-end and high-end applications, while there may be options for incorporating extended functionality. As every customer tends to be different (and no rough categorisation will be perfect) more customisation may be needed than can be achieved through these means**.**

Even for one customer (say a certain radiology department), many people may be involved with the system (like radiographers, technicians, administrators, radiologists, surgeons, nurses, etc.). The satisfaction of the customer then depends upon the satisfaction of a group of people, who may have different (sometimes even conflicting) requirements and preferences. Depending on their rôle, they will want to use the same system for different tasks, they may have a different level of education and experience, etc. To a certain extent, it may be possible to use stereotypical roles, and customise the system to these roles. However, these roles and their corresponding characteristics (like level of education) may differ per customer group (for instance, the education level of radiographers and the tasks they are permitted to perform differ according to national regulations).

Differences may also exist between individuals having the same role. There is a lot of literature on individual differences between people that may affect their operation of systems. Straightforward differences are, for instance, the level of experience, and the level of domain knowledge. Other differences mentioned are differences in psycho-motor skills, learning ability, understanding, expectations, motives, cognitive strategies, cognitive abilities, preferences, etc. These individual differences may impact the best way to present information for a certain user on the screen, the amount of information needed, the input device preferred, etc.

Over time, the system will be used to perform different tasks under different conditions. For instance, in the medical domain, different patients are examined, and different types of examinations take place. Customisation can also take place to these different situations over time.

In summary, a system can be customised to the environment in which it is used, to the (group of) users that operates it, to the tasks that these users perform in general and at a certain moment in time, to the situation at hand.

What is customised: Many aspects of a system can be customised. For instance, in the Appeal tutoring system (ref. 5), various aspects were adapted at run-time to the individual student. Examples are the navigation path, timing and amount of instruction, form of information (e.g., using formulae or pictures), (amount, tone and timing of) feedback to the student, and the content of exercises. In an adaptive component of a fighter aircraft (ref. 6), the system adapted task allocation on the basis of the pilot's state, system state and world state.

In general, a system gives the user the opportunity to give input, i.e., to make choices, and presents output depending on the choices made. For instance, an X-ray system gives the user the opportunity to set several parameters (like dose, geometry position, shutters), and the system produces X-ray images accordingly. As such, a system can be said to consist of:

- *Parameters* that the user can set, such as Brightness and Contrast, each with an associated default and range.
- *Functions* that the user can apply to change the value of parameters.
- A *Navigation Structure* determining which functions are directly available to the user at a certain moment in the interaction and how the functions are grouped (this includes the menu and dialogue structure).
- *Ways of Interaction* determining how the user can operate a function and how output or information is presented.
- *Information* regarding which functions the user can apply and how to apply them, and which tasks to perform without the system (like what to tell/ask the patient).
- Features determining the *appearance* of the system such as colours, shapes, icons, sounds, labels, and fonts.

All of these aspects can be customised. For instance, the preferred value of a parameter (such as Brightness) can depend on the task (namely the particular examination type), individual preferences of the user, the environment (such as the lighting condition in the room) and the values of other parameters (like Image Polarity). The default offered to the user could reflect this. Similarly, the preferred way of interaction may depend on the availability of hardware (like microphones), the examination type, and the experience and preferences of the individual user. For instance, during an intervention the radiologist has to be sterile and the use of speech input for non-critical tasks may be preferred. On the other hand, users who have operated a certain function in a certain way for a long time will not always like to change.

The preferred navigation structure may depend on the user's rôle (for instance, a nurse normally does not need advanced functions), the current task, and the experience of the individual user.

How is it customised: There are several ways in which customisation can take place (see also reference 7):

- System designers can search for a general design that is customised to most of the users' needs
- the interface can be designed to incorporate the means for users to customise the interface themselves
- mechanisms can be devised and built in that are able to determine a user's particular needs and preferences and automatically customise the interface accordingly
- it can be attempted to "customise the users to the interface" by providing support for training users.

Mostly, a combination of these options is preferable.

In every *run-time* customisation process four phases can be distinguished (ref. 8): in the first phase a customisation is suggested, in the second phase a number of alternatives are offered for the customisation, in the third phase a choice is made from the alternatives, and in the fourth phase this choice is carried out. Each phase of the customisation can be performed by the system or the user.

If the system performs all phases, the process is called Self-Customisation (e.g., ref. 9). In case the user performs all phases, the process is called User-Controlled Customisation (e.g., ref. 10). We are most interested in hybrid situations, in which depending on the situation, the phases can be performed by either the system or the user. A primary reason for intermediate situations is that the user may have more information than the system in some areas, especially concerning his or her own preferences, while the system may have more information in other areas, especially concerning its options. The system may, for instance, suggest alternative available methods for increasing Brightness from which the user can make a choice. A second reason for intermediate situations is that the user may feel more at ease knowing that he or she is in control (ref. 5).

Constraints: There are several issues to be considered when designing self-customising systems (see amongst others ref. 11).

One potential problem is that it might become more difficult for the user to learn to operate a system, if it changes over time. The acceptability of the system may also decrease, as the user might not understand the system's behaviour. To compensate for this it has been suggested that the user be given an explicit model of the customisations being undertaken by the system. For instance, the user may be notified that the position of items on a list may change according to their most frequent choices. In this way the user can anticipate, or at least rationalise the actions being taken by the system.

Norcio & Stanley point out that users may feel a lack of control if the system starts to a make choices on their behalf (ref. 12). This may be prevented by giving them at all times the chance to override the system's decisions, or by only giving them suggestions. An alternative is to give the users the opportunity to turn off automatic customisation without prior user confirmation. In this way they have ultimate control, the knowledge of which may be sufficient for them to accept customisations. Also, a system should know when and when not to consult the user about a certain customisation.

Related to the previous issue is the issue of safety and legal responsibilities. Especially in the medical domain, these issues are very important. The operator should have the final responsibility for the settings chosen, if these settings have an impact on the safety of a patient.

In order to self-customise, a system has to gather and store data about the user, his tasks and preferences. There are, of course, privacy issues related to this. For instance, not every radiographer might appreciate it if a supervisor could inspect the way they normally use the system. So, there have to be rules about who can access which data.

Users may also not like to be categorised. Great care has to be taken when using stereotypes.

Concept generation

After the analysis phase (in which we had already identified some options for improvement), we started to generate concepts for new systems. For this, we needed individuals skilled in ergonomics, industrial design, interaction design, and graphic design. We were able to call on colleagues in Philips Design, for this skills set.

First we organised a workshop involving four designers (three with experience in consumer systems like audio systems, one with experience with medical systems), three researchers, an application specialist of PMS, two people from PMS Predevelopment, and a very experienced radiographer. The purpose of the workshop was to come up with far-out ideas. This is typical for a research project. In a more development-like project, you would also come up with concepts, using focus groups etc, but these concepts would (intentionally) have shorter time horizons for realisation.

Next, five concepts were worked out in more detail, into storyboards, and made more realistic. The designers involved (both without previous experience in the medical domain) were brought up to speed by reading the hospital visit reports and studying the draft generalised task model. The task model was used as a basis for this work: we identified which sub-tasks were closely related to a concept and visualised how the concept could impact these tasks. The storyboards were presented to PMS and we used some criteria (like technical feasibility and marketing potential) to decide which concepts to proceed with.

Figure 8 shows some fragments of storyboards.

Ways to reduce human error

From our hospital observations, it can be concluded that current systems are being used sub-optimally. This is a cause for concern, as trends in the clinical domain indicate that sub-optimal use will continue to increase in line with system functionality and complexity. According to our definition of human error in the clinical domain, sub-optimal use is a main source of errors. In order to ensure optimal use, the clinical users need to know:

- *Which functions can be useful in a certain situation*. For instance, when performing a Subtraction task (i.e. subtracting a reference mask image from another image), motion artefacts can occur due to patient movement. The user needs to know that the ‘Pixelshift’ function can correct this (by shifting the relative positions of the mask and source images). This relates to observations Obs1, Obs2, and Obs3.

- *When to apply a certain function*. For instance, Brightness can be set whenever viewing an image, but it is advisable to set Polarity, electronic shutters, and optionally perform subtraction, beforehand.

- *How to operate a certain function in a certain situation.* This concerns actions, their sequencing, and timing. For instance, in a future system, one could envisage that a Zoom function could be activated by a voice command and then, touching the area that is to be the centre of the zoomed image. There may be alternative ways of operating the same function. During some interventions, hands-free or even eyes-free operation may be needed (cf. Obs7 and Obs5). If the background noise is too high, speech recognition is impaired and another way of operating the function may be more appropriate. Additionally, the experience and preferences of users is a factor (cf. Obs2 and Obs3).

- *What the scope of a function is.* For instance, when changing Brightness while viewing an image, the effect can be a change in Brightness of all images in that set, all subtracted images, or only that single image.

- *How to undo the effect of applying a certain function.*

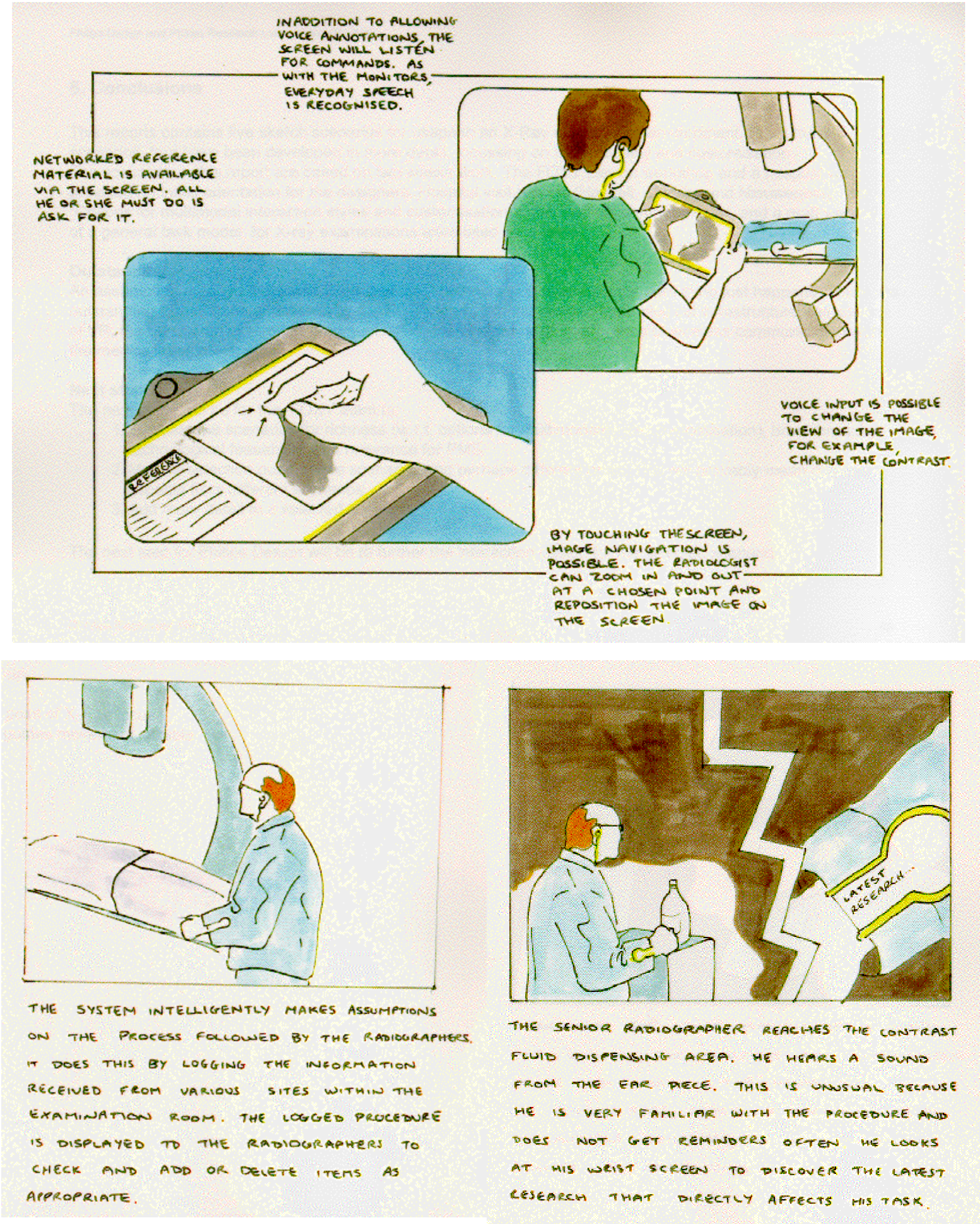


Figure 8 - Fragments of two Story Boards

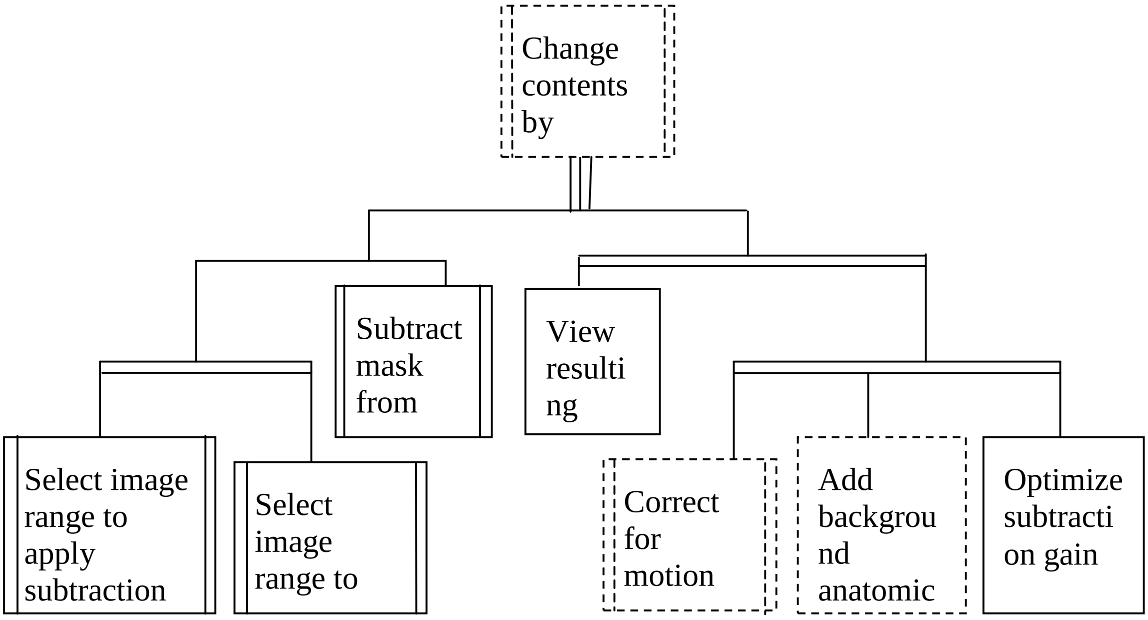


Figure 9 - Fragment of a Task Model for a Target Application

The system needs to support the various users in performing their tasks and obtaining the knowledge relevant in a certain situation. We try to achieve this by a combination of:

(1) a well designed user-interface (including consistent operation, cf. Obs6),
(2) customised default values minimising the amount of effort required from the user, and
(3) active support by the system for the learning process (needed as training takes place on the job, cf. Obs4).

These are now discussed in turn.

<u>Improved design of the user-interface</u>: Designing a user-interface requires us to answer the following questions: (1) which user tasks to support, (2) which functions to provide in order to support these tasks, (3) what navigational structure and screen layout to use to make these functions available, (4) how to enable the user to operate these functions, and (5) what information to present to the user about the functions.

The first question leads to the construction of a task model of the target application (as discussed in the task modelling section). This task model is based on (improvements of) the generalised task model of existing systems and the concept generation activities. A fragment of a task model for a target application supporting Viewing and Post- processing of images is presented in Figure 9. The notation is similar to the notation above, except that location and person indications have been omitted, as all the tasks are intended to be performed in the same location and by any user type. We are currently in the process of validating the task model with application specialists and clinical users.

The second and fourth question lead to the determination of which functions to provide to a user to assist with the leaf-tasks of the task model and which interaction to provide for these functions. To make this decision, we have analysed the way in which these tasks have been implemented in existing systems (such as existing X-ray, surgery and cardiology systems, but also completely different systems like drawing packages). Such an analysis leads to a choice inspired by best practise across many domains, as much as the need to balance consistency with existing practise in the medical domain.

The third question (about navigation structure and screen layout) can be answered on the basis of the task model, the functions and interactions chosen, and the available space on the interface. The ordering in the task model influences the navigation structure and screen layout of the system as follows. A task that has to be performed before other tasks, might be better represented on an earlier screen. Tasks that can be performed in any order should preferably be present on the same screen. In western systems, tasks on the same screen will normally be ordered from top to bottom and from left to right. For example, in the case of a sequential ordering of tasks A and B, the functions to enable task A should either be on a screen that precedes the screens with the functions of task B, or they should be to the top (and left). It should be noted that even in the case of free ordering, the layout of the task model drawing may lead to conclusions from, say, the screen-layout designers: there is a high probability that in a free ordering of a number of tasks, the order from left to right will be used as a normal (though not required) ordering. Therefore, we have ordered the tasks in the task model within a free ordering from left to right according to a likely sequence. Notes have been added to the task model to indicate more than one likely sequence, or to state conditions under which a certain sequence is likely to occur.

The grouping of subtasks in the task model, reflects the fact that certain tasks are conceptually more related to each other than to other tasks. In the user interface this could result in visually grouping functions related to those tasks (for instance by proximity).

The fifth question (what information to present about the functions) is addressed below in the section on active support. In principle, all relevant information, which the user does not know has to be presented. The different types of information have been discussed above.

<u>Customised default values:</u> The main input of a user to a system can be seen as the setting of parameter values, such as, for instance, choosing brightness and contrast values, determining the current image by image navigation, etc. Often, a system will provide the user with default values (these may also be the last values used), which the user can modify to his own choice. The closer the default values are to the preference of the user, the less effort is required from the user to change them. Ideally, the user does not have to do anything, as the values offered by the system already reflect his preference. This is very difficult to achieve, as the situation in which the parameters have to be chosen changes continuously, for instance, as the user is viewing different types of images. However, as a step in this direction, we have developed a so-called Customisation Engine, which determines the default values (and ranges between which the parameters can be modified) on the basis of the current situation, and the preferences of the (individual) user.

The Customisation Engine uses (see Figure 10):

- *Rules* specifying the range of parameter values regarded as good in a certain situation. We distinguish two types of rules: constraints and suggestions. Constraints specify the allowed values of a parameter in a given situation, while suggestions give good default values.

  The situation in which a rule is applicable can consist of a number (concatenation) of situational conditions. Each situational condition specifies a range for a certain parameter. For instance, the following can be situational conditions: “Examination type in {Barium, ERCP, Skillet, Myelo}”, “12<=Brightness<=25”, “Subtraction=Off”, and “Physician =Dr Spock”.

  It is possible that more than one rule is applicable in a given situation. In that case, the intersection of the ranges of suggested or allowable values for the parameter is used. If this intersection is empty, the rule that uses the most situational conditions is applied.

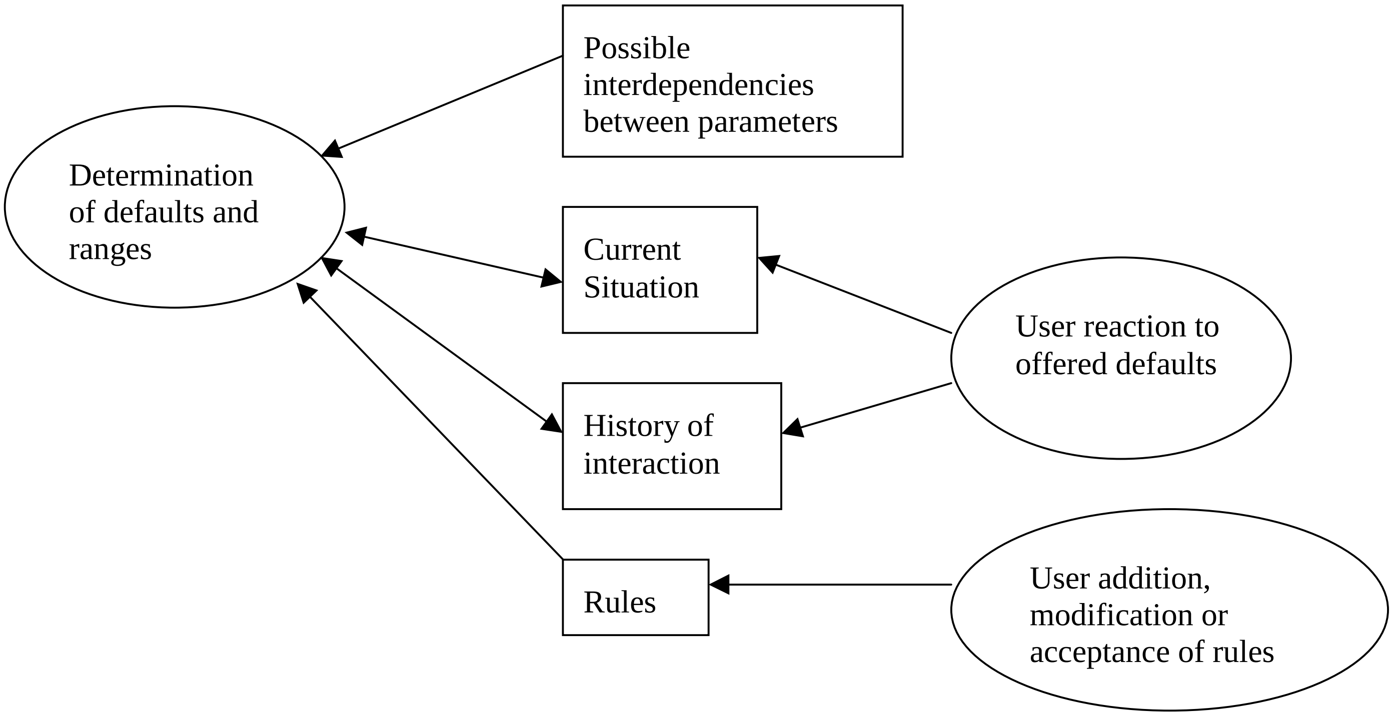


Figure 10 – Schematic Representation of the Customisation Engine

- *The History of the Interaction* reflecting how the user has reacted to suggested default values in the past. For instance, when an image is presented to the user with a default Brightness of, say, 5, then it is stored whether the user has kept the Brightness at that level (i.e., implicitly agreed to the offered value; this is called a "match"), or to which value he has changed it (i.e., did not agree to the offered value; this is called a "mismatch"). The values of other relevant parameters at the time when the user made his decision (like the Image Polarity and the type of examination) are also stored.

  The history is used to learn the preferences of the user at run-time. For instance, when the user repeatedly increases the Brightness value of images of a certain type then it is concluded that the user may have a preference for a higher value in that situation. The speed at which the system learns is directly related to the number of history instances it takes into account (the more instances the lower the speed). If the learning speed is high then customisation takes place rapidly but is prone to misunderstand the user. If the learning speed is low then customisation is slow but more accurate. The learning speed should be customisable by the user (or the system could learn it itself based on observation of correct and incorrect customisations in the past).

  It should be noted that no deep modelling of the user takes place, but that decisions are almost directly based on the data stored in the history.

- The *Current Situation* specifying the current values of all the parameters. The current situation determines whether a rule is applicable or whether a certain piece of history information is relevant.

- *Possible Interdependencies between parameters*. In order to make the use of the history information for learning easier to realise, information is available on which parameters are likely to depend on which other parameters. This makes it possible for the designer to exclude highly improbable dependencies from being considered, like the preference for an input modality depending on the Brightness of the image.

The use of rules described above is not only applicable when determining the default values and ranges of the parameters. It can also be used to customise other types of parameters. For example, interaction style parameters (e.g., does the user prefer touch or speech), implementation parameters (e.g., what is the maximum time between two touches for the system to regard them as part of one action), and navigation structure parameters (e.g., "if Subtraction is off then Availability of Pixelshift function is false"). Learning the preferences of the user through observation is, however, more difficult and requires different mechanisms, because preferences of the user can only be determined as above if the user is presented with a choice.

Interaction for customisation: As mentioned above (see sub section "Constraints"), it is very important that the user keeps feeling in control, understands the adaptation taking place, and is not hindered in his learning of the interface by continuous changes. Therefore the interaction style used for the customisation – i.e. the way customisation is presented to the user and the way the user is involved- is crucial.

In the description of the Customisation Engine above, various moments can be seen at which such an interaction might occur:

- Parameters are presented to the user, with certain defaults. After some observation of the user's actions in the past, the system may conclude that the preference of the user is different to what is programmed in its rules. The following options are now possible:

  (1) The user is presented with a new default, with an indication that the default has changed by making it very conspicuous
  (2) The user is notified that a deviation in his behaviour from the standard settings has been observed, and asked whether he would like to change the standard settings to a newly suggested setting reflecting his recent behaviour,
  (3) The user is presented with the old default, but it is made very clear that the system has observed a new default, and the user can get this default if he wants it (without being explicitly asked).
  (4) The user is presented with the old default, but at a moment when the user is not busy, like between patients or at the beginning or end of day, he is notified about the observed deviation and given the opportunity to add a new rule, or modify an existing one.

  A problem we have found (during informal evaluations) is that it is difficult for users to realise that changing a rule during a certain examination may have implications on other similar examinations in future (even though this was clearly stated in the question asked to the user). Therefore the option to change rules when the user is not busy with a specific examination seems preferable. We still have to conduct user-evaluations to see which option or combination of options is the best.

- When the user changes the value of a parameter, then the best default for another parameter may change because of interdependencies. The following options are now possible:

  (1) The defaults (and perhaps even the ranges) of the dependent parameters are changed automatically, with a clear indication that they have changed
  (2) The user is notified about the interdependencies and given the opportunity to accept or reject them (this holds only for defaults, ranges have to be accepted).

(3) Nothing is changed automatically, but the user is informed that the system sees an opportunity for optimisation and can obtain this optimisation if he wants it (without being explicitly asked).

The acceptability of automatic interdependency changes has to be investigated, especially when it concerns automatic changes of a parameter that the user has already explicitly set. (For instance, the user sets the Brightness value, then determines to change the Polarity, which has a possible implication on the Brightness value.)

We have the hypothesis that the third option is the best, but we have to evaluate this in practice.

In addition to these two moments of interaction, there is also the explicit changing of rules by the user. We have explored a way to enter knowledge that may be more familiar to clinical users than the use of rules: the user is presented with prototypical images and can change the appearance of these images, or chose between alternative appearances of images. Rules are deducted on the basis of the results. This work is still in a preliminary phase.

Active support for the learning process: It is impossible to give a user all the knowledge mentioned above during a short training period. However, more training time is unlikely: users generally have to learn on the job. Therefore, we are working on active support by the system of the learning process. The support has to be highly customised, depending on the hospital, the experience and preferences of clinical users, the examination being conducted, and user roles.

There has been quite some research into the design of intelligent help (see, e.g., reference 13), focusing on providing the users with good ways to ask questions and on volunteering answers when the user is observed to have a problem. In that research, the user-interface is seen as a constant factor. We have taken a different approach, and are focussing on changing the user-interface to increase learnability. We see both approaches as being complimentary.

Learnability is mainly related to efficiency and effectiveness over time. Preferably, the system should be highly intuitive to use the first time, i.e., have an immediate high effectiveness without much effort on the side of the user, and efficiency should increase over time, leading to a highly efficient system for experienced users. Moreover, systems should support the rapid progress of users from initial to experienced level.

In the system we consider (and this holds for most systems), only a limited amount of space is available for the presentation of functions and information. Nevertheless, we know for the X-ray viewing and post-processing domain that the users want many functions to be available at the same time, without having to take extra navigational actions (so, in manner of speaking “within one button click”). By using a good user-interface design, and making only those functions available which are appropriate in a certain situation (see section above and Customisation section below) it is to a certain extent possible to meet this user wish. If we include the use of speech commands to operate functions, it is even possible to have functions available that are not immediately visible on the system. However, to present all allowable functions on the system or hide them by using speech commands, has a rather negative impact on learnability. There is insufficient room to explain what all the functions do, and how to operate them, and this is particularly acute when the user is presented with a wide choice. So, our concern for novice users conflicts with our desire to provide highly efficient systems to experienced users.

One option may be to provide separate interfaces for novice and experienced users. However, this has only limited advantages: a user who is familiar with the use of certain functions (such as setting Brightness and Contrast) may never have heard of other functions (like Pixelshift) which can be very useful for his current situation (say, performing a Subtraction with motion artefacts). This is something we have actually observed in our hospital visits. The option of two separate interfaces also poses the danger that the gap between novice use and experienced use becomes too large.

What we would like, is that, for instance, when the user is performing a Subtraction, and has not used Pixelshift before, the system indicates that this function can be used to remove motion artefacts, and how to use it. This means that we will not use two separate interfaces, but one interface which gradually changes to reflect the experience of the user (i.e., the navigation structure and information is customised to the experience of the user).

The (simplified) example illustrates how this could work (see Figure 11, note that it uses a touch sensitive screen). If the user is not familiar with the concept of subtraction then this concept is introduced (image 1). Next, the user is guided through the subtraction task. At each step,

information is given on what he should do, how he should do it, and how he can correct mistakes (images 2-4).

Note that the initial sequence presented (images 2-4) is directly based on the task model (see fragment of the task model in Figure 10). Initially the user is forced to do the tasks in the task model from left to right, even though the model allows free ordering. When the user gets more experienced, the free ordering is introduced in the user interface (compare images 2,3 and 6). Note also that the information presented reflects the types of information mentioned above. What information is presented to the user depends on his experience. If the user has never used a certain function before (and not seen the information related to that function recently), then information regarding the purpose of the function is given, and the user is instructed how to operate it. More detailed information regarding the steps to take is given only when needed. For instance, if the user already has experience with selecting images, as part of another task, then no information is given anymore on how to (de)select an image (compare images 2 and 5). The user can get this next level of information if he wants it after all, by explicitly asking for it with the help function (available from the button marked "?"). We envisage that users will also be able to explicitly select the level of information they like for certain functions. This enables experienced users to still use more structured interfaces if they prefer that.

Conclusions

Our main interest is in exploring how new user-interface technologies can be applied to the design of medical systems, so as to reduce sub-optimal system usage in the X-ray examination and control rooms. This has a direct impact on patient throughput, which is expected to continue to be an important measure in the healthcare industry. We have described how we have applied user-centred design methods basing our work on first-hand observations conducted in hospitals, and on Task Analysis and Modelling.

Various problems have been observed during field studies. In this paper, we have suggested a number of ways these could be addressed:

(1) Improved user-interface design based on task analysis and customisation can make it easier for people with different backgrounds to operate the same system. The increasing amount of functionality can be shielded. By basing the design on the task model, it can be prevented that users have to run from one place to the other in order to operate the system. By relating functionality and interaction styles to tasks, consistency over different systems can be better achieved. Customisation is also an option to allow the operation of systems to be customised to what the user is accustomed to.

(2) customised default values minimising the amount of effort required from the user. The customisation uses a combination of pre-programmed rules, and on-line system learning. A prototype system has been implemented. We have thought about the interaction styles needed for customisation, but evaluations have to be conducted to test our hypotheses.

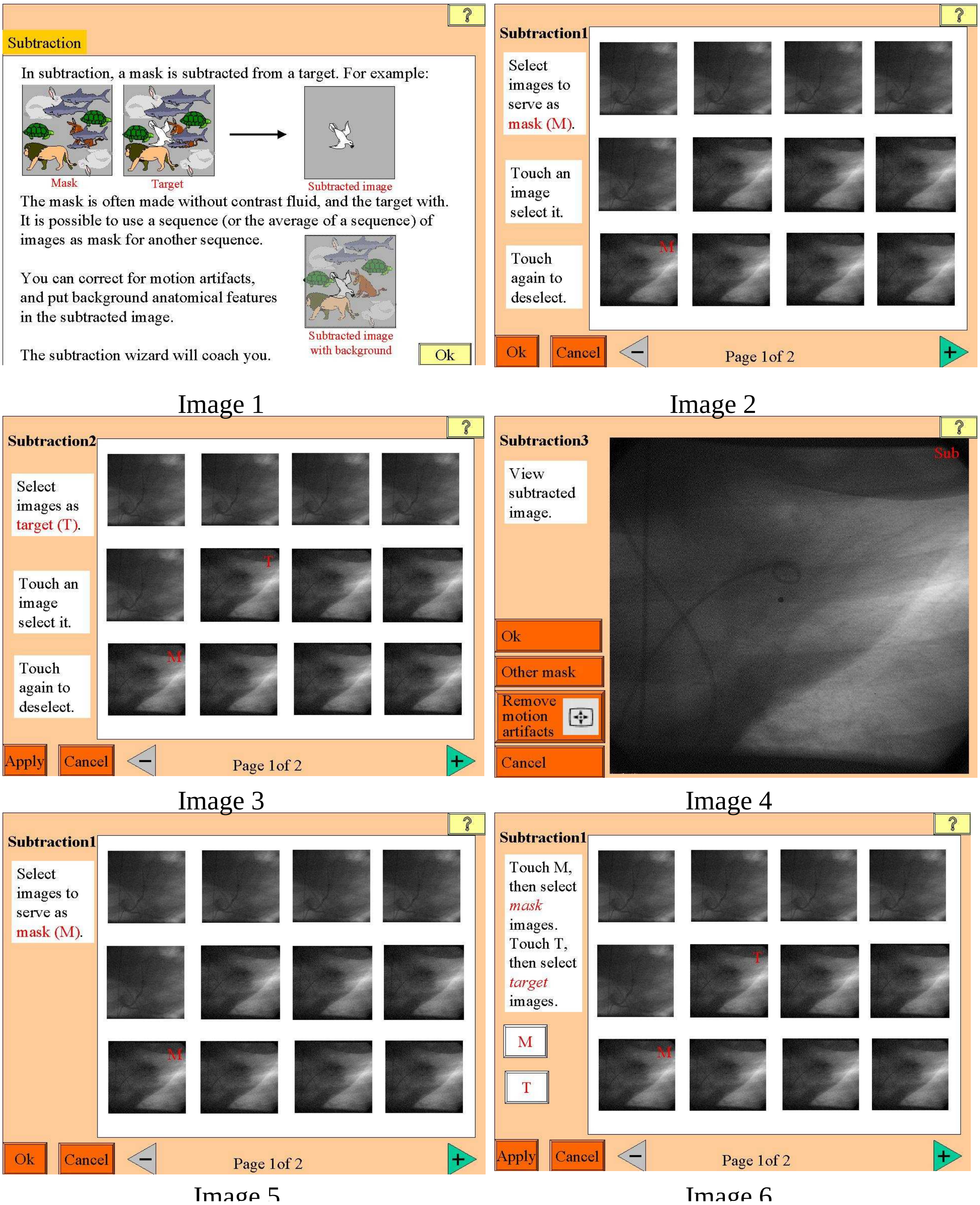


Figure 11 - Example of Active Support for the Learning Process

(3) active support by the system for the learning process (needed as training takes place on the job). We have sketched a way in which the learning process of the user is supported by gradually changing the interface to reflect the user's growing experience.

Our work is ongoing; we are presently in the midst of building an integrated research prototype, which will be tested in usability evaluations.

Acknowledgements

We would like to thank the application specialists who accompanied us in the hospital visits: Erik Vossen, Theo Loohuis, Jacques Lukassen, and Peter Sutton. Thanks are due to the following designers: Horland Hudson for organising the workshop, Job Rutgers and Jude Rattle for their work on the scenarios, and Sue Coles for her work on interaction styles for customisation. Boris de Ruyter was involved in some discussions on the notation for the task model. Paul Gough gave comments on an earlier version of this paper. This project was supported in PMS by Eric von Reth and Franklin Schuling.

References

1. Bés, M-O., Analysis of a human error in a dynamic environment: The case of air traffic control. Proceedings of the workshop on human error and systems development, Glasgow, 1997.

2. Reason, J. Human error. Cambridge: Cambridge University Press, 1990.

3. Sackett, D.L., Rosenberg, W., Gray J.A., Haynes, R.B., & Richardson, W.S. Evidence-based medicine: what it is and what it isn't. Centre for Evidence-Based Medicine, http://cebm.jr2.ox.ac.uk/ebmisisnt.html, 1996.

4. Lim, K.Y, & Long. J. The MUSE method for usability engineering. Cambridge series on human-computer interaction, 8, Cambridge: Cambridge University Press, 1994.

5. Masthoff, J. An agent based interactive instruction system. PHD thesis. Eindhoven University of Technology, 1997.

6. Rouse, W.B., Geddes, N.D., & Curry, R.E. An architecture for intelligent interfaces: outline of an approach to supporting operators of complex systems. Human-Computer Interaction, 3, 87-122, 1987.

7. Totterdell, P., & Rautenbach, P. Adaptation as a problem of design. In D. Browne, P. Totterdell, & M. Norman (Eds.) Adaptive user interfaces. London: Academic press, 1990.

8. Dietrich, H., Malinowski, U., Kuhme, T., & Schneider-Hufschmidt, M. State of the art in adaptive user interfaces. In M. Schneider-Hufschmidt, T. Kuhme, & U. Malinowski (Eds.), Adaptive user interfaces.

   Amsterdam, 1993.

9. Dang, W. (1988). Intelligence in a user interface management system. Proceedings of ERGO-IA 1988. CNRS, Université de Paris-Sud, Orsay.

10. Sherman, E.H. A user-adaptable interface to predict user's needs. In M. Schneider-Hufschmidt, T. Kuhme, & U. Malinowski (Eds.), Adaptive user interfaces. Amsterdam, 1993.

11. Browne, D. Conclusions. In D. Browne, P. Totterdell, & M. Norman (Eds.) Adaptive user interfaces. London: Academic press, 1990.

12. Norcio, A.F., & Stanley, J. Adaptive human-computer interfaces. Naval Research Laboratory Report 9148, 1988.

13. Mallen, C. Designing intelligent help for information processing systems. International Journal of Human-Computer Studies, 45, 349-377, 1996.

Biography

Ashok Gupta, Philips Research Laboratories, Cross Oak Lane, Redhill, RH15HA, UK. Telephone: +44 (1293) 815647, Fax: +44 (1293) 815500. Email: gupta@prl.research.philips.com.

A. Gupta is a Principal Scientist at Philips Research Laboratories where he has been working since 1988. His primary research interests are in user-centred design for professional applications, user-interface process assessment, requirements specifications, and software process improvement. In addition to medical, he has worked in the consumer electronics, and semiconductor domains.

Judith Masthoff, Philips Research Laboratories, Cross Oak Lane, Redhill, RH15HA, UK. Telephone: +44 (1293) 815383, Fax: +44 (1293) 815500.
Email: masthoff@prl.research.philips.com.

J. Masthoff has experience in artificial intelligence, situated cognition, interactive instruction, user-centred design and experimental evaluation.

Both A. Gupta and J. Masthoff are working in the Software Engineering and Applications Group, focusing on research into mobile and personal user-interfaces.

Paul Zwart, Philips Medical Systems Nederland B.V., Veenpluis 4-6, 5680 DA Best. Tel: ++31 (0)40-2763024, Fax: ++31 (0)40-2765657. E-mail: P.Zwart@best.ms.philips.com

P.Zwart has been with Philips Medical Systems since 1985. His interests are in image compression, medical image processing and user interfaces to medical systems.

Appendix: Checklist for hospital visits

1. Background

Obtain or develop a global organisation model.

2. Workflow in X-ray examination

- Who are the people carrying out tasks? (Position, level of experience)
- What objects are used when performing the tasks?
- How are different systems used together?
- What are the actions - what is actually happening?
- Which information is needed at each stage of any process?
- How long does each stage of the process take? (Both own observation and actors opinion)

3. Examination environment

- Make a plan of the examination room and control room.
- Where are the domain objects located?
- Where are the actors located at each stage of the process?
- In what direction are the actors looking at each stage of the process?
- Where is the patient at each stage of the process?
- What can the patient perceive at each stage of the process?
- What kind of background noise is there (causes, level)?

4. Information

- How do the actors have access to relevant information (in particular about the patient, but also about the process)?
- Is sound used to convey information? If yes, for what kind of information?
- How is patient or other information exchanged between different systems (like RIS, HIS, etc)?
- What information is recorded during each stage of the process and how is it recorded? (Particularly interesting would be information on the procedure used and parameter settings.)
- Try to get samples of forms and annotate the meaning of terms on the forms!
- Are computer logs of the various systems available (in particular reflecting the use of functionality)? (Question to ask the application specialist)
- Do such logs also have an actor code on them? Or is there another way to find out who the actor was?
- What kind of information can we get from the help desk/service desk? (Question to ask the application specialist)

5. X-ray acquisition procedures

- For which X-ray acquisition procedures is the equipment used? (Both now, in general in that particular room, and in the hospital.)
- Do all the actors perform all the procedures, or do actors specialise in certain procedures?
- How do actors use the possibility to adjust parameters and sequences manually?
- Are there fixed scenarios (parameter settings) in this hospital for specific circumstances, meaning that default values are changed on a formal basis?
- How much freedom do actors have (within the hospital policy) in adjusting parameters to their own taste or the taste of the image reviewer?
- Do the actors believe there are patterns in the changes of parameters, and what kind of

patterns are they thinking of?

6. The way the system can be operated

- What are good features of the way the systems can be operated?
- What would be good things to change in a future product (future meaning in a couple of years)
  - with respect to the way systems can be operated?
  - with respect to the way information is provided by the system at each stage of the process?
- How easy was it to learn to use the systems? What can still be improved?
- How easy is it to use the systems? What can still be improved?
- How efficient is it to use the systems? What can still be improved?
- Are there functions/buttons on the systems that the actors never use? Which? Why don't they use them?
- Would the actors prefer to obtain some information/messages from the systems in audio which are currently on a display? Which messages? Why?
- Would the actors prefer to enter some commands in speech instead of by pressing buttons? Which commands? Why?
- How do the actors think about customisation? Can they see a use for it? What would they like to see customised?
- General comments?

7. Patients

- How do patients experience the various X-ray procedures?
- What information do the actors give them before or during the process?
- What information would the actors like the patient to have, but have difficulty providing?
- What information would the patient like to have but is not getting?
- How regular does it happen that the same patient is examined again (e.g., after a year)?
- Would the parameter settings of the previous X-ray acquisition be useful in such a case?

8. Introduction of new equipment

- What kind of process is used for introducing new equipment in the hospital?
- Who are the main persons involved in an early stage?
- Who are the first people using it?
- Who are influential in a hospital in making new equipment successful?
- How long does it take before new equipment is in full operation?
- What would be good things to change in the process of introducing a future product in the hospital?
- What is the role of Service in the introduction process?
- How does customisation by Service of pre-set system parameters take place?

9. Training

- What kind of training do the radiographers/ radiologists receive?
- How long does it take?
- What can be improved in the training process?

10. General

- What are the trends in this hospital ?(e.g., with respect to digitalisation and use of radiology by other departments)

- What is the opinion of the radiologists/ radiographers on these trends? What potential problems do they see? What opportunities?

- Look for any opportunity for introducing existing/new UI techniques, plus opportunities for customisation and multi-modal interaction!

11. List systems in the hospital (full names and versions) and the software running on these.

(When talking to an operator, always ask what system is referred to.)

Additional questions related to Image viewing and post-processing

During which parts of the examination are images viewed?
When do they use the monitors in the Examination room, when do they use the monitors in the Control room?

What image processing and viewing functionality is used, why, where, when, and by whom? Ask about:

- Zooming (on center of image?)
- Changing image contrast
- Changing image brightness
- Changing polarity
- Masking image edges
- Changing edge enhancement
- Subtraction
- Correction for motion artefacts (pixelshift)
- Adding background anatomical features
- Quantative analysis
- Annotation of images: left/right, examination data, observations

Are images stored during the viewing process?

Hardcopies versus digital images

- Are hardcopies still used in this digital hospital? If yes, for what?

- What are the main advantages of having digital images compared to hard copies with respect to possibilities for viewing and post-processing?

- What are the main disadvantages of having digital images compared to hard copies?

Ideas for improvement

Mention/request any other ideas.

Problems caused by human error in the operation of a clinical data collection system

Barbara McManus,

Dept of Computing and Dept of Health Informatics, University of Central Lancashire, Preston

Background

Within the UK, the NHS has been rather slow in its use of technology. However this attitude has had to change in line with the UK government's publication of IM&T strategies. The first IM&T Strategy was produced in 1992 but it was not until the later publication of the new IM&T strategy that clinician orientation rather than the management needs came to the forefront. The system described in this paper, although considered by the designers to be clinician-oriented, when considered from a usability perspective, is certainly not the case.

The software under consideration was written for a NHS Trust in England. It is a simple data collection system and as such should be very straightforward to use. This is not the case. The designers of the systems have taken no account of the users, the tasks they need to perform or their ability (both physically and technically). As a consequence of these oversights, a simple system has become a very unusable piece of software, causing the user to make far more errors than would have been expected. This paper aims to examine the system and analyse the reasons for these problems.

There are a multitude of reasons for the limited success of the data capture system, but many of them can be linked back to the designers' failure to consider the users as human beings with all the inherent faults which human beings have.

Human error

There appears to have been no consideration made of human beings' potential for making mistakes. Although errors cannot be eliminated completely, the designer should aim to minimise those errors.

On-line help

Users are not expected to make mistakes since there is no on-line help available at all in this system. This obviously indicates that the designers considered the software to be so easy to use that no on-line help was needed.

Error messages

Error messages are at best unhelpful and at worst confusing. Many of the error messages read like a programmer's test statement as seen in figure 1. The users should not need to know that fld is an abbreviation for field, nor should they need to know what a field is. According to Schneiderman' s Error message guidelines (Schneiderman, 1998), the designer should choose user-centred phrasing as well as being constructive, indicating what the user needs to do. What seems obvious to an IT literate individual is not so obvious to a panic-stricken novice. Error messages do not conform to a standard either in the format of the message or in the action taken by the program on acknowledgement of the message having been read. At one point in the system, on acknowledging the fact that a field has been omitted, the program takes the user to the screen containing that field in order that the user can correct the error. However this is not always the case. This leaves the user confused by the inconsistency within the system.

Error correction

Error correction is made very difficult with this system. Common Windows standards are not adhered to, that is, double tapping on a field highlights the field, but on numerous occasions does not allow the field to be altered by overtyping. Standard keyboard editing techniques, for example CTRL X to cut, do not work in all fields. A date field, accessed when the rest of the form has been completed, reacts differently when the rest of the form is empty. Anomalies such as these cause even IT literate users a great deal of confusion.

One of the greatest difficulties faced by the user is when a new patient has to be entered. Any patient, whose details are not on that clinicians hand held computer, has to be entered onto the machine. This entails working through four linked screens with no way of backtracking when a mistake is made. This is made more hazardous by the nature of the clinician's job; they are frequently disturbed by phone calls, appointments etc, all of which interfere with their train of thought. Dropping out of the screens before completion loses all the data previously entered, giving an error message and creating an 'orphan' record that the user has no rights to delete. The last of the four screens contains a date that references back to a date on the previous screen. This fails on Don Norman's suggestions on Designing for Error (Norman, 1988). He states that the designers should 'make it possible to reverse actions or make it harder to do what cannot be reversed' and 'make it easier to discover the errors that do occur and make them easier to correct'

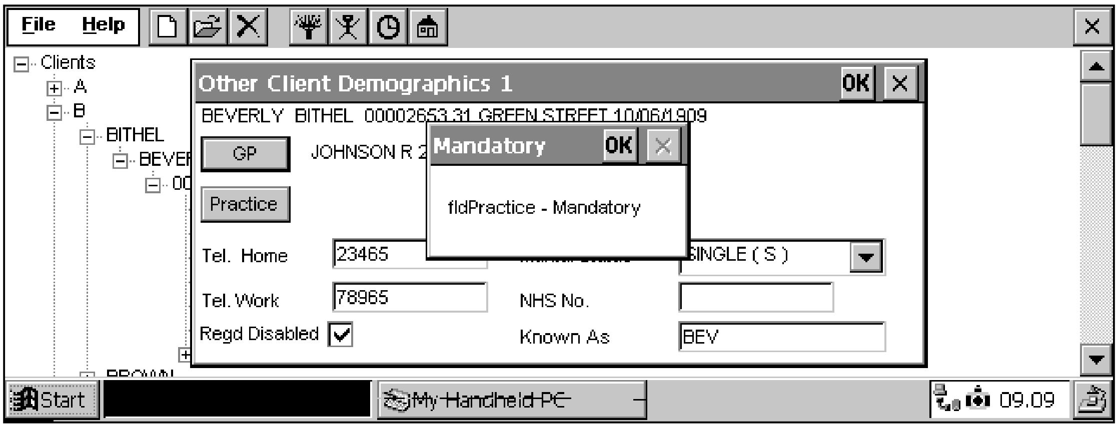


Figure 1.

Human limitations

There is little evidence that allowance has been made for human limitations since if a user profile had been performed, it would have been obvious that the software would face major problems.

Mixed Ability users

There is a very mixed ability level of users ranging from complete novices (many of whom have not even used a typewriter) through to highly IT literate users. Many of these IT literate users have gained their literacy skills through the use of computer systems within the Trust with a large proportion of these systems being DOS based. As a consequence of this, these users expect the RETURN key to be used for navigating around the form on screen and a function key to be used to save the form away.
Since this system was Windows CE based the TAB key could be used for navigation and the RETURN was the equivalent of tapping OK on the screen, enabling the form to be saved away. The novice users did not use the keyboard, but struggled with the use of the stylus on the small touch sensitive screen.

Age Range of Users

The users cover a wide age range from those in their twenties to late fifties, with a large proportion over forty. It is a well-known fact that eyesight starts to fail after forty years of age so the use of such a tiny screen and small font size with no enlargement facility, caused problems to many users. However, it was the combination of the small font together with the stylus-based touch screen that caused the greatest difficulty to the older users. The small font size meant that there was a limited screen area around which the user could tap to make a selection. Added to this problem was that of the user ensuring that they did not touch elsewhere on the screen or inadvertently rest their hand on the keyboard before touching the screen with the stylus.

Had the designers performed a full user profile perhaps they would not have chosen the hand held computer or would have thought more carefully about the screen design and input method. As Bergman and Johnson point out 'All users have a range of capabilities that vary across many dimensions depending on the user and his or her life stage, task and environment. As more of the population approach their middle forties, there are increasing numbers of users peering through their bifocals at the screens.' (Bergman E, Johnson E, 1995) Had the designers observed trial users interacting with the system, and recorded the errors made and inefficiencies exhibited (Fields et al, 1995), they may have realised the users' difficulties and changed the system design accordingly. These should have led the designers to have considered the Microsoft Accessibility guidelines for disabled users, offering the users the ability to consistently use the keyboard as an input device, a function to enlarge the font size and ensuring that commonly used Windows standards were adhered to.

User Tasks

Perhaps one of the more fundamental oversights is the failure of the designers to consider the users' main task. It was written to replace a paper-based system whereby clinicians wrote down details of face-to-face contacts with their patients. The computerised version can be seen in figure 2.

The designers obviously thought that the fact that this system was 'tap to select' would make it easy to use. However considering each of the fields on this form in turn, it will become obvious that vast improvements could have been made to this system for very little extra programming work.

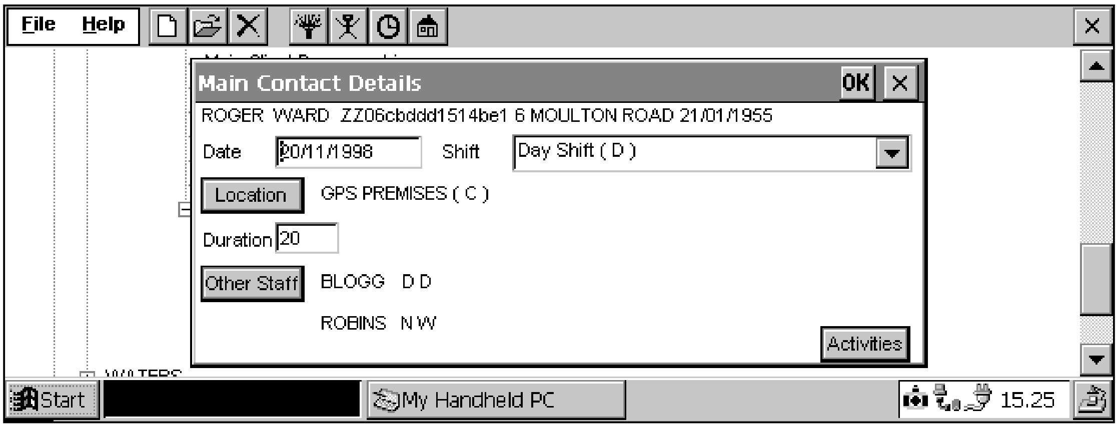


Figure 2

Date field

To save the user time, the date field is set with that day's date as its default. But, the technique for selection of that field to alter the date neither adheres to Windows standards nor reacts consistently in the same way at all times.

Shift field

The majority of the clinicians work only on one shift, so rather than forcing the user to choose one from the three in the list, a default shift tied to that user would save the clinician time.

Location field

The user is offered approximately three hundred locations to choose from. Had the designers talked to the clinicians they would have found that the majority of clinicians use only five locations. A far better design would have been one offering, at the start of the list, the last five locations used by that user.

Other Staff

This field, like the location field, offers hundreds of names to choose from yet the majority of the clinicians only perform joint visits with the same clinicians. A facility with, at the start of the staff list, the last five staff chosen by that user would help speed up the selection process.

Activities

At present, the user is only offered four activities to choose from so there is little chance of them making mistakes in choosing the activities. However when the user saves these activities away, their choices are not echoed onto the Main Contact Details screen. Since all other choices made in the same way are fed back to the user, there is an expectation that this will also feedback. When it does not, the user thinks they have made a mistake and wastes time tapping the activities button again.

As can be seen by the problems above, the user wastes time and causes more errors than need be the case. The more searching and entering the user must do the higher likelihood there is of them making mistakes.

Not only were the tasks themselves not considered, neither was the time taken to complete those tasks. It took the user nine taps on the screen to reach the form that performed the most frequently used task (entering in face to face contacts). So although it is more often better to select than to type, this system is an example where this is not the case. The metaphor chosen was that of a directory structure, which forces the user to tap their way through the levels before reaching the point at which they can perform their most frequently used task. Therefore the design has not been considered with respect to the time taken to complete the most commonly performed task.

Discussion

One could question that if mistakes such as these mentioned can be made on a small system, it does not augur well for larger more complex systems. Although it could be argued that errors made in this system are not catastrophic, the effect of repeated errors caused the user's confidence to reduce and their stress levels to rise, causing them to make even more mistakes.

Systems similar to the data collection project discussed above are becoming more common in the NHS with clinicians entering and maintaining their own patients' records. Both designers of these systems and managers considering the use of systems such as this have much to learn about designing to minimise errors. Human-Computer Interaction has an important place in these systems.

References

Department of Health (1992) Getting Better with Information: an IM&T strategy for the NHS in England, HMSO, London

NHS IM&T (Information Management and Technology) strategy, 1998

Bergman, E., Johnson, E. 'Towards accessible

Human-Computer Interaction.' In Jakob Nielsen (Editor), Advances in Human-Computer Interaction, Volume 5, Ablex Publishing Corporation, New Jersey. Also at www.sun.com/tech/access/updt.HCI.advance.html

Fields, B., Wright, P. and Harrison M. Applying Formal Methods for Human Error Tolerant Schneiderman, B. Designing the User Interface; Strategies for effective Human-Computer Interaction, Third Edition, Addison-Wesley.

Design.' In Taylor and Coutaz (1995), pp185-195.

Microsoft Accessibility Guidelines, at www.microsoft.com/enable/dev/guidelines.htm

Norman, D.A. (1988), The Psychology of Everyday Things, Basic Books.

Safety and Computer Aided Design of Chemotherapy Plans

P. Hammond, S. Modgil,
Dept of Informatics, Eastman Dental Institute for Oral Health Care Sciences,

J.C. Wyatt,
Health Knowledge Management Centre, School of Public Policy,

University College London, U.K.

Keywords: chemotherapy, computer-aided design, planning, protocol, safety reasoning.

Abstract

Previously, generic principles were derived empirically from examples of reasoning about the efficacy and safety of chemotherapy. Some of the principles were used in a prototype decision support system for managing patients through chemotherapy. It reminded clinicians which, how and when drugs were to be given, suggested treatment modifications as a result of detected adverse events, and warned how unplanned actions undermine efficacy or exacerbate hazardous side-effects. Here, we investigate the use of the safety principles to generate symbolic and textual representations of new chemotherapy plans. In clinical trials of chemotherapy, a written protocol is essential, especially in obtaining ethical approval. The symbolic representation, from which textual extracts are derived, can be used in conjunction with software for managing chemotherapy.

Introduction

Oncologists, clinical specialists in cancer treatment, employ chemotherapeutic drugs in quite complex combinations often over considerable time periods. Their side effects can be severe, permanent and even life threatening, so it is essential to monitor for adverse reactions as well as review their intended effect on a malignancy. Hence, chemotherapy is a fine balance between maximising efficacy and minimising risks and hazardous side effects. The monitoring of patients during chemotherapy can also be complex, and once more takes place over extended periods. It is likely, therefore, that the use of computerised reminders of treatments, patient monitoring and associated hazards will become more widespread and offer potential for improving data capture, statistical analysis of treatment results and patient safety.

Human error does occur in the management of cancer patients, often with disastrous results, for example if a drug is administered by an incorrect route (ref. 1). There is also concern that some patients are being treated by non-specialists (ref. 2) or are not receiving optimal treatment for which there may even be substantial published evidence (ref. 3).

We focus mainly on in-vivo adverse effects, but it is important to make a thorough analysis of all safety and efficacy issues since there are numerous ways in which hazardous situations can arise in chemotherapy treatment, including the following:

- exceeding maximum values for single or cumulative dosages, for treatment cycles or even over an individual’s lifetime

- achieving correct concentrations in administering solution or in blood plasma

- the administering route affects efficacy and toxicity, and because of localised effects a site may need to vary

- rates of administration can be hazardous

- drug interactions occur, often subtley

- drugs interact with equipment used for storage or administration

- the preparation of drugs and their accidental spillage can be hazardous.

Following the inspection of over 100 chemotherapy treatment protocols, nine generic safety principles were previously extracted (ref. 4). Table 1 contains a summary of the principles and a single illustrative example for each. In all, more than 200 examples of such reasoning have been collected. The *Exacerbation* and *Diminution* principles described were included in OaSiS, prototype chemotherapy management software (ref. 5), to inhibit additions to treatment that might exacerbate a known hazard or undermine efficacy. The latter is considered a hazard given the potentially life threatening nature of malignancy. The two principles were used in OaSiS as integrity constraints on updates to a database representing the state of the treatment, the patient monitoring data and relevant knowledge of chemotherapeutic drugs and their side-effects. In addition, the *Reaction* principle was used, not as an integrity constraint, but as a filter so that drug doses, for example, could be modified when reported monitoring data heralded adverse reactions.

Chemotherapy Plan Design

When devising new chemotherapy plans, oncologists make extensive use of publications such as 'The Cancer Chemotherapy Handbook' (ref. 6) or specialist sections of the 'British National Formulary'. Rather than capture such efficacy and safety information as many ad hoc examples, we believe it is advantageous to capture them as generic safety principles.

In the remainder of this section, we demonstrate how the safety principles can contribute to the description of a chemotherapy plan involving the drugs *cisplatin* and *taxol*, both regularly used in cancer treatments. In practice, this will require a graphical or form-based interface more user friendly than the linear presentation given here. We show how information supplied by the plan designer and that present in an appropriate oncology knowledge base might interact with the safety principles to produce simple facts characterising the design of the chemotherapy plan, at what we refer to as "build-time". Facts defining a particular plan (always shown below in italics) are also then available for "run-time" use to configure software for the management of chemotherapy. In order to simplify the presentation, our Prolog implementation may be more complex than the descriptions suggest[1].

<u>Use of the Efficacy principle:</u> The design of a plan might begin with the selection of drugs with a desired effect on the target malignancy. Details of drug efficacy in the supporting oncology knowledge base could be expressed in terms of a general predicate

efficacy(Action,Focus, Efficacy)

with instances

```
efficacy(admin(cisplatin,_,_),
      lesion(ovarian,carcinoma), fda_approval).

efficacy(admin(cisplatin_,_),

      lesion(brain,cancer), may_be_useful).
efficacy(admin(taxol,_,_),

      lesion(ovarian,carcinoma), fda_approval).

efficacy(admin(ifosfamide,[daily,5,days],_),
      lesion(ovarian,carcinoma),may_be_useful).

efficacy(admin(ifosfamide,single_dose,_),
      lesion(ovarian,carcinoma),not_so_useful).
```

There are a number of interpretations of the EFFICACY principle *"Ensure that overall plans are efficacious in pursuing stated objectives"*. For example, at build-time, a plan designer would want to be informed that a suggested effect of a drug is unknown or might have previously been refuted. Of course, in some situations, a special class of users may seek, and be granted, permission to override such a veto. Alternatively, if there is a more effective way of administering a drug for a particular objective, then once again the designer should be informed.

---

[1] In Prolog clauses, variables are denoted by strings with an initial upper case letter. Predicates in italics are used to indicate either a generic form or in some cases simple semantics. The clause *a :- b, c.* should be read as *conclude a if b and c*; the condition free clause, or fact, *a.* should be read as *assume a*.

Table 1- Generic Safety Principles with Illustrative Examples

SAFETY PRINCIPLE		EXAMPLE
Exacerbation	Avoid exacerbating anticipated hazards	Nephrotoxic antibiotics such as gentamicin should be avoided during & immediately after cisplatin infusions
Diminution	Avoid undermining the benefits of essential actions	Aspirin reduces the efficacy of Interferon
Reaction	React appropriately to ameliorate detected hazards	Administer diuretics and digoxin and withdraw doxorubicin if signs of cardiac failure are noticed
Warning	Warn about hazards due to incorrect execution of essential actions	Bleomicin is not vesicant [blistering] but direct contact with the skin should be avoided
Monitoring	Monitor responses which herald hazardous situations	Measurement of methotrexate levels is essential ... at 24 and 48 hours (after its administration)
Efficacy	Ensure that overall plans are efficacious in pursuing stated objectives	Tumour necrosis factor is not a useful antitumour agent if administered IV or IM
Sequencing	Order (essential) actions temporally for good effect and least harm	Taxol is given before cisplatin. The reverse order can produce severe neutropenia
Critiquing	Critique proposal of certain hazardous actions even if they are well motivated	Doses of etoposide should not be reduced for elevated serum bilirubin concentrations
Prevention	Prevent or ameliorate hazards before executing an essential action	Folinic acid rescue helps ameliorate methotrexate-induced bone marrow suppression

We might formulate the generic efficacy principle as the rules (1-3) below. The first two reject a proposed action for inclusion in a plan; the respective reasons being absence of acceptable evidence of efficacy, and evidence suggesting poor or inferior efficacy:

(1)

```
invalid(Action,Plan,Focus,unsubstantiated_effect) :-
    not (efficacy_value(Efficacy),
         efficacy(Action, Focus, Efficacy)).
```

(2)

```
invalid(Action,Plan,Focus,insufficient_effect) :-
     efficacy(Action, Focus, Efficacy),
     insufficient(Efficacy).
```

In rules 1 and 2, *invalid(Action, Plan, Focus, Reason)* can be read as "*Action* is ill advised in *Plan* with aim *Focus* because of *Reason*" [2].

Given specific facts describing what is more efficacious in chemotherapy, the third rule modifies the plan on the basis of evidence suggesting a more effective alternative:

---

[2] Throughout the paper, all rule-based versions of the safety rules are expressed in their most general form, with no reference to their roots in chemotherapy treatment plans. This has always been an important feature of our approach and has already proven potentially useful in other domains, for example HSE regulations for handling hazardous chemicals (ref. 7).

(3)

```
advise_plan(Action1,Plan,Focus,Action2,Reason) :-
    efficacy(Action1,  Focus, Efficacy1),
    efficacy(Action2,  Focus, Efficacy2),
    more_efficacious(Efficacy2, Efficacy1),
    Reason = better(Action1,Focus,Action2).
```

where we might interpret *advise_plan(Action1, Plan, Focus, Action2, Reason)* as "In *Plan* with objective *Focus* action *Action1* should be replaced or supplemented by *Action2* for *Reason*". For example, a combination of rules (1-3) and the efficacy clauses will allow the nomination of *cisplatin* to treat ovarian cancer. But, they will question the use of *ifosfamide* if a single dose is proposed (because of its inferior efficacy) and suggest a more potent regime. Clearly, for this principle to be applied here it is necessary to know the intended effect (focus or objective) as well as the mode of drug administration.

Knowledge of typically used dosages can be used to check the acceptability of, or even guide, a user's suggestions. For example, with the addition of the following instances in the knowledge base

```
typical_dosage(cisplatin, [20-40,mg/m2],
            [daily,3-5,day], [cycle, 3-4, week]).
typical_dosage(cisplatin, [20-120,mg/m2],
            single_dose, [cycle, 3-4, week]).
```

of the predicate *typical_dosage(Drug, Dose, Frequency, Cycle)*, the following facts could be generated, with some help from the user, for inclusion in the representation of the chemotherapy plan:

focus(plan_1256, lesion(ovarian,cancer)).

part_of_plan(admin(cisplatin,single_dose, slow_iv_infusion), plan_1256).

dosage(cisplatin,plan_1256,[75,mg,m2], single_dose,[cycle,3,week]).

As stated earlier, the facts depicted in italics will form part of the symbolic description of the chemotherapy plan and can be used to configure a computerised system for the management of patients. They can also be translated to a textual form for inclusion in a protocol document associated with the chemotherapy plan.

<u>Use of the Prevention, Monitoring and Sequence principles</u>: Anticipated hazards such as *dehydration* and *nephrotoxicity*, and associated monitoring procedures, arising from use of the chemotherapeutic drug *cisplatin* might be described in clauses as:

(4)

```
anticipated_effect(admin(cisplatin,_,_),dehydration,[]).
```

(5)

```
anticipated_effect(admin(cisplatin,_,_),nephrotoxicity,
        [monitor(creatinine_clearance)]).
```

These two clauses are instances of a predicate *anticipated_effect(Action, Effect, Monitoring)*. Their hazardous nature might be recorded by two facts:

```
hazardous(nephrotoxicity).
```

(6)

```
hazardous(dehydration)
```

We represent the combination of *prehydration* and *posthydration* to prevent *cisplatin induced dehydration* in clause 7

(7)

```
prevention(admin(cisplatin,_,_),dehydration,
        [prehydration(P),posthydration(P)]) :-
        default_hydration(P).

default_hydration( [(normal_saline,1-2,litre),
                (kcl,20,mEq,per_litre),
                (mgso4,8,mEq,per_litre)]).
```

Clause 7 is an instance of the more general form *prevention(Action,Effect,Prevention)*.

In the previous section, we demonstrated how the efficacy principle can be used during the design of a plan to reject inappropriate actions (rules 1 and 2). We also described how changes to a plan might be advised (rule 3). One can consider the latter as a specific instance of a general scheme in which a user's suggested addition to a plan is assessed, the plan is modified and a justification for the changes is given. We show how this scheme might be used for animating the prevention, monitoring and sequencing principles.

Prevention: The **PREVENTION** principle "***Prevent or ameliorate hazards before executing an essential action***" might be expressed in rule form as:

(8)

```
advise_plan( Action, Plan, Focus, Prevention,
        ameliorate(Action, Effect) ) :-
    anticipated_effect(Action,  Effect,_),
    hazardous(Effect) ,
    prevention(Action, Effect, Prevention).
```

From clauses 4, 6, 7 and 8 and the query

```
?- advise_plan(admin(cisplatin,_,_), plan_1256,
 lesion(ovarian,carcinoma),PlanChange,Reason).
```

one can derive:

PlanChange = [*prehydration(P), posthydration(P)*]

where

P=[(normal_saline,1-2,litre), (kcl,20,mEq,per_litre),
(mgso4,8,mEq,per_litre)].

and Reason = *ameliorate(admin(cisplatin,_,_),*
dehydration).

This solution to the query would then be automatically translated to facts forming part of the description of the plan (which is why its components are italicised). Of course, in a fully operational system, the user will not enter Prolog queries, but interact with a more friendly interface where such complexities are hidden.

Monitoring: The **MONITORING** principle "***Monitor responses which herald hazardous situations***" can be represented in clausal form:

(9)

```
advise_plan( Action, Plan, Focus, Monitoring,
        monitor(Action, Effect) ) :-
    anticipated_effect(Action,  Effect, Monitoring),
    hazardous(Effect).
```

So the same query that was used for the animation of the prevention principle above will generate an additional solution:

PlanChange = *[monitor(creatinine_clearance)]*
Reason = *monitor(admin(cisplatin,_,_),*
nephrotoxicity)

Once again, this solution would be used to generate components of the description of the plan.

Should the user attempt to add the drug *taxol* to the plan for a combination therapy, then as before efficacy and dosage components of the knowledge base can help generate facts such as

part_of_plan(admin(taxol,single_dose,slow_iv_infusion), plan_1256).

dosage(taxol,plan_1256,[135,mg,m2],single_dose,
[cycle,3,week]).

If the knowledge base also includes details of the hazardous effect of the drug taxol on blood cell counts:

anticipated_effect(admin(taxol,_,_),neutropenia,
[monitor(cell_count)]).

hazardous(neutropenia).

then the *monitoring* principle (clause 9) and the query

?-advise_plan(admin(taxol,_,_),plan_1256, lesion(ovarian,carcinoma),PlanChange,Reason).

yield:

PlanChange = *[monitor(cell_count)]*

Reason = *monitor(admin(taxol,_,_),neutropenia))*

An appropriate formulation of this query solution will be also added to the description of the chemotherapy plan.

Sequence: The **SEQUENCE** principle "***Order (essential) actions temporally for good effect and least harm***" can be represented as the rule:

(10)

```
advise_plan(Seq,Plan,Focus,PlanChange,Reason) :-
   combined_effect(Seq,Effect,PlanChange),
   hazardous(Effect),
   Reason = warning( Seq,Effect).
```

The sequencing restriction on the order of administering *taxol* and *cisplatin* (see sequencing example in Table 1) is captured as:

(11)

```
combined_effect(
      sequence([cisplatin,taxol]), neutropenia,
      [sequence([taxol,cisplatin])]
              )
```

The suggestion of using cisplatin followed by taxol, given clauses 10 and 11:

?- advise_plan(admin(sequence([cisplatin,taxol]),

_,_),plan_1256,lesion(ovarian,carcinoma),

PlanChange,Reason).

produces the response

PlanChange = *sequence([taxol,cisplatin]*

Reason = *warning(sequence([cisplatin,taxol]),*

neutropenia).

From this solution two new facts will be generated. The first describes the inclusion of the correct sequence for using the two drugs. The second adds a warning about the hazardous nature of the incorrect ordering.

Generation of Protocol Text Extracts

We now focus on the *exacerbation* principle and discuss both its run-time and build-time use. One might expect its typical use to be at run time, when for example a veto is placed on a suggested treatment of an adventitious medical condition by the managing clinician (see *exacerbation* example in table 1). At build time, it is important to anticipate such run time application. In this way, the protocol document describing the designed plan can demonstrate awareness of possible hazards arising during patient management, as well as appropriate responses to them. This not only enhances safe patient management, but is important in gaining ethical approval for a clinical trial of a new chemotherapy treatment.

We can represent the **EXACERBATION** principle in an expanded textual form ***An action is ill advised in a plan if there is another action in the plan with a hazardous effect that the proposed action exacerbates***.

Corresponding to this might be the rule:

(12)

```
invalid(Action1,Plan,Focus,exac(Action1,Action2,Effect
)) :-
  part_of_plan(Action2, Plan),
  anticipated_effect(Action2,Effect, _),
  hazardous(Effect),
  exacerbation(Action1, Action2, Effect).
```

Suppose that the oncology knowledge base includes clauses describing the exacerbation of cisplatin induced nephrotoxicity by antibiotics that are themselves nephrotoxic (clauses 13-15 below) and in particular by the antibiotic *gentamicin* (clauses 16 and 17 below):

(13)

```
exacerbation(admin(Drug,_,_),admin(cisplatin,_,_),
                    nephrotoxicity) :-
  antibiotic(Drug),
  anticipated_effect(admin(Drug,_,_),nephrotoxicity,_).
```

(14)

```
anticipated_effect(admin(cisplatin,_,_),
        nephrotoxicity,
        [monitor(creatinine_clearance)]).
```

(15)

```
hazardous(nephrotoxicity).
```

(16)

```
anticipated_effect(admin(gentamicin,_,_),
        nephrotoxicity,
        [monitor(creatinine_clearance)]).
```

(17)

```
antibiotic(gentamicin).
```

We need to suppose also that the user has declared the inclusion of cisplatin in the treatment plan. This is recorded as:

(18)

part_of_plan(admin(cisplatin,single_dose,

slow_iv_infusion), plan_1256).

We can obviously use clauses 12-18 at **run-time** to veto a suggestion to use the antibiotic *gentamicin* in *plan_1256*. We could do this by solving an appropriate query as in the earlier examples using a standard Prolog interpreter. However, in the **build-time** situation, we want to make use of the exacerbation principle in its most general form to generate a form of words that captures its use in this particular treatment plan. For example, we might want to generate the general exhortation *"Using nephrotoxic antibiotics is ill advised"*, or alternatively, the more specific *"Using an antibiotic such as gentamicin is ill advised"*. The question is how to get from the symbolic form in clause (12) to these abbreviated but more appropriate textual abbreviations.

We have yet to consider fully the natural language generation requirements for producing protocol document extracts from the symbolic representation of the chemotherapy plan. However, we are confident that even definite clause grammars with some simple extensions can produce the majority of the extracts shown in table 2. Here, though, we confine ourselves to the problem outlined above, namely, the production of simple abbreviations of a generic safety rule for inclusion in the protocol document. In particular, we want to be able to generate the phrase labelled (*) in table 2.

We have implemented a natural language translation of the generic safety rules using a combination of simple templates for rule components and a variable referencing scheme. The sentence in (20) is the result of using this translation process on the exacerbation principle defined in clause (12).

(20)

An action$_1$ is ill advised in a plan if

another$_3$ action is part of the plan and

the other$_4$ action causes an effect and

the effect is considered hazardous and

the action$_2$ exacerbates the effect of the other$_4$ action

The subscripts are not generated by the process, but are markers for the following explanation. Variables in rules are allocated a type (e.g. drug, effect, action, plan) and an indication of their first or other appearance in the rule. If the variable being referenced appears for the first time, then its translation (according to type) is preceded with the indefinite article$_1$. Otherwise it is preceded with the definite article$_2$. We use a variation of Definite Clause Grammar rules to define natural language templates for each atom (conclusion or condition) appearing in a rule:

```
clause_read_as((Head :- Body)) ➔
    head_read_as(Head),
    {new_line(NL)},
    body_read_as(Body,NL,' if ').

head_read_as(Head) ➔ atom_read_as(Head).

body_read_as(true, A, B) ➔ [].
```

```
body_read_as((Goal,Body),A,B) ➔
    [A,B],
    atom_read_as(Goal),
    body_read_as(Body, ' and ',A).
```

with for example a template for the atomic predicate *hazardous(Effect)* :

```
atom_read_as(hazardous(Effect)) ➔
    variable_read_as(Effect, effect),
    [is, considered, hazardous].
```

These grammar rules keep track of all variables in a clause so that references to different occurrences of the same type of variable can be differentiated. For example, a second distinct variable of a given type can be qualified by "another"[3], or "the other"[4]. For further occurrences, an ordinal qualifier can be used.

In theory, (20) could be included in the protocol document of every chemotherapy plan. Of course, it would not be welcome because of its unspecific nature; it obviously needs specialisation. The problem is to decide just how specialised to make the rule. In the next section, we demonstrate how we currently use a metaprogram to create these more specialised forms.

Table 2 - Examples of textual extracts generated at build time for inclusion in a protocol document

STUDY:		plan_1246:
FOCUS:		ovarian carcinoma
STUDY DRUGS:		cisplatin and taxol
DRUG ADMINISTRATION:		cisplatin: 75 mg/m^2 single dose every 3 weeks by slow intra-venous infusion in 100-500 mL of normal saline over a minimum period of 30 mins
	ADVERSE EFFECT:	cisplatin induced dehydration
	AMELIORATION:	prehydration and posthydration 1-2 litres of normal saline with 20 mEq/litre of Potassium Chloride and 8 mEq/litre of Magnesium Sulphate
	ADVERSE EFFECT:	cisplatin induced nephrotoxicity
	MONITORING:	creatinine clearance rate
	DOSE REDUCTION:	25% if creatinine clearance rate is between 10 & 50 mL/min & 50% if less than 10
(*)	WARNING:	Using a nephrotoxic antibiotic (such as gentamicin) is ill advised in this plan (it exacerbates the nephrotoxicity of cisplatin)
DRUG ADMINISTRATION:		taxol: 135 mg/m^2 single dose every 3 weeks by continuous IV infusion over 24 hours
	ADVERSE EFFECT:	taxol induced neutropenia
	MONITORING:	cell count
	WARNING:	taxol MUST be given BEFORE cisplatin to avoid severe neutropenia

Generation of Specialised Safety Principles

The **demo** logic program defined below is a typical metaprogram for theorem proving with logic programs in the form of Horn clauses (KB denotes the knowledge base).

```
demo(KB, Goals) ← empty(Goals).
demo(KB,Goals) ←
   select_goal(Goals,Goal,RestGoals),
   select_clause(KB,Clause),
   rename_variables(Clause,Goals,RenamedClause),
   parts(RenamedClause,Head,Body),
   match(Goal, Head,Substitution),
   add_body(Body,RestGoals,InterGoals),
   apply_sub(InterGoals,Substitution,NewGoals),
   demo(KB,NewGoals).
```

Intuitively, given a list of goals A, B, C…, the selection of a goal A and its subsequent resolution with the clause A` :- A1, A2 (where A and A` unify), yields the new set of goals A1``, A2``, B``, C`` with appropriate unification and variable renaming. We can use **demo** as the basis for a metaprogram that is able to execute (pure) Prolog programs. If we allow the unification, variable renaming, explicit knowledge base reference and resolution steps to be carried out by the underlying Prolog system we obtain:

```
demo(Goals) :-  empty(Goals).
demo(Goals) :-
    select_goal(Goals,Goal,RestGoals),
    select_clause((Goal :- Body)),
    add_body(Body,RestGoals,NewGoals),
    demo(NewGoals).
```

In a Prolog system, the built-in selection strategies are to choose the first goal in a sequence (from left to right) and the first clause whose head (or conclusion) can be unified (matched) with the selected goal. The clause selection is made by a built-in predicate and goal selection is defined by the meta-rule:

```
select_goal((Goal,RestGoals), Goal, RestGoals).
```

We use the term **unfolding** to refer to the steps

```
   select_clause((Goal :- Body)),
   add_body(Body,RestGoals,NewGoals)
```

where a matching clause (**Goal :- Body**) is selected and **Goal** is replaced in the original conjunction of **Goals** by **Body**. Our simple unfolding metaprogram **unfold** is a variation of this **demo** theme:

```
unfold(stop, Rule).
unfold(continue, (Head :- Goals)) :-
     select_goal(Head, Goals, Goal, RestGoals),
     select_clause(Goal, (Goal :- Body)),
     add_body(Body, RestGoals, NewGoals),
     next_step(Choice),
     unfold(Choice, (Head :- NewGoals)).
```

In **unfold(_, Rule)**, we maintain a record of the latest, unfolded representation of a safety principle. The goal selection and the selection of a clause from those whose head matches the goal to be unfolded is left to the user to decide.

Although, the **unfold** metaprogram manipulates clauses and their components, the explanations below are more understandable if we use natural language forms to illustrate the unfolding process.

If the first condition of sentence 20) (is unfolded using the record of the intended use of cisplatin (clause 18) we obtain:

```
an action is ill advised in plan_1256 if
  using  cisplatin causes an effect and
  the effect is considered hazardous and
  the action exacerbates the effect of using cisplatin
```

Selecting the first condition of this unfolded rule and the matching clause (14) concerning the nephrotoxic effect of cisplatin, we end up with the sentence:

```
an action is ill advised in plan_1256 if
  nephrotoxicity is considered hazardous and
```

the action exacerbates nephrotoxicity of using cisplatin

The goal "nephrotoxicity is considered hazardous" uniquely matches clause 15, so we actually unfold this automatically and obtain:

an action is ill advised in plan_1256 if

the action exacerbates nephrotoxicity of using cisplatin

Given clause 13, then unfolding the only condition left we further generate:

using a drug is ill advised in plan_1256 if

using the drug causes nephrotoxicity and

the drug is an antibiotic

This is still not as close as we would like to the target phrase (* in table 2). To get even closer versions, we need to extend our use of types and subtypes so that properties of types can be inherited by their subtypes, and subtypes can be used in the same way as we have already used types in the generation of textual summaries. For example, we can extend resolution to allow inferences of the form:

$$\{f(X{:}T) \leftarrow g(X{:}T),\ X \in S;\ S < T\}\ \vdash\ f(X{:}S) \leftarrow g(X{:}S)$$

where in general, $X{:}U$, $X \in U$ and $W < U$ respectively denote "X is of type U", "X is restricted to type U" and "W is a subtype of U", and $\vdash$ denotes logical inference. For example, since *antibiotic* is a subtype of *drug* then we can rewrite the last specialised textual form of the exacerbation principle as:

using an antibiotic is ill advised in plan_1256 if

using the antibiotic causes nephrotoxicity

A similar extension to handle properties of types (such as a drug being nephrotoxic) enables us to produce the final form:

using a nephrotoxic antibiotic is ill advised in

plan_1256

which is much closer to our target phrase. Work is also currently in hand on generating the additional qualifying phrases that appear in parentheses in (*) of Table 2.

Conclusions

Generating symbolic descriptions of chemotherapy plans and related textual components of protocol documents is certainly a useful service to provide clinicians when they are preparing new chemotherapy schemes. Recent surveys have identified considerable anxiety amongst clinicians about their incomplete and incorrect knowledge of drugs and contraindications. The inclusion of generic safety principles to monitor the formulation of chemotherapy plans may discourage errors being made both in the design phase and subsequently in their application in the management of patients.

Even with our current very naïve approach to natural language generation, we can produce useful textual summaries of a chemotherapy plan and so reduce the burden on compiling a protocol document. However, much work remains to be done if the tools we are developing are going to be practicable and acceptable to clinicians. In particular, we must develop an interface that can be used effectively even with the kind of irregular exposure that many of its potential users are likely to experience.

References

1. 'You'll die' doctor tells patient after mistake. The Sunday Times, Nov. 21st 1993.

2. Doctors fail to refer cancer patients to specialist centres. The Guardian, Dec. 1st 1993.

3. Breast cancer care inquiry call. The Guardian, Jan. 14th 1994.

4. Hammond P, Harris A L, Das SK and Wyatt JC. Safety and decision support in oncology. Methods of Information in Medicine. 1994; 33(4): 371-381.

5. Hammond P and Sergot MJ. (1996) Computer support for protocol-based treatment of cancer. Journal of Logic Programming, 26(2), Feb 96, 93-111.

6. Fischer DS, Tisch Knobf M and Durivage H. The cancer chemotherapy handbook. (1993) 4th Edition. Mosby, St Louis, USA.

7. Hammond P, Sergot MJ and Wyatt JC. Formalisation of Safety Reasoning in Protocols and Hazard Regulations (1995) 19th Annual Symposium on Computer Applications in Medical Care.

Biography

Prof. Peter Hammond, Department of Informatics, Eastman Dental Institute for Oral Health Care Sciences, University College London, 256 Gray's Inn Road, London, WC1X 8LD, UK. Telephone & Fax : +44 (171) 915 2303. Email: p.hammond@eastman.ucl.ac.uk .

Prof. Hammond is a computer scientist with an interest in the application of logic programming, computer graphics and image processing in biomedicine.

Dr. Sanjay Modgil, Department of Informatics, Eastman Dental Institute for Oral Health Care Sciences, University College London, 256 Gray's Inn Road, London, WC1X 8LD, UK. Telephone: +44 (171) 915 2343. Email: s.modgil@eastman.ucl.ac.uk

Dr. Modgil recently completed a PhD in Logic and is currently a postdoctoral research assistant on the "Design-A-Trial" project.

Dr. Jeremy Wyatt, Health Knowledge Management Centre, School of Public Policy,

University College London, U.K. Telephone: +44 (171) 504 4986.

Email: uctqjwy@eastman.ucl.ac.uk

Dr. Wyatt leads the Health Knowledge Management Centre. He is a qualified doctor of medicine with a longstanding involvement in medical informatics.

Acknowledgement

This work has been carried out within the Design-A-Trial project and is supported by the UK Engineering and Physical Sciences Research Council under the Healthcare Informatics Programme.

Computer-Aided Mammography: A Case Study of Error Management in a Skilled Decision-making Task

Mark Hartswood and Rob Procter,

Institute for Communicating and Collaborating Systems,
University of Edinburgh, Scotland.

Keywords: breast screening, prompting aids, error management, ethnographic studies

Abstract

The practice of breast screening calls for radiologists to exercise a combination of perceptual skills to find what may be faint and small features in a complex visual environment, and interpretative skills to classify them appropriately -- i.e., as benign or suspicious. Radiologists make errors, however, and evidence suggests that these can be reduced by employing computer-based image analysis techniques to find, and prompt for, target features.

Computer-aided breast screening provides an interesting case study of error management issues raised by the introduction of computer support for a skilled decision-making task. Of course, since the rationale for computer-aided screening is that humans make errors, it is essential that system designers understand the nature of these errors. Equally, since the image analysis techniques themselves are not error free, it is important that radiologists understand the nature of the errors that the prompting system makes. More generally, this requires that radiologists be able to account for the behaviour of a technically complex system.

To investigate these issues, we have performed an ethnographically based study of clinic and reading practices, and controlled studies of the effects of prompting on radiologists' performance. These studies have enabled us both to understand better how radiologists approach the problem of error management in current practice, and their needs with respect to error management when using computer aids. Our results also show that the ways in which a computer-based tool actually gets used may be quite different from what was originally envisaged by its designers.

Finally, we outline what may be required in terms of training and practice to ensure the safe and effective use of computer prompting systems in screening applications.

Introduction

Breast cancer is the commonest form of cancer in the UK. Each year there are about 24,000 new cases and 15,000 deaths from the disease, accounting for one-fifth of deaths among women from all forms of cancer. Mammography (radiological imaging of the breast) remains the only method of detecting early stages of breast cancer, and preventative breast or mammography screening programmes operate in many countries.

We have been working as members of a team which is developing PROMAM (Prompting for MAMmography), a computer-aided mammography system designed for use in the UK breast screening programme (refs. 9-11, 20, 26). PROMAM is a prompting system which aims to improve radiologists' detection performance by drawing their attention to possible ill-defined lesions and micro-calcification clusters, and so reduce errors.

Prompting systems are designed to improve observer performance in visual search tasks by employing image analysis techniques to highlight areas that the observer should examine. In principle, observer errors arising from inattention, fatigue, etc. can be reduced. However, the practical realisation of improved observer performance is not easy. First, prompting systems are not infallible, so observers must be able to recognise system errors if their performance is not to be adversely affected. Second, the introduction of such a system in e.g., breast screening, may involve changes in work practices which reduce the overall effectiveness of error management procedures in the workplace. Investigating these issues has been a key part of our contribution to the PROMAM project.

We begin with an overview of the UK breast screening programme, followed by a review of the literature on visual search errors and the role of prompting as an aid to improved observer performance. Next, we present the results of field studies of breast screening work. Our focus here is on how radiologists manage their work both individually and collaboratively, and formally and informally, so as to minimise errors in their performance. We then present the results of a study of radiologists using the PROMAM prompting system, with a special emphasis on how they make sense of its behaviour and the role that this sense-making has in the management of system errors. Finally, we consider the implications for prompting systems on the informal mechanisms for error management.

Screening Mammography

The basic principle behind mammography as a screening test is that signs of malignancy are sufficiently distinct to achieve a reasonable sensitivity and specificity. Types of feature that are indicators of malignancy include:

> **Microcalcifications** are small deposits of calcium visible on a mammogram as tiny bright specks. They can be due to benign processes: e.g., it is common for vessels to calcify, giving a characteristic 'tram line' appearance on the mammogram. Small clusters of calcification can be indicative of early breast disease. Typically, the number, shape and distribution of calcifications within a cluster are used to determine the likelihood that they are the result of a malignant process.
>
> **Ill-defined lesions** are areas of radiographically-dense tissue appearing as a 'bright patch' on the mammogram that might indicate a developing tumour. Typically, lesions that are well-defined are the result of benign processes: e.g., they may be cystic. Lesions that do not have a well-defined edge are considered suspicious.
>
> **Stellate lesions** are visible as a radiating structure with ill-defined borders. The radiating components (or spicules) are the result of malignant processes infiltrating the breast tissue.
>
> **Architectural distortion** may be visible when breast tissue around the site of a developing tumour contracts. In the absence of other signs this might give a subtle clue to the presence of a tumour.
>
> **Asymmetry** between left and right mammograms may be the only visible sign of some hard to detect features. Asymmetry can be difficult to interpret as there is often a natural asymmetry in the distribution of breast tissue.

The UK Breast Screening Program (UKBSP) is a national service with a regional organisation. Each region is served by a number of screening clinics, each with two or more radiologists. The initial screening test is by mammography, where one or more X-ray films (mammograms) are taken of each breast by a radiographer. Each mammogram is examined for evidence of abnormality by two experienced radiologists. Women between the ages of 50 and 64 are invited to attend a clinic for screening every three years.

The goal of screening is to achieve a reliable and controlled cancer detection rate. Two performance parameters are particularly important: specificity and sensitivity. A high specificity -- i.e., low false positive (FP) rate -- means that few women will be recalled for further tests unnecessarily; a high sensitivity -- i.e., a high true positive (TP) rate -- means that few cancers will be missed. Achieving high specificity *and* high sensitivity is difficult.

The radiologists' task is a difficult one, not least because the small number of cancers is hidden amongst a large number of normal cases. It is a task which demands a high level of perceptual and interpretative skill: under certain circumstances normal tissue can have an abnormal appearance -- and vice versa. Figures from the UKBSP show that in the prevalent round (first screening visit) 6.4% of women screened are recalled for assessment, and in the incident round (later screening visits) this figure falls to 3.0%. More cancers are detected in the incident round -- 6.3 per thousand, compared with 3.4 per thousand in the prevalent round (ref. 2). FPs are not life-threatening errors, but they do cause stress and anxiety for those women who are recalled unnecessarily, and they waste resources. Far more serious are interval cancers, i.e., false negative (FN) errors, which are life-threatening.

The UKBSP is continually investigating ways of reducing FP and FN errors. For example, in many clinics, current practice involves each mammogram being 'double read' (i.e., examined independently by two radiologists). In the past five years, interest has grown in the possibility of employing computer-based image analysis techniques to enable a single reader (radiologist) to achieve performance equal to that achieved by double reading. The goal of the PROMAM project is to develop and apply these techniques to a prompting system.

Errors in mammography: Psychological approaches to decision-making have been widely used to provide a conceptual framework for understanding observer performance in radiological search tasks such as screening mammography. There are two basic approaches. One examines the nature of perceptual skill, the other is concerned with cognitive reasoning.

Studies of perceptual skill as applied to radiological tasks often involve eye-tracking studies or ROC (receiver operator characteristic) methodologies. Eye tracking involves using a mechanical apparatus to determine where a reader's gaze is directed when examining an image. Using this approach, Kundel, Nodine and Carmody (ref. 14) classified three types of error that can result in a FN decision for radiological search tasks. These are search errors, detection errors and classification errors. Search errors occur if the lesion does not enter the radiologist's 'useful field of view'. Detection errors are said to occur if the visual dwell time for an unreported lesion falls below some empirically determined threshold, where as classification errors occur if visual dwell time exceeds this threshold. Savage, Gale, Pawley and Wilson (ref. 21) define search and detection errors as occurring where a radiologist fails to report the presence of a lesion, and classification errors where the lesion has been reported, but inappropriately acted upon.

Studies of cognitive processes involved in radiological search tasks tend to be concerned with the radiologist's orientation to, and interpretation of, an image. Gale (ref. 8) posits a conceptual model whereby an observer selects from a pool of possible hypotheses according to their expectations about the mammogram. Each considered hypothesis is confirmed, or replaced in response to the information gathered by a visual inspection. Similarly, Lesgold, Glaser, Rubinson, Knopfer, Feltovich and Wang (ref. 15) equate the accurate interpretation of radiological images with the selection of appropriate schemata by drawing a distinction between the approach of novice and expert radiologists. They suggest that errors of interpretation made by novices can be explained in terms of both inefficient schemata selection, and the inappropriate testing of selected schemata.

Computer-aided Mammography

The prevailing view is that computer-based prompting aids are designed to address a different problem from those that assist with classification. The goal of the former is to improve sensitivity by cueing readers' attention to features that they may overlook, while the latter are designed to improve specificity by assisting the decision-making process (ref. 7).

The goal of a prompting system is to reduce the occurrence of search or detection errors. It is not intended that radiologists should attach any clinical significance to the presence -- or absence -- of a prompt. In contrast, classification aids offer an interpretation (either in terms of a probability value, or some explicit reasoning) designed to support a radiologist's judgement about the significance of a lesion already detected.

There are number criteria that have to be satisfied for a prompting system to be acceptable in the role of an attention cue in a screening environment. Current mammography image analysis techniques have a poor specificity compared with a radiologist -- prompting systems will therefore generate a relatively large number of FP prompts. The value of a prompting system lies in the fact that it can offer a complementary synthesis of system and radiologists' strengths: the former is more consistent in its visual search performance, the latter has interpretative skills which the system cannot match. If prompt system specificity is too low, however, radiologists will have to attend to too many FP prompts, and the costs of using the system may be perceived as outweighing its benefits. So, a prompting system must have sufficient specificity to maintain FP prompts at a manageable level. To achieve this, a prompting system must necessarily have some capacity to discriminate between features, if not to classify them.

Mammography image analysis techniques do not achieve 100% sensitivity, so prompting systems will also generate FN prompts (i.e., fail to prompt for some cancers). It is important that system and radiologists' FNs are independent. A correlation between system and radiologist FNs would preclude any improvement in overall sensitivity.

Finally, the prompting system approach assumes that prompting does not have an adverse effect on reader performance, i.e., that in attending to prompting information, no systematic bias is introduced into the radiologist's decision-making.

<u>The PROMAM system:</u> PROMAM's prompts consist of a hard copy, low resolution image of the mammogram pair with prompt information superimposed (ref. 20). The choice of this very 'low tech' prompt interface design was determined after an initial requirements investigation. Radiologists liked the simplicity of paper prompting interfaces which have, in addition, the virtue of fitting in easily with current reading practices; paper is handled routinely during the reading session.

The PROMAM system is capable of detecting microcalcifications and ill-defined lesions (refs. 12, 16). Prompts for microcalcifications consist of an irregular outline of the potential cluster. Prompts for ill-defined lesions consist of an ellipse surrounding the suspect region (see Figure 1).

<u>An Overview of Breast Screening Work</u>

An ethnographic style investigation of work practices in two Scottish and four English breast screening centres was conducted over a six month period. The centres are referred to here by the letters A through to F to preserve anonymity. Both observational and interview data were collected during a 2 month period of investigation at centre F, and during one week period in each of the other five centres. Each of the six centres studied had agreed to participate in clinical trials of PROMAM, and access for the purposes of this study was negotiated as a contribution to the on-going development of the PROMAM system. Thus centre selection was governed by suitability for clinical trials, rather than representativeness of screening practice.

Where data is presented, the mode of data collection is indicated (e.g., interview, field notes). Observations of reading sessions were conducted by asking the reader to indicate and explain their reasoning when they encountered something 'interesting' while reading. Where a comment or observation is attributed to the statement or activity of a reader, the reader is identified by a number and the screening centre by a letter (A-F). Thus FR1-C refers uniquely to a particular reader in centre C.

Breast screening practice in the UK has undergone important changes since the programme's inception. The autonomy given to individual screening centres has enabled innovation to proceed by the independent adoption of new practice. Practice innovations may be regularised (with the concomitant national resourcing implications) in the light of formal studies demonstrating clinical and/or cost effectiveness. An example of this is the adoption of two view mammography for incident round screens. The practice was initially adopted ‘unofficially’ by a number of screening centres in England, and a prospective study followed that demonstrated its effectiveness in reducing both FN and FP errors (ref. 23). This precipitated the adoption of two view mammography as standard practice by the UKBSP.

At the time of this investigation, the clinics studied were involved in more or less formal studies of practice innovations. For example, clinic E was examining the logistics of performing two view mammography in the second and third round by dedicating one of its mobile units to this procedure, and had just completed its involvement in a trial to study the efficacy of reducing the screening interval.

The practice of double reading represents an innovation of uncertain status. It has been adopted as standard practice in Scottish screening centres (ref. 4), and by a number of English centres (ref. 25). This situation is partly due to methodological problems in ascertaining performance gains attributable to double reading, resulting in widely varying estimates for its effectiveness (ref. 27), and also because of local shortages of trained readers (ref. 25).

Differences in practice may also emerge as a response to variations in local circumstances. For example, clinic A changed from a system of worst opinion recalls to a system of third reader arbitration because their recall rate became unmanageable. In contrast, centre D discontinued a policy of discussion of recall disagreements because individual readers held out for their own recall decisions. In centre E, only around 50% of cases are double read. This is due, in part, to the fact that two of its readers ‘got used’ to single reading during a period when they were the only two readers available. Centre E is also the only centre to operate a system of blinded double reading (where the opinion of the first reader is not available to the second reader).

In contrast with inter-clinic practice differences, intra-clinic practice is surprisingly homogeneous. This can be explained in part by how reading is organised. Many of the activities involving readers require co-ordination with other clinic staff (e.g., supervising clinics, attending meetings etc.), and so are difficult to re-schedule or to interrupt. In contrast, reading demands less commitment, and so can be more flexibly attended to. In consequence, reading tends to be organised around other activities; it is often done ‘in a quiet moment’ or at the beginning or end of the working day. The availability a given reader at a given time often cannot be guaranteed, nor can the time they have available for reading be predicted. One consequence of this is that the artefacts used in reading have to be arranged in a way that is neutral with respect to who will be reading the films, leaving little scope for tailoring the selection and organisation of artefacts to suit individual preferences.

Of course, reading is not entirely commitment free. Obligations associated with reading include ensuring that women are informed of the outcome of screening in a timely fashion. The level of commitment associated with reading may therefore increase if there is a backlog of cases to be read. For example, in centre B cases may be single read if there is more than a week’s delay in reading.

Thus a combination of local circumstances and innovation can lead to inter-clinic variations, but the nature of screening work usually demands that within clinics homogeneous practice is established by consensus between readers.

Reading practices: Each mammogram is examined by at least one medically qualified reader who typically is also a trained radiologist. However, members of other medical specialities may also be employed as film readers if given the appropriate training, so in this paper the generic term ‘reader’ is used. On average, between 50 and 150 mammograms may be read in a single session. A reader will work through the cases on the viewer and mark his/her decision on the screening form. The decision of a reader may be one of:

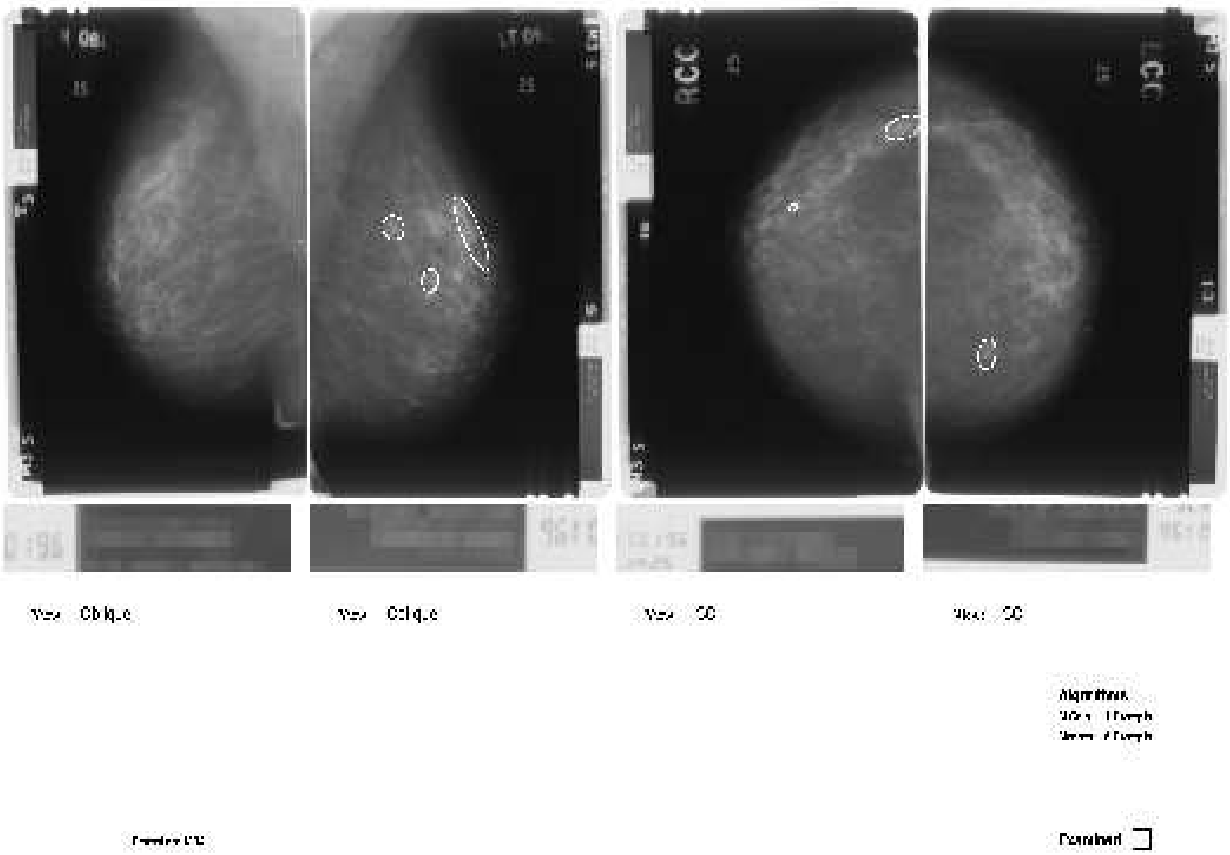


Figure 1: An Example Prompt Sheet

Return to routine recall When the reader decides that the case is normal the woman will be invited again for screening after a three year interval.

Recall for assessment When a possible abnormality has been detected the reader will recommend attending an assessment clinic for further tests.

Technical recall When the reader decides that a diagnosis may be inaccurate because of imperfect mammography then repeat screening films may be requested.

The degree of certainty about whether a feature indicates malignancy can vary considerably. Some are unequivocally malignant, whereas others might be only mildly suspicious. There are also various natural processes in the breast that can give the appearance of malignancy to varying degrees, and there are malignancies that are mammographically 'occult', i.e., they do not appear at all on the mammogram. It is common practice for reader to classify the features they find according to the probability that they indicate malignancy. For instance, at one clinic readers use a five point classification scale: C1 (normal), C2 (benign), C3 (equivocal), C4 (suspicious), and C5 (malignant), and set the recall threshold at C3.

Some readers mark each decision as it is made on the screening form, others defer marking decisions until they reach a case they wish to recall. In the latter case, intervening normal decisions are then marked as a 'batch'. The number of cases examined consecutively in this way will vary as recalled cases are randomly distributed. Some film readers might 'batch up' an arbitrary number of cases, rather than waiting for the next recalled case. If the cases are being double read, it falls to the second reader to ensure that the recalled cases are removed from the viewer before the normal cases are taken down.

Double reading involves the separate

examination of each case by two readers, who each give their opinion. A final decision to either recall the woman for further tests, or to return the case to routine screening is made later by combining the decisions of the individual readers. In the clinics studied, three different strategies were employed for deciding the final outcome. Clinics B, D, E and F use a system of 'worst opinion recalls'. By this method, if either, or both, readers recommend that the case requires assessment, then the case is recalled. Centre A uses a system of 'third reader arbitration', where a reader not involved in the initial reading decides cases where there is disagreement. In centre C, disagreements are resolved by discussion between the two readers.

Where screening decisions are recorded on a paper screening form, double reading is generally done without any formal blinding of the second reader to the first reader's decision. This is because the screening form is routinely attended to as a source of various types of evidence, (e.g., HRT status, radiographers comments etc.) and so that the second reader may record his/her own decision. An exception to this is centre C, where the first reader's decisions are written on the back of the 'batch slip'. Although this method of recording the first reader's decision could easily serve as a blinding mechanism, a second reader was observed to examine the first reader decisions before reading the batch themselves, so apparently it is not always used in this way. However, double reading in centre C is viewed as serving a particular purpose, and this is discussed in more detail below. In centre E reading is blinded. This is facilitated by the use of an electronic system for recording screening decisions.

Comments made by readers suggest that they are alert to the possibility that access to the first reader's decision may bias the decision of the second reader:

> "Sometimes the second reader suppresses a potential recall on the basis that the first reader thought it was nothing. Therefore recalls go up with blinding." (Comment made while reading: field notes FR1-D)

Reader FR1-E stated that in a double reading team there is a tendency for readers to converge in their performance characteristics, maybe due to personality dominance. One reader suggested that the degree of influence that access to first reader decisions has might be a function of experience:

> "... and the fact that X had already first read it normal, you see that should not make a difference ... I think that perhaps did when I first started, but in (...?) I've been at it for a while so I've never thought it would, it shouldn't do." (Comments made while reading: transcript FR1-A)

There is some empirical evidence to suggest that this lack of independence can affect recall decision-making. Data taken from a published double reading study (ref. 27) indicates a strong relationship between the reported sensitivity of readers and the percentage of time they read second. One interpretation is that the second reader is 'prompted' by the first reader and thus picks up cancers that they would otherwise have been overlooked. However, this relationship might also be accounted for by inter-observer variation.

In the absence of a blinding procedure, readers may seek to maintain independence in their decision-making by employing strategies that decrease the accessibility of the first reader's decision. The simplest approach involves "trying not to look" (FR1-A) at the comments made by the first reader before making their own. Another approach is made possible by the practice of 'batching up' cases when reading:

> "When reading first, [I] maybe batch 7 or 8 films before scoring them, when reading second only batch three or four. This is because if the first reader has written something then have to go back and examine the films to see what they were referring to." (Comment made while reading: field notes FR2-A)

In centre D there is a particular incentive for reading first. Assessment clinics are organised so that the radiologist conducting the clinic is usually given cases recalled from batches for which they were the first reader. This provides an opportunity to receive feedback on screening decisions:

> "This makes assessment clinics more interesting. For example, if see something unusual, and don't know what it is, then get a chance for this feedback. [I] may not see anything similar for a number of years." (Comment made while reading: field notes FR1-D)

In contrast, two readers in centre A (FR1-6 and FR4-A) maintain an informal arrangement whereby they contrive to be first and second reader an equal number of times. Furthermore, when second reading one reader (FR4-A) reads the batch in reverse order. This is done under the assumption that readers are likely to be more fatigued towards the end of the reading session and so may be more likely to make mistakes on cases towards the end of the batch.

A double reading system involving discussion of recall decisions demands a greater degree of commitment from readers. Readers acknowledged its logistical difficulties, and also expressed concerns that explicit collaboration could bias decision-making. For example, when a reader in centre E was asked if they ever discussed cases, he replied that this was only done at the review session and at the interdisciplinary meetings. He stated that they were "worried about the effects of dominant personalities" (field notes FR3-E). A reader in centre B expressed similar concerns.

In clinic D, a system of discussing recall decisions had been in place, but this practice was discontinued. One reader suggested why:

> "[Discussion meetings] rapidly became a waste of time as each reader has a particular feature that they are able to detect well (patchy asymmetry, distortion, microcalcs are my own) and would hold out for recalls that they are convinced are something (usually falling into these categories)." (Comment made while reading: field notes FR1-D)

Published studies of double reading are typically concerned with its effects on the sensitivity and specificity of the screening test (refs. 1, 3, 5, 22, 24). Only one makes a passing reference to other potential roles (that it places an emphasis on teamwork (ref. 4)). However, in the centres studied, the practice of double reading appears to fulfil a number of roles that are arguably as important as any direct performance effects. These include:

1. a means of training film readers, and
2. to monitor, provide feedback and reassurance on reader performance.

Although there are acknowledged disadvantages to unblinded double reading, and readers appear to value their autonomy as decision-makers, the unblinded nature of double reading as typically practiced facilitates the achievement of the above roles. We will now discuss them in more detail.

<u>Training:</u> In clinic C double reading is used primarily as a mechanism for training. Typically a trainee will be paired with an experienced reader and disagreements about recall decisions are decided by discussion. For the purposes of training the potential for 'bias' inherent in a system that relies on discussion is actually desirable -- here the aim *is* to influence the decision of the trainee. Use of discussion enables the degree of autonomy given to the novice reader to be actively managed:

> "With the locum reading, the recall rate has gone up ... [I feel] that is important not to always override the decisions of junior readers as this can be a learning experience." (Comment made while reading: field notes FR1-C)

In centre C, an experienced reader (FR1-C) was observed to be reading second following a trainee. The trainee had flagged a case for recall, but had left a comment stating that the case was 'probably OK'. After examining the film the senior reader scribbled the request out, and the case was returned to routine recall (i.e., it was not removed from the viewer for discussion). Thus managing novice decision-making may be effected before the discussion stage is reached.

Centres B and E were also involved in training readers at the time of this study. Both centres employ a system of 'worst opinion recalls' and both have a similar policy of incrementally introducing novices into the reading process. Trainee readers initially attend a recognised training course. They may then spend a period of time reading films in the screening centre and discussing their opinions with experienced readers, but they do not at this stage contribute to recall decisions. Novices are introduced into reading proper as a first reader, and then as either a first or second reader as they gain experience. A number of reasons were suggested for limiting novice readers to reading first initially. These included:

- Providing a learning environment where the novice has to make decisions independently.
- So that any 'unnecessary' recalls can be stopped by the second reader.
- So that the second reader can act as a check and detect any missed abnormalities.

Thus training is organised to take advantage of the structure of double reading to provide a safe an supportive environment where novice readers can be encouraged to make independent decisions. The second reader is able to monitor and manage the novice reader's decision-making, and this also serves to provide a degree of reassurance that any cancers overlooked by the first reader may still be detected. One experienced reader from centre C suggested that she was "particularly careful when reading following a registrar [trainee]" (FR1-C).

<u>Monitoring and feedback:</u> One advantage of double reading is that the responsibility for decision-making is not shouldered entirely by a single reader. Thus, as one reader from centre C suggests:

> "Double reading can take away some pressure. People can have 'off' days." (Response to questionnaire: FR1-C)

Readers are often concerned to ensure their performance is consistent on a day by day basis -- that they are not unduly affected by fatigue, distractions etc. For example, after returning from maternity leave, one reader from centre E asked the clinical director for all the cases that she read to be double read (field notes: FR2-E). She sought to monitor her performance by comparing her cancer detection rate with others. She stated that after returning from maternity leave she 'missed' three that another reader had detected, but has only 'missed' one since then. She stated that double reading provides "reassurance" in these circumstances.

Another reader from centre E uses a similar mechanism to monitor his day to day performance. When reading second, he compares his decisions with those of the first reader to see if he has missed a lesion, or classified one differently. He states that when there is a difference of opinion then 3/4 of the time he has also seen the lesion and has dismissed it, and in the remaining 1/4 he has overlooked the lesion. Where there is disagreement, it is usually over less suspicious features, and that typically there is a large degree of agreement between readers over "actually malignant" features (those with a 4 or 5 classification). He recalled an occasion when he discovered (by checking the first reader recalls) that he had missed an "obvious" spiculated lesion (field notes: FR3-E).

The work of clinic staff is formally monitored through procedures for quality assurance and work documentation. Clinic staff hold regular meetings in which feedback may be given and received:

- multi-disciplinary pathology meetings where radiological appearance and pathology data are compared;
- review of interval cancers which may be evidence of FNs, and
- informal (and at some clinics, formal) discussion about differences in recall opinions.

Such meetings provide an opportunity for readers to articulate aspects of their work which they perform as individuals, such as their reasons for giving a 'recall' or 'no recall' opinion. Our studies suggest, however, that readers use double reading itself as a less formal mechanism for monitoring and feedback (ref. 10).

The first reader's opinion on each of a set of cases effectively provides a standard against which the second reader might compare his or her decisions. Even in centre C, where double reading is seen primarily as having a training function, one experienced reader commented:

> "... the two consultants like to read against each other as well as against the inexperienced radiologists." (Comment made while reading: field notes FR1-C)

Feedback gained by the second reader in this way may fulfil a number of functions. Readers can monitor their performance on a session by session basis and gain some reassurance that intra-observer variations are compensated for. This informal monitoring activity may also have a role to play in maintaining readers' recall thresholds within a manageable range by establishing and reinforcing normative interpretations. In two of the clinics visited it was evident that this informal monitoring has evolved further: first readers sometimes annotate the reporting form in cases they decided are normal (see Figure 2). In effect, this extends the set of cases over and above those recalled for which evidence about the reasoning of the first reader is available:

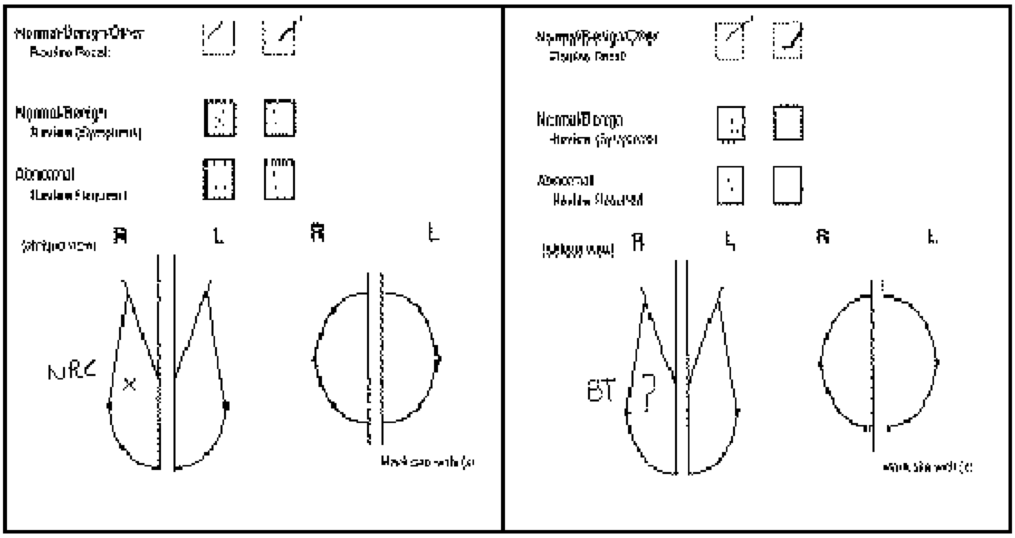
"Leaving messages for the second reader is useful -- to let them know that you've seen it -- the second reader might want to know whether you've seen it and what your opinion is." (Comment made while reading: field notes FR2-A)

It is possible that annotation enables the first reader to assert their competence, and the second reader to assess the specificity of their decisions in the context of those made by their colleagues. In addition, annotations may be used to make inferences about the degree of suspicion in particular cases:

> "If the first reader flags a 'composite shadow' and the second reader goes straight past it, then it probably isn't significant." (Comment made while reading: field notes FR2-A)

The act of annotating a feature implies that the feature is worthy of annotation. That is, some characteristic of the feature appears to be sufficiently suspicious to warrant particular attention by the reader so that this suspicion may be discharged. Annotation may thus serve to demonstrate a reader's accountability to the decision-making process.

A common (and arguably the simplest) type of annotation is to label a feature on the schematic by writing "Benign" or simply "B". Nothing is said about the reasoning behind the decision, indicating a tacit assumption that this will be readily apparent to other readers. Another common annotation is "BT" (Breast Tissue). Here some interpretation is offered: that the presentation of the feature is ascribed to normal breast tissue, but no reason for this ascription is given. Both these types of annotation appear to suggest that there is little doubt in the reader's mind that his or her opinion is correct, the annotation seems intended to reinforce this opinion and to demonstrate vigilance. On occasions, however, the use of "I think" and "?" is used in association with the description to express, and draw attention to, the reader's uncertainty.



1 2

Figure 2: Example of first reader annotations

More complex annotations are also used. These typically make explicit information about a reader's reasoning by referring to the evidence used to mitigate the initial suspicion. Examples include: "Comp CC Ok" -- not visible in the CC view, so is a composite shadow; "NRC" (No Real Change) -- the feature has not changed over time, and thus is less suspicious. Readers were also observed to annotate changes they thought were due to different projections, new microcalcifications clusters that had a benign appearance, calcification clusters that hadn't changed and clusters of benign microcalcifications embedded in a background of vascular calcifications.

Agreement between readers is higher for recalled cases that actually turn out to be cancers, than for recalled cases that turn out to be normal. This 'virtuous' difference between readers' recalls accounts for the performance gains reported for double reading. In fact, double reading serves to compensate for both intra- and inter-observer variations (ref. 16). If these differences are too large then assessment clinics may be overwhelmed and changes in procedure may follow, like changing from a 'worst case' to a 'third reader' arbitration recall decision-making policy. It is possible that feedback from both assessment clinics and from first reader annotations may play a role in maintaining readers' recall thresholds within a manageable range. It is interesting to note that many annotations are for features that fall on the benign side of the recall threshold, that this is the region were most FP and FN decisions are likely to occur, and thus where differences in opinions are likely to have the greatest significance. Annotations are made where there is likely to be some uncertainty -- where decisions can be 'open to interpretation', thus articulation about such cases may serve to communicate and establish norms about the significance of particular kinds of presentation.

Using double reading as an informal or formal (if part of training) mechanism for obtaining information about a reader's performance may be useful, but readers also recognise the possibility that their judgements may be influenced by the first reader's decision. Thus there is a tension between the decision-making and monitoring aspects of a reader's work, where access to a first reader's decision is recognised both as useful as a metric of performance, but also as potentially harmful if it serves then to bias decision-making.

Summary: The results of this study suggest that readers are very aware of the extent and limitations of their expertise and apply these insights routinely in their work. Readers not only possess self-knowledge of performance in terms of measures like sensitivity and specificity in respect of particular feature types and circumstances, they also demonstrate a more general understanding of the psychology of the decision-making process. This understanding relates how particular conclusions are drawn from particular types of evidence, and suggests biases and errors to which a reader may be subject.

The results also show how, through an informal, emergent extension of formal monitoring procedures, readers may use each other to gain reassurance about their performance and to help themselves ensure that it remains within acceptable limits.

Studies of the Effects of Prompting

The key question about prompting in breast screening is what effect will it have on readers' performance? Owing to the low incidence rate of breast cancer, it is impossible to answer this question quantitatively without doing a large scale clinical trial (ref. 26). However, there are several issues that can be addressed using smaller scale investigations. These include: assuming that prompting system performance is not error-free, is there an upper limit to prompting errors -- FPs and FNs -- before a prompting system becomes useless and, if so, what level of error is tolerable by its users? A related question is, are some kinds of errors more tolerable than others? Finally, how do readers make sense of how the prompting system behaves and does this have a bearing on readers' assessment of its usefulness, and on how they use it?

Prompting information represents an additional source of evidence that a reader can draw upon when deciding a case. As with conventional sources of evidence, it is important to examine the possibility that use of prompting information may bias decision-making. It is also important to explore how access and interpretation might be managed to reduce any such effects.

In contrast to a human observer, a computer-based image analysis system will typically make use of only a subset of the available evidence, and will be limited in the ways in which it can access and combine evidence from different sources. Consequently, such a system is unlikely to match the performance of trained human observers in terms of both sensitivity and specificity, and will exhibit behaviours that might be considered naive by human observers. These limitations imply that a reader cannot use their knowledge of the behaviour of readers (as they do to interpret the decisions and annotations of colleagues) to reliably account for the behaviour of the system.

As part of PROMAM's development programme, we conducted three investigations designed to further understanding of how readers would make use of prompts under clinical conditions. The first was designed to elicit readers' subjective responses to system performance. Earlier work had suggested that prompting can improve visual search performance, but only if the FP prompt rate was no more than 1.5 times the TP rate (ref. 13). However, there are problems with extrapolating directly from such studies to the clinical setting. First, heavily biased test sets were employed and so the results may not be directly applicable to the circumstances in which reading is performed in the clinic. Second, it was unclear whether the FP prompts were representative of the types of FP that a prompting system might actually produce. Part of our programme of work has been to investigate these issues further.

The results of our first study indicated that, under realistic reading conditions, readers' tolerance for FP prompts was significantly higher than the upper limit established by Hutt (ref. 9). Subjects commented that they found FP prompts useful as an aid for making sense of the system's behaviour. We concluded that prompts for candidate[3] features may be acceptable in clinical use because they afford the development of their understanding of how the system works, and what its capabilities are (ref. 10).

Second study: To examine in greater detail readers' use of prompts, we devised a second study. All 144 prompted cases from sets 1 and 2 used in the first investigation were employed. In these sets, the ill-defined lesion detection algorithm prompted 155 features, and the microcalcification detection algorithm 188 in a total of 144 cases. A modified prompt form was devised to capture rating information and subjects' classification of each prompted feature. Three screening radiologists were recruited from a Scottish breast screening centre as subjects. Each subject examined the entire series of mammograms individually over two sessions, with 74 cases reported in the first session, and 70 in the second. The appropriate copy films were made available, and arranged sequentially on a Rad X style viewer. Subjects were asked to rate each of the prompted features in the following ways:

1. Indicate whether the prompt would be acceptable in a screening environment (yes/no).
2. Rate the prompt as 'useful' to 'distracting' on a five point scale.
3. To state whether they would recommend recall on the basis of the prompted feature (yes/no).
4. To classify each prompted feature (free text response).
5. To rate the significance of each feature on a five point confidence scale: C1 normal, C2 benign, C3 equivocal, C4 suspicious, C5 malignant.

In addition to supplying details about prompted features, subjects were also asked to annotate and describe additional features in the mammogram if they felt that prompting for that feature would be useful. They were also asked to rate these additional prompts in the same way as the actual prompts. Finally, subjects were encouraged to give a verbal commentary on both their interpretation of the mammograms in the test set and of the system's response using a 'think aloud' protocol. Subjects' commentary was tape recorded and transcribed. In the following, transcript extracts are labelled according to the subject (H, J or R), the session (1 or 2) and the case examined (1-74 or 1-70). Thus the label (H-1.23) identifies the abstract as belonging to subject H reading case 23 in her first session.

The results we present here focus on the question of how subjects used prompts as an error management tool, and on how they made sense of system behaviour. They show how even FP prompts may serve a useful purpose for the reader.

Accountability: Mammograms are information rich artefacts that are examined by trained observers for abnormalities that sometimes have a very subtle presentation. Given the reading workload generated by the breast screening programme, it would be impossible for readers to approach their task by exhaustively examining and analysing each part of each mammogram. Moreover, readers' attention is a limited resource, and human observers are prone to fatigue. Studies have shown that visual search in radiology is often incomplete (ref. 18) and that experienced readers are able to quickly 'zero in' on significant features (ref. 19), attending to each mammogram in a way that is dependent on content.

Mammograms can be more or less difficult to interpret for a number of different reasons, including variations in tissue type, tissue distribution, and the effectiveness of the image acquisition process. Similarly, there can be variation in the degree of difficulty associated with the interpretation of individual features. Some features may be obviously benign or malignant, others may be ambiguous because either they are in the early stages of

[3] Features that have some properties in common with those readers interpret as suspicious.

development, or because they are imperfectly represented within the image. Thus it is not necessary to attend equally to every feature within the mammogram -- some can be cursorily dismissed, others require more protracted thought and examination.

The approach taken by readers involves selectivity in the application of effort to produce an acceptable level of performance under particular resource constraints. Selectivity is mediated by heuristics for deciding what is 'worthy' of examination and in what detail. Readers may expend greater effort in examining dense breasts, and may examine regions in detail with the aid of a magnifying glass. They will examine closely features that 'catch their eye' and, depending on the characteristics of the presentation, may resort to other information sources, such as additional views or previous films, or to particular strategies, such as 'undressing' lesions, or use of a bright light source. They also may pay greater attention to particular regions of the breast known to be sites where cancers can be missed -- the so-called 'danger areas'. In short, readers employ a more strategic approach to render the reading task tractable.

Readers do not have to account for every feature within an image, but they do have to account for features that satisfy generally accepted heuristics for significance. Readers also have to account for features in a particular way -- that is, according to the most appropriate strategy for analysing a given type of presentation. Thus accountability to the process is bound by what an experienced reader might reasonably be expected to notice, the lengths that they might be reasonably expected to go to establish the status of some noticed feature, and by what analytic strategies it might be most reasonable to select given the type of presentation. Accountability demonstrates an approach to the management of selective attention by driving a continual series of reflections about courses of actions available and the certainty of any conclusions. The end results include both a decision and its rationale.

On several occasions subjects appeared to be using PROMAM to maintain their accountability to specific presentations:

> "So it's brought ... for some reason it's decided on that one, but ... I suppose it's valuable in that it makes you look a bit more closely at it. But I think it's breast tissue and I would not be bringing the women back -- I don't think that it's unreasonable to prompt it." (J-1.42)

> "So I think that is useful to prompt. If at the end of the day if we then analyse and say well that's benign that's fair enough. But that's useful to have brought to your attention." (H-2.33)

In each of these cases the prompted feature was judged to be benign and had no influence on the final decision. However, subjects believed the prompt had served a useful function by 'bringing [it] to their attention' and by doing so encouraging a 'closer' or 'proper' inspection. In so doing, it is possible that prompting has a psychological benefit by reducing the anxiety a reader may have about whether a thorough visual search has been made. Prompting may also improve readers' capacity for self-awareness and reflection. When confronted by a prompt it is natural for a reader to reflect on whether they saw the feature, how much attention they gave to it, and the interpretation they reached. One benefit of encouraging reflection might be to prevent the reader from too hastily narrowing the possible interpretations that might be attributed to the mammogram.

Overall, subjects' views on the reasonableness of prompts appear to be highly contingent and dependent on interrelated factors. In addition to the importance of a feature's character as an indicator of suspicion, the context of its presentation also plays a role in determining how much effort should be invested in its investigation. Thus prompts may be judged reasonable because they attend to contextual considerations, sometimes to the extent that they may be judged reasonable even where the feature prompted has little or no significance.

Context: One of the ways readers' orientate themselves to their task involves attending to perceived shortcomings in their abilities. Film readers are aware that there are regions within the breast where lesions are more often missed, the so called ‘review areas’, and thus may pay greater attention to these regions, may imbue lesions presenting in these regions with a greater degree of significance:

> “I think we probably would recall on that, so it’s an asymmetry in a review area -- therefore it’s a bit more sinister, it will probably be nothing but it’s an area we would want to see.” (H-1.25)

Subjects occasionally judged the reasonableness of prompts against the criteria of location. Subject H was particularly keen for the system to prompt for features in the review areas:

> “What I think it would be useful to prompt is this asymmetry up here in the left. Erm, I’ll circle this area up here -- the reason I think that’s useful is although you get a lot of normal asymmetries up there, its also a common site, or a relatively common site of cancers.” (H-1.3)

In this case, subject H draws attention to her detailed, region specific, knowledge. She demonstrates a sensitivity to the importance of examining a specific region for abnormal presentations, and also to the possibility that in doing so there may be a danger of misinterpretation.

> “Well sometimes we see wee cancers down there, and it’s just, it’s just you know the sort of thing you just, you (don’t?) attend, I’m not saying you don’t look, it’s the kind of thing you can miss, because it’s just at the edge of your field of vision as it were, and I’ve seen a few missed there just when they’ve just been at the lower, at the infra-mammary fold. And that’s not one, but I mean it’s perfectly reasonable to be prompted to have a second look at it.” (R-2.7)

Here subject R suggests that it is possible to miss features that present in a particular region of the mammogram. He suggests that the reason for this is not that readers ‘don’t look’, but because it is in an area that they may not be inclined to attend to so readily. Subject R’s interpretation of the system’s response is particularly interesting because although the prompt is for microcalcifications, it is entirely clear that there are no microcalcifications present. The subject finds the prompt tolerable because of the effect it has in drawing his attention to region of the breast that he believes deserves attention.

One particular difficulty encountered by readers is the interpretation of dense, or patchy, breast tissue:

> “These are a nightmare when I’m doing them, because I think you could hide Moby Dick in there and not know. And these are the ones where we have a high error in that there can be opacities in there which you don’t really appreciate, and there can be some micro-calc which you don’t appreciate.” (J-2.23)

Dense breast tissue complicates the task of interpretation and readers are aware that their judgements may be less reliable. In addition to the anxiety associated with the possibility of misinterpretation, there may be also a wider professional concern that the decision that a breast is normal is made with less confidence, and thus may not have the same significance for a woman with lucent breasts.

Subjects also commented that a prompting system might have a useful role to play in addressing this difficulty:

> “I know that when we read them our sensitivity and specificity goes down the denser the breast, and you would hope that the computer would be able to iron that out and have the same sensitivity.” (H-2.23)

It may be that the prompting system’s value is perceived as being dependent on specific weaknesses readers identify in their own abilities. For example, more notice may be taken of, and greater significance attributed to, the presence or absence of a prompt in dense breasts as opposed to lucent breasts.

In the next case, subject H attends to breasts rich with features that have suspicious characteristics, and is thus faced with the problem of differentiating between these confusing, attention grabbing, benign presentations and any actual malignancies. If there was only a single presentation of this type, then resources could be efficiently, and less ambiguously, allocated to its consideration. In the case of multiple presentations, additional effort is required to organise how the lesions might be considered:

> "The area that have been prompted erm, are up here -- now that has been quite useful because ... in a breast like that they are difficult to assess because it's so patchy and you can imagine asymmetries all over the place -- what the prompt has made me do is go back and look particularly at that one -- I think that's actually quite useful -- I don't think it's worrying, but out of all the patches that are in front of me it's said look again at these two -- and that's quite useful, I think." (H-1.9)

Subject H uses the prompt to focus her analysis in a situation where there are many regions in the breast demanding attention. In doing so, she makes a tacit assumption that prompted features are more likely to be more significant than unprompted features. In the following extract, subject J demonstrates greater tolerance to prompts for vascular calcifications because of the occurrence of a prompt for a suspicious cluster:

> "This is for micro-calc. Oh (...?), this lady's got a cluster, a cluster of micro-calc on the left -- which is A, and on the right -- that's B. And let's see what C is ... Now, I think C is, I'm looking at the diagram, I think C is actually vascular but B is definitely not and none of it is ... or it's not definitely not, it's probably not. And neither is it on the right. So A and B are micro-calc, which actually look ... and I would be recalling. C I think -- probably vascular. And I wouldn't recall for that. Definitely helpful. In this case actually it's not distracting, so it's helpful to look (...?). In the situation where they've got other clusters, then of course this could be another cluster of the same. So I'm going to give it a C4, (...?) C3 (...?)."

The presence of suspicious clusters appears to heighten subject J's alertness towards microcalcifications more generally. She is pleased to have her attention drawn towards other instances of calcification so that they might be accounted for.

<u>Accounting for prompts:</u> For readers to benefit from using a prompting system, they must be confident that it is capable of detecting cancers, and that the system's specificity is sufficient to make its use worthwhile.

> "Now, there's one prompt that's been put all the round the left breast. And there's nothing there -- breast tissue. Deciding that I should look again and make sure there's not a mass, but -- very slightly different projection from the right, and it's breast tissue, and I would not be bringing this lady back." (J-1.40)

In the extract above, subject J makes a point of re-examining the prompted regions to confirm her initial analysis. She takes reasonable steps to ensure that the system has not detected something that the she did not initially apprehend and finally identifies a characteristic of the mammogram as a possible reason for that region being prompted.

> "Now looking at the other bit, that's what caught my eye to start with, it's gone for another area here, sort of (...?) oblique linear. I think that's breast tissue, I would be recalling the women anyway so I will see what it's like." (J-1.61)

Subject J entertains the hypothesis that the prompted feature may be significant, although she doubts this. She maintains her accountability to the prompt by suggesting that she should investigate the region to the limit allowed by current circumstances, and she is initiating an act that will further her understanding of the capabilities of the system.

If a reason for a prompt is not readily apparent, then this can pose problems. Readers are aware that minimal signs of cancer can be overlooked or misinterpreted. Thus using the system is not a simple matter of examining, or re-examining, the prompted region for signs of cancer: the prompt itself demands interpretation. A plausible explanation for the presence of the prompt, in terms of both image properties and system behaviour, has to be sought. The following extracts demonstrate how the interpretation of FP prompts for 'subtle' features can be problematic:

> "Right, so A -- I can't really see -- so well should I be saying, 'Oh, there's calcium there -- recall the patient' -- and obviously I'm overriding this (thing) -- can't see it -- you know, it can't be that worrying." (H-1.65)

The prompt in case 1.65 presents subject H with a dilemma: has the system detected something significant that she cannot herself see? Her discomfort is in part due to the lack of an obvious cause for the prompt that can be used to account for its presence -- there is no good reason for discounting the prompt other than that she cannot see 'what it is for'.

A prompt does not, in itself, indicate what it is for, other than in the broadest sense of being produced by either the microcalcification or for an ill-defined lesion algorithm. It simply highlights a region for examination by a film reader. Thus the onus is on the reader to discover a rationale for the prompt. This process can be time consuming and inconclusive without an understanding of how the feature detection algorithms work -- often a rationale is not obvious from the examination of the prompted region alone. We will return to this issue later.

<u>Influencing interpretation:</u> A prompt should not increase a reader's suspicion of a feature simply because of its presence, but there were a number of occasions where subjects reported that their interpretation of a feature was affected by the presence of a prompt:

> "So it wouldn't be unreasonable at all to bring this woman back and I probably ... with the prompt I probably ... it would make me think 'yeah maybe we should get reviews on this'. That's probably nothing though. So I think that's acceptable and useful." (J-1.36)

Common to subject J's appraisal of these cases is the influence of the prompt on her decision -- the prompt is 'making her think' about recalling, or she is recalling 'because of the prompt', for features that in all probability are 'nothing'. Subject R also reports heightened suspicion due to the presence of a prompt:

> "(...?) That's fair enough to make you look more closely at that particular area, maybe (...?) that's quite useful actually. Would you recall having been prompted to it? I think that once I had been prompted to it I probably would recall it, it's a bit like seeing it as a second reader. If you saw it the first time you might let it go, but if someone has seen it before you wouldn't let it go, so I think we would recall it." (R-2.40)

The features considered for recall by subjects J and R in the above cases all appear to have borderline significance -- they fall on or around the readers' recall thresholds. Readers face a dilemma because they know that some cancers will present minimal signs on the mammogram, but recalling for all features presenting a minor degree of suspicion would overwhelm resources available for assessment clinics. In case 2.40, subject R demonstrates an approach to managing recall decisions where the evidence from the image alone is ambiguous by using the decision made by a first reader as an additional source of evidence. Similarly, he suggests that the presence of a prompt could be used as evidence of abnormality for ambiguous cases. Conversely, subject H suggests that the lack of a prompt can be significant:

> "It hasn't really picked up on the asymmetries but they're not worrying in any way ... would I rather it prompted or didn't? I don't seem to be very consistent do I? Because on the one hand you've got the comfort factor -- oh it's seen it and dismissed it. I think they're not in anyway worrying, if they were more striking asymmetries then perhaps I would want it -- in that case I think I would let them go." (H-1.33)

Here the lack of a prompt is seen as 'comforting', precisely because subject H equates the lack of a prompt as indicating that the system has assessed a region and found it to be benign.

<u>Making sense of PROMAM:</u> In subjects' repeated attempts to understand what a particular prompt 'means' and how it should be 'interpreted', we find evidence of their active engagement in making sense of the system's behaviour. In this way, subjects develop the capacity to assess PROMAM's capabilities -- what it might reliably detect, and what might be overlooked, and also to explain its responses.

Subjects' strategies for making sense of the system included:

1. Comparing the system's responses for similar types of feature.
2. Comparing their ideas of significance with the systems.
3. Assuming purposeful behaviour.
4. Considering what might be indicated by the shape, size and location of prompts.

Often subjects made comparisons between the system's response for similar types of feature, either in the same breast, or between cases. Subjects often found the tendency of the microcalcification detection algorithm not to prompt all benign microcalcifications to be a source of confusion:

> "I'm surprised that it hasn't, that it hasn't picked up on the vascular calcification on the right. (...?) really quite surprised about that, since it's gone for things on the left." (J-1.39)

Similarly, subjects often attempted to make sense of the ill-defined lesion algorithm by making comparisons between prompted and unprompted features:

> "I'm kind of struggling to see what they are prompting for. It's just asymmetrical breast tissue. If it prompted for that -- why did it not prompt for that. So I'm writing 'why not prompted?' because it's more of the same -- (plus)? a bigger area." (H-1.64)

Subjects expect the system to be consistent and so are puzzled when it doesn't prompt for features that are to them (diagnostically) similar to prompted features. This is partly due to their lack of familiarity with the details of system behaviour and also because of the effect on the system of variations in image properties that may seem insignificant, or are difficult for a reader to perceive.

This can be seen most clearly in the operation of the micro-calcification detection algorithm where a simple clustering rule is the system's criteria for prompting. In case 1.39 (above), subject J expresses confusion because of the system's inconsistent response to seemingly similar regions of benign vascular calcification. In its early stages, vascular calcification can be discontinuous or fragmented, and it is this type of presentation that satisfies the algorithm's simple clustering rule. Perceived inconsistencies in prompting may sometimes stimulate a search for more sophisticated explanations of system behaviour:

> "There's a vessel running down there -- and isn't that strange? Well this is it again because we've got other bits of vascular calcification which it hasn't prompted on the same vessel with it coming down here, and that's the bit it's gone and highlighted, I don't know why. So that it is a bit of a cause for concern I think. Just why has it gone for that bit, is it because it's in the bit of black breast ... you know, fat, that's standing out a wee bit more." (J-1.59)

In the above case subject J is able to identify 'low level' differences between prompted and unprompted features -- improved contrast between calcifications and background tissue. In another example, subject J was able to account for the tendency of the ill-defined lesion detection algorithm to produce FP prompts for a "linear increase in density":

> "Now this is another area, it seems to pick up areas like this of linear increase in density which it is calling a mass, I'm sure it's not. It's just the way the breast tissue has involuted. We're left with fibrous strands and just vaguely increased density." (J-1.24)

Much of subjects' sense-making is driven by the assumption that not only is there some reason for the presence of a prompt, but also that there is some *good* (i.e., diagnostically relevant) reason:

> "Why has it prompted that lymph node, and not others, I wonder? [...] Because I mean if it's ... if there's some particular reason it's because if it's margins or something like that, and that's fair enough, just to make sure that it is a lymph node, but if it's going to pick up every lymph node then it's completely unacceptable, it would prompt every second film just about. But it hasn't been doing that, so there must be a reason why it's prompted that, so I'll say that's Ok. But I'm happy (...?) it's a lymph node. Some women will get cancers there as well which, I suppose." (R-2.25)

However, an assumption of purposeful behaviour can be misleading. A striking example of this is where subjects associate ill-defined lesion prompts with asymmetries:

> "... an elliptical prompt there round something that I'm sure is breast tissue -- composite. [break] ... (done?) the same again. But I mean, I suppose, looking at it, there isn't an equivalent area over here, so it's reasonable enough to have prompted that. But I wouldn't have recalled it for that." (J-2.31)

In fact, prompts for asymmetry are chance occurrences. Breasts are naturally asymmetric, and the ill-defined lesion algorithm will tend to produce FPs on denser patches of tissue, which may just happen to correspond to regions of differential brightness or distribution.

Memory for both prompts previously encountered, and candidate explanations for those prompts, accumulate as part of a compiled biography of the system's behaviour that may be used to account for current prompts. This working understanding of the system's behaviour is subject to incremental -- or sometimes radical -- revision as the reader is exposed to more evidence.

The form of the prompts, and their relationship to the prompted feature, were deliberately chosen to deliver certain types of information. Prompts from the system components are distinctive, allowing the responses made by the microcalcifications and ill-defined lesion detection algorithms to be easily distinguished. Prompts produced by the micro-calcification algorithm delineate the shape of the cluster that has been detected. Similarly, ill-defined lesion prompts consist of an ellipse that circumscribes the detected lesion with an additional margin of ten percent by area. However, subjects suggested that they might learn to recover additional information from prompt characteristics:

> "Multiple prompts on this one. They're all micro-calc. I'm not dismissing them out of hand, but just even looking at the prompt, at the way it's outlined on here, it looks like vascular calcification, and indeed that's what it looks like on first looking at the film." (J-2.11)

Subjects' responses to case 1.61 provide an interesting example of how subjects' interpretation of the meaning of a prompt can differ. The system has produced two ill-defined lesion prompts. Prompt B circles a large region of dense tissue in the upper part of the right breast. Prompt C highlights a smaller region and lies wholly within the region circumscribed by B.

> "Now, there is an asymmetry -- but I wonder if there is a cyst in the middle of all that. So I would recall this patient. Now, what they've done is they've prompted the whole area -- which I don't think has been helpful. That bit ... B, the whole area, not helpful (...) C, again it's, I think it's not the right area -- so that's not helpful." (H-1.61)

> "... so B, an increase in density and it does merit ... it's helpful ... it will be relevant -- because I can't see the margins of it, that's fine, it's suspicious. Now looking at the other bit, that's what caught my eye to start with, it's gone for another area here, sort of (...?) oblique linear. I think that's breast tissue, I would be recalling the woman anyway so I will see what it's like, but if that wasn't there, then I wouldn't be impressed by that ... C is in the middle of this bit. I'm not quite sure whether it thinks there is a second mass within ... I don't quite know what ... whether it thinks there is a separate mass, because if I was drawing a line I would draw it round here ... So I'm not quite sure what it's getting at." (J-1.61)

Subject H identifies a significant feature within the region highlighted by prompt B, but doubts its relevance because it prompts too wide an area. Subject H also dismisses prompt C as being for the wrong area and proceeds to annotate the extent of the feature as she sees it. In contrast, subject J feels that prompt B is relevant, but has difficulty finding an adequate explanation for prompt C. Prompt C poses a problem for subject J because it both misses the focal region of significance, and occurs within the region of prompt B, which is seen as significant. This poses several questions simultaneously: If the system has identified the region annotated by H as significant (although by a wide margin), why has C been prompted? If a small focal area such as C can be prompted, why is B not more focal? If the entire region given by B is significant, then why bother prompting smaller regions within B at all?

Such questions can be settled by an understanding of the ill-defined lesion algorithm. Processing is essentially done in two stages. In the first stage features within the mammogram are extracted and segmented according to four different scale sizes. This can be thought of as a sieving operation which allows features falling into broad categories of size to be treated separately. Features conforming to each of these sizes are then classified according to known properties of malignant lesions. Each of these steps is independent and the prompts generated are completely independent of one another. The regions highlighted by prompts B and C thus belong to different scale sizes and their significances are unrelated.

Summary: The results of this second study confirm readers' tolerance of FP prompts and provide interesting pointers as to how readers may use prompts as an error management tool in the clinical setting. They also show that readers began to develop quite a detailed understanding of PROMAM as evidence of its behaviour accumulated.

This result might be purely an effect of the study, and may not be relevant to PROMAM's clinical use. Notionally, the user of a prompting system does not need to know how it works, merely to attend to those areas in the image that it prompts. However, our data suggests that readers benefit from investing effort in making sense of the PROMAM's behaviour because this facilitates more efficient and effective use.

The PROMAM system is designed as an attention cue -- its role is to ensure that features with possibly malignant characteristics are not overlooked by a human observer. The presence of a prompt should merely imply that attention is required (because the system is sensitive), but should not imply that a recall decision is appropriate (because the system is not very specific). The responsibility for assessing the significance of a prompted feature, and thus for making a recall decision, should rest entirely with the reader. However, the data shows that sometimes readers were influenced in this way, so using the system in an unintended manner.

Small Scale Clinical Evaluation of PROMAM

To investigate whether the issues raised by the first two studies were relevant to understanding PROMAM's use in the clinical setting, they were followed up by a small-scale clinical trial.

Five subjects were recruited from radiologists at a Scottish breast screening centre. Two thousand and two archive cases (including 102 pathology proven cancers) were digitised and analysed by the PROMAM system (ref. 26). The films were then divided into twenty sets of approximately one hundred films and double read, once by a subject in a prompted condition and once by a subject unprompted. In the prompted conditions, subjects were asked to first examine the films, then examine the prompt sheet, and then to record their decision. Subjects were given an overview of how the system worked, including the types of feature it was capable of detecting, and trained in its use. Training included specific instructions regarding how to use PROMAM, i.e., that prompts were simply attention cues with no classification significance. A full account of methodology and results can be found elsewhere (refs. 11, 26).

Outputs from two of PROMAM's feature detection algorithms were used to generate prompts for microcalcification clusters (ref. 12) and ill-defined lesions (ref. 17). Representative film sets were selected at random from four average days' screening at one clinic and balanced with respect to number of recalled cases, density of breast tissue and modularity.

Data collection methods included observation of all the experimental sessions. Subjects were interviewed and asked to complete a questionnaire immediately following the prompted sessions; the interviews were tape recorded and subsequently transcribed. Further questionnaires were administered prior to starting the trial, and after each subject had completed all their allocated sessions.

Dealing with FP prompts: Ideally, readers should give all prompts equal consideration, and only dismiss prompts after careful examination of the prompted region on the mammogram. However, interview data indicates that subjects developed strategies to determine the significance of system information based on an *a priori* assessment of the prompt sheet. For example, one subject indicated that the shape of prompts for vascular calcifications, and the location of prompts for ill-defined lesions, could give a clue as to their cause:

> "I think now you'll start dismissing masses at the back, you're dismissing the calcification at the back and maybe you don't look as (...?) carefully as maybe -- you do look carefully but maybe not to the same degree when you clearly see that it is vascular calcification it's prompting on."

Another indicated that some prompts could be assessed from her examination of the prompt sheet alone:

> "I mean, if it's the one particularly along the edge of the pectoral and the bottom, lower, inner aspects, yes ... then the vascular calcification is one (...?) those are very obvious, yes."

These comments indicate that subjects learnt to recognise patterns in shape, frequency and location that characterise FP prompts, and used this to determine how much effort they invested in further scrutiny of the mammogram. In such cases, consideration of possible explanations is not deferred until all the evidence has been gathered (ref. 8). Two subjects, for example, indicated that they might not look back as carefully -- or at all -- depending on their initial assessment.

Predicting prompts: Subjects reported that they were able to develop quite quickly a capacity to predict which features in the mammogram would be prompted.

> "I find myself sometimes thinking 'well, I bet it's going to prompt for that'. Erm, and that actually makes it easier, if the prompt is there then I can forget about that straight away. But sometimes, when it prompts something out of the blue, then there is nothing you can do ... [I think I know what it's going to prompt for] about 50% of the time."

One subject volunteered an explanation for why predicting prompts was useful:

> "At times I'm definitely anticipating that that's going to be prompted. And sort of already decide I'm not going to look at it again almost, you know, you're kind of expecting prompts on certain things so I think you sort of ... very quickly dismiss it as (harmless?) without looking again."

There is a cognitive cost associated with this strategy as it requires that readers must form a more accurate model of system behaviour. However, checking whether prompts meet with expectations appears to be an intuitive reaction for readers, and is perhaps essential for establishing and maintaining trust in system performance. We argue also that prediction is a valuable strategy because it implies that the reader has actually made an assessment based on the evidence in the mammogram.

The success of prediction is dependent upon prompting system consistency as *perceived* by the reader. Image analysis algorithms can be sensitive to variations in appearance which are too subtle for the radiologist to appreciate without close examination -- if at all. Though system behaviour may be *strictly* deterministic, it may not be *observably* deterministic if it doesn't respond in the same way to features that readers would classify as being similar.

Impact on decision-making: In each of the post-prompted session interviews, subjects were asked if the prompts had had some influence on their recall decisions. Out of a total of sixteen interviews held after prompted sessions, subjects indicated that their recall decisions had been affected *one or more* times in a total of eleven of those sessions. Subjects reported a number of occasions where the prompts had drawn significant features to their attention which they had overlooked, sometimes resulting in a recall decision.

Despite making instructions given in pre-trial training, both questionnaire data and responses given in post-session interviews indicate that subjects sometimes used prompts as classification decisions aids. Subjects referred to occasions where they had found the absence of a prompt 'reassuring':

> "Yes, yes, I think that that is reassuring. It might just be falsely reassuring sometimes."

The quote above indicates that the absence of a prompt was viewed as 'reassuring' only, merely confirming a decision that had already been made. However, subjects also reported cases where the presence of a prompt had made them more inclined to recall:

> "There was one where I was undecided, and it was prompted ... 'I will bring it back, yes' ... otherwise I probably would have said 'oh,

> forget it', whether that's right or not I don't know."

Overall, subjects' comments suggested that the presence or absence of a prompt was most likely to influence a decision when the evidence available from the mammogram alone was ambiguous. It is possible that in these situations subjects attempted to use whatever evidence was to hand, including prompts, to resolve uncertainty:

> "Maybe it was highlighting something that I wasn't seeing in a dense breast, so that's why it needed confirmed. Erm ... I (...?) with it you go with the prompt."

One subject drew an analogy between prompts and heightened suspicion when another radiologist asks her to examine a case:

> "... it's like when someone shows sets of mammograms and they'll say, you know, it's always nice for someone not to say, point out what they are worried about, because if you do, then immediately you heightened suspicion because someone else is suspicious about it."

Summary: The results of this third study reveal how readers adapted to using PROMAM in the clinical setting. They revealed that subjects spontaneously improvised their use of prompts in a number of ways which helped them economise on the effort required to deal with prompt system errors. First, they began to apply strategies for determining the significance of prompts based on prompt -- rather than image -- features. Second, they began to actively predict where prompts were likely to appear.

The study also confirmed that not only did subjects use prompts as attention cues, but as decision-making aids when other available evidence was ambiguous.

Conclusions

The use of prompting systems in breast screening is intended to reduce observer error by helping readers to avoid errors of attention. While unambiguous, quantitative evidence of performance improvement must necessarily await the outcome of large scale clinical trials, our studies provide some support for the achievement of this goal. They also suggest that readers can learn how to manage the potentially undesirable effects of prompting system errors on their own performance. However, we also find that presumptions of prompting systems being used purely and simply as attention cues may be misplaced.

One problem is that the design rationale for prompting systems assumes generic difficulties -- i.e., that readers sometimes have difficulty ensuring that the entire mammogram is examined. However, the problems readers face when examining a set of mammograms are actually very specific and highly contingent. For example, the reading of dense, or feature rich, breasts poses demands very different from those posed by lucent, or uncomplicated, breasts. Furthermore, although readers have general concerns that they might, for example, overlook a malignancy, they also have a more specific understanding of particular deficiencies in their expertise. For example, they might perceive themselves to be more or less able to detect and correctly classify particular feature types.

It would be a mistake to believe that error-free and effective use of prompting systems in breast screening can be achieved if the user is expected to treat the system as a mere "black box", with no understanding of how it behaves, even if the user is a highly skilled reader. On the contrary, our studies show that efficient and accurate use of prompting systems depend on the system's behaviour being accountable to its users. Readers maintain accountability of their own work in the context of an understanding of their performance characteristics (knowledge about their skills, limitations and expected behaviours). This is often a poor model for accounting for prompt system behaviour, especially where erroneous prompts are simply artefacts of the methods used to analyse the image.

One solution may be training. Prior to the small scale clinical trial of PROMAM we sought to provide an account of its behaviour through training. However, the training material (though informed by the earlier studies) has still to mature. Developing and evaluating training materials and protocols may prove to be as complex as the development of the system itself.

To address the issue of how prompting systems *should* be used by readers, we may learn from the preparation and use of evidence in current reading practice. For example, readers often organise the ordering of attending to evidence to minimise bias. Though we enforced a protocol in the small scale clinical trial whereby subjects examined each mammogram before examining the prompt sheet, our evidence suggests PROMAM still influenced classification decisions. Further innovations in reading protocol might be appropriate, such as requiring readers to reach a decision before examining prompts, and only allowing 'routine recall' decisions to be amended in light of evidence from the system.

Finally, the goal of replacing double reading with a single reader aided by a prompting system may raise wider problems concerning reading practices. Our studies show that readers may exploit of double reading as a way to monitor their performance. There is an informal, collaborative dimension to double reading, and to readers' management of their performance, which has so far been largely ignored. We argue that the implications of this for single, prompted reading require careful consideration.

References

1. E D C Anderson, B B Muir and A E Kirkpatrick, The efficacy of double reading mammograms in breast screening, Clinical Radiology, 49, 248-251, 1994.

2. J Chamberlain, S M Moss, A E Kirkpatrick, M Michell and L Johns, National Health Service breast screening programme results for 1991-2, British Medical Journal, 307, 353-356, 1993.

3. S Ciatto, M R D Turco, D Morrone, S Catarzi, D Ambrogetti, A Cariddi and M Zappa, Independent double reading of screening mammograms, Journal of Medical Screening, 2, 99-101, 1995.

4. H E Deans, D Everington, A E Kirkpatrick and E Lindsay, Scottish experience of double reading in the National Breast Screening Programme, The Breast, 7, 75-79, 1998.

5. E R E Denton and S Field, Just how valuable is double reporting in screening mammography?, Clinical Radiology, 52, 466-468, 1997.

6. J Fox, Decision-support systems as safety-critical components: Towards a safety culture for medical informatics. Methods of Information in Medicine, 32, 345-348, 1993.

7. M Giger, Computer-aided diagnosis. In A Haus and M Yaffe, (Eds.) A categorical course in physics: Technical aspects of breast imaging, 283-298. RSNA, 1993.

8. A Gale, Human response to visual stimuli. In W Hendee and P Wells, (Eds.) The perception of visual information, 115-130. Springer-Verlag, 1995.

9. M Hartswood, R Procter and L Williams, Subjective reaction to prompting in screening mammography. In C Taylor, (Ed.) Proceedings of Medical Image Analysis and Understanding '97, Oxford, 1997.

10. M Hartswood, R Procter, L Williams, R Prescott and P Dixon, Drawing the line between perception and interpretation in computer-aided mammography. In E Fallon, L Bannon and J McCarthy, (Eds.) Proceedings of the 1st International Conference on Allocation of Functions, 275-290. Dublin: IEA Press, 1997.

11. M Hartswood, R Procter and L Williams, Prompting in practice: How can we ensure radiologists make best use of computer-aided detection systems in screening mammography?. In N Karssemejer, (Ed.) Proceedings of the 4th International Workshop on Digital Mammography, 1998.

12. A Hume, P Thanisch, M Hartswood and R Procter, On the evaluation of microcalcification detection algorithms. In Proceedings of the 3rd International Workshop on Digital

Mammography, Chicago, 1996.

13. I Hutt, The Computer-Aided Detection of Abnormalities in Digital Mammograms. Unpublished Ph.D. Thesis. Manchester University, 1996.

14. H Kundel, C Nodine and D Carmody, Visual Scanning, Pattern Recognition and Decision-making in Pulmonary Nodule Detection. Investigative Radiology, vol 13, 175-181, 1978.

15. A Lesgold, R Glaser, H Rubinson, D Knopfer, P Feltovich and Y Wang, Expertise in a complex skill: Diagnosing X-Ray pictures. In M Chi, R Glaser and M .J Farr, (Eds.) The Nature of Expertise. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1988.

16. C E Metz and J Shen, Gains in accuracy from replicated readings of diagnostic images: prediction and assessment in terms of ROC analysis, Medical Decision Making, 12, 60-75, 1992.

17. L Miller and N Ramsay, The detection of malignant masses by non-linear multiscale analysis. In Proceedings of the 3rd International Workshop on Digital Mammography. Chicago, 1996.

18. C F Nodine and H L Kundel, Eye Movements: From Physiology to Cognition. In J K O'Regan and A Levi-Shoen (Eds.) The cognitive side of visual search in radiology. Elsevier Science Publishers, 1987.

19. C F Nodine, H L Kundel, S C Lauver and L C Toto, Nature of expertise in searching mammograms for breast masses, Academic Radiology, 3, 12, 1000-1006, 1996.

20. R Procter, P Thanisch, A Hume, S Astley and A Kirkpatrick, User interface design and data management for digital mass mammography. In A G Gale, S M Astley, D R Dance and A Y Cairns, (Eds.) Proceedings of the 2nd International Journal of Digital Mammography, 405-414. Elsevier, 1994.

21. C Savage, A Gale, E Pawley and A R Wilson,.To err is human to compute divine? In A G Gale, S M Astley, D R Dance and A Y Cairns, (Eds.) Proceedings of the 2nd International Journal of Digital Mammography, pp. 405-414. Elsevier, 1994.

22. E L Thurfjell, K A Lerneval and A S Taube, Benefit of independent double reading in a population-based mammography screening programme, Radiology, 191, 241-244, 1994.

23. N J Wald, P Murphy, P Major, C Parkes, J Townsend and C Frost, UKCCCR multicentre randomised controlled trial of one and two view mammography in breast cancer screening, British Medical Journal, 311, November, 1189-1192, 1995.

24. R M Warren and S W Duffy, Comparison of single reading with double reading of mammograms, and change of effectiveness with experience, The British Journal of Radiology, 68, 958-962, 1995.

25. J C Wells and J Cooke, Film reading practice of the UK breast screening units, The Breast, 5, 404-409, 1996.

26. L Williams, R Prescott and M Hartswood, Computer-aided cancer detection and the UK National breast screening programme. In N Karssemejer, (Ed.) Proceedings of the 4th International Workshop on Digital Mammography, 1998.

27. L Williams, M Hartswood and R Prescott, Methodological issues in mammography double reading studies, Journal of Medical Screening, 5, 202-206, 1998.

Biography

Mark Hartswood, Institute for Communicating and Collaborative Systems, Division of Informatics, University of Edinburgh, Edinburgh EH9 3JZ, Scotland. Telephone: +44 (131) 650 5168. Facs: +44 (131) 667 7209. Email: mjh@dcs.ed.ac.uk

Mark Hartswood has spent five years working on the PROMAM project researching human factors issues in computer-aided mammography. Currently he is completing his Ph.D. and working as a Research Associate on a project applying participatory design methods in the development of an EMR system at Edinburgh Royal Infirmary.

Rob Procter, Institute for Communicating and Collaborative Systems, Division of Informatics, University of Edinburgh, Edinburgh EH9 3JZ, Scotland. Telephone: +44 (131) 650 5177. Facs: +44 (131) 667 7209. Email: rnp@dcs.ed.ac.uk

Rob Procter is a senior lecturer in the Division of Informatics, University of Edinburgh. His research interests lie principally in the field of human factors, particularly methodologies for the design and implementation of IT systems. His current research includes computer-aided mammography, participatory design in the clinical setting, and social learning processes in IT systems implementation.

Critical Incident Reporting in an Intensive Care Unit: 10 Years' Experience

David Wright,

Intensive Therapy Unit; Western General Hospital, Edinburgh

Keywords: Intensive Care Unit; Incident Reporting

Introduction

When accidents or mistakes are looked at in detail, certain key factors can be seen to have influenced events. The technique of identifying such factors – critical incidents - was developed by Flanagan (ref. 1) and others in the 1940s to help analyse aircraft accidents.

By the 1970s, critical incident studies had been used to investigate anaesthetic mishaps (ref. 2). Modelled on this work, we have been looking at critical incidents in an intensive care unit (ICU) since 1989. We have used a questionnaire which was based on which had been developed by Williamson and colleagues (ref. 3) for use in anaesthetic practice.

We defined a critical incident as an occurrence that might have led (if not discovered in time) or did lead, to an undesirable outcome. The incident must have been caused by an error made by a member of staff or by a failure of equipment. It had to be described in detail by a person involved in or who had discovered the incident. It had to have occurred while a patient was under the care of the ICU staff (though not necessarily in ICU). It had to be clearly preventable.

Methods

Such studies need a form for describing the incidents, a system for collecting the forms, classification and analysis of the incidents, a review of the incidents and action as a result of this, and feedback to staff, which includes summaries of the incidents or any action suggested to lessen the likelihood of future incidents.

The essence of such studies is anonymity and absence of criticism or blame, and these should be constantly emphasised.

Results

The study started in 1989 and our first year's experience was published in 1991 (ref. 4). By early 1999, over 100 incident forms had been completed, and this paper reviews 710 reports from the period of January 1989 to February 1999.

The majority of incidents was due to errors by staff, and was not caused by equipment failure. These human errors fell naturally into four groups: those related to ventilation, those related to vascular lines, those related to drug administration and a miscellaneous group.

When the incidents were analysed and classified, a number of contributing factors emerged. Such factors were inexperience, shortage of staff, night time, poor communication. Also, a number of factors could be seen as contributing to the detection of incidents, such as regular checking, the presence of alarms on equipment, and the presence of experienced staff.

Nurses completed more than 90% of the forms and the vast majority of incidents recorded had no serious sequelae for patients.

Discussion

Earlier studies confirm how important human error is in the generation of critical incidents (ref. 2; 3). Other studies show that most incidents reported do not have serious long-term sequelae. ICU studies have previously reported 12 (ref. 4) (ref. 5) or 24 (ref. 6) months of incidents. Our current report of 10 years of incidents allows certain points to be considered over a longer period of time. These points include variation years by year in the number of reports submitted, variation year by year in the overall groups of incidents (ventilation, vascular line, drugs, miscellaneous) and whether some incidents occur time and time again despite their being highlighted and preventive action suggested.

In addition, the type of preventive action that is suggested can be classified and analysed.

It is important to appreciate that such studies do not record the true incidence of critical incidents. What they do do is to provide a relatively simple way of drawing attention to things that have gone wrong. They allow these individuals who submit forms an opportunity to analyse their own behaviour and attitudes, and they provide a stimulus for action to try and prevent further incidents. They encourage mutual trust and a mature approach to improving quality, which enhance the way staff feel about themselves and their unit.

References

1. Flanagan, J.C., *The Critical Incident Technique.* Psychological Bulletin, 1954. **51**: p. 327-358.
2. Cooper, J.B., *et al.*, *Preventable Anaesthesia Mishaps: A Study of Human Factors.* Anaesthesiology, 1978. **49**: p. 399-406.
3. Williamson, J.A., R.K. Webb, and G.L. Pryor, *Anaesthesia Safety and the "Critical Incident" Technique.* Aust Clin Rev, 1985: p. 57-61.
4. Wright, D., *et al.*, *Critical Incidents in the Intensive Therapy Unit.* The Lancet, 1991. **338**: p. 676-78.
5. Beckman, U., *et al.*, *The Australian Incident Monitoring Scheme in IntensiveCare ICU. An Analysis of the First Year of Reporting.* Anaesth Intens Care, 1996. **24**: p. 320-329.
6. Hart, G.K., *et al.*, *Adverse Incident Reporting in Intensive Care.* Anaesth Intens Care, 1994. **22**: p. 556-561.

Biography

Dr David Wright, Intensive Therapy Unit, Western General Hospital, Edinburgh, UK. Tel +44 131 537 1666

Dr David Wright is an anaesthetist, and an ICU consultant, at the Western General Hospital, Edinburgh.

Identification and Analysis of Incidents in Complex, Medical Environments

Daniela K. Busse and Chris W. Johnson

Department of Computing Science, University of Glasgow, UK.

Abstract

Medical risk management is often seen as lagging behind other safety-critical industries, where there has been considerable research into safety and accident causation models. Accident analysis models used in, for instance, aviation and process control recognise the importance of formalised root cause analysis, and the multi-level nature of incident causation. Latent factors, such as management and organisational issues, are stressed as underlying structural precursors for incident occurrence. Also, the constraints of the human cognitive system and their relationship to task performance are taken into account. These consideration are reflected in the use of incident reporting schemes, the anaysis if the collected incident data, and the generation of remedial action recommendations. In this paper, we will illustrate how these concepts can be applied in a clinical setting. An incident reporting scheme implemented at an Edinburgh Intensive Care Unit will serve as a case study.

Introduction

Clinical Adverse Events; In 1990, the Harvard Medical Practice Study (Harvard_University, 1990) investigated the occurrence of patient injury caused by treatment - so-called adverse events. It found that nearly 4% of patients suffered an injury that prolonged their hospital stay or resulted in measurable disability. Leape (1994) pointed out that, if these rates are typical of the US, then 180000 people die each year partly as a result of iatrogenic ('doctor-caused') injury. Since most of the precursors to iatrogenic injuries are perceived to be 'Human Error' (Runciman et al., 1993), the possibility of negligence causes great concern. This is mirrored in the litigious climate in the US (Leape et al., 1991), and led to a considerable increase in interest in the causes of adverse events (Bogner, 1994). The cost of adverse events is high; not only in human suffering, but also in compensation claims and the need for prolonged treatment of afflicted patients.

Learning from Adverse Events; In the UK, the drive towards clinical effectiveness (Dawson et al., 1998) has led investigators to concentrate on increasing the quality of care while lowering the costs associated with our current health care system. In the course of the clinical effectiveness program, the idea of clinical audits has gained in strength – a "professionally led initiative which seeks to improve the outcome of patient care as a result of clinicians examining their practice and modifying it appropriately" (Wedderburn, 1998). However, the data collected for audits concerns itself only with factors peripheral to adverse events – cost effectiveness being the focus of the investigation (see e.g. Van der Schaaf, 1996). Formal analysis of the causes of adverse events does not take place. Rather, a variety of committees and commissions are typically set up locally, meeting regularly to review any cases of iatrogenic injury that had occurred and had been brought to attention by the staff involved (Leape, 1994). The reporting as well as the analysis of these events, however, are subject to local convention, and thus expresses the self-regulating policy of the health care community.

In some cases, incident reporting schemes are in place. This concerns "near miss" adverse events, i.e. cases in which iatrogenic injury was likely to have occurred, but the hazardous situation could be recovered from successfully. However, it was noted that even under the clinical reporting schemes, in-depth analysis and search for root causes of adverse events does not take place (Leape, 1994).

Human Error Analysis in Aviation and Process Control; In safety-critical domains other than health care, accidents (i.e., like clinical adverse events, non-intended events which lead to negative outcomes and loss) have received a great deal of attention. Research into their causes and prevention has made considerable advances. Accident causes formerly described solely as 'Human Error' have come under close scrutiny, notably with the work of Rasmussen (e.g. Rasmussen et al., 1987), Reason (1990), Hollnagel (1991), and Hale (Hale et al., 1997). They have systematically analysed cognitive mechanisms underlying the various phenotypes of human error. Also, latent contributing factors are taken into account. These concern organisational as well as managerial influences on the course of the accident. Thus, there was a shift from 'blaming the human' (such as the oft-cited 'pilot error') to the insight that error invariably occurs in complex systems. The aim now is to create error-tolerant systems that absorb errors through 'system defences' and provide redundancy and possibilities for error recovery. The move away from the blame culture also made possible the introduction of institutionalised, anonymous, and non-punitive incident reporting schemes. Subsequent detailed comparison and analysis of the identified root causes is carried out by organisations such as the US National Transportation Safety Board (NTSB), or the US Federal Aviation Administration (FAA) and NASA. In the clinical domain the argument has also been put forward for the use of incident reporting to complement post-hoc accident investigation with the inherent problems for scant information, altered perception and outcome bias (Runciman et al., 1993). Furthermore, analysis of the reported events should extend beyond the investigation of proximal causes to include latent system failures.

Structure of this Paper

This paper looks at incident management practices in a clinical setting, as compared to the approaches in other safety-critical domains. The main issues we will investigate in this paper concern the problems associated with incident reporting, categorisation and subsequent analysis. We will illustrate existing risk mangement in medicine by an incident reporting scheme employed in an Edinburgh Intensive Care Unit (ICU). In Section 2, we will introduce current applications of the Critical Incident Technique (Flanagan, 1954) in medicine and give an outline of the Edinburgh implementation. Section 3, 4, and 5 will look at incident reporting, categorisation, and analysis in turn, mirroring the stages of the incident investigation process. Section 6 investigates the generation of action recommendations from the prior analyses. In each section, we will compare and contrast theory and methodology 'lessons' from safety-critical domains such as aviation with the implementation of the Edinburgh incident reporting scheme. Conclusions are given in Section 7.

Incident Reporting Schemes in Medicine

Incident reporting schemes and error analysis methods in medicine are two of the lessons that can be drawn from research in domains such as aviation and process control (Leape, 1994). The US Federal Aviation Administration (FAA) established a confidential reporting system (the Aviation Safety Reporting System) for safety infractions as early as 1975. The program's success was found to depend on its non-punitive confidential approach. It also supported implemented feedback of data analyses and implemented remedial measures. The analysis of the reports and subsequent actions appear as a regular feature in several pilots' magazines to ensure a closed feedback loop.

A successful incident reporting scheme is not deemed possible with a punitive 'perfectibility model' – i.e. the belief that if staff are properly trained and motivated, there will be no mistakes (Leape, 1994). This, however, is seen to be a predominant cultural factor in the medical domain (op.cit.). The focus on motivation in assigning blame to individuals has been under heavy attack in other domains (e.g. Battles et al., 1998). Punishment of individuals as a means to prevent future adverse events is considered a dead-end approach to error management (Reason, 1997).

Thus, there is a need in medicine to recognise the inevitability of error and adverse events. Safety culture that takes this into account in clinical system design is still lacking (Leape et al., 1998). There have been some notable exceptions in the recent past where incident reporting schemes were implemented and the identified incidents analysed. Runciman et al. (Runciman et al., 1993) studied anaesthesia incidents in Australian hospitals. This study was subsequently extended to also investigate ICU incidents. Also, Battles et al. (1998) extended their incident reporting and analysis scheme from process control to fit medical systems. These are some examples of large-scale, systematic effort to transfer some of the insights and methods of other safety-critical domains to medicine. Thus, the use of incident reporting schemes and theories on accident causation are the two main contributions so far to safety applications in the medical domain.

The Edinburgh ICU Incident Reporting System

The Edinburgh incident reporting scheme was set up in an adult intensive care unit in 1989. It has been maintained by Dr David Wright, who is an anaesthesist and one of the ICU consultants. The unit has 8 beds at its disposal, and there are roughly 3 medical staff, one consultant, and up to 8 nurses per shift on the ward. Equipment in an ICU ranges from monitors displaying life sign data, such as heart rate and intra-cranial pressure (ICP), to drug administration equipment, automatic breathing machines, and oxygen humidifier masks. Patient mangagement involves tracking and transcribing monitored vital data, laying and maintaining lines such as endotracheal tubes, and chest drains, and handling equipment, such as three-way taps for drug administration, ventilators and defibrillators.

Incidents reported over ten years (see Wright, 1999) fell mainly in four task domains: relating to ventilation, vascular lines, drug administration, and a miscellaneous group.

The incident scheme employed reporting forms that encouraged staff to describe the event in narrative form, as well as noting contributing factors, detection factors, grade of staff involved in the event and that of the reporting staff.

One crucial factor in the implementation of the scheme was its anonymity. Dr Wright, in his role of the scheme manager, was the only person who had access to the completed forms. The collected data was coded into categories (see Section 4), summarised. It was then collated into frequency tables. Information that may identify staff was removed, and action recommendations were proposed. The results of this initial analysis were then iterated over by Dr Wright and the Senior Nurse of the unit. Together, the data was again inspected and final revisions of action recommendations were carried out. The conclusive results were disseminated regularly among the staff of the unit, and thus an effective feedback loop was created.

For a more detailed description of the implementation and findings of the incident reporting scheme, see (Wright, 1999).

Incident Reporting

There are several definitions of what constitutes an ‘incident’. Incidents might be considered adverse events only, near miss events only, or both. In the Edinburgh study, staff are asked to report ‘critical incidents’, which are defined as any occurrence that might have led (if not discovered in time) or did lead, to an undesirable outcome. In consequence, each recorded incident:

1. was caused by an error made by a member of staff, or by a failure of equipment
2. can be described in detail by a person who was involved in or who observed the incident
3. occurred while the patient was under our care, though […] it could be patients in transit
4. was clearly preventable

Also, complications which occur despite normal management are noted not to be critical incidents.

Notable points in this definition are the preventability of the event, and the term ‘normal management’. Also, by this definition, human involvement in the occurrence must have been characterised by an ‘error’ of staff. All these three concepts can be argued to be open for interpretation, and thus as being subject to reporting biases.

Reporting Biases

What constitutes an ‘error’ has been subject to intense debate in the accident analysis community. The Australian Intensive Care Unit Incident Monitoring (AIMS-ICU), for instance, settles for the description of an incident as "any *unintended* event or outcome […]" (our emphasis). However, Reason often noted the important role played by so-called violations, rather than error, in accident causation (Reason, 1997). As an example, work procedures that present safety barriers for infrequent adverse events but are time consuming or awkward to carry out, might frequently be omitted to suit current, seemingly more urgent, task demands. These violations might be common place, and not be considered an 'error' as such, since they were 'intended'. Management attitudes, safety culture, and factors such as time pressure and fatigue impact on the perspective taken on such violations.

Also, cases often go unreported for reasons such as fear of punishment, or not being convinced of the usefulness of the scheme. But incidents might also not be conceived by staff to be 'accidents waiting to happen' (Reason, 1990), but mere task characteristics, with 'error' detection and recovery taken for granted. It has also been put forward, that sometimes reporters are not aware of 'upstream precursors', such as underlying system faults (Chappell, 1996). They also might not appreciate the significance of local workplace factors. For instance, if they have been accustomed to working with substandard equipment, they may not report this as a contributing factor; if they habitually perform a task that should have been supervised but was not, they may not recognise the lack of supervision as a problem (Reason, 1997). The implicit nature of such tacit task and problem knowledge has been widely recognized in work practice research, and is being addresses for instance by knowledge acquisition methods, and direct observation methodologies (see for instance Jordan (1993); Woods, (1994); Billings, (1998)).

These are biases in what is *not* being reported. There are also biases in what *is* being reported. As frequently noted, the absolute incidence of incidents cannot be inferred from incident reporting data. There are many factors confounding the representativeness of the data sample obtained. The people who volunteer to report are not representative of the population at hand. In the Edinburgh study, reporters were also often not involved in the incident themselves. In the most recent documented reporting period (May-November 1998), only one third of the reporters had been involved in the incident (10 out of 27 reports). Also, there were hardly any medical staff represented among the reporters (2 out of 27).

<u>The Importance of Incident Detection Factors</u>

Noting staff grade can provide interesting information about the distribution of staff involved in the incident and those reporting it. Wright (Wright et al., 1991), in an initial presentation of the scheme's results, pointed out that it is usually experienced staff who correct the incident, and thus prevent an adverse event occurring. Establishing the factors that contribute to the incidents' detection is a highly important aspect, which is often neglected in incident reporting schemes. It has been suggested that provisions for incident detection and recovery provide more effective safety measures than an approach solely targeting accident prevention or avoidance (Rasmussen and Vincente, 1989; Reason, 1990). This mirrors the above-mentioned insight into the universality of human error. The Edinburgh study encourages staff to detail factors that contributed to the detection of the incident. Given the high rate of people reporting who detected the incident (rather than being involved), this is especially appropriate.

However, even if detection factors are noted, they are typically not being analysed in depth. The analysis should include system factors as well as cognitive aspects of the task and work environment. We will briefly illustrate this issue in Section 4.3.

<u>Incident Categorisation</u>

The relationship between accidents (adverse events) and (near miss) incidents is often represented in an 'iceberg metaphor' (Van der Schaaf et al., 1991; Battles et al., 1998). It is assumed that near miss incidents and accidents involving negative outcomes share the same root causes. In accident causation research, the categorisation and analysis of human error has received considerable attention. Several taxonomies exist, ranging from behavioural classifications (Hollnagel, 1991) to classification of error according to underlying cognitive mechanisms (Reason, 1990). Each of these presents a first step towards the analysis of an incident, since choices as to the nature and characteristics of the occurrence are made. However, behaviourist classifications such as

Hollnagel's do not aid the understanding of the underlying mechanism of erroneous actions.

Moreover, behavioural classifications quickly become large and unwieldy. Rasmussen et al. (1987) also observe, that if the analysis is based only on the consideration of human error in terms of their external manifestation (such as omission, commission, and inappropriate timing) a priori error identification will be hindered by a combinatorial explosion.

The distinction between varieties of human error according to their congitive origin plays a significant role in accident analysis because they require different methods of error management and remediation (Maddox, 1996). Rasmussen's Skill, Rule, Knowledge (SRK) framework, and Reason's Generic Error Modelling System (GEMS) have been widely applied in safety-critical domains to investigate the nature of accidents.

Incidents, according to Rasmussen (1987), should be considered occurrences of a human-system mismatch, which can only be characterised by a multifaceted description (see Figure 1). Thus, faults and errors cannot be defined objectively by considering the performance of humans or equipment in isolation. This insight needs to be reflected in the error categorisation system.

Reason (1990) pointed out that an error is not to be seen as the accident or failure itself, but as preceding the failure. 'Error' can lead to 'active failure' as well as 'latent failure'. Reason maintains that active failure is usually associated with the performance of 'front-line' operators (such as pilots and control room crews) and has an immediate impact upon the system. Latent failure is most often generated by those at the 'blunt end' of the system (designers, high-level decision-makers, managers, etc.) and may lie dormant for a long time. Rasmussen's model (Figure 1) also mirrors these hierarchy of incident causation. The 'Causes of Human Malfunction', 'Situation factors', 'Factors affecting Performance' and 'Personnel Task' can be seen in relationship to Reason's latent failures. These, together with the constraints of the human cognitive processing system, produces an 'External Mode of Malfunction', leading to an accident or incident.

Thus, the categorisation of 'causes', meaning precursors, of the incident needs to reflect the hierarchical nature of the causal chain. This chain involves latent system failures that influence work conditions, which in turn provide the context for the proximal incident cause to occur. This perspective on accident causation is mirrored in most accident causation and analysis models in e.g. the transport domain.

For instance, the NTSB accident/incident database contains sections with factual elements ("who, what, where"), sequence of events, and a narrative description of the incident. The NTSB uses a multiple-cause classification system with allowance for up to seven "occurrences", and several "findings" for each occurrence. Any of the findings may be prescribed as a cause or contributory factor. The probable cause statement is usually an integrated text. Each "finding" may, in turn, be divided into "subjects". The subjects come from a coded list and refer to whether the finding was person-related, and whether it was a proximal cause or not (NTSB, 1998). Therefore, the database allows for a causal tree being constructed by ascribing the level of proximity of the findings to the reported occurrence. Thus, it can be distinguished between causal factors and conditions facilitating the occurrence, and suggested root causes. However, the available constructs largely identify who was involved in a finding but often stop short of an assessment of why the error was made. Even the probable cause statements are largely descriptive in nature, without reference to latent failures or human information processing. This information, however, is important in developing preventive strategies (Luxhj et al., 1997).

The Edinburgh Categorisation Scheme

In the Edinburgh study, information drawn from the incident reports were categorised into 'causes', 'contributory factors', and 'detection factors' (see Wright, 1999). The categories were arrived at through informal coding of the narrative incident data. This bottom-up approach led to a domain-specific, behavioural categorisation scheme.

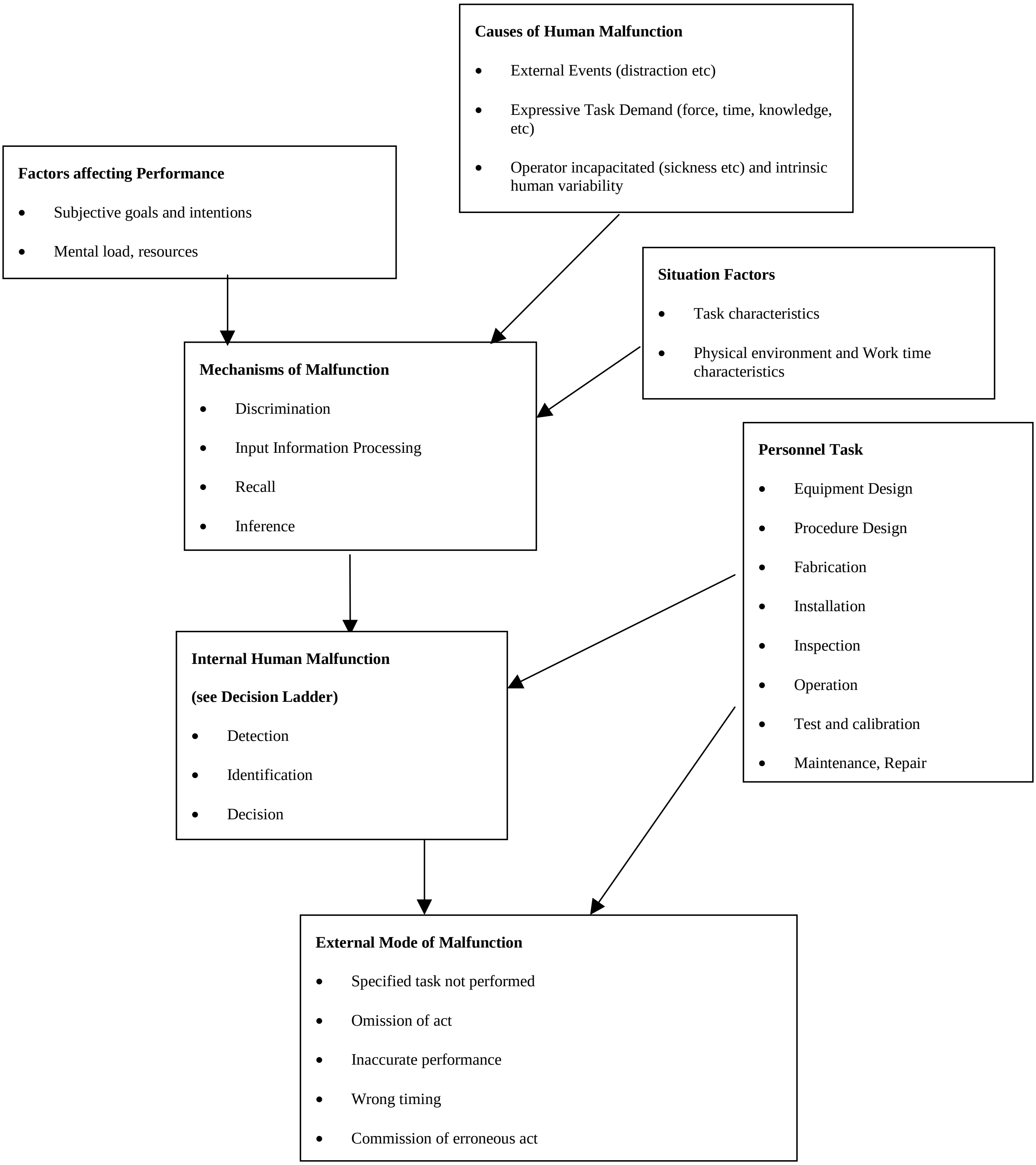


Figure 1 - Taxonomy for Description and Analysis of Events involving Human Malfunction (Rasmussen, 1982)

'Causes' offers the subcategories of Human Error and Equipment Failure. Any incident that has some degree of human involvement is considered a Human Error. Furthermore, the human error incidents are classified as to the various task and equipment domains these refer to, such as "vascular lines related", "drugs-administration-related", or "ventilator-related". Thus, the categorisation mainly labels the incidents without providing a step towards causal analysis. Rather, it points to where in the patient management task sequence the incident occurred. 'Cause' here refers to the task domain of the proximal causal factor. Mostly, the actual proximal 'cause' of the incident cannot be inferred from this categorisation per se. A summary of the narrative description of the occurrence is referred to in order to reconstruct the proximal cause.

However, the categorisation of the contributing factors sheds some light on more distal, and less domain-dependent, 'causes'. Both, the task/-equipment domain and the contributory factors together can be seen as representing pointers as to in which, and where in the task sequence the underlying problems that led to the incident might be found. Thus, they can focus further enquiry. The initial categories (Wright et al., 1991) were

- Inexperience with equipment
- Shortage of trained staff
- Night time
- Fatigue
- Poor Equipment Design
- Unit Busy
- Agency nurse
- Lack of Suitable equipment
- Failure to check equipment
- Failure to perform hourly check
- Poor Communication
- Thoughtlessness

The 'contributing factors' categorisation scheme evolved since the time of creation, and nearly doubled from 12 categories to 23. This was the result of the ongoing iterative coding of the collected incident data, and of experience with the reporting scheme. The added categories are:

- Presence of students/teaching
- Too many people present
- Poor visibility/position of equipment
- Grossly obese patient
- Turning the patient
- Patient inadequately sedated
- Lines not properly sutured into place
- Intracranial Pressure Monitor not properly secured
- Endotracheal tube not properly secured
- Chest drain tube not properly secured
- Nasogastric tube not properly secured

In the initial version of the taxonomy, mainly so-called Performance Shaping Factors (see also Rasmussen, 1982) are listed as contributing causes, such as Fatigue, Unit Busy, and Night Time. Poor Communication can also be considered a performance shaping factor. The one factor notably not a PSF is 'Thoughtlessness'. The only factor that clearly denotes a latent failure is 'Poor Equipment Design'.

The initial categorisation scheme focussed on factors that created the situation precipitating the incident. However, the refined categories are increasingly task and domain specific, and do not denote generic Performance Shaping Factors. Instead, a behavioural and task domain dependent taxonomy is introduced, see especially the last seven factors above. These categories can provide the basis for descriptive statistics on relative frequency of occurrences and for denoting trends in the distribution and combination of incidents. However, analysis of the underlying causes of the incident is not facilitated.

'Cause' Occurrence '89		**'Contributing Factors' Occurrence '89**		**'Detection' Occurrence '89**	
'Ventilator':	**10**	Poor Communication:	**14**	D1 Regular Checking:	**11**
'Vascular line':	**6**	Poor Equip. Design:	**11**	D2 Alarms:	**11**
'Miscellaneous':	**5**	Inexperience with Equipment:	**5**	D3 Experienced Staff:	**8**
'Disposable Equipment':	**4**	Lack of Suitable Equipment:	**4**	D5 Patient Noticed:	**1**
'Drug-administration':	**3**	Night Time:	**3**		
'Non-disp. Equipment':	**2**	Fatigue:	**3**		
		Unit Busy:	**2**		
		Failure to Perform Hourly Check:	**2**		
		Thoughtlessness:	**2**		
'Cause' Occurrence '98		**'Contributing Factors' Occurrence '98**		**'Detection' Occurrence '98**	
'Drug-administration':	**10**	Thoughtlessness:	**11**	D1 Regular Checking:	**9**
'Ventilator':	**8**	Poor Communication:	**8**	D3 Experienced Staff:	**8**
'Vascular line':	**4**	Inexperience with Equipment:	**4**	D2 Alarms:	**2**
'Miscellaneous':	**4**	Night Time:	**3**	D4 Unfamiliar Noise:	**1**
'Non-disp. Equipment':	**1**	Failure to Check Equipment:	**3**	D5 Patient Noticed:	**1**
		Failure to Perform Hourly Check:	**2**	D7 Handover Check:	**1**
		Endotrach. Tube Not Properly Sutured:	**2**		
		Poor Equipment Design:	**1**		
		Patient Inadequately Sedated:	**1**		
		Turning the Patient:	**1**		

Table 1 - Causal Categorisation Sample '89 and '98

The evolution of the categorisation scheme itself provides valuable information. The novel factors were created by filtering them out of the incident data reported and analysed over the years. Such a bottom-up approach establishes and clarifies problem areas within both the task domain and the handling of the provided equipment. Thus, insights into task characteristics and performance can be gained. This can be put to use, for instance, for providing training focus, while offering strong empirical support. For instance, action recommendations in the period form August 1995 to August 1998 pay heed to the recurring problem of dislodged endotracheal tubes. Initially, reminders are repeatedly puplicised about this common problem. Reasons for dislodgement are given. Then, a list is devised that summarises "reasons for endotracheal tubes coming out". This is disseminated, and later 'suggested actions' recommend to revise this list and publicise it further.

In comparison of the revised categorisation scheme with the categorisation approaches discussed above it can be noted that a hierarchical causal chain can still be constructed by dividing the data according to, for instance, Rasmussen's Taxonomy for Description and Analysis of Events involving Human Malfunction (see Figure 1, and Section 5.3 for an example). For instance, contributing factors on the AIMS-ICU form are divided into 'system-based factors', which detail work condition factors as well as latent failures, and 'human factors', which note factors that impact on the human cognitive processing levels. There is a trade-off, however, since the provision of fixed categories lessens the flexibility of the data reported, and might stifle creativity for staff attempting to explain how and why the incident occurred.

Multi-causal classification

In the classification of data into the contributing causes categories, combinations of factors are allowed, and are noted frequently in the data sample. The data analysis might thus also be based on noting the frequency or likelihood of certain factors correlating. Trends could be established, and conclusions drawn that are justified by a richer data set than only noting the occurrence of single factors. To illustrate this, we took two data samples (see Table 1), one sample covering the first categorisation interval, January and February 1989 (sample89), and the other covering a more recent interval from May to November 1998 (sample98). Both samples cover 25 incident reports.

In sample98, the predominant factors are 'Thoughtlessness' (10 occurrences), 'Poor Communication' (9 occurrences), and 'Inexperience with Equipment' (5 occurrences). In contrast, in sample89, factors 'Poor Communication' (14 occurrences), 'Poor Equipment Design' (11 occurrences), 'Inexperience with equipment' (6 occurrences), and 'Lack of suitable Equipment' (4 occurrences) were most implicated in incidents.

A closer look at the data reveals that in the 1989 interval, half of all 'Poor Equipment Design' incidents and one third of 'Poor Communication' are not single factor categorisations, but are placed in combinations. 'Poor Equipment Design' is predominantly (four out of five incidents) paired with 'Lack of Suitable Equipment'. 'Poor Communication' is camobined with a variety of factors, such as 'Fatigue', 'Thoughtlessness', and 'Unit Busy'. This use of combinatorial categorisation embeds behavioural and person factors (e.g. Failure to Check Equipment, Thoughtlessness) within latent system and work condition factors such as Poor Equipment Design, Fatigue, and Poor Communication.

In sample 98, 'Poor Communication' is paired with other factors in 6 out of 9 incidents. 'Thoughtlessness' is shown with other factors (such as Inexperience with Equipment) in 4 out of 10 incidents. 'Inexperience with Equipment' is only ever mentioned in combination. In sample 89, 'Inexperience with Equipment' is left as sole contributory factor only twice out of a total of six incidents. Three times it is mentioned in combination with 'Poor Equipment Design' and 'Lack of Suitable Equipment', respectively. Again, this shows how factors that warrant further explanation can be placed in context by considering the multi-combinatorial categorisation. It also shows, however, that descriptive statistics neglecting this facet of analysis shed a slightly misleading light on the collected incident data.

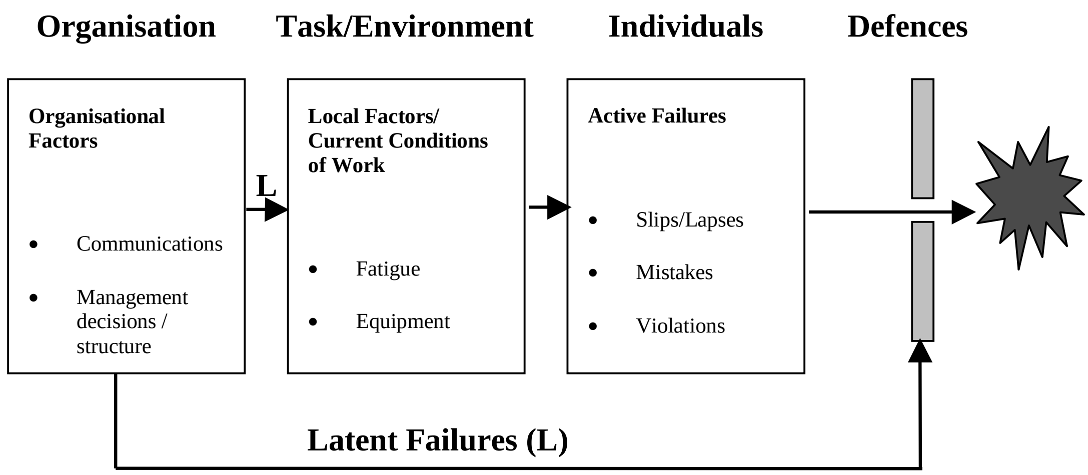


Figure 2 Reason's Organisational Accident Model [Stanhope, 1997 #57]

Incident Detection

The crucial role of detection factors is being recognised in the Edinburgh scheme. This is reflected on the incident form, as well as in the conclusions that are drawn form the data (see Table 1). Not only needs to be observed that although humans cause incidents, it is also humans who detect it and either remedy the consequences or prohibit the course of the incident to proceed. Staff is encouraged specifically to note which factors are believed to have aided detection. This is not only the data that can with significant confidence be assumed to be the reporter's own experience, but also that what ultimately can assist in finding ways of reducing the number of incidents and consequently, accidents.

The detection factor taxonomy evolved alongside the iterative development of the contributory factors taxonomy. Initially it consisted of the categories 'Repeated Regular Checking', 'Presence of Alarms on Equipment', 'Presence of Experienced Staff', 'Hearing Unfamiliar Noise', 'Patient Noticed', and 'Relative Noticed'. A task specific factor 'Having Lines or Three Way Tap Visible' was added, as well as the factor 'Handover Check'. These added factors point to possible system improvements to facilitate detection. They can be actively influenced by system factors such as work procedures, design, training, and staffing levels.

This is pointed out in the initial presentation of the incident scheme (Wright, MacKenzie et al., 1991). It states that "regular checking by experienced staff is critical in detecting errors, but this may be adversely affected by nurse staffing policies where agency staff are commonly used or where little time is available for handovers".

The iteration over the collected incident data thus clarified two more detection facilitating conditions. The importance of handover checks to make up for contributing causes 'Failure to Check Equipment' and 'Failure to Perform Hourly Check' is pointed out. Without iterative revision and coding of the data categorisation, these factors might have gone neglected. A formalised framework of analysis, such as those used in the aviation domain, can aid the recognition of detection factors and the generation of suggested actions (see Section 5 and 6).

Incident Analysis

In many medical incident reporting schemes, in-depth analysis and a search for the root causes of adverse events does not take place (Leape, 1994; Chappell, 1996). In contrast, the formal investigation of adverse events in industry is an increasingly well-established concept. Studies of accidents in industry, transport and military spheres have led to a much broader understanding of accident causation, with less focus on the individual who makes the error, and more on pre-existing organisational factors.

For instance, Reason's organisational accident causation model (see Figure 2) shows how latent failures can pave the way for an accident to occur. To produce effective recommendations, the information collected must be analyzed in a way that reveals the relationships between the human error that occurred, the design, and characteristics of the systems.

The benefits of analysing a case using a formalised model are that it allows an analysis in a structured format based on theories of accident causation and human error. This type of analysis allows analysts not only to identify the active failures, which they are accustomed to do, but also the potentially more important latent failures which create the conditions in which people make errors. Stanhope et al. (1997) suggested that a more systematic approach dealing with a smaller number of cases in more depth is likely to yield greater dividends in understanding incident causation and generating action recommendation than the 'many' cases currently analysed quite briefly and hence less effectively.

Barnhart et al. (1975) present a seminal paper on NASA's accident investigation method. Combined with Reason's accident causation model (Eagle et al., 1992), it can be described as consisting of three steps:

"When":

- Assembly of facts and generation of timeline
- Embed event within task sequence

"What":

- Identification of active failures
    - Description of all behaviours
    - Root Cause Analysis, assembling proximal and distal causes and work conditions factors

"Why"

- Identification of latent failures, for instance:
    - "Information Processing Analysis", e.g.:
        - sources of information available (latent)
        - what information perceived, used, and in what ways
        - which decisions made, listing all choices
        - whether actions matched decisions (latent and active)
        - consideration of managerial and organisational failures; follow-up of root cause analysis

Therefore, this model extends the NTSB model by including the analysis of the "why" of the incident. The incident analysis proceeds from identification of 'active failures' and local working conditions that precipitate those to the identification of latent system failures. For instance, Reason (1990) has proposed a list of General Failure Types of organizations. They include: incompatible goals; organizational deficiencies; inadequate communications; poor planning; inadequate control and monitoring; design failures; inadequate defenses; unsuitable materials; poor procedures; poor training; inadequate maintenance management; and inadequate regulation.

Root Cause Analysis Methods

'Formalised root cause analysis methods can be used within incident and accident investigation in order to guide the process of arriving at more distal causal factors precipitating the occurrence. One example is the Management Oversight and Risk Tree (MORT) Analysis method. The US Department of Energy developed this root cause analysis method (DoE, 1992) for the investigation of e.g. Nuclear Power Station incidents. It defines a 'direct cause' of an accident as the immediate events or conditions that caused the accident. 'Contributing causes' are said to be events or conditions that collectively with other causes increase the likelihood of an accident but that individually did not cause the accident. Contributing causes may be based on longstanding conditions or a series of prior events that, while not important in and of themselves, collectively increased the probability that an accident would occur. 'Root causes' are the causal factors that, if corrected, would prevent recurrence of the accident. Root causes are derived from and generally encompass several contributing causes. They are higher-order, fundamental causal factors that address classes of deficiencies, rather than single problems or faults. These are identified using root cause analysis. Root causes can include system deficiencies, management failures, inadequate competencies, accepted risks, performance errors, omissions, non-adherence to procedures, and inadequate organisational communication. In the advanced root cause analysis method MORT (Johnson, 1995), these are listed in a tree-shaped diagram, which guides the collection, interpretation, and analysis of the data. Thus, it provides a 'tool for thought', a structural framework, and a documentation tool for causal chain analysis. The final aim is identify management weaknesses, or in Reason's terminology, latent organisational failures.

The Edinburgh Study: Incident Analysis

The narrative given by the reporting staff on the incident report form provides the first level of interpretation of what happened. In the case of the person reporting the incident not being the same as the one having 'caused' it, the reporter provides a second-level interpretation of the events.

The narrative, together with the contributory and detection factors mentioned, typically lays out a timeline of the events. Otherwise, this is inferred when analysing the data. One method of establishing a task-related timeline is to embed the erroneous task event into a sequential, high-level task model. This is partly carried out by classifying occurrences according to task aspects, as shown in Section 4.

The classification of events involves informal (and non-documented) analysis followed by the above noted categorisation into 'causes', contributory, and detection factors. This can be seen as representing an informal root cause analysis process (see below). However, to repeat, the categorised data often seems to present behavioural descriptions or proximal 'causes'.

Following Reason's accident causation and analysis model (see Figure 2), Table 2 shows a tentative classification of the provided 'contributing factors' categories into latent failure types (distal causal factor), work conditions failure types (distal causal factor), and active failures (proximal causal factor). The latter constitute in our case task and behaviour oriented categories, rather than the error types based on cognitive theory as suggested by Reason. This classification can be compared to Rasmussen's event description scheme (see Figure 1), where latent factors and work conditions are mirrored in 'Personnel Task', and the 'Causes of Human Malfunction', 'Situation Factors', and 'Factors Affecting Performance' respectively. In order to be able to categorise the proximal failure types into Reason's cognitive Error Types, more detailed incident data is required than could be accessed from the samples. There is no one to one relationship between contributing factors and their underlying cognitive mechanisms (see for instance Busse and Johnson, 1998), and a more in-depth analysis of the error types is required.

Contributing Factor Categories	**Failure Type**
1. Inexperience with equipment	1. Distal: Training, Protocol on new equipment
2. Shortage of trained staff	2. Distal: Management, shift arrangement
3. Night time	3. Distal: e.g. Fatigue/Staffing level
4. Fatigue	4. Distal: Shift
5. Poor Equipment Design	5. Distal: System Design
6. Unit Busy	6. Distal – refined by #13 and #14
7. Agency nurse	7. Distal: Contract work, training
8. Lack of Suitable equipment	8. Distal: management of equipment maintenance
9. Failure to check equipment	9. Proximal, behavioural
10. Failure to perform hourly check	10. "
11. Poor Communication	11. Distal - Team Communicatin or Training
12. Thoughtlessness	12. Proximal
13. Presence of students/teaching	13. Distal (refining #14)
14. Too many people present	14. Distal (refining #6)
15. Poor visibility/position of equipment	15. Distal – detection facilitating factor
16. Grossly obese patient	16. Patient characteristic
17. Turning the patient	17. Proximal – Task/behavioural
18. Patient inadequately sedated	18. Proximal "
19. Lines not properly sutured into place	19. Proximal "
20. ICP monitor not properly secured	20. Proximal "
21. Endotracheal tube not properly secured	21. Proximal "
22. Chest drain tube not properly secured	22. Proximal "
23. Nasogastric tube not properly secured	23. Proximal "

Table 2 - Failure Type Categorisation

Currently, the Edinburgh study does not proceed much beyond the "what" phase within the above mentioned analysis model. However, given the contributing factors classification above, root cause analysis can be used to reflect the variety of levels in the incident causation tree. The consideration of latent and work condition factors draws attention to the deficiency of single-cause categorisation. Multi-causal categorisation can be used to reconstruct a possible root cause analysis as an example. For instance, the above-mentioned combination of factors 'inexperience with equipment', 'poor equipment design', and 'lack of suitable equipment' can be illustrated in a causal tree as shown in Figure 3. The incident (e.g. ventilator related) was 'caused' by staff 'inexperience with equipment'. This is then hypothesised to be mediated by 'lack of suitable equipment', which in turn pointed to 'poor equipment design'. Thus, a causal chain is established, formalised, and documented. It could also be argued that 'lack of suitable equipment' contributed directly to the occurrence of the incident. This hypothesis, again, has different implication for potential system redesign. Thus, this kind of analysis, taking several levels of causation into account, can aid precise and structured reasoning about the incident occurrence. Factors to be considered in lower levels of causation are work conditions and latent system failure types, for instance Training, and alternative contributing factors why a failure to check equipment occurred.

A structured, formalised analysis framework is also necessary to prevent hindsight from biasing error analysis. Attribution of error is a social and psychological judgement process rather than a matter of objective fact. Hindight view is fundamentally flawed because it does not reflect the situation confronting the practitioners at the scene. Thus, rather than being a causal category, human error should be seen as representing a symptom, and a starting point for investigation (Woods et al., 1994).

Action Recommendation

In order to arrive at sound and relevant action recommendations, a systematic and structured way of bridging the result of the analysis process to remedial measures is needed. The documentation of this process plays an important role, for instance to allow monitoring of the effect of the measure.

In industry domains such as aviation and process control, cognitive analysis of error occurrences is often used to point towards remedial actions. For instance, Reece et al. (1995) investigated human error in radiation exposure events. The proximal cause to the incident was situated in the task sequence and, additionally, cognitive failure analysis was carried out. Then the relationship between them was analysed, and thus it could be identified

- where training might be most effective;
- where equipment interface enhancements may be most appropriate;
- where job aids might help performance (e.g. checklists).

In a similar vein, van der Schaaf (1996) proposed the Eindhoven Classification Scheme for classifying events and identifying incident causes in process control. The main categories represent Technical Factors, Organisational Factors, and Human Error categorised according to Rasmussen's Skills Rules and Knowledge (SRK) framework. The translation into proposals for effective, preventive, and corrective action can then be guided by means of a proposed Classification/Action Matrix. Thus, the action categories relate back to the SRK error types, and include Equipment, Procedures, Information & Communication, Training, and Motivation.

This shows how error categorisation, when done according to cognitive level of performance and latent factors, can provide the basis for sound, structured, and theory-based remedial recommendations. Without error categories being based on sound psychological theory, systematic and relevant action recommendation generation is not possible.

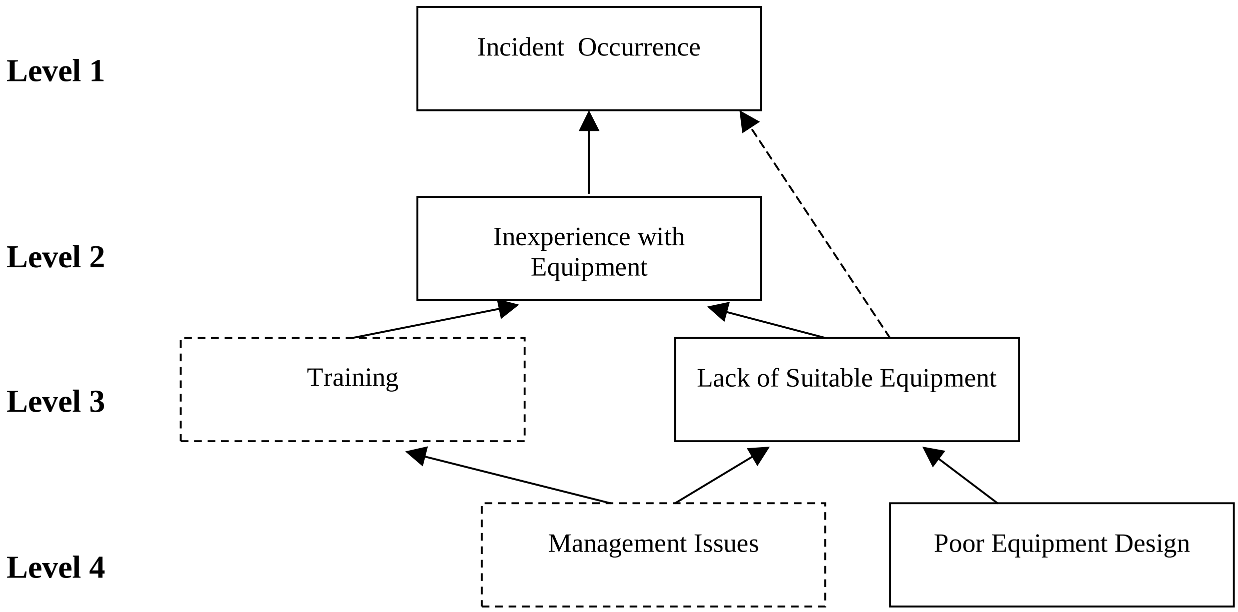


Figure 3 – Example of Incident Root Cause Analysis

The Edinburgh Study: Action Recommendation

In the Edinburgh study, the incident data was categorised and summarised by the scheme manager. Action recommendations were arrived at in an iterative process. The scheme manager suggested remedial actions and presented those together with the summary data to the senior nurse of the ICU. Together, the data was discussed and the rationale for the action recommendations reviewed. This led to a final version of suggested actions for each incident analysis period. Table 3 and 4 show a categorisation of the suggested actions for our two samples (sample89 and sample98).

First, we related the recommendations back to the initial 'cause' categories, with the scheme manager's assistance. Then we re-interpreted the suggested actions in the light of system safety concepts, such as presented by van der Schaaf or Reason. In sample89, Thoughtlessness and Poor Equipment Design featured most often as contributing factors. This is mirrored in the action recommendations falling in the 'remind staff…', 'change equipment' and 'create protocol for equipment use'. Entries under 'remind staff…' typically are in the form of a reminder statement, drawing attention to problematic task or equipment characteristics, for instance "Remind all staff of the importance of careful, correct use of 3-way taps on central venous and arterial lines" (February 1989). 'Change equipment' is represented by recommendations such as "Particular sort of disposable ventilator tubing used on trial should no longer be used". 'Create protocol for equipment use' mentioned for instance "Consider use of small Graseby syringe drivers with smaller volumes of solution".

'Cause'	**Contributing Factors**	**Detection Factors**	**Action Recommendations**
Ventilator **10**	Thoughtlessness: **14**		
Vasc. Line **6**	Poor Equip. Design: **11**	Reg. Checking: **11**	**4** Ventilator
Misc. **5**	Inexperience with	Alarms: **11**	**3** Vascular Line
Disp. Equip. **4**	Equipment: **5**	Exp. Staff: **8**	**2** Drugs
Drug-admin. **3**	Lack of Suitable	Pat. Noticed: **1**	**1** Miscellaneous
Non-disp. Equip. **2**	Equipment: **4**		**1** Equipment
	Night Time: **3**		
	Fatigue: **3**		Revised classification:
	Unit Busy: **2**		
	Failure to Perform		**3** Remind Staff…
	Hourly Check: **2**		**2** change equipment
	Thoughtlessness **2**		**2** create protocol for
			equipment use
			2 review protocol for
			equipment use
			1 create protocol for
			equipment maintenance
			1 review equipment

Table 3 - Action Recommendations for the reporting period Jan/Feb 1989

Incident ‘Cause’ 98	**Contributing Factors 98**	**Detection Factors 98**	**Action Recommendations 98**
Drug-admin. **10**	Thoughtlessness **11**	Reg. Checking: **9**	**3** Ventilator
Ventilator **8**	Poor Communication **8**	Exp. Staff: **8**	**2** Vascular line
Vasc. Line **4**	Inexperience with Equipment: **4**	Alarms: **2**	**1** Drugs
Misc. **4**	Night Time: **3**	Unfamiliar Noise: **1**	**2** Miscellaneous
Non-disp. Equip **1**	Failure to Check Equipment: **3**	Patient Noticed: **1**	
	Failure to Perform Hourly Check: **2**	Handover Check: **1**	Revised Classification:
	End. Tube not Properly Secured: **2**		**4** Remind Staff…
	Poor Equip. Design: **1**		**1** training viz new equipment
	Patient Inadequately Sedated: **1**		**1** equipment maintenance (management)
	Turning the patient: **1**		**1** create protocol for equip. use
			1 review procedure viz home patients' safety

Table 4 - Action Recommendations for the reporting period May/Nov 1998

In the period of May to November 1998, a marked increase in reminder statements can be noted. Following inspection of recommendation data, the dissemination of reminder statements were noted to be the single most often suggested action. In the period August 1995 to August November 1998, 82 "Remind Staff…" statements out of a total number of 111 recommendations coud be noted. The 29 other recommendations concerned procedure creation or change suggestions (e.g. "produce guidelines for care of arterial lines - particularly for femoral artery lines post coiling"), or were equipment related (e.g. "Obtain spare helium cylinder for aortic pump to be kept in ICU").

Reminder statements as potential error prevention mechanism have come into disrepute in domains such as aviation (Reason, 1990). Rather than further burdening operators’ and pilots’ memory capacity, indirect safety methods such as reduced complexity, standardisation, proceduralisation, and work aids such as checklists have been introduced.

However, on closer inspection, the nature of the Edinburgh reminder statements proves to be interesting. Reminders seem to target either very common, but still error-prone, details of tasks and practices, or problem points that occur very infrequently, or otherwise problematic parts or uses of procedures.

Distinguishing thus between types of recommendation, a link to Rasmussen's SRK framework can be created. Skill (S) level performance concerns automatic behaviour routines, such as the common but still error prone details of tasks. Rule (R) level performance concerns the conscious but practiced following of procedures and protocols, and Knowledge-based (K) performance relates to potentially effortful, fully conscious problem solving and decision making. In aviation and process control, it has been realised that performance on the rule-based level is the least error-prone. Therefore, design methodologies such as Ecological Interface Design (EID, (Rasmussen and Vincente, 1989)) emphasise proceduralies tasks, and ensure that task features that relate to S or K level performance are assisted accordingly. For instance, K based performance can be supported by careful information design.

The Eindhoven classification/action matrix (Van der Schaaf, 1996) is also based on a SRK style cognitive classification of error. It details, as described above, that R-based error is best targeted with with Training measures, K-based error with improvements in the Information & Communication domain, and S-based error with change in equipment. Thus, this is at odds with the Edinburgh results of the recommendation generation. Instead of reacting with reminder statements indiscriminately of cognitive performance level, these could be taken into account when suggesting remedial actions. The categorisation of error according to cognitive mechanisms will also further the understanding of performance problems.

Conclusions

We have illustrated in this paper how methods and insights from safety-critical domains other than medicine can be applied in a clinical setting. This concerns the use of incident schemes, as well as the application of accident causation models in the analysis of incidents, and in the generation of action recommendations. These recognise the importance of latent failure as part of the causal chain of an incident, which is reflected in the incident analysis process.

Incident investigation schemes often neglect formalised, in-depth anlaysis of single incidents in favour of a quantitative surface analysis. Also, the crucial role of detection factors and the need to support those is often underestimated. The Edinburgh incident scheme caters for those in the data collection process as well as in the generation of action recommendations. Thus, the analysis process and its results of the Edinburgh study showed how not only theoretical 'top-down' approaches can inform incident analysis, but also how practical incident avoidance can be supported by a 'bottom-up', detailed (albeit non-formalised) analysis process.

Acknowledgements

This work was supported by UK EPSRC Grant No GR/L27800. Thanks goes to Dr David Wright for data access and his support, and also to the Glasgow Accident Analysis Group.

References

Barnhart, W., C. Billings, et al. (1975). A method for the study of human factors in aircraft operations. Washington, D.C., National Aeronautics and Space Administration (NASA).

Battles, J. B., H. S. Kaplan, et al. (1998). "The Attributes of Medical Event-Reporting Systems." Arch Pathol Lab Med **122**.

Billings, C. (1998). What can and What can't be Learned from Self-Reported Errors? Fourth Annual Symposium on Human Interaction with Complex Systems - HICS'98, Fairborn, OH, USA.

Bogner, M. S., Ed. (1994). Human Error in Medicine, Lawrence Erlbaum Associates.
Busse, D. and C. Johnson (1998). Modeling Human Error within a Cognitive Theoretical Framework. ECCM '98, Nottingham, UK, Nottingham University Press.

Chappell, S. L. (1996). Using Voluntary Incident Reports for Human Factors Evaluations. The International Society of Air Safety Investigators.

Dawson, S., P. Garside, et al., Eds. (1998). Organisational Change - the Key to Quality Improvement. Quality in Health Care, BMJ Publishing Group, RCN Publishing Company.

DoE (1992). Root Cause Analysis Guidance Document. Washington, D.C., US Department of Energy.

Eagle, C. J., J. M. Davies, et al. (1992). "Accident Analysis of large-scale technological disasters applied to an anaesthetic complication." Canadian Journal of Anaesthesia **39**(2): 118-122.

Flanagan, J. C. (1954). "The Critical Incident Technique." Psychological Bulletin **51**: 327-358.
Hale, A., B. Wilpert, et al. (1997). After The Event - From Accidents to Organisational Learning, Pergamon Books.

Harvard_University (1990). Medical Practice Study: Patients, Doctors, and Lawyers: Studies of Medical injury in New York. Boston, Harvard University Press.

Hollnagel, E. (1991). The Phenotype of Erroneous Actions: Implications for HCI Design. Human-Computer Interaction and Complex Systems. G. R. S. Weir and J. L. Alty. London, Academic Press Ltd.

Johnson, C. W. (1995). The Application of Petri Nets to Represent and Reason about Human Factors Problems During Accident Analyses. The Design, Specification And Verification Of Interactive Systems. P. Palanque and R. Bastide. Berlin, Springer Verlag**:** 93-112.

Jordan, B. (1993). Ethnographic Workplace Studies and Computer Supported Cooperative Work. Interdisciplinary Workshop on Informatics and Psychology, Schaerding, Austria.

Leape, L. L. (1994). "Error in Medicine." JAMA: The Journal of the American Medical Association **272**(23): 1851-57.

Leape, L. L., T. A. Brennan, et al. (1991). "Adverse Events and Negligence in Hospitalised Patients." Iatrogenics **1**: 17-21.

Leape, L. L., D. D. Woods, et al. (1998). "Promoting Patient Safety by Preventing Medical Error." JAMA: The Journal of the American Medical Association **280**(16): 1444-47.

Luxhj, J. T., D. N. Arendt, et al. (1997). An Application of Advanced Information Technology for Assessing Aircraft Accident Causation. The International Society of Air Safety Investigators.

Maddox, M. E. and J. Reason (1996). Human Factors Guide for Aviation Maintenance. Washington, DC, Federal Aviation Administration/Office of Aviation Medicine.

NTSB (1998). National Transportation Safety Board: Revised Accident/Incident Database Coding Manual. Washington, DC.

Rasmussen, J. (1982). "Human Error: a Taxonomy for Describing Human Malfunction in Industrial Installations." Journal of Occupational Accidents **4**: 311-333.

Rasmussen, J. (1987). The Definition of Human Error and a Taxonomy for Technical System Design. New Technology and Human Error. J. Rasmussen, K. Duncan and J. Leplat, John Wiley & Sons Ltd.

Rasmussen, J., K. Duncan, et al., Eds. (1987). New Technology and Human Error. Wiley Series: New Tecnology and Work. Berlin, John Wiley & Son Ltd.

Rasmussen, J. and K. Vincente (1989). "Coping with Human Errors through System Design: Implications for Ecological Interface Design." International Journal of Man-Machine Studies **31**: 517-534.

Reason, J. (1990). Human Error, Cambridge University Press.

Reason, J. (1997). Managing the Risks of Organizational Accidents, Ashgate Publishing Limited.

Reece, W. J. and S. G. Hill (1995). Human Performance Analysis of Industrial Radiography Radiation Exposure Events. Human Factors and Ergonomics Society 39th Annual Meeting.

Runciman, W. B., A. Sellen, et al. (1993). "Errors, Incidents, and Accidents in Anaesthetic Practice." Anaesthesia and Intensive Care **21**(5): 506-519.

Stanhope, N., C. Vincent, et al. (1997). "Applying Human Factors Methods to Clinical Risk Management." British Journal of Obstetrics and Gynaecology **104**: 1125-1232.

Van der Schaaf, T. W. (1996). PRISMA: A Risk Management Tool Based on Incident Analysis. International Conference and Workshop on process Safety Management and Inherently Safer Processes, Orlando, Florida, American Institute of Chemical Engineers.

Van der Schaaf, T. W., D. A. Lucas, et al., Eds. (1991). Near Miss Reporting as a Safety Tool, Butterworth-Heinemann Ltd.

Wedderburn, D. (1998). "Where Does the Clinical Audit fit in?" Quality in Health Care **7**: 1-4.

Woods, D. D. (1994). The Critical Incident Technique - 40 Years Later. The Human Factors and Ergonomics Society 38th Annual Meeting.

Woods, D. D., L. J. Johannesen, et al. (1994). Behind Human Error: Cognitive Systems, Computers, and Hindsight, CSERIAC Crew Systems Ergonomics Information Analysis Center.

Wright, D. (1999). Ten Years of a Medical Incident Reporting System. First Workshop on Human Error and Clinical Systems, Glasgow, UK.

Wright, D., S. J. MacKenzie, et al. (1991). "Critical Incidents in the Intensive Therapy Unit." The Lancet **338**: 676-78.

Biography

Daniela Busse, 17 Lilybank Gardens, Glasgow G20 8RZ, Scotland UK, Telephone +44 141 3398855 x0918, email bussed@dcs.gla.ac.uk.

Daniela Busse (MA) is a Research Assistant and a PhD student at the Computing Department, University of Glasgow. Her thesis concerns Cognitive Models in Human Error and Accident Analysis. She received her MA in Computing and Psychology from the University of Glasgow.

Prof. C.W. Johnson, DPhil, MSc, MA, FBCS, CEng., Dept. of Computing Science, Univ. of Glasgow, Glasgow, G12 8QQ, Scotland. Tel: +44 (141) 330 6053. Fax: +44 (141) 330 4913, e-mail: johnson@dcs.gla.ac.uk.

Chris Johnson leads an accident investigation group that includes human factors experts; software developers and systems engineers. He chairs IFIP Working Group 13.5 on Human Error and Systems Development and has authored over ninety papers on the human contribution to systems failure.

Learning From Errors and Recoveries in Medical Chains

Tjerk W. van der Schaaf
Eindhoven University of Technology, Safety Management Group.

Abstract

Most, if not practically all, systems for reporting and analysing medical incidents are based on and restricted to single, mono-disciplinary, hospital departments. Patients (or patient-related products such as blood or medicine) however usually "flow" through a series of (sub-) departments and healthcare-organisations outside of the hospital in the course their treatment, thus forming a "chain" . As shown in a dozen projects by our research group in a wide variety of medical domains during the last few years, some 15 -30 % of the causes of incidents in a single department may be "imported" from others, located earlier in the medical chain. This fact may then severely restrict a department's capability for learning lessons based on the analysis of incidents reported in that part of the chain only.

Examples are given of three recent or ongoing projects in the Netherlands, in transfusion medicine, care of asthmatic (COPD) patients, and medication errors. These projects were carried out explicitly to explore the benefits of "Chain-based" systems for reporting and analysing medical incidents, as opposed to the current Department-based ones. The prelimenary results show improved insight on the part of the medical staff ,as well as the necessity to consider new methods of analysis, fit to handle the sequential, time-related nature of these data.

Application of a System Model to Obstetric and Neonatal Audit

J.M. Davies,

Department of Anaesthesia, University of Calgary, Canada

Keywords: audit, database

Abstract

This paper describes the development of a database to facilitate Quality Assurance review of obstetric and neonatal care. The architecture of the database follows that of the Winnipeg model, developed in part for the Pediatric Cardiac Surgery Inquest in Winnipeg, Manitoba. The Winnipeg model was based on the modification and integration of three previously existing models, those of Professors Donabedian, Reason, and Helmreich. The neonatal component of the new database was developed in an attempt to produce an alternative classification system to that of Wigglesworth, which groups deaths according factors including weight of the infant and timing of death. Use of the Winnipeg model allows reviewers to determine not only 'what' happened and 'how' but also 'why'. This provides a starting point for educational endeavors for all those involved in the care of the patient, with the aim of preventing future deaths.

Introduction

Investigation of accidents and critical incidents is important to future safety initiatives in both aviation and medicine, to allow determination of the nature and contribution of the active factors and latent conditions (ref. 1). From these investigations, both aviation and medicine (in particular, anaesthesia) have established databases to record results, analysis and evaluation. These databases are like electronic filing cabinets. They contain data - elements of description, of people, places and things, such as the details of the accident, the morbidity or mortality. Databases then provide information - the ordering and interpretation of data, to impose or extract meaning (ref. 2), such as a chronologically-derived mechanism and even a classification of error into 'human' versus 'technical'. The information so derived may then be used to produce structural changes in health care (ref. 3).

However, data and information in health care have been described as *"unreliable, fragmented, imbalanced and incoherent*" (ref. 2) Such poor quality information, *"confusing, disturbing, and often misleading or flatly false"* (ref. 2), has led to a call for *a "health information template"*. This *"map of the territory"* should allow for diversity and complexity of information and should have an organising structure that is *"coherent & intuitively plausible"*. The goals of the template should be to provide a comprehensive method of classification, to *"highlight gaps"* in the compilation and analysis of data, and to serve as a teaching device (ref. 2).

Could there be a better way to collect data and present information? Rather than being assembled alphabetically or even chronologically, data and information should be presented in a way that conforms to a model of the system. By definition, the term 'system' will represent a grouping of inter-related components that act together in an environment to achieve a particular outcome. In medicine, for example, the system includes the patients, personnel, equipment and environment, organization and regulatory agencies (ref. 4).

By using the system as a template, questions asked during an investigation and resulting answers should help to show where deficiencies lie in the system. This applies both to reactive investigations (the 'retrospective' view of problems that have occurred) and to proactive investigations (the 'prospective' view of problems that might occur). In addition, all components of the system and their inter-linking should be transparent to all who work in it or who are influenced by it. For the aviation and medical systems, even the general public (i.e., either passengers or patients) should understand their position in the system because they are ultimately the customers of the services provided. In addition, use of such a model should provide a better understanding of the role of human error in any system and how errors might be prevented or avoided, detected or absorbed, and their effects mitigated or reduced (refs. 5-6).

The aim of this paper is to present application of a new model to audit or Quality Assurance in obstetrics and neonatology. A brief description is given of development of the model. Then the reasoning behind the perceived need for a new model follows. Finally, application is presented.

Development of the Model

History: In 1995 – 1996, a new model applicable to both the aviation and health care systems was developed (ref. 4), based on three existing models (refs. 7-10) which were modified and then integrated. Initial work on the model was carried out in part for a Quality Assurance project undertaken with the Department of Obstetrics and Gynaecology at the Foothills Medical Centre. The model was further refined for the Pediatric Cardiac Surgery Inquiry, Winnipeg, Manitoba and has been used in both aviation and medicine. The aviation application was to the SIAM Project for the Bureau of Air Safety Investigation, Australia. The medical applications have included a number of investigations of both accidents and critical incidents, for the Patient Care Committee of the Foothills Medical Centre, for the Critical Review Committee of the Calgary Regional Health Authority, for medico-legal and coronial reviews, and for medico-legal cases.

Components: In developing this new model (ref. 4), three existing models - those developed by Professors Donabedian (ref. 7), Reason (refs. 8-9) and Helmreich (ref. 10) were incorporated.

Donabedian component: The first component of the new model was that defined by Professor Avedis Donabedian more than three decades ago, when he was a Professor of Public Health Economics in the School of Public Health at the University of Michigan. Donabedian detailed the components of quality in a triad that has stood the test of time. The components are Structure, Process and Outcome (ref. 7). The Structure aspects of any system include the administration, personnel, equipment and environment. The Process aspects are the tasks performed and the specific methods followed. The Outcome aspects represent the final results of any system and an evaluation of those results. Table 1 shows the triad of Quality Assurance (QA) and its components.

Table 1 - The Quality Assurance Triad and Components

Structure	where with what by whom	environment equipment personnel, administration
Process	what is done how it is done	tasks methods
Outcome	how one knows what is done how one knows how well a task is done	audit evaluation

Taken at their most basic, the triad of Structure, Process and Outcome constitutes a simple method of determining if a system is organised, working and producing to its optimum level (ref. 11). In effect, any activity that improves the quality of care can be defined as 'QA'. Thus, continuing education of personnel and evaluation of new trends in clinical practice are all 'QA' activities as they are presumed to increase the quality of care delivered. However, the primary goal of any QA program should be to provide both excellent patient care and optimal working conditions for the providers of that care. Although health care workers may be able to work well under pressure, e.g., during national or environmental crises, long-term results will suffer if the pressure is maintained. Thus, quality requires optimal workplace conditions.

The Reason component: The second component of the new model was that defined by Professor James Reason of Manchester University (refs. 8-9). In the Reason model, complex socio-technological systems (such as anaesthesia or the operating room) are afflicted with two different but complementary types of flaws or failures, "*active*" and "*latent*". The first type, "*active failures*", represents the actions of the operator, in particular those which occur when humans interact with the other components of the complex system in which they work. These "*front-line operators*" often appear to be the "*principal instigators of system breakdown*". (This is where the term, 'human error', was often applied.) However, rather than just instigators, their role is catalytic, providing many of the "*local triggering conditions*" necessary to expose the second type of system flaws. Reason has recently stated that *"errors are the symptoms that reveal the presence of latent conditions in the system at large"* (ref. 12). These crucial weaknesses, or "*latent failures*", represent flaws already present or built-in to the system.

Although the role of human factors is well recognized in the genesis of system failure, this will not originate solely from human error. Rather, systems will fail when active failures combine with or produce a particular set of circumstances necessary to reveal the latent failures (also described by Reason as "*resident pathogens*".) When active failures combine with latent failures, then the potential for system dysfunction will depend on the degree of 'malignancy' of the latent failures - defined by their number, complexity and position in the system. All systems contain some pathogens and the more complex a system, the greater the number (refs. 8-9).

Reason now uses the term, *"latent conditions"* (ref. 12), rather than *"latent errors"* or *"latent failures"* (ref. 8) as he believes that *"in the causal sense, the term 'condition' is much more appropriate"*. He bases this change on a definition of *"mere conditions"*. These are *"just those (things) that are present alike both in the case where accidents occur and in the normal case where they do not; and it is this consideration that leads us to reject them as the cause of the accident, even though it is true that without them the accident would not have occurred*" (ref. 13).

The Helmreich component: The third component of the new model was that defined by Professor Robert .L. Helmreich of the University of Texas at Austin. The Helmreich model was developed as part of evidence presented to a Commission of Inquiry into the crash of an Air Ontario Fokker F-28 at Dryden, Ontario in 1989 (ref. 10). The model was developed to show how *"the behaviour of flight crews in any given situation is determined by a number of simultaneously operating factors"*.

The four factors in Helmreich's model were:

1. *"the regulatory environment - operational standards and supervision"*;

2. *"the organizational environment - the culture and behavioural norms of the organization including morale, policies and standards, organizational stability and change, and available resources";*

3. *"the physical environment - meteorological and operating conditions and the aircraft, including its condition and capabilities";* and

4. *"the crew environment - interpersonal coordination and communications including cockpit, cabin, and ground personnel, and individual characteristics of crew members - training, experience, motivation, personality, attitudes, fatigue, and stress both from the immediate operational situation and significant personal life events"* (ref. 9).

Description: The Donabedian triad has been represented as a triangle, with Structure, Process and Outcome at each of the three points. The Reason model has been likened to a series of cards, each with various Swiss-cheese defects, which, if aligned, allow the trajectory of catastrophe to pass unimpeded. The Helmreich model was illustrated as a serious of three concentric spheres, sitting within the box of the regulatory environment.

Arrow of time: In contrast, the Winnipeg model has been structured along the arrow of time. A time-based model was chosen for several reasons. First, an incident or accidents does not have a single or sole cause. Like small streams, each of which feed into an increasingly large and turbulent river, so multiple human errors and factors contribute to the evolution of critical incidents (CI) and accidents. Several studies have shown the increase in numbers of factors as a critical incident evolves into an accent. For example, in Cooper and colleagues' 1978 study of CI, those that did not progress to actual complications had 2.5 associated factors, whereas those that were associated with a complication had 3.4 associated factors (ref. 14).

The second reason for basing the model on time was because of the concept of 'causal relationships' which are intrinsic to science and medicine. To delineate causal processes, one needs to specify the time intervals and sequences, as well as identify the *"temporal order of postulated causes and effects"*. This may be difficult to do. There is an *"intuitive bias"* which favors the concept that the closer in time and space an activity and an outcome exist, then the more confident we are about the causal relationship between that activity and outcome. Part of this relationship relates to the fact that the closer in time that two events exist, the less opportunity there is for other factors to influence outcome (ref. 15).

In support of this relationship are the results from research into the 'psychophysics' of time. This research has provided values for certain time-related experiences. For example, an 'interval' must last for slightly over 0.01 second (for sound) and between 0.07 and 0.10 second (for light) to be *"perceived and reported"*. An 'interval' can last for 2 to 12 seconds (for *"simple stimuli"*) and still be perceived as a *"unified event"*. The minimum separation needed for two stimuli to be perceived as separate entities is 0.0001 second for noise, 0.0002 second for tones, and 0.16 second for visual flicker. The minimum separation needed for two stimuli to be perceived as separate and for their order to be determined (correctly 75% of the time) is 0.02 second. The optimal duration for accurate judgment of the length of time of *"(very short) intervals"* is between 0.50 and 0.75 second (ref. 15).

Questions: Thus the Winnipeg model represents a system that exists in time. At the end of this time line, something will have happened. Classically one first asks, 'What happened?'. This question is followed by 'How did it happen?' and finally, 'Why did it happen?'. Thus, the model has three main components: 'What?', 'How?' and 'Why?'. These three questions conform to Donabedian's triad of Outcome, Process, and Structure, respectively. 'What?' represents the Outcome of the system. This may be an accident or an 'occurrence'. 'How?' represents the Process of the system and explains the mechanism of the Incident / Accident. 'Why?' represents the Structure of care. These three components can also be considered 'phases' in time in any system, as the system starts with its Structural components, moves through the Procedural phase, and ends with the Outcome phase.

Reason's model indicates that 'active failures' (or factors) represent actions of the workers - their tasks and methods. These, therefore, are found in the Process area of the system. This leaves 'latent conditions' (and resident pathogens) to exist in the Structure area of the system, as there is nowhere else for them to lurk.

Architecture: The architecture of the starting Structure is derived from the Helmreich model. A first revision of Helmreich's model yielded a series of five concentric circles, with passengers as the central, and most important, component. In a second revision, patients were substituted for passengers, to conform to the system under discussion. Thus, there are five components to the architecture of Structure. These are patients (or passengers), personnel, environment/ equipment, organization, and regulatory authority, with the latter at the outer limits of the system. Each of the component circles can be subdivided. For example, 'Personnel' can be divided into those characteristics that apply to the individual and those that apply to the work groups or teams to which the individuals belong.

Defenses and envelope: The system's 'defenses' and 'envelope' were also considered, in particular in relation to the concept of 'exceeding the envelope'. The 'defenses' are those components of the system which function to make workers aware of hazards, to warn them of danger, to help them to control problems to restore the system to (normal) operations, to protect them from hazards, to help them escape from danger, and to help them to rescue other workers should containment of the hazard fail (ref. 12). The progression from incident to critical incident to accident is said to occur when system defenses are weakened or absent. The 'envelope' is a term often used in aviation and has to do with how far the limits of the system can be pushed before something happens, e.g., the wings fall off the airplane. Thinking about defenses and the envelope together led to the idea that each part of the system contributes to defense of the system and that each component of the system provides defenses for the system. That is, the system is its own defense. This led to the idea that the system envelope is the outer boundary of the system itself. That is, the system forms its own envelope. Just as the skin of any individual forms the envelope and the boundary line of defense of that individual, so does the outermost ring of the system form the envelope or boundary line of defense of the system.

Degradations and improvements: However, no system is static. Most systems will degrade, given a long enough timeline. In the course of a lifetime we may not see some systems degrade because we cannot perceive the length of eternity. In contrast, during the course of an eight-hour shift or an eight-hour flight, we might see a marked decrease in the robustness of the system, particularly in the 'Personnel' and 'Equipment' components. Some systems may increase in robustness, particularly if efforts to ameliorate or eradicate latent conditions are identified and initiated. It is the sum of each of the components (which start and flow from Structure) that determines the overall degree of strength or robustness of the system (and therefore of the system defenses). However, should there be a marked reduction in the diameter of most or all of the components, then there will be an overall narrowing of the (envelope of) the system. At a time specific to each system, a point of criticality will be reached and the system will collapse.

Although such collapse may occur because of the intrinsic nature of the system, abrupt changes in a system may be triggered by specific local factors. (These are Reason's *"local triggering factors"*.) When proactively examining a system, attempts can be made to identify the presence or potential presence of such local triggering factors. (Some of these may have been identified through previous accident investigations.) However, although one can predict the presence, effect and importance, of local triggering factors, one can only suspect, but one cannot know. For example, the presence of blood on the floor of the Operating Room is recognised as a factor in falls. However, until someone does fall, then one cannot know that the blood on the floor was indeed a local triggering factor for that accident, nor the degree of that contribution, nor which other factors might have contributed to the accident, e.g., new floor wax, leather-soled shoes, speed of individual walking across the floor, individual's ability to balance. (The inability to predict exactly which local triggering factors will be implicated in any incident or accident until after the occurrence is analogous to Heisenberg's Uncertainty Principle, which states that one cannot know the position and velocity of a particle at the same time.)

Chronological dimensions: Finally, the chronological dimensions of the model must be considered. That is, the model exists in any time until temporal dimensions are applied. At some point, a 'real-time' clock must be started for the system under consideration. The Structure of the system is that which exists at the time the clock is started, with Process flowing on from that point. Similarly, a time must be given for when the clock is 'stopped' in the model so that the model may be interrogated. At that time, Process ends and Outcome begins.

The distinction between Process and Outcome was described by Donabedian as *seeming "to imply a simple separation between means and ends*". He added that "*perhaps more correctly, one may think of an unbroken chain of antecedent means followed by intermediate ends which are themselves the means to still further ends"* (ref. 7). Thus, within that time span there may be more than one outcome observed in the system and these different outcomes may apply to any of all of the five components. The determination of 'which' outcome is considered 'the' outcome will depend on when the reviewers determine that the 'clock is stopped'. To do this, reviewers must specify the questions they pose in the course of the review, e.g., 'how and why did the patient die' versus 'how and why did the patient arrest' versus 'how and why did the patient hemorrhage'.

Perceived need for a new method of audit

Currently in the Province of Alberta, neonatal deaths are classified according to a system devised by Wigglesworth (ref. 16) in 1980 and then revised by Wigglesworth and colleagues (ref. 17) in 1989. The aim of the original Wigglesworth classification was to *"find means of reducing the number of perinatal deaths"* (ref. 16). In doing so, Wigglesworth hoped "*to initiate cost-effective improvements in perinatal care"*. His starting premise was that doctors should *"select for analysis those aspects of perinatal mortality which we can hope to influence by alterations of management"*. He stated that perinatal deaths in infants weighing more than 2500 g were *"influenced by obstetric management"*. In contrast, neonatal deaths in infants weighing less than 2500 g were *"modified by the quality of neonatal care"*. He therefore reasoned that a *"breakdown of perinatal mortality by birth weight should thus constitute the first stage of any analysis"*. Wigglesworth also argued that classification by clinical cause was *"inadequate"* as the definitions were *"highly subjective"* and there was lack of information *"relating to management"*. In addition, he stated that classification by primary necropsy findings was limited by the rate of post-mortem examinations and depended *"on the uniformity of decision by the pathologist as to which of a number of findings was the major one"* (ref. 16). Table 2 shows the 1980 classification developed by Wigglesworth.

Table 2 - Wigglesworth's (1980) Suggested Classification of Perinatal Deaths by Pathological Subgroups

Group 1	normally formed macerated stillbirth
Group 2	congenital malformations (stillbirth or neonatal death)
Group 3	conditions associated with immaturity (neonatal death only)
Group 4	asphyxial conditions developing in labour (fresh stillbirth or neonatal death)
Group 5	specific conditions other than above

In a paper describing the validation of this classification, Wigglesworth and colleagues *"set out amendments to Wigglesworth's original definition to clarify certain ambiguities"* (ref. 17). This was necessary because of substantial disagreement between the classifiers when using the original classification. A decision tree analysis was also presented. The analysis led the reviewer through a set of questions, including the lethality of a malformation, timing of death, weight of the infant, time of death (hours), presence or absence of birth trauma or asphyxia, and gestational age. This revision is shown in Table 3.

Table 3 - Clarification (1989) of the (Wigglesworth) classification

Group 1	deaths before start of labour
Group 2	congenital malformations
Group 3	conditions associated with preterm birth or immaturity
Group 4	all fetal deaths in labour
Group 5	specific conditions other than above

However, review of this classification suggests some inherent difficulties. First, the decision tree does not follow a gestational timeline. For example, using the decision tree shows that groups occur in the following sequence: Groups 2, 5, 4, 1, 3, 4, 4, 3, and 5. Second, assignment of groups appears subjective. For example, Group 1 would include *"minor isolated lesions (such as small ventricular septal defect)"* but would exclude *"multiple minor anomalies (for example, cleft lip and a skeletal defect)"* (ref. 17). In addition, the classification system does not seem to lend itself to determining any of the underlying contributing factors that should be addressed to prevent future problems, e.g., why gestational age was < 37 weeks.

Even the developers of the classification had some difficulty in applying it. After incorporating the modifications, reviewers required a substantial initial learning curve and only reached 85% agreement between all reviewers. This was apparently *"largely the result of failure to comply with the rules laid down for classification"*. The question of using *"trained clerks rather than doctors"* was considered (ref. 16). Interestingly, Amar and colleagues described use of the classification by trained nursing staff. Each death was then reviewed by a clinician and only 2.9% required reclassification (ref. 18).

In 1986, two publications documented further attempts to clarify the classification of perinatal death (refs. 19-20). They offered variant classifications, by Cole and colleagues for obstetric factors (ref. 19) and by Hey and colleagues for fetal and neonatal factors (ref. 20). Settatree and Watkinson later reviewed 451 deaths from a perinatal mortality survey, with a view to examining problems encountered in classifying deaths using these two methods. The comparative *review "highlighted the complexity … of the classifications, the uneven distribution of deaths within their categories, and the variable levels of agreement between professionals classifying deaths"*. These authors went so far as to question the validity of individual maternity units in generating data that would later be used for comparative purposes in regional and national statistics (ref. 21).

Most recently, Cole described three methods of classifying perinatal deaths (ref. 22). These were (1) from the obstetricians' point of view, the Baird classification, (2) by Wigglesworth and (3) by the International Federation of Obstetricians and Gynecologists (FIGO). The Baird classification consisted of eight categories:

congenital anomaly, isoimmunisation, pregnancy hypertension, antepartum hemorrhage, trauma, maternal disorders, miscellaneous, and unknown (< 2500 g, > 2500 g). The FIGO classification recommended that rates for stillbirth, perinatal and neonatal mortality rates should be shown by birth weights >/= 500 g and >/= 1000 g. In addition, classification should differentiate *"normally formed infants"* from those with *"lethal congenital anomalies"*. Cole noted that none of these three classifications was considered *"uniquely superior"* and all had *"their own purposes and uses"*. She emphasised that the *"most important feature"* was that there should be clear definitions for groups within each classification and that there should be *"rules"* to help with allocation if multiple factors were noted in the death.

Application of the Winnipeg model

Using the Winnipeg model, a new classification was developed. This classification includes both mother and fetus / baby, not only for mortality review but also for audit of morbidity, satisfaction, critical incidents and incidents. The form for data collection is designed to be easily accessible and to consist (where possible) of objectively based questions and Yes / No answers. Less emphasis has been placed on autopsy and other post-mortem test results, without negating their importance (ref. 23). More emphasis has been placed on system-wide factors, by the simple fact of their inclusion. These factors include: maternal age, parity, both of which influence maternal mortality (ref. 24); maternal occupation, and maternal and paternal education (ref. 25); antenatal care (refs. 26-27); smoking history and "*unmarried at delivery status*" (ref. 28), all of which have been shown to increase perinatal morbidity or mortality. Also included are the other four elements of the Winnipeg model – personnel, equipment / environment, organisation and regulatory agencies.

The need for standardised definitions is well-recognised (ref. 29). For example, gestational age is defined by the duration of gestation as measured from the first day of the last menstrual period. However, the data should allow calculation of statistics to facilitate comparison to some external standard. For this reason, the new database will not group birth weight by weight intervals of 500 g, e.g., 500 – 999 g, but will simply list the actual weight. This format will then allow any number of groupings by any weight.

Taking a systems wide approach is not new in obstetric audit. For example, Ifenne and colleagues demonstrated a marked improvement in obstetric care at the teaching hospital in Zaria, Nigeria. Both deaths and complications decreased (85% and 80%, respectively) over the five years of the programme. The components of this programme can be classified into the five basic elements of the Winnipeg model. Patients were educated to decrease delays in seeking care and their families were encouraged to donate blood. Resident doctors and midwives were trained in emergency obstetric care. An emergency drug pack was developed. A surgical theatre was restored to working order and a maternity ward was renovated. The culture of the organisation became more positive, with an increase in staff morale. The organisation's budget was expanded to cover some of the costs of renovation. The government provided salary support for additional staff and 65% of renovation costs (ref. 30).

The Winnipeg model database is now being used for a retrospective review of major obstetric complications, such as postpartum hysterectomy. In addition, all maternal deaths in the Calgary region over the past ten years are undergoing re-review. The definitions being used for this review are those used in the Report on Confidential Enquiries into Maternal Deaths in the United Kingdom 1991 – 1993 (ref. 31).

- Maternal deaths are deaths of women while pregnant or within 42 days of termination of pregnancy, from any cause related to or aggravated by the pregnancy or its management, but not from accidental or incidental causes.

- Direct deaths are those resulting from obstetric complications of the pregnant state (pregnancy, labour, puerperium), from interventions

- Indirect deaths are those resulting previous existing disease, or disease that developed dying pregnancy and which was not due to direct obstetric causes, but which was aggravated by the physiologic effects of pregnancy.

- Fortuitous deaths are those deaths from unrelated causes that happen to occur in pregnancy or the puerperium.

- Maternities are the number of mothers delivered of registrable live births at any gestation or stillbirths of 28 weeks gestation or later prior to 1st October 1992 and 24 weeks or later thereafter.

The United Kingdom Report is one of the external standards against which the data will be compared. However, potential difficulties in comparing a small population with a large one will need to be considered. For example, each year in Calgary there are an estimated 12,000 maternities, for a total of 36,000 over the comparable triennium 1991-1993. During the same period in the United Kingdom there were 2,315,204 maternities. In addition, the problem of potential bias from selective hospital referrals has been noted (ref. 32). For example, the Foothills Medical Centre is the tertiary referral centre for obstetric care in the Calgary Regional Health Authority, the southern half of the Province of Alberta, and parts of southern Saskatchewan and British Columbia. Some patients have even transferred themselves to Foothills from neighbouring American states.

Educational emphasis: The purpose of all of these reviews is not to pick out any 'bad apples' but to identify items of educational importance. One problem identified was that of delay, in either recognition of complications, in transfer of the patient, or in treatment of the problem. Although delay is well recognised as a major contributor to maternal mortality (refs. 33-34), further educational activities are required to re-emphasize its role in maternal complications.

In fact, this emphasis on education and not on discipline was designed to conform to the Critical Review Policy of the Calgary Regional Health Authority. This Policy was developed from one originally proposed and implemented in the Department of Anaesthesia at Foothills Hospital in 1986 (ref. 35). The purpose of the Policy is to provide immediate, multidisciplinary identification, management and investigation of critical incidents and events, and to prevent similar or related events through correction of any factors that contributed to less than optimal outcome. These tasks are carried out using a non-judgmental approach, to avoid impeding communication and to facilitate fact-finding. Investigations include any of the involved medical, nursing, or other staff. Consideration is given, not only to the actions of the individuals involved, but also to faults in the system in which those individuals work and which contributed to the problem (ref. 36).

The results of the investigation lead to the development of recommendations which may be in the form of educational activities, the aim of which is to promote behavioural changes, and the formation of policies with respect to personnel, equipment, the environment and the organization. These recommendations may be released to the patient and or the patient's family. Responsibility for ensuring that recommendations are implemented lies with the hospital and regional health authority. The same level of authority also tracks the implementations, at a department, hospital or regional level, to ensure compliance. In addition, the identification of new problems, secondary to the recommendations, is vigorously sought (ref. 37).

The goal of the Critical Review Policy has been to ensure improvement in patient care, and not to reduce the motivation of patients to engage in litigation. The initial work (ref. 35) on which the policy was based was developed out of a sense that *"patients and their families have a right to know the facts about a complication or even a near hit"* (ref. 37). Proper communication with the patient and/or the family may result in a lessening of anxiety and fears that 'there is a cover-up' or that they are not being told the truth. More than a decade of experience with this policy in the Department of Anaesthesia at Foothills Hospital has revealed that patients and their families want to know what happened, and that steps have been taken to ensure that it does not happen to anyone else. As a secondary benefit, this policy may then indeed potentially lessen the risk of litigation (ref. 37).

Further development: At this stage of development, plans for the database do not include assigning a single ‘cause’ of death. In part, this decision is a reflection of the difficulties shown by others in attempts to classify deaths (ref. 38), particularly when clinical and autopsy information do not agree (ref. 39) or when the quality of the perinatal autopsy does not adhere to defined standards (ref. 40). A larger part of the decision is based on the appreciation of the multiplicity of factors that contribute to the occurrence of an accident. This appreciation has led to rejection of the term 'cause' by some workers. Aviation investigators have led the way in moving from the position of seeking 'cause' to the position of attempting to identify the *“factors that lead to an occurrence or safety deficiency "* (ref. 41). For example, the Australian Bureau of Air Safety Investigation (BASI) seeks ’significant factors’, where a ’factor’ is an ’element in the circumstances of an incident/accident without which the incident/accident would not have occurred’. BASI discarded ’cause’ as it is often difficult (or impossible) to derive single cause-effect relationships in incidents and accidents. For example, bad weather may have been a significant factor in an accident, but it could not have been ‘the’ cause. Although some countries still require safety investigations to nominate a single 'most probable cause' of accidents, BASI does not because it is not productive from the perspective of achieving prevention of similar events. Nominating a single 'cause' tends to focus corrective action on that one factor only, although many others typically are identified in an investigation. In addition, the term 'cause' has legal implications, including the concepts of blame and liability, which are not helpful. This is particularly so in the investigation of an incident or accident in aviation or in anaesthesia, as the purpose of investigation should not include apportioning blame or liability. However, there will be an attempt to identify ‘causal relationships’, through recognition of the role of multiple factors and their inter-relationships. The importance of the identification of the specific mechanisms through which such *“social, economic and cultural factors lead to the event of interest”* (e.g., mortality) has previously been emphasised (ref. 42). The result of this identification should be the production of recommendations to make changes in the structure and process of care.

The next step in the development of the maternal and neonatal audit programme is application for a grant to convert the paper-based questionnaire into a web-compatible computerized database. The current paper version is comprehensive but requires a road map. This has been provided in the form of tables or grids for each of ‘what’, ‘how’ and ‘why’. These grids are used to guide the reviewers through the hospital record, which is not organised in the same way that the database is!

One important factor to be considered in the computerisation is data security, one aspect of which is medico-legal litigation. In Alberta, Section 9 of the Alberta Evidence Act provides protection from civil suit subpoena for those medical quality assurance activities carried out by doctors. This Act is currently under review and it is hoped that revisions will allow inclusion of the activities of non-physicians. If so, then the new database could be used by nurses and midwives to assist with audit of labour ward practice, as has been described by the developers of an on-line perinatal data collection and information system (ref. 43).

While awaiting the results of the grant application, the questionnaire will be used (and further refined) for both retrospective review of previous maternal morbidity (post-delivery Caesarean Section) and neonatal deaths (in those infants with a cord pH < 7.0 at the time of delivery) and concurrent review of any new problems. Once the computerized version is developed and accepted, region-wide implementation by the Department of Obstetrics and Gynecology and the Perinatal Mortality Review Committee will be undertaken. Interest in using the new database has also been expressed by the Alberta Medical Association Committee on Reproductive Care. This would allow comparison of system-based data from different regional health authorities, such as gestational age distributions for geographically defined populations (ref. 44). For example, Calgary has the highest umber of pre-term births, which might relate to the number of cigarette-smoking others, but which could also reflect the influence of other factors, such as altitude.

In summary, development of a system-oriented database should allow all those who care for mothers to be to provide the best possible structure, process and outcome of care, to learn from problems, and to continue to make changes to improve the system.

References

1. Reason J. Managing the Risks of Organizational Accidents. Aldershot, United Kingdom: Ashgate Publishing Company, 1997.
2. Wolfson M.C. Social proprioception: Measurement, data and information from a population health perspective. *In:* Evans RG, Barer ML, Marmor TR (Eds). Why are some people healthy and others not? The determinants of health of populations. New York: Aldine de Gruyter, 1994.

3. Moodley D, Payne AJ, Moodley J. Maternal mortality in Kwazulu / Natal: need for an information database system and confidential enquiry into maternal deaths in developing countries. Tropical Doctor 1996;26:50-4.

4. Davies JM. From Dryden to Winnipeg and all points beyond. *In:* Hayward B, Dwyer A (Eds.) Beyond Human Factors. Proceedings of the Fourth Australian Aviation Psychology Association Conference. Aldershot, United Kingdom: Ashgate Publshing Ltd., in press.

5. Davies JM. Critical incidents during anaesthesia. *In:* Healy T, Cohen PJ (Eds.). Wylie and Churchill Davidson's A Practice of Anaesthesia. London: Edward Arnold, 1995.

6. Helmreich RL, Merritt AC. Culture at Work in Aviation and Medicine. National, Organizational and Professional Influences. Aldershot, United Kingdom: Ashgate Publishing Company, 1998.

7. Donabedian A. Evaluating the quality of medical care. Part 2. Milbank Memorial Fund Quarterly 1966;11:166-203. (Discussion: 204 – 206).

8. Reason J. The contribution of latent failures to the breakdown of complex systems. Philosophical Transactions of the Royal Society of London 1990;B327:475-84.

9. Reason J. Human Error. Cambridge: Cambridge University Press, 1990.

10. Helmreich RL. Human factors aspects of the Air Ontario crash at Dryden, Ontario: Analysis and recommendations to the Commission of Inquiry into the Air Ontario crash at Dryden, Ontario. *In:* Moshansky, The Honourable VP (Commissioner). Commission of Inquiry into the Air Ontario Crash at Dryden, Ontario. Final Report. Technical Appendices. Ottawa: Ministry of Supply and Services Canada, 1992.

11. Davies JM. Quality assurance: learning from the American experience. Hospimedica. The International Journal of Hospital Medicine 1991;Jan-Feb: 14-6.

12. Reason J. Managing the Risks of Organizational Accidents. Aldershot, United Kingdom: Ashgate Publishing Company, 1997.

13. Hart HLA, Honore A. Causation in the Law, 2nd Ed. Oxford: Clarendon Press, 1985

14. Cooper JB, Newbower RS, Long CD, McPeek B. Preventable anesthesia mishaps: A study of human factors. Anesthesiology 1978;49:399-406.

15. McGrath JE and Kelly JR. Time and Human Interaction. Toward a Social Psychology of Time. New York: The Guildford Press, 1986

16. Wigglesworth JS. Monitoring perinatal mortality. A pathophysiological approach. The Lancet 1980;ii:684-6.

17. Keeling JW, MacGillivray I, Golding J, Wigglesworth J, Berry J, Dunn PM. Classification of perinatal death. Archives of Disease in Childhood 1989;64:1345-1351.

18. Amar HSS, Maimunah AH, Wong SL. Use of Wigglesworth pathophysiological classification for perinatal mortality in Malaysia. Archives of Disease in Childhood 1996;74:F56-9.

19. Cole SK, Hey EN, Thomson AM. Classifying perinatal death: an obstetric approach. British Journal of Obstetrics and Gynaecology 1986;93:1204-12.

20. Hey EN, Lloyd DJ, Wigglesworth JS. Classifying perinatal death: fetal and neonatal factors. British Journal of Obstetrics and Gynaecology 1986;93:1213-23

21. Settatree S, Watkinson M. Classifying perinatal death: experience from a regional survey. British Journal of Obstetrics and Gynecology 1993;100:110-21.

22. Cole S. Classification of perinatal death. European Journal of Obstetrics & Gynecology and Reproductive Biology 1991;41:17-26.

23. Wigglesworth JS. Investigation of perinatal death. Archives of Disease in Childhood 1987;62:1207-8.

24. Salanave B, Bouvier-Colle MH. The likely increase in maternal mortality rates in the United Kingdom and in France until 2005. Paediatric and Perinatal Epidemiology 1996;10:418-22.

25. Fiktree FF, Gray RH. Demographic survey of the level and determinants of perinatal mortality in Karachi, Pakistan. Paediatric and Perinatal Epidemiology 1996;10:86-96.

26. Ibrahim SA, Babikert AG, Amin IK, Omer MIA, Rushwan H. Factors associated with high risk of perinatal and neonatal mortality: an interim report on a prospective community-based study in rural Sudan. Paediatric and Perinatal Epidemiology 1994;8:193-204.

27. McCaw-Binns A, Greenwood R, Ashley D, Golding J. Antenatal and perinatal care in Jamaica: do they reduce perinatal death rates? Paediatric and Perinatal Epidemiology 1994;8 Supplement 1:86-97.

28. Golding J, Henriques J, Thomas P. Unmarried at delivery. II. Perinatal morbidity and mortality. Early Human Development 1986;14:217-27.

29. Chiswick ML. Commentary on current World Health Organization definitions used in perinatal statistics. British Journal of Obstetrics and Gynaecology 1986;93:1236-8

30. Ifenne D, Essien E, Golji N, Sabitu K, Alti-Mu'azu M, Musa A, Adidu V, Mukaddas M. Improving the quality of obstetric care at the teaching hospital, Zaria, Nigeria. International Journal of Gynecology & Obstetrics 1997;59 Supplement 2:S37-S46.

31. Hibbard BM, Anderson MM, Drife JO, Tighe JR, Gordon G, Willatts S, de Swiet M, Shaw R, Thompson W, Gwyneth L, Botting B. Report on Confidential Enquiries into Maternal Deaths in the United Kingdom 1991 – 1993. Department of Health Welsh Office, Scottish Office Department of Health, Department of Health and Social Services Northern Ireland. London: Her Majesty's Stationery Office, 1996.

32. Barron SL. How can we improve perinatal surveillance? British Journal of Obstetrics and Gynaecology 1986;93:1201-3.

33. Maine D. The strategic model for the PMM Network. International Journal of Gynecology & Obstetrics 1997;59 Supplement 2:S23-S25.

34. Djan JO, Kyei-Faried S, Twum S, Danquah JBO, Ofori M, Browne ENL (The Kumasi PMM Team). Upgrading obstetric care at the health center level, Juaben, Ghana. International Journal of Gynecology & Obstetrics 1997, 59 Supplement 2:S83-S90.

35. Strunin L, Davies JM. Doing what is right in anesthesia. Canadian Medical Association Journal 1986;134:1232.

36. Eagle CJ, Davies JM, Reason J. Accident analysis of large-scale technological disasters applied to an anaesthetic complication. Canadian Journal of Anaesthesia 1992; 39: 118-22

37. Davies JM. Risk assessment and risk management in anaesthesia. *In:* Aitkenhead AR (Ed.). Quality Assurance and Risk Management in Anaesthesia. Bailliere's Clinical Anaesthesiology International Practice and Research 1996;10, 357-72.

38. Cole SK. Accuracy of death certificates in neonatal deaths. Community Medicine 1989;11:1-8.

39. Duley LMM. A validation of underlying cause of death, as recorded by clinicians on stillbirth and neonatal death certificates. British Journal of Obstetrics and Gynaecology 1986;93:1233-5.

40. Wigglesworth JS. Quality of the perinatal autopsy. British Journal of Obstetrics and Gynaecology 1991;98:617-23.

41. Lee RB. How BASI operates. Asia-Pacific Air Safety 1998;March:4-16.

42. McCarthy J. The conceptual framework of the PMM Network. . International Journal of Gynecology & Obstetrics 1997;59 Supplement 2:S15-S21.

43. Maresh M, Dawson AM, Beard RW. Assessment of an on-line computerized perinatal data collection and information system. British Journal of Obstetrics and Gynaecology 1986;93:1239-45.

44. Macfarlane A, Cole S, Hey E. Comparisons of data from regional perinatal mortality surveys. British Journal of Obstetrics and Gynaecology 1986;93:1224-32.

Biography

Prof. Jan Davies, Department of Anesthesia, Foothills Medical Centre, University of Calgary, Calgary, Alberta, Canada T2N 2T9. Telephone: 403 670 4707. Facs: 403 670 2425. Email: jdavies@ucalgary.ca.

Prof. Davies has a long-standing interest in all aspects of anaesthetic safety, including unexpected postoperative deaths, quality assurance and risk management. She also has a strong research and consulting relationship with aviation and has collaborated on a number of projects with various workers in the field of aviation safety.

Human Factors Engineering is the Basis for a Practical Error-in-Medicine Curriculum

J.W. Gosbee,

Center for Applied Medical Informatics,
Michigan State University Kalamazoo Center for Medical Studies, USA

Keywords: error curriculum, human factors engineering, error countermeasures, teaching

Abstract

It is essential that health care practitioners and students learn the underlying reasons for their mistakes and appreciate appropriately designed countermeasures. During training, medical practice, and administrative duties, health care personnel need to have a constructive attitude toward and knowledge about error-in-medicine. However, many clinicians are led to believe that humans are perfectible, and the "blame and train" approach is the optimal route to improving patient care. In contrast, a practical error-in-medicine curriculum — based in human factors engineering — has been taught to over 300 medical, nursing, and pharmacy students during health care informatics courses. The three main aims of the "error" curriculum are to have the students: 1) become an asset to organizational quality improvement activities; 2) become better evaluators and selectors of medical software and devices; and 3) assist and encourage feedback to medical software and device companies. Over five years, the teaching sessions have become more problem-based, simplified, and fun. Short and long term student evaluations (interviews and questionnaires) have been very positive. Many practical lessons have been learned which can support dissemination of this curriculum. The main points of this paper are to show that teaching about error-in-medicine is necessary, effective, and relatively easy.

Introduction

There are significant preventable adverse events throughout the health care system. Large retrospective chart reviews in the eastern United States revealed 3.7 percent of patients admitted to the hospital suffer an adverse event (ref. 1). Prospective studies in an Israeli intensive care unit uncovered more than one preventable error per patient, per day (ref. 2). Almost half of the fifty-three family physicians interviewed admitted to a mistake that led to a premature patient death (ref. 3). Books have been written on the subject of error in medicine (ref. 4) and a myriad of organizations created with the purpose of improving patient safety (refs. 5-7).

There is no lack of ambitious and comprehensive ideas to identify and reduce errors in health care settings. Leape (ref. 8) has suggested that error interventions need to occur at nearly every stage and level of clinical practice. From studies in many high risk industries, including medicine, Vincent (ref. 9) cites the limited usefulness of safety initiatives that are too narrow. Vincent proposes a framework for analyzing risk and safety in health care that also recommends a wide assessment. Finally, continuous quality improvement methods and systems thinking are touted throughout the United States and the world (ref. 10).

Will any of these efforts come to fruition without error-in-medicine education of researchers, educators, and, most importantly, health care personnel? If we are going to "attack" error and its root causes on "all fronts" will we not need thousands of recruits within all aspects of medical care? Most agree that awareness of causes and solutions of error in medicine will encourage and improve participation in quality improvement activities.

However, two other important reasons for this education relate to the human factors engineering discipline. First, the learner needs to gain an interest and ability to assist and encourage feedback to medical software and device companies. This is actually required by organizations like the United States Food and Drug Administration (ref. 11); and done voluntarily through organizations like ISMP (ref. 5). Second, the learner needs to become a better evaluator and selector of medical software and devices. In other words, the training will need to improve the learner's consumer skills to "push" industry to adopt human factors engineering methods during design, which results in fewer latent errors (ref. 12-13).

Error-in-medicine training could fall in many places within the student's required course work. However, many aspects of understanding human error involve information handling and decision making. Therefore, it makes sense to include this training within a health care informatics curriculum. Since 1995, the educational activities described below have been used for teaching over 150 students and residents during a month-long medical informatics rotation at Michigan State University Kalamazoo Center for Medical Studies (ref. 14). The curriculum has also been used for the "Drug Information" month-long rotation for doctor of pharmacy students at Ferris State University (Big Rapids, MI, USA). Finally, the human error curriculum is tailored to be an integral part of a "Nursing Informatics" course at Western Michigan University (Kalamazoo, MI, USA).

The main points of this paper are to show that teaching about error and medicine is necessary, effective, and relatively easy.

Similar Curricula

Teaching about the nature and reduction of human error within the context of human factors engineering is already done in universities, government, and industry. Many universities offer an introductory course for students seeking degrees in psychology, engineering, and computer science (ref. 15). Some have tried to identify the commonality in teaching to semi-interested students (i.e., human factors engineering is not their major area of study) (ref. 16). Organizations and industry in high risk domains like aviation and nuclear power provide ongoing training courses relating to error and error countermeasures for many of their workers. Some of this curriculum and format for teaching could be applied to health care students.

Studies on teaching human error and human factors engineering principles to health care personnel and students are limited. In the United Kingdom, a study was done with 129 nurses to see how their coping mechanisms with errors affected their learning mechanisms (ref. 17). Pilpel and his colleagues in Israel aimed their teaching program at medical students to "impart a tolerance of error" (ref. 18). They posited, but had no outcomes data, that teaching about the inevitability of error and how it can be reduced would lead to acceptance and, subsequently, to candid reporting. They also theorized that the curriculum would help the students realize their doubts and fears were shared by peers and superiors. From the interview data of 53 physicians, Ely and his colleagues concluded that physicians who understand common reasons for mistakes may be better equipped to prevent recurrence (ref. 3). Gaba and other groups are using anesthesia simulators and principles of human factors engineering to teach residents and practicing anesthesiologists (ref. 4). The continuing education is often done over a few days of simulating critical incidents with informed reviews of the videotaped sessions. Finally, Wu and his colleagues have written several articles on their study of medical residents and their errors, as well as strategies for educators to teach about and cope with those situations (refs. 18-19).

Requirements and Regulations in the United States

Professional groups and governing bodies for training medical doctors, nurses, and pharmacists have addressed the need to teach about the nature and reduction of error in medicine. Governing bodies for graduate and undergraduate medical education in the United States have several requirements for teaching quality assurance, continuous quality improvement, and systems thinking. The enumeration of these is beyond the scope of this article. The American Nurses' Credentialing Center, partnered with American Nurses Association, lists human factors engineering (which would include error) as one of seven areas of competency (ref. 20). The American Society of Health-System Pharmacists cites the human factors engineering training of pharmacy directors as one of seven key strategies to reduce adverse drug events in hospitals (ref. 21). The Joint Commission on Accreditation of Healthcare Organizations may require the contribution of all these health care personnel in root cause analysis of sentinel (significant) events (ref. 22). Lastly, the United States Food and Drug Administration comes just short of requiring the training of physicians and nurses to participate in reporting and analyzing errors that occur with medical devices (ref. 23).

Overview of the Error Teaching Sessions

The optimal aims of an error-in-medicine curriculum are to have the students: 1) become an asset to organizational quality improvement activities; 2) become better evaluators and selectors of medical software and devices; and 3) assist and encourage feedback to medical software and device companies. Within the resources and practical time constraints at Michigan State University Kalamazoo Center for Medical Studies (MSU/KCMS) and Western Michigan University (WMU), the objectives of their error-in-medicine teaching sessions are for students to:

1) Understand the scope and gravity of error in health care settings

2) Gain a familiarity with human perceptual limitations and cognitive biases, and learn that they are uncontrollable, yet very predictable

3) Know theoretical and practical reasons why “blame and train” and “bad apples” approaches fail

4) Understand importance of discovering root cause towards proper countermeasures

5) Become familiar with human factors engineering and continuous quality improvement techniques that determine root causes and help design countermeasures

6) Understand major categories of error countermeasures, especially the role of computers

7) Understand limitations and pitfalls of automation as a countermeasure

8) Understand that some latent errors and systemic problems are exacerbated by poor design

The fourth-year medical students and residents began to receive error-in-medicine training in 1994 within a month-long medical informatics *elective* rotation at MSU/KCMS. The doctor of pharmacy students began taking their *required* month-long drug information month along with the medical students and residents in 1997. Groups of two to six students have reading assignments, home work, and two two-hour teaching sessions early in the month. The first session is an introduction to human factors engineering with an emphasis on error and user-centered software design. The second two-hour session is devoted entirely to error, and is described in full below. The entire month of the medical informatics curriculum has been described previously (ref. 14), and will soon be located on the Internet (www.kcms.msu.edu). Almost none of our students receive education regarding errors previous to this elective, except some brief exposure to quality improvement principles.

Since 1996, the nursing students have been taught about error-in-medicine as part of a course in “Nursing Informatics” at the School of Nursing, Western Michigan University (Kalamazoo, MI, USA). In summary, they receive three consecutive one-hour sessions, including: 1) human factors engineering and nursing computer systems and devices; 2) small group system evaluation exercises; and 3) human error and nursing. As with the month-long rotation, the first session is an introduction to human factors engineering with an emphasis on error and user-centered software design. The small group exercises support concepts in both the first and third hour. Most nursing students have had several months of clinical experience.

Early Evolution of the Curriculum

When the error-in-medicine teaching sessions began in 1994 for the medical residents and students, much of the curriculum was adapted from introductory human factors textbooks and mostly didactic (ref. 24). The lecture started with a detailed presentation of three different types of error-in-medicine journal articles (ref. 1-3). For the remaining lecture, an overview of error theories, applicable cognitive psychology theories, taxonomies of errors, and a listing of countermeasures were presented.

Some of the material adapted from textbooks seemed to be based on the needs and mind sets of psychology or engineering students.

Early Feedback and Adult Learning Principles: Most of the early feedback from course surveys and informal interviews was positive. For example, one of the students said, *"[It] gave me a better understanding of medical tools and analysis."* However, some learners found the lecture material interesting, but not relevant. Others did not ask many questions or interact during learning sessions. This experience and literature on adult learning principles (ref. 25) prompted a change towards more relevancy, realness, simplicity, and added fun. Some of this evolution is similar to problem-based learning changes that were being made at that time in the Michigan State University College of Human Medicine curriculum (East Lansing, MI, USA). Finally, lessons on format, readings, and exercises were borrowed from studies of introductory human factors course effectiveness (ref. 26). With these considerations, experimentation with elements of the course and improvement continued through several sessions, since the material is taught 10-12 times per year.

Curriculum Description

Self Study: Required chapters from Norman (ref. 27) and Casey (ref. 28) provide very readable introductions. These books are very popular in introductory human factors courses, and highly recommended to any audience new to these principles. Both books use everyday examples. Norman weaves them into the basic principles of user-centered design and human factors engineering. Casey tells several riveting short stories about all sorts of disasters, including a few medical ones, to implicitly illustrate the human cost of error-exacerbating system design.

Common Highlights of the Teaching Sessions: The session begins with a review of two Internet sites that contain interesting and obvious attempts to catalogue and mitigate errors in high risk situations: kayaking and SCUBA diving (refs. 29-30). Then, some students are asked to recount their everyday encounters with slips, mistakes, and errors. These are noted for discussion at appropriate times during the session. Next, data from Brennan (ref. 1), Ely (ref. 2) and Gopher (ref. 3) are used to prompt a sobering discussion on the scope of error in health care. The students then give their definition of human error, these definitions are written for the whole class to see, and the instructor highlights key words and phrases for comparison and contrast. To stimulate discussion, the instructor shows how some people can claim that error doesn't exist, and that "error" is simply normal human performance. To further these ideas, a radar screen and vigilance graph are presented to show error is part of normal performance. The vigilance graph is a simplified result of several subjects watching a radar screen for "enemy ships" in a laboratory situation, and performance decreases over several hours. Following this, the graph is redone so the "y" axis shows errors (missed ships), rather than per cent of ships detected, to demonstrate that error can be considered the "flip-side" of human performance.

The major portion of the teaching occurs when four to five students list and comment on actual experiences with error in health care. They are asked not to identify their role in the error, but they need to know enough detail to describe the circumstances leading to the event and the clinical and personal consequences. After each description is listed for all to see, the instructor helps the class determine the root cause(s) and possible countermeasures. During this analysis, s/he interjects several principles, including: findings from human error/performance theory and taxonomy; methods for root cause analysis; countermeasure types (e.g., warnings, interlocks); and the limitation of error theories and study methods.

An example would be: Incident description: ordered Clonidine 1.0 mg, instead of 0.1 mg in middle of night, which led to the patient falling down and pharmacist alerting physician that dose was x10 in the morning. Clinical and personal consequence: patient was nervous with a few bruises, but no lasting effects; the resident was embarrassed and reprimanded, but the rest of "team" ignored it. Root causes: fatigue; inexperience; and missed steps and checks in getting and giving medications. Possible countermeasures: computerized order entry that checks for unusual doses; and the resident must be standing when giving phone orders at night.

In these courses, a major point of emphasis is the limitations and pitfalls of using computers as countermeasures (fixes) for errors. Many in health care have identified the role of computers in addressing errors, including medication errors (ref. 21) and as reminders for oft forgotten clinical interventions (ref. 31). Unfortunately, many computer systems solve the wrong problem or do not address the error in a usable manner. At best, the computer does nothing to eliminate target errors. At worst, it introduces insidious new problems (ref. 32). A recent empirical study showed that only 4 out of 307 pharmacy computers correctly identified all unsafe medication orders in a field test (ref. 33). The researchers suggested the root causes of the failures were: complex programming, hard-to-use human-computer interfaces, and unrealistic time commitments needed to properly to maintain and use the systems. Most importantly, the students are given those rare medical informatics articles where the investigators have appropriately used human factors methodologies to develop information systems (ref. 34).

At the end of the session, all the lessons are summarized using the student's error stories and a "Swiss cheese" diagram adapted from Reason (ref. 35). All of the countermeasures at the level of the device (or software), the clinician, and the organization are represented as three pieces of swiss cheese (i.e., barriers), with holes of various sizes, and arranged in a row. The "arrow" of an observed adverse event can only make it through if all the holes line up. The sizes of the holes are redrawn smaller to demonstrate how improved device design, improved clinician knowledge and error awareness, and continuous improvement of the organization can all foil adverse events. In contrast, the holes for device or software are redrawn much larger to show that many existing systems were not designed with human factors methods. This, in turn, shows how the "pressure" is on individuals and organizations to keep "holes" small. In addition, it illustrates how they should be "pushing" manufacturers to use human factors engineering methods to create "Cheddar" instead of "Swiss" devices and software.

Finally, a customized group of case studies and hand-out materials are provided, depending on class composition. For example, a recent newsletter from the Institute for Safe Medication Practices is provided to doctor of pharmacy students (ref. 5). Or, during the nursing courses, stories encountered directly by the instructor are recounted, such as catastrophic errors with IV pumps used by home health nurses.

<u>Another View of the Curriculum:</u> Even with customization and progressive changes, there are several common elements to the teaching of error and health care.

A useful way to view the error and medicine teaching activities is to organize them across four important considerations: real, relevant, illustrative and simple, and fun. Most activities or elements of the curriculum are organized in this manner in Table 1. As mentioned above, these considerations arise from the experimentation on the content and format (the basic session has been taught over 50 times).

<u>Additional Exercises and Teaching Sessions:</u> Although they are briefly covered in the error teaching sessions, there are many topics and skills that are reinforced. Decision making errors and use of clinical decision support systems are also covered in later sessions (ref. 36). Projects assigned for the month or semester are mainly evaluation of information technology or devices with human factors methods. This can range from field observation to usability testing with low-fidelity prototypes. The students are also assigned to ongoing projects as observers or data collectors. In some cases, the students are given self-study lab exercises, such as short human factors evaluation of an IV pump prototype. In the nursing classes, all of the students do a brief, yet real, evaluation of a familiar system (e.g., Legos(TM) toy, calculator, pocket-sized lens cleaning kit). Lastly, the pharmacy students are specifically required to learn about processes to analyze and report medication adverse events (ref. 23), and occasionally medical students and residents learn with them.

<u>Ongoing Evaluation of the Curriculum</u>

<u>Medical Student and Residents Comments:</u> Given the small size and frequency of the teaching sessions, constructive comments are the most common and useful feedback mechanism. During class, many of the students comment that this was the first time they understood <u>why</u> they might be involved in quality improvement efforts at the hospital or clinic. Some of the students and residents "breathe a sigh of relief" when they see the theory behind the mistakes they make, how some of the errors are exacerbated by poor

design, and the potential for properly designed information systems to help them. In surveys at the end of the rotation, some have noted that the error or human factors teachings were the most unique and surprisingly useful. Unsolicited, a few have even said that the information should be a required part of early medical school courses, and that they were "mad" that it took so long to hear about the topic. It should be noted that formal surveys after each teaching session have recently been implemented, and the data so far are very positive.

Real	**Relevant**	**Simple** (illustrative)	**Fun**
• The students tell their stories of clinical errors • The students are given current case studies of medication and device mishaps from the ISMP newsletter • Presentation of current case studies (medical journals or popular press) • Students are usually involved in system evaluation projects where errors occur during testing	• The students tell their stories of clinical errors • There is purposeful emphasis on case studies that end in death or maiming • Presentation of case studies from literature within their specialty •Students are usually involved in system evaluation projects where errors occur during testing	• The students are asked for recent, everyday examples of error • "Best-of" case studies, that are not current or relevant (e.g., coffin-shaped pills not for internal use) • Swiss cheese diagram adapted (simplified) from Reason • Performance and error graphs from the radar example	• Small group exercises to evaluate errors with common items • Humor by instructor and class • Review of Internet databases on kayak and SCUBA diving deaths and injuries

Table 1 - Elements of Curriculum Organized Across Four Important Parameters.

<u>Follow-up Interview with Medical and Pharmacy Students:</u> An in-depth semi-structured phone interview was conducted with a convenience sample of six residents, pharmacy and medical students who completed the month long rotation. The six students had taken the rotation six, eight, nine, ten, twelve, and twenty-two months previous to the interview. Two physicians and two pharmacists were now full-fledged practitioners. Two were now resident physicians. All remembered the error and medicine lecture and class exercises, but not to any level of detail. All but one of them said that the change in attitude about system design had persisted. The most vivid memory for four of them was their new understanding about the lack of user-centered design, and how this deficiency led to hard-to-use systems and errors. Some of the more illustrative quotes from the interviewees were: "it helps you attack the problem [of error], instead of avoid it"; "I think I was very impacted by your course"; "I remember thinking it [the course] was very helpful...stuff that was thought to be common sense does need study"; and "computers and technology needs to be usable and user-friendly or [it will be] rejected". The participants' permission to use these quotes is on file.

<u>Nursing Informatics Students:</u> A semi-structured questionnaire was completed by the four nursing informatics classes that received human factors and human error teaching sessions in 1998 and 1999 (N=67). The intent of the questionnaire was not research, but for

continuous improvement of the sessions. Tables 2 and 3 below summarize the number and per cent of responses for the two questions: “What was the most useful portion of the class tonight?”; and “What was the least useful portion of the class tonight?”

Responses	**Frequency**	**Per Cent of Total**
Class exercises were a good learning experience, allowed clarification of the ‘human error concept’, or provided insight into the human factors process	36	54%
Case studies helped to analyze why systems are not working in health care, they gave insight into problems that occur, or showed that systems often fail	24	36%
The first lecture helped apply human factors to nursing practice	5	7%
Slide handouts aided understanding or helped students follow the slides	2	3%
Total Comments	67	100%

Table 2 - Most Useful Portions of the Class

It is clear from Table 2 that class exercises and case studies are most preferred. The results from Table 3 are of such low frequency and scattered to make much of a claim. It is of note that 54 of 67 (80%) nursing students either filled in nothing or explicitly stated that there was no “least useful” part of the training.

Anecdotally, other nursing faculty have told the nursing informatics instructor that they can tell when students have taken the nursing informatics course because the students will use phrases like, "that's a human factors problem." Students were also heard to say that the course helped them recognize that they, themselves, are not the problem when they are slow to learn computers. Instead, they recognize that the problem is very likely poor design (ref. 37).

Major Issues

The toughest issue for most new curriculum ideas in health care education is the problem of how to fit them into an already full curriculum. As described above, error-in-medicine teachings can be added where it best fits into a fourth-year medical school elective, a residency elective, or a required course. Without better learning outcomes data, it is hard to claim this approach is sufficient. People that champion other new medical curriculum ideas, like computer training and evidence-based medicine, are trying to include new courses, or alter required ones, early in training. It would make the most practical sense to offer error-in-medicine training before students began working with patients in hospitals and clinics. Given the number of hours devoted to such training in other domains, the curriculum should probably be longer than three or four hours.

Another recently highlighted criticism by Casarett and Helms is the possibility that the systems approach to error may teach irresponsibility or laziness (ref. 38). The authors suggest that if residents are able to put blame on “systems”, they may not work as hard to modify their lack of knowledge. First, this assumes that the main reason for this systems error and medicine training is to encourage and improve participation in quality improvement activities. In the curriculum above, supporting continuous quality improvement activities is only a small part of the overall aims (goals). Second, the authors describe a limited root cause analysis, which only addresses organizational and other flaws outside the control of the resident. Human error is not either the system or the person, it is almost always a combination of root causes that need to be carefully analyzed, and deficiencies appropriately fixed. In the error-in-medicine teaching at MSU/KCMS, the aim is to teach residents to see when errors are due to inadequate skills and knowledge versus when they are due to *inherent* cognitive limitations and biases. Finally, it is not the intent to have error-in-medicine training and causal analysis stop at self or organizational improvement. The learners should get improved consumer skills to "push" industry to adopt human factors engineering methods during design (ref. 39). This, more than any regulation or journal article, will lead to less error-prone medical software and devices.

Responses	**Frequency**	**Per Cent of Total**
Nothing was “least useful”; all portions of the class were useful	26	53%
I either couldn’t hear or see the video or didn’t understand what it was about and needed a better example of observer evaluation	5	11%
History portion in the first lecture	4	9%
Class Exercises were geared more towards engineers and not health professionals; Didn’t need 30 minutes to evaluate the [Legos(TM)] toy	2	4%
Not used to having multiple breaks in one class	2	4%
Sometimes difficult to follow the outline, repetitive	2	4%
Inadequate explanation of Human Factors and the concept was difficult to understand	2	4%
Miscellaneous (one each): improve exercises, ‘dry’ notes, discussion of examples, too long, human factors information was unnecessary	5	11%
Total comments	48	100%

Table 3 - Least Useful Portions of the Class

Practical Lessons Learned

There are several things one should know to prepare for and consider while teaching error-in-medicine to almost any health care student or practitioner. You need to change the learner’s perspective and preconceptions first, or many will not listen to the remaining lessons. This is usually not too difficult because the existing perspective is not ingrained. As described above, the case studies, examples, and references have to be as relevant as possible for the type of health care student, and their specialty (e.g., medication errors for pediatric ICU for a pediatric resident). Be prepared for learners who are resistant or arrogant. Focusing on how improving the system helps their “less fortunate” colleagues is one method. With more time, you can carefully make them use error-prone devices and software, and then show them how a human factors redesign diminishes such errors. Be prepared for learners who get stuck on worrying about litigation, and how “nothing”, even human factors engineering redesign, will help stop the lawyers. You can address this helplessness by noting the fact that there is little overlap in the subsets of malpractice suits and preventable adverse events (ref. 1, 4). You can also state that systems thinking and human factors engineering analysis are their duty as practitioners as keys to offering the very best care. Finally, you will be tempted to overwhelm students with too many references, complex theories, and tangential examples. Stick with a few examples and theoretical points that you know very, very well. Unless they are getting an advanced degree in this area, the textbooks on error or human factors engineering are too much.

If you are training the educators who will train students, or are preparing to teach error-in-medicine yourself, please consider these recommendations. The educator will get negative reactions from some, so novice teachers should prepare themselves with on-hand references (see above). It is easy to get confused about the major types of errors (slips and mistakes) and countermeasures (e.g., warnings, training, interlock, etc.). The teachers must have a user-centered, human factors engineering philosophy themselves or the lessons will not ring true. Teachers will have to recall their own "attitude change" to foster one in the students. The teacher must be able to deal with ethical and legal issues (ref. 18-19). Specifically, the student might ask how to now report an error or work through a root cause analysis in their hospital or clinic. The contact person or standard procedures will need to be readily available. The teacher should practice being able to coach students on how to approach superiors about incidents or near misses. Finally, plan for a few learners who will have significant emotional reactions from their memories of incidents or the discussion of the material, in general.

Conclusions and Next Steps

The error-in-medicine training for doctors-, nurses-, and pharmacists-in-training is easy, fun, effective, and desperately needed. The author of this paper offers his support to those that want to try it. In turn, some feedback is appreciated. Some health care organizations might not realize they have people with human factors engineering or error-in-medicine backgrounds in their midst (e.g., anesthesia, pharmacy, or radiology departments). Others might try asking faculty or graduate students from their local (friendly) university department (listed at ref. 15, 40, 41). It is hard to imagine how all the ambitious quality and safety programs around the world will succeed unless health care personnel have a constructive attitude toward and knowledge about error-in-medicine.

References

1. Brennan, T.A., L.L. Leape, N.M. Laird, et al., 1991. Incidence of Adverse Events and Negligence in Hospitalized Patients: Results of the Harvard Medical Practice Study I. *New England Journal of Medicine* 324: 370-6.

2. Gopher, D., M. Olin, Y. Badhi, et al., 1989. The Nature and Causes of Human Errors in a Medical Intensive Care Unit. In *Proceedings of the Human Factors and Ergonomics Society 33rd Annual Meeting.* Santa Monica, CA: Human Factors and Ergonomics Society.

3. Ely, J.W., W. Levinson, N.C. Elder, and A.G. Mainous, 1995. Perceived Causes of Family Physician's Errors. *Journal of Family Practice* 40(4): 337-44.

4. Bogner, M.S. 1994. *Human Error in Medicine.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

5. Institute for Safe Medication Practices. 1999. [online]. Warminster, PA. Available from World Wide Web: www.ismp.org.

6. Anesthesia Patient Safety Foundation. 1999. [online]. Pittsburgh, PA. Available from World Wide Web: http://apsf.med.yale.edu.

7. National Patient Safety Foundation. 1999. [online]. Chicago, IL. Available from World Wide Web: www.npsf.org.

8. Leape, L.L., 1994. Error in Medicine. *Journal of the American Medical Association* 272: 1851-57.

9. Vincent, C., S. Taylor-Adams, and N. Stanhope, 1998. Framework for Analyzing Risk and Safety in Clinical Medicine. *British Medical Journal* 316: 1154-1157.

10. Berwick, D.M., 1996. A Primer on Leading the Improvement of Systems. *British*

Medical Journal 312: 619-622.

11. Food and Drug Administration Safe Medical Devices Act of 1990, 1995. Medical Device User Facility and Manufacturing Reporting: Certification and Registration (21CFR Part 803). *Federal Register* 60(237): 63578-63607.

12. Burlington, D.B., 1996. Human Factors and the FDA's Goals: Improved Medical Device Design. *Biomedical Instrumentation Technology* 30(2): 107-109.

13. Lin, L., R. Isla, K. Doniz, H. Harkness, K.J. Vicente, and D.J. Doyle, 1998. Applying Human Factors to the Design of Medical Equipment: Patient-controlled Analgesia. *Journal of Clinical Monitoring and Computing* 14: 253-263.

14. Stahlhut, R.W., J.W. Gosbee, and D.J. Gardner-Bonneau, 1997. A Human-centered Approach to Medical Informatics for Medical Students, Residents, and Practicing Physicians. *Academic Medicine* 72(10): 881-887.

15. Human Factors and Ergonomics Society. 1999. [online]. Santa Monica, CA. Available from World Wide Web: www.hfes.org.

16. Gibb, R.W., 1998. Undergraduate Human Factors Curriculum and Introductory Course Content. In *Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting.* Santa Monica, CA: Human Factors and Ergonomics Society.

17. Meurier, C.E., C.A. Vincent, and D.G. Parmar, 1997. Learning from Errors in Nursing Practice. *Journal of Advanced Nursing* 26(1): 111-119.

18. Wu, A.W., S. Folkman, S.J. McPhee, and B. Lo, 1991. Do House Officers Learn From Their Mistakes? *Journal of the American Medical Association* 265: 2089-2094.

19. Wu, A.W., S. Folkman, S.J. McPhee, and B. Lo, 1993. How House Officers Cope With Their Mistakes. *Western Journal of Medicine* 159: 565-569.

20. Healthcare Information Management Systems Society. 1996. *Guide to Nursing Informatics.* Chicago, IL: Healthcare Information Management Systems Society.

21. American Society of Health-System Pharmacists, 1996. Top-priority Actions for Preventing Adverse Drug Events in Hospitals: Recommendations From an Expert Panel. *American Journal of Health-System Pharmacists* 53: 747-51.

22. Joint Commission on Accreditation of Healthcare Organizations. 1999. [online]. Oakbrook Terrace, IL. Available from World Wide Web: www.jcaho.org.

23. Food and Drug Administration, 1995. Medical Devices: Medical Device User Facility and Manufacturing Reporting: Certification and Registration (21CFR Part 803). *Federal Register* 60(237): 63578-63607.

24. Sanders, M.S., and E.J. McCormick. 1993. *Human Factors in Engineering and Design.* 7th ed. New York: McGraw-Hill, Inc.

25. Nowlen, P.M. 1988. *A New Approach to Continuing Education for Business and Professions.* New York: Macmillan.

26. Sojourner, R.J., A.J. Aretz, and K.M. Vance, 1993. Teaching an Introductory Course in Human Factors Engineering: A Successful Learning Experience. In *Proceedings of the Human Factors and Ergonomics Society 33rd Annual Meeting.* Santa Monica, CA: Human Factors and Ergonomics Society.

27. Norman, D.A. 1988. The Design of Everyday Things. In *To Err is Human.* New York: Basic Books, Inc.

28. Casey, S. 1993. *Set Phasers On Stun and Other True Tales of Design, Technology, and Human Error.* Santa Barbara, CA: Aegean Publishing.

29. *American Whitewater Safety Database.* 1999. [database online] Lakewood, CO. Available from World Wide Web: www.awa.org/awa/safety/index.html.

30. Divers Alert Network (DAN). 1999. [online] Durham, NC. Available from World Wide Web: http://dan.ycg.org.

31. McDonald, C.J., 1976. Protocol-based Computer Reminders, The Quality of Care and

the Non-perfectibility of Man. *New England Journal of Medicine* 295: 1351-5.

32. Louie, C. and A. Luber. 1996. *Automated Drug Delivery Systems and the Potential for Medication Misadventures: A University's Experience.* New York: American Association for the Advancement of Science.

33. Cohen, M. 1999. Over-reliance on Pharmacy Computer Systems May Place Patients at Great Risk. *ISMP Medication Safety Alert!* 4(3): 1.

34. Thull, B., U. Janssens, G. Rau, and P. Hanrath, 1997. Approach to Computer-based Medication Planning and Coordination Support in Intensive Care Units. *Technology in Health Care* 5(3): 219-233.

35. Reason, J.T. 1990. *Human Error.* Cambridge, England: Cambridge University Press.

36. Riegelman, R.K. 1991. *Minimizing Medical Mistakes: The Art of Medical Decision Making.* Boston, MA: Little, Brown and Co.

37. "Personal Communication with Kathy Young." 1999. [electronic mail]. [cited February 1999].

38. Casarett, D., and C. Helms, 1999. Systems errors versus physicians' errors: Finding the balance in medical education. *Academic Medicine*; 74: 19-22.

39. J. Gosbee, 1998. Communication among health professionals: Human factors engineering can help make sense of the chaos. *British Medical Journal* ;316: 642.

40. International Ergonomics Association. 1999. [online]. Available from the World Wide Web: www.iea.org.

41. The Ergonomics Society. 1999. [online]. Available from the World Wide Web: www.ergonomics.org.uk.

Biography

Dr. John Gosbee, Director, Center of Applied Medical Informatics at Michigan State University Kalamazoo Center for Medical Studies, 1000 Oakland Dr., Kalamazoo, Michigan, 49008, USA. Telephone (616) 3370-4435. Email: gosbee@kcms.msu.edu. Web site:

www.kcms.msu.edu/cami/camihome.html

John Gosbee, MD, MS has been involved in education, research, and development aspects of error in medicine and patient safety for eleven years. As an expert in human factors engineering and health care, he has provided leadership to interdisciplinary professional societies in these areas (e.g., Human Factors and Ergonomics Society), as well as consultation to government agencies (e.g., U. S. Food and Drug Administration), trade organizations (e.g., Canon Communications) and medical technology companies (e.g., Baxter Healthcare).

Fragmentation in Care and the Potential for Human Error

Andrew Shepherd and Olga Kostopoulou,

Cognitive Ergonomics Research Group, Loughborough University, UK
Keywords: neonatal intensive care, information exchange, teamwork, task analysis.

Abstract

In neonatal intensive care medical and nursing staff must coordinate care and treatment to facilitate a baby's recovery and development. Staff must monitor the baby's health and treatment in order promptly to detect deviation from a satisfactory state and identify when treatments are not working as planned. Detected problems require a swift response — to predict the consequences of the deviation, to diagnose a problem, or to revise the current care and treatment plan. Such decision-making requires knowledge of principles of care and treatment, the operation of equipment, the significance of biological indicators, rules of discretion, as well as knowing something of the baby's history.

While the care and treatment of babies must be continuous over days or weeks, members of staff change. All staff are subject to shift-changes. Nurses and doctors also have different sorts of expertise and responsibility and are, therefore, engaged with the baby in different ways at different times. Collaborating effectively entails communication — through records, formal handovers, and informal verbal exchanges.

This paper considers issues of task fragmentation and problems associated with communication systems designed to minimise its effect. It describes strategies for understanding neonatal intensive care tasks aimed at identifying sources of latent human error and overcoming the potential problems of fragmenting care.

Introduction

The beginning of independent life is a dangerous time for a baby. A complication during labour or an illness can have serious or even life-threatening consequences. Moreover, to survive, babies born prematurely must be given special support and care, above and beyond what parents alone can offer. Such care is provided in Neonatal Intensive Care Units (NICUs). The majority of patients in NICUs are *premature* babies and it is in this group that 60% of all neonatal deaths occur (Merenstein & Gardner, 1993) . NICU care and treatment entails keeping those babies alive and well, so that their organ systems mature with time and the babies can function autonomously.

Neonatal Intensive Care is carried out by highly qualified medical and nursing staff. Despite their commitment, their working arrangements embody features that can give rise to human error. Some latent errors arise from the nature of individual tasks. Several tasks are intrinsically difficult due to their high information content, the probabilistic nature of some information, the uncertainty of outcome, the criticality of error, and the stressful context in which work is done. However skilled individual members of staff may be, their effectiveness will be blunted if the information upon which they base their judgements is inadequate. Problems arise in this respect because of the inherent difficulties of the need to collaborate and communicate effectively within a team. We argue that difficulties that arise as a consequence of inadequate information exchange between team members can only be dealt with by exploring team functions and more effectively understanding the respective duties of colleagues.

Research Strategy

Methods for investigating real behaviour in real contexts are not clear-cut. Clearly, behaviour is influenced by the information available to personnel and the resources that can be used to make changes. It is also affected by the expertise of staff, their attitudes and perspectives, their individual and collective experience and the factors which affect and have affected the circumstances that emerge to be dealt with, including the costs and consequences of actions.

Cognitive psychology has a role to play in understanding how human beings use information to carry out decision making tasks, but its methods tend to preclude a proper understanding of the range of factors that affect performance in real contexts. Cognitive psychologists tend to focus on testing hypotheses about the structures and processes of cognition and are usually motivated to devise tasks that suit their interests, rather than explore real tasks. Ethnographers take a different perspective, observing behaviour in a real context, whether or not this behaviour is directed towards the goals of the system of which they are a part. This approach may fail to take account of infrequent events if these do not occur during periods of observation (Room, 1984). This would be a critical omission since infrequent events often yield serious outcomes and have not been anticipated. Therefore, contingency plans may not have been devised or appropriate responses practiced by staff.

Where a researcher's motivation is directed towards establishing hypotheses for improving aspects of a system, the system's goals must be understood. However, so too must those activities that do not appear directly related to meeting systems goals. Understanding some personal goals can point to strategies that staff adopt for coping with the main goals, or with the tedium or other stresses of the work context. These may be vital for a workplace to be effective. Thus, two consultants from different units engaged in diagnosing the problems of a seriously ill baby were seen to talk, from time to time, about common acquaintances and organisational aspects of their respective hospitals. This was done in a light-hearted manner. When one of these consultants was questioned afterwards he admitted that he was deliberately trying to inject some normality into proceedings. In some cases, people may do things that are counter-productive and these too need to be understood in terms of the purpose that these activities serve.

Neonatal intensive care exists for a purpose and staff are employed to prosecute that purpose. However, as with all work contexts there are many things going on besides the pursuit of the stated goals of the group. These can include social tensions and rivalries. Interactions between staff can be influenced by a desire to promote relationships as well as the need to communicate information pertinent to care and treatment. In the case of neonatal intensive care, the goals of staff can change from that of promoting life to that of managing a dignified death. Staff are cognisant about the well being of families as well as the babies and adapt their behaviours to balance these objectives in an appropriate manner. Issues such as these must be picked up by any researcher as a prime focus for research or as part of the context that needs to be understood in order to investigate cognition in an appropriate manner.

In resolving the interactions between contextual factors and cognitive factors we question the validity of any simple approach. We have sought an understanding of what staff are employed to do. We have identified the tasks that staff must carry out to achieve these goals and the contextual factors that affect performance. We have identified the factors that could be redesigned or modified to improve performance. We have then examined these constituent tasks to establish how information is used. We are currently investigating design hypotheses concerned with helping staff communicate more effectively.

Fieldwork in NICUs:

We have undertaken fieldwork at 4 different NICUs. The aim of these visits was to observe the complex NICU environment, to provide a general perspective on how tasks are carried out, and to identify when, why and from what sources, information is obtained. These observations showed how units were organised and drew attention to the fact that units are all different. Discussion with staff also placed in perspective the purpose of neonatal intensive care, which was crucial to a proper understanding of how units function.

Initially, our intentions were to record observations faithfully. With greater familiarity with units, we became more reticent about this. We were reticent about filming and recording in front of families. Families observed at the point of greatest crisis were distraught and vulnerable. We were equally concerned about risking the trust of staff in recording utterances that they might subsequently feel compromised their positions. Part of this problem rested with the nature of shiftwork. We made substantial effort to explain the purpose of our work to staff, yet continually encountered staff who had not been at our briefings or who had not read the documentation we had left for other staff. Moreover, work pressures in units were so heavy it was difficult to arrange satisfactory briefings. Any researcher entering this field must confront these questions of ethics and trust. A researcher would need to feel confident that the benefits of such recording justified these intrusions.

We attempted a number of formal approaches to recording data. We attempted a form of activity classification both at shift handover and during the shifts. Activity charts were used to establish the extent to which a nurse's attention switched between patients and between the various care activities. We also attempted to obtain estimates of the nurse's perception of risks to the baby as a shift progressed. However, all of these efforts proved to be of little benefit during our initial investigations and little insight about tasks was gained. Part of the problem was due to the nature of the different cases. Common events, such as nursing routines that occurred every hour, were easy to document and did not need any formal observational methods to identify what was happening. Non-routine events, however, were difficult to observe because we did not fully understand terminology and motivation. In an attempt to understand what was going on, we asked questions to staff while they were working. We took care to do this only when staff had time to answer and avoided interrupting during obvious critical activities. Most of our questions concerned the purpose of a current activity, how tasks fitted together and the information links between them. Questioning was also directed towards documentation issues. To understand the documentation better, we looked at records and talked to staff. We looked at how records were structured, when they were filled in and how they aided care decisions.

As we became more familiar with these units, we became more focused and sought answers to specific questions, concerning the information recorded on the nursing and medical notes, how nursing and medical notes differ, how doctors communicate treatment plans to nurses and protocols for shift handover. We explored these issues by studying situations of interest and questioning the persons involved. We employed retrospective analysis in cases where we observed something interesting and asked the person what went on. We confronted nursing staff with notes to enable them to talk us through what happened and why certain decisions were taken. Input from doctors was also sought to complement these observations.

The outcome of these interventions was a clearer understanding of the purpose of neonatal intensive care, how staff worked with one another and how staff carried out their duties. From this we were ably to obtain greater insight into organisation of units, team activity, the potential risks to care and the manner in which tasks are carried out and information is used.

The Organisation of NICUs

Location: NICUs are situated in large tertiary care centres near the maternity ward. They are divided in three areas: maximal intensive care, high-dependency intensive care and special care. Babies are allocated to each area according to the severity and instability of their condition. Most babies requiring maximal intensive care are ventilated and their vital signs are continuously monitored. Our research focuses mainly on maximal and high-dependency intensive care because these areas involve higher information load due to greater number of interventions, monitoring a greater number of parameters, and dealing with more unstable patients requiring more frequent assessments, changes in treatment and documentation.

Staff and their Duties: The staff of units serve both the care and treatment of babies and the organisation of the unit itself. Thus staff are concerned with working within a budget and meeting a range of organisational targets.

Medical staff in NICUs include *Consultants*, *Registrars* and *Senior House Officers* (SHOs). Consultants and Registrars are involved more in treatment planning than in direct provision of care, while SHOs carry out invasive procedures, prescribe drugs and fluids and make a limited number of treatment decisions. Consultants and Registrars are specialists in neonatal intensive care. SHOs are not specialist but are on the unit to undertake a medical placement that might last only 6 months before moving to another placement in a different area of medicine. Thus, there is no reason to assume that SHOs have any specific career aspirations in neonatal intensive care.

Senior Consultants have a managerial role too. This includes ensuring the unit is adequately resourced and interacting with the wider managerial structures of the hospital.

Most of the nurses in NICUs are qualified in neonatal intensive care. Nurses are allocated 1 to 3 babies during each shift and are constantly with them. They monitor them, keep them comfortable, feed them according to the care plan, administer medication and record their observations routinely and also when changes occur. They bring such changes to the attention of doctors who decide on treatment modifications. Some of the more senior nursing staff have managerial roles. Nurse Managers have an important role in staff deployment and maintaining standards. Ward sisters provide a supervisory role in which they ensure that day to day performance of staff is satisfactory.

While it is useful to separate managerial and organisational aspects of care from the direct care and treatment of babies, these aspects often interact. For example, decisions concerned with admitting babies or allowing them to go home affect both the best interests of the baby and their families and organisational factors such as the availability of cots and the use of other resources. Equally, if an SHO were undertaking a routine test on all babies in the unit a ward sister may choose to limit access in order to minimise disturbances to the babies. This may be inconvenient to the SHO carrying out the tests but judged to be in the best interests of the babies.

The general patterns of duties described above applied in all units, but it was clear that there were several variations regarding how these activities were carried out. The manner of communication and collaboration between staff varied for a variety of reasons. Some staff were new to this form of nursing. Some staff were experienced at this form of nursing yet had little experience of working in their current unit. Other staff had considerable experience of colleagues and of the unit. They understood the strengths and weaknesses of colleagues and reflected this is the way in which they were able to share descriptions of babies and conditions. It meant that they could communicate with one another in more coded and more efficient ways than colleagues without such familiarity. For example, staff could refer back to cases where they had a common experience.

Formal differences between staffing on units also existed with regard to the employment of specialist categories of staff. For example, *Nurse Tutors* were experienced nurses who also had unique unit wide experience and responsibility. Their seniority enabled them to obtain a broad perspective that was beneficial to the unit in ways other than simply providing educational support. *Nurse Practitioners* were particularly interesting. These were usually senior nurses who had been trained to undertake responsibility for some of the more invasive treatments as well as taking responsibility for care. Thus, they were able to provide a more seamless support for the baby, without the need to communicate between professions. They were specialists and were able to build up substantial experience. They were, therefore, able to compensate for the comparative lack of experience of SHOs. Moreover, Nurse Practitioners provide excellent insight into the benefits of integrating care and treatment.

The Working Goals of NICU Staff

Critical to our understanding of NICUs was properly understanding the goals of staff. Our perception of their tasks shifted from initially assuming a model of ‘diagnose and treatment’ to recognising that, in general, the role of staff in NICUs was to sustain life systems until the baby was able to mature and become self-sufficient. Premature, and some full-term babies who experience problems prior to birth and during labour, need special support after they are born. The aim of staff in NICUs is to sustain life of such babies by appropriately augmenting immature organ systems and treating disturbances, for example, in supplementing oxygen and controlling the thermal environment. When the baby's condition is stabilised, staff aim gradually to reduce and eventually remove this augmentation in order to promote self-sufficiency. Then the baby is safe to go home.

Documentation

Formal documentation is essential within NICUs as a means of communicating information and as a means of recording what action was taken. Such records of activity can serve in cases where a review of treatment is necessary and in cases of litigation.

Two of the most important forms of documentation are the *Treatment Plan* and the *Care Plan*. The Treatment Plan is created by the doctor in charge of a case. It will specify treatments to be given and conditions to be maintained. This provides guidance for later medical staff who might be concerned with modifying treatments in the light of new circumstances, it can provide targets and it provides guidance for nursing staff concerning parameters for monitoring. The Care Plan is the province of the nursing staff. It will set out the pattern of care activities to be followed and will record systematically the baby's response to care and treatment.

Treatment and Care Plans look forward. Staff are also required to maintain records, specifying the nature of each intervention, current status of key health parameters and any observations made concerning the nature of treatment. Thus, exceptional circumstances would be recorded here, including the kinds of things that ought to be conveyed to colleagues.

Litigation is the reason that nurses are required to record practically everything, irrespective of whether its importance. This is not only time consuming but also results in large amounts of information making it difficult to for staff integrate it and to single out the most important events. Moreover, there is extensive duplication of information both between and within medical and nursing records which is prone to transcription errors. Gopher et al. (1989) found that 28% of the overall activities in a respiratory ICU were recorded in both medical and nursing notes, while 18% were not recorded at all, suggesting incomplete documentation. Indeed, staff were aware of these problems and the possibility of combining records in a more rational way was raised on a number of occasions.

Neonatal Intensive Care as a Team Activity

Neonatal intensive care is a complicated activity that requires several interdependent individuals collaborating for a common goal, performing different tasks and contributing their individual expertise. Nurses need to be constantly with the care babies to monitor their vital signs, respond to changes and provide routine care. Doctors assess babies, decide on their treatment and sometimes determine the conditions for care and monitoring. The contribution of other experts, such as physiotherapists, lab workers and social workers, may also be required. At the level of action, however, neonatal intensive care is essentially an individual activity, as a single nurse looks after one or more babies during a shift; it is less the case that several people co-operate to carry out a single task. Gopher et al. (1989) estimated that 84% of the activities carried out in a respiratory ICU were performed by a single nurse. The doctor is called upon when the nurse cannot or is not allowed to deal with a non-routine situation. This constitutes a difference with other supervisory tasks, such as process control, where operators work. In power generation or transport control, for example, each operator may be engaged in controlling different parts of the same system and, therefore, should be able to form a representation of what each other is doing in order not to interfere with what their colleagues are trying to do. This contrasts with NICU nurses who may not need to know what goes on with the rest of the babies in the unit (although they often have some general idea which helps them to cover for their colleagues for short breaks). They do need to know what went on before they took over caring for a particular baby, and they need to know the goals of the current treatment plan.

<u>Concurrent Activity:</u> When doctors are involved with the baby, the onus remains on nurses to maintain their basic nursing duties. This was brought into focus during one incident observed. The seriousness of the baby's condition engaged the attention of 3 Consultants. Normal routines were severely interrupted and the baby's nurse seemed to focus her attention towards supporting the doctors. After observing this incident for about 2 hours, the nurse seemed then to start working independently of the doctors to carry out normal recording and make more typical nursing interventions. When questioned afterwards, she admitted that the incident had prompted them to suspend their normal activities and then she realised that the basic nursing functions were being neglected.

<u>Information Collaboration:</u> Since carers (nurses and doctors) work independently, they gain different information on the basis of their involvement in a baby's care. The amount and

content of information gained will be a function of the events to which the carer has been exposed and the tasks carried out. As a result of continuous contact with the babies, a nurse is likely to know a great deal more about a particular baby than doctors or other practitioners who have only short and intermittent contact. Information gained at any one stage of care can influence execution of care tasks at another stage. Since different carers carry out different tasks, they need to have all the appropriate information available. This has implications for complete documentation and efficient communication.

Doctors rely on information provided by nurses to make treatment decisions. This information is usually documented but most of the time, according to our observations, it is provided verbally by the nurse either voluntarily or upon the doctor's request. At the same time, doctors may possess some types of information that need to be shared with the nurse. This includes, diagnostic or prognostic information and information about treatment goals and reasons underlying treatments, which could help the nurse to deliver care more efficiently and be more alert to changes in the baby's condition. Cases were reported where Doctors had stipulated care and treatment actions without notifying nurses of their motivations. For example, to enable the baby to develop self-sufficiency, it is sometimes necessary to withdraw some aspects of treatment to prompt the baby to cope. However, this will subject the baby to temporary discomfort. Where the nurse is unaware of the Doctors motivation in this respect, he/she may act in a way to alleviate discomfort, by adjusting sedation, for example. In this way, the Doctor's strategy for treatment is compromised.

Information communication is therefore a very important aspect of neonatal intensive care and a determinant of its quality. Gopher et al. (1989) found that the dominant cause of errors in a respiratory ICU were problems related to complete and clear documentation and transfer of information between staff members. Staff in the NICUs that we visited characterised communication problems as one of the most common problems in the everyday operation of the units.

<u>The Traditional Roles of Doctors and Nurses</u>

Relationships between staff appear to owe more to the traditional roles of doctors and nurses than to any rational assessment of how such relationships and lines of communication should best be organised and supported in neonatal intensive care. Thus, certain demarcations in responsibility are assumed, with nurses 'nursing and caring', and doctors 'diagnosing and planning treatment'. In some cases, the distinction is explicitly stated, with the implication that nurses have little discretion.

In reality, these distinctions are blurred, with both types of staff engaging in both types of activity. For example, nurses may facilitate diagnosis and treatment planning by anticipating and collecting relevant information in order to inform doctors. In one incident reported to us, an experienced nurse was prompted to question the validity of a 'normal' blood oxygen reading. Taken at face value, the baby's health was entirely satisfactory. However, this nurse had taken note of prior treatment interventions and had noted that four hours earlier, on the previous shift, the baby had been given an increase in oxygen and no change to this had been made since then. The Nurse extrapolated from the current condition to the next occasion for routine monitoring and was concerned that the baby would be at risk from excess oxygen. She had no discretion to act directly, but was at liberty to call the Consultant who then made the appropriate adjustment.

Equally, doctors must appreciate the importance of 'minimal handling' and ensure that treatment interventions do not compromise recovery. Minimal handling is a care principle of utmost importance for nurses because it is this that allows the baby to rest — this is one of the most effective means by which a baby can recover. Specialist Consultants and Registrars appreciate this, but the less experienced SHOs may not. It is clear that there can be conflict between the inexperienced SHO and the experienced Nurse. Effective senior nurses are strong characters indeed!

If this anticipation and co-operation did not occur, then it would be difficult for NICU teams to work so effectively. Therefore, examination of the organisation of NICU teams and the information infrastructure that supports them is important in resolving performance problems.

Communications

Cooperation depends upon communication. There are a number opportunities or channels for communication.

Shift handover: Shift handovers occur between nurses going off shift handing over responsibilities to other nurses and doctors going off shift handing over responsibilities to other doctors. The role of shift handover is to facilitate a smooth transition from one shift to the next without interruption in care. It is intended to help oncoming staff quickly to update their knowledge of the babies and to identify changes that occurred since they were last on shift. It can helps them to put the babies in order of severity and to prioritise attention. It can provide information on things to look out for, on unusual and important events that occurred during the previous shift, events that may have been recorded but may be missed amidst the detail of other recorded information. It can provide a summary of what needs to be done and of any changes in the care plan that were decided during the previous shift.

However, there are risks inherent in handover. For example, some important information may be omitted, there may be ambiguity in the verbal communication leading to misperceptions, erroneous inferences and suppositions. Biases may be introduced in the information search of the oncoming shift (Grusenmeyer, 1991). The outgoing shift may fail to direct the oncoming shift's attention to important events and aspects of care. The oncoming shift may fail to memorise or take into account some transmitted information or may fail to check the validity and completeness of the transmitted information before starting care activities and continue with a wrong practice 'inherited' from the previous shift (Grusenmeyer, 1995).

In some of the NICUs that we observed, handover was done from memory and neither the outgoing nor the incoming shift kept notes. Moreover, oncoming nurses often started their care activities without consulting the records. Shift change is indeed a critical phase of work in terms of safety; a higher number of accidents have been found during shift changeover in other contexts than during the shift or at the end of the shift (Godimus et al., 1977, cited in Grusenmeyer, 1991). Gopher et al. (1989) found peaks of errors immediately following each shift change of nurses in a respiratory intensive care unit. According to Grusenmeyer (1991), there is a period when the oncoming operator is trying to gain information and is not "fully operational" for his tasks which increases the likelihood of incidents.

Our observations of handovers revealed a vast range of personal styles that were difficult to assess from the viewpoint of their adequacy. A major problem that arose was that staff were conscious of being watched and tended to be more comprehensive that they might otherwise have been — this was as much to help inform us than simply that they were on their best behaviour. We witnessed a wide range of styles. In some cases, a nurse would systematically work through a standard template of observations, while in other cases the nurse would be briefer. Sometimes the on-coming nurse would ask questions. Sometimes the nurse coming on shift was already familiar with the baby and little had changed. As babies become more stable and staff are more confidant of their prognosis, the handover communication centred more around the families and their increasing involvement in treatment. Sometimes, very little information would be conveyed and the on-coming nurse would move into a nursing routine that revealed nothing of the previous exchange. This suggested, that in many cases, the handover served little purpose in terms of obtaining information. Indeed handover could serve simply as a ritual to enable people to engage and disengage with the task or with coming to work. As a means of providing information, it could exist to confirm that all is as would be expected for a baby of these characteristics. It could provide specific information about crises or any observations concerning the baby's health or response to specific treatments.

Ward rounds: Ward rounds are a different context for information exchange in which all staff concerned with a baby are present. They occur daily and involve doctors going to each cot, reviewing and discussing the case, examining the baby and deciding on whether revisions to the treatment plan are warranted. The nurse in charge of the baby is usually present to provide information and to comment on the baby's progress. The doctors write the new treatment plan in the medical notes and the nurse updates the care plan in the light of the doctors' latest instructions.

Substantial verbal exchange of information takes place during ward rounds but only part of it is recorded. Doctors rarely record the reasons behind their decisions, although they may communicate them verbally to the nurse in charge of the baby or other doctors present. The nurse may not record the doctor's reasons in the nursing notes (it is not a type of information that she is required to record) and may not inform her shift successor about them (because she forgot or did not consider it important). Verbal information can easily get lost and staff often commented on this. Of course, people cannot record everything and read everything. It would be too time consuming and inefficient. Therefore, judging which information is important and would be useful to the next person coming on shift seems to be an essential skill in intensive care.

Fragmentation of Care and Treatment

NICU babies require continuous monitoring, as they can quickly become very unstable and their appearance may not sufficiently indicate an impeding deterioration. Moreover, only a limited number of invasive tests can be made on neonates, which makes continuous, non-invasive monitoring even more important (Cunningham et al., 1992). The interdependency of staff in the delivery of care suggests that if the forms of communication exchange were in some way deficient, then care could suffer. In the description of NICUs given above it is clear that care is fragmented as staff change shift and as they engage at different stages of the baby's treatment and care.

Fragmentation of care between doctors and nurses can also reflect their respective traditions and philosophies. For example, some doctors are not concerned with how the baby is establishing feeds; this is considered one of the nurses' main activity. Nurses are interested in maintaining the baby's comfort, whilst doctors are more interested in delivering treatment. Delivering treatment often compromises the baby's comfort, providing a constant source of tension between doctors and nurses. This gap between doctors and nurses is reflected in separate handovers and separate records which result in overlapping information but also in valuable information not getting across, simply because staff think that colleagues are not interested. Some members of both the nursing and medical staff have pointed out that it is necessary to have joint doctor-nurse treatment decisions. This suggests that they acknowledge the fragmentation in care and its potential consequences.

Functional Analysis of NICU teams and the tasks of team members

In cases where demarcation within teams has evolved through local practices, it is difficult to predict where performance failures due to inadequate staff interactions may arise. What constitutes an individual task is unclear, as is the relationship between individual tasks. This is made more complicated due to local variations in team organisation and relationships. To take account of these variations, we have adopted the strategy of first examining the general function of delivering neonatal intensive care and then seeing how this may be resolved into different tasks. This has been done using Hierarchical Task Analysis (Annett et al. 1971; Shepherd 1995, 1997). Hierarchical Task Analysis is a method of describing tasks in terms of the *operations* that aim to meet functional *goals* and *plans* which describe the conditions under which goals must be met to satisfy the conditions of their superordinate goal.

Figure 1 shows the top level of this analysis illustrating how staff must take information from the previous shift (1) to inform their principle activities of maintaining equipment (2) and maintaining care and treatment (3). At an appropriate stage of the day, the wardround is conducted to review each babies development and to enable information exchange to take place between different members of the team. Finally, responsibility must be handed over to the next shift (5).

0.Maintain functions of NICU

plan 0:

When taking over from previous shift - 1;

throughout shift - 2 & 3;

routinely each day — 4;

at end of shift — 5.

1. Get information from previous shift
2. Maintain equipment and supplies
3. Care for and treat babies in NICUs
4. Carry out ward round
5. Hand over to next shift

Figure 1: Main Tasks in NIC

The importance of this part of the analysis is to set the context in which care and treatment are delivered. Indeed, the elements of handover (1 & 5) cannot be understood without reference to what goes on within care and treatment (3). The general nature of the model at this level reflects how different members of staff may be engaged. Thus all NICU staff are concerned with handover, care and treatment and being engaged in ward rounds. We have included the operation concerned with maintaining equipment and supplies. If this is not done to a satisfactory standard, then the standards of treatment and care are at risk. We observed a critical event where a piece of equipment was called for during an emergency. First the technician bringing it went to the maternity ward instead of the NICU, thereby causing delay. Then, when the equipment arrived it was found to be unservicable and an alternative strategy for dealing with the problem had to be pursued. The potential consequences of this sort of event do not need to be spelled out. Clearly all staff must issue unambiguous instructions to each other and certain staff must be given responsibility to make routine investigations to guarantee the availability of such crucial equipment. This applies whether or not these staff are members of the NICU or other staff within the hospital.

The central activities that we wish to address are described in Figure 2. This shows how the tasks involved in ‘Care for and treat babies in NICUs’ (3) are set out. These functions are reasonably standard and would apply to most NICUs irrespective of their specific organisation.

<u>The Principal Goals of Managing Care and Traetment:</u> We have focused on the aspects of treatment and care provided in NICUs. This is seen as the external control of immature organ systems to facilitate development to a state where the baby may be safely discharged, transferred to another unit or be terminally cared for. For example, the babies may not be able to sustain their cardiorespiratory function and regulate their temperature and fluid homeostasis and, therefore, require assistance. Such control should be administered in a way that does not damage the baby's organ systems (e.g., lung or eye damage due to excessive mechanical ventilation is a common side effect of NICU treatment). External control should be gradually reduced and eventually removed, so that the baby can regulate its own bodily functions. Both provision of and weaning from support require careful assessment, continuous monitoring and adjustment. Moreover, they require dealing with routine deviations as well as tackling unexpected problems.

Figure 2 shows how care and treatment depends upon a cycle of activities including monitoring key indicators of the baby's condition and monitoring the adequacy of current care and treatment plans; delivering routine care and treatment; treatment planning; and communication. Some task elements, such as the monitoring tasks must be conducted throughout. Others are done in response to certain events, for example, the baby entering the ward or various unacceptable conditions arising.

For the purpose of discussion, we have categorised the types of tasks identified by the task analysis as *procedural* tasks, *monitoring* tasks, *care and treatment planning* tasks, *care management support* and *communication* tasks.

3. Care for and treat babies in NICUs

Plan 3:

When baby enters unit — 1.
When baby has been settled — 2.
Throughout shift — 3, 7, 8, 9 & 10.
If current treatment plan is judged as no longer suitable — 4.
At start of shift — 5.
If current schedule requires changing — 6.
At end of shift — 11.

1. Settle the baby and make initial measurements
2. Specify treatment/care plan
3. Monitor adequacy of current treatment/care plan
4. Revise treatment/care plan
5. Prepare schedule of treatment/care activity for baby
6. Review/revise schedule of activities for each baby
7. Carry out treatment/care and specified monitoring activities according to current schedule
8. Maintain baby's comfort
9. Review/evaluate whether baby is progressing satisfactorily in accordance with current treatment plan.
10. Record/relate significant observations, events and inferences.
11. Provide information to the next shift.

Figure 2 – Principle tasks identified in delivering care and treatment. Each of these is further developed in the full task analysis.

Procedural : These tasks are 'settle the baby and make initial measurements' (1) and 'carry out treatment/care and specified monitoring activities according to current schedule' (7). Care and treatment planning tasks are proceduralised to a large extent, due to the difficulty of diagnosing and treating highly unstable, highly interacting organ systems. Most of the time, treatment consists of selecting an established method and appropriately adapting it to accommodate individual differences and ongoing responses (Shalin & Bertram, 1996).

The word 'procedural' denotes tasks that are generally executed in a fixed way and are often straightforward, requiring little discretion. Procedural tasks in the context of neonatal intensive care include obtaining routine measurements such as temperatures, obtaining blood samples, administering drugs, delivering oxygen, giving fluids, inserting a catheter, intubating/extubating, suctioning etc. These are the 'normal' routines of nursing and treatment delivery. On further analysis it becomes clear that each of these is subject to variation, some of which requires judgements to be made by the nurse or SHO carrying out the task and some of which require information to be supplied by the doctor responsible for the treatment plan. Invasive treatments or the administration of drugs to alleviate pain, for example, will yield different responses from different babies. If a nurse recognises that one method of treatment is unsuitable for a particular baby, then an alternative method for achieving the same nursing goal must be found. The nurse will then know what to do on the next occasion and should inform colleagues to ensure that the baby is not discomforted unnecessarily. Thus the *team* can learn about the characteristics of individual babies. It soon becomes obvious that these 'procedural' tasks all rely on information being provided by colleagues to ensure they are dealt with most effectively. It is also clear that each task provides opportunities for staff to learn new things about caring for and treating the baby to communicate to other members of the team.

It must also be acknowledged, though, that some treatments are intended to be followed irrespective of some temporary discomfort to the baby. In these cases, the routes to follow must be made explicit to staff and reasons given. This must be done in a way in which instructions and motivations can be conveyed beyond the members of staff who were present when the original decision was taken. If this is not done, then nursing staff may promote comfort at the expense of the treatmentplan.

<u>Monitoring:</u> Three types of monitoring are identified. First, 'maintain baby's comfort' (8), is concerned with nursing decisions. Nurses will ensure that babies are not too cold or have been discomforted by an item of equipment slipping out of position, for example.

Second, nurses are engaged in reviewing and evaluating whether the baby is progressing satisfactorily in accordance with the current treatment plan' (9). Here, their job is to make general observations of the baby and to take various measurements from which judgements are made about the baby's progress. They will record the information or draw problems to the attention of a doctor. At one level this task appears straightforward. The nurse needs to know the target parameter, the target value and the tolerances to be accepted. It can be assumed that there is little judgement and expertise involved, a view held by some doctors. In practice, monitoring can be a complex task. Where a nurse is heavily loaded, he or she must prioritise and, therefore, needs to understand factors associated with criticality and recoverability in order to direct attention in an optimal way. Where a parameter is seen to go outside acceptable tolerance, the nurse must investigate the problem. Is a poor pulse signal the result of the pulse sensor being dislodged by the baby moving or is there general concern? Is the current acceptable reading of oxygen in the blood the sign of a healthy baby or is it a transitory state following an earlier intervention? If the nurse failed to anticipate these problems or failed to account of the factors that would direct an effective schedule of monitoring behaviour, then the nurse would be less effective, despite what doctors might say about the level of discretion a nurse has.

The third type of monitoring is 'monitor adequacy of current treatment plan' (3). If the baby's response to treatment is unsatisfactory, then the doctor might decide to be more cautious and revise the treatment plan. Alternatively, the baby might be progressing well and the doctor decides that treatment must move forward. For example, the nurse might notice that the baby is making spontaneous attempts to breathe and so the doctor might decide the baby should be weaned off the ventilator. Where such action is taken, the treatment plan is revised. The nurse needs to know this so that greater attention can be given to the baby during this critical phase of change and because other vital signs will be disturbed caused by anxiety brought on by the change.

<u>Care and treatment planning:</u> These tasks include '*s*pecify treatment/care plan' (2) and 'revise treatment/care plan' (4). The importance of the process of prescribing a treatment plan has been made clear. This is the

responsibility of doctors. Their task is to make medical judgements, either to identify a problem or to prescribe what to do next. Much neonatal care treatment planning is done in conditions of uncertainty. Because of the fragility of the organism and the idiosyncratic nature of some problems, doctors are often uncertain. Their task is not necessarily one of diagnosing, then confidently prescribing. They are often making their best judgement about what treatment will help stabilise the baby and allow it to settle before progressing the treatment. As a consequence, they must rely upon the effectiveness of the nursing staff in obtaining information and of nursing staff and SHOs in delivering care and treatment. Paramount, of course is the effective monitoring in between to ensure that treatment is progressing well. This means that unless nurses and SHO fully understand the motivations of doctors in prescribing a course of treatment and unless they understand the parameters to which they must now attend, the effectiveness of treatment might be compromised. The interelationship between task elements and, hence, staff, is again emphasised.

Treatment planning directs the strategy for care and treatment. It may involve diagnosis or simply prescription of an established programme of treatment on the basis of assessment of the baby's condition. Sometimes treatments are prescribed with little confidence, but something has to be done. Treatment planning is invariably followed by treatment actions, and then careful monitoring and assessment to determine treatment effectiveness. Thus, a treatment plan should inform monitoring and testing activities in terms of parameters to be monitored and diagnostic tests, routine or non-routine, to be carried out.

<u>Care management support</u>: This is seen as the management of the care and treatment interventions. It was emphasised to us that one of the prime treatments that babies can be given was to allow them to rest. Unnecessary disturbances will arrest progress. Exercising some degree of forward planning by care staff will ensure that everything that needs to be done will be done, but that it is coordinated to minimise upset. Thus, preparing and revising the schedule of treatment/care activity for the baby (5 & 6), needs to be taken seriously and respected as a task. We have seen strong-willed senior nurses drive this process very effectively, because they know that unless they do so, other staff may cause interruptions to care which could have been avoided.

<u>Communication:</u> Communications are vital. Both 'record/relate significant observations, events and inferences' (10) and 'provide information to next shift' (11) depend upon staff understanding what is of significance to promote treatment and care, above and beyond the information that will be routinely monitored. The information judged to be significant must be formally recorded or committed to memory. Where information is obtained during a complex procedure or decision making task, then the carer may simply forget to make a record of it.

<u>Errors and Error Handling</u>

Our observations and task analysis have revealed a number of opportunities for error inherent in the NICU system. Errors are frequent and often relate to unavailability of information. Fortunately, the system can often prevent such errors from developing into incidents. Defence mechanisms include duplication, staff memory and shiftwork. Thus, information that has been omitted from one record may exist in a different record. Information that has not been recorded may be remembered because it was verbally expressed at some point. Information that may be missed by one shift may be picked up by the next shift. However, it is evident that such defences are informal and sometimes may not work or may work to the detriment of the system. For example, duplication in records is time consuming and may lead to transcription errors resulting in inconsistencies. Memory of verbal exchanges is notoriously unreliable. NICU staff repeatedly comment on how often information gets lost or altered through verbal communication. The involvement of various members of staff in the care of an individual baby increases the likelihood of information being picked up but at the same time increases the likelihood of information getting lost. There is a big difference between a member of staff obtaining a piece of information and using it to perform a task and a member of staff obtaining a piece of information and communicating to a colleague who will then use it to perform that task.

The Experimental Programme

The potential sources of error revealed by our fieldwork are plausible yet remain hypothetical. The logic of the care task, observations made of events and incidents and comments made by various NICU staff give credence to these suggestions, but they do not consitute categorical evidence.

Our current phase of research is to investigate these potential weaknesses further and to investigate ways in which problems might be overcome. Thus, we have developed baby simulations whose health can be monitored, where problems to be introduced in their development and treatments can be administered. Our current simulation has been programmed on Stella™, a modelling package available to run on desktop computers. Individual babies are represented by running the baby simulation on different iMac™ computers. Different 'babies' are given different initial states, different recovery characteristics and different problems to contend with. We are using the simulations to invistigate how engagement with different tasks influences the manner in which people handle information and how they communicate with colleagues. Thus, we are investigating, for example, the benefits of different ways of training people to think about their task, different forms of representing records and different protocols for briefing colleagues. We also plan to investigate the development of teams over time, observing how their skills and relationships develop and how, as a consequence, they learn to share information.

We hope to establish how various elements, including computer displays, recording formats, training and protocols for verbal communication, can be designed to improve the handling of information in contexts such as these. The outcome of such work can never be in the form of a categorical finding that we can be confident of applying to real situations and we are fully aware of the constraints of this approach.. However, through such methods we would seek support for more viable design hypotheses than can be obtained through observational studies alone.

Observations and Conclusions

Our studies of neonatal intensive care have revealed a complex working environment in which different people with different expertise and responsibilities and working across different shifts, must collaborate in order to maintain the care and treatment of babies at extreme risk. Through observations and task analysis, we have identified a wide range of variations of practice, even though the basic processes of care are common between units.

A particular observation has been the complex interdependency between different sorts of care task, especially from the perspective of sharing information. The need for staff to collaborate is paramount. Yet there are clear indications that this collaboration can be compromised by problems of communication brought about by the fragmentation of tasks across staff.

Several aspects of the NICU context lend themselves to redesign, including the format of records, the effectiveness of training and protocols for shift handovers. Making recommendations in this respect depends upon further research into the design of information artifacts. This is one reason why simulation studies are recommended to supplement observational studies, even though the simulation of these tasks severely compromises the realities of the real NICU context.

One area that has been noted, which cannot be studied in a task simulation, concerns the traditions of the medical and nursing professions. The staffing of NICUs owes much to the traditional relationships that have emerged between doctors and nurses. Yet, if NICU staffing were to be specified from the perspective of good team and job design, the recommended allocation of responsibilities would probably look very different. There is scope for more clearly identifying how nurses undertaking monitoring and care tasks can more effectively anticipate the requirements of medical staff undertaking treatment planning. And there is scope for prescribing better ways for medical staff to convey their plans to the staff who will be responsible for carrying them through.

References

Annett, J., Duncan, K.D., Stammers, R.B. and Gray, M.J. Task Analysis. HMSO, 1971.

Cunningham, S., Deere, S., Elton, R. A., & McIntosh, N. Neonatal physiological trend monitoring by computer. International Journal of Clinical Monitoring and Computing, 9, pp 221-227, 1992.

Gopher, D., Olin, M., Badihi, Y., Cohen, G., Donchin, Y., Bieski, M., & Cotev, S. The Nature and Causes of Human Errors in a Medical Intensive Care Unit. Paper presented at the Perspectives. Proceedings of the Human Factors Society 33rd Annual Meeting, Denver, Colorado, 1989.

Grusenmeyer, C. La relève de poste - Une phase critique du travail en équipes successives. Cahiers de notes documentaires, 144, pp 407-418, 1991.

Grusenmeyer, C. Shared Functional Representation in Cooperative Tasks-The Example of Shift Changeover. The International Journal of Human Factors in Manufacturing, 5(2), pp 163-176, 1995.

Merenstein, G. B., & Gardner, S. L. Handbook of Neonatal Intensive Care. (3rd ed.). St. Louis: Mosby-Year Book, Inc., 1993.

Room, R. Alcohol and ethnography: a case of problem deflation? Human Organization,. 25, no. 2, pp.169-78,1984.

Shepherd, A. Task Analysis as a Framework for Examining HCI Tasks. Perspectives on HCI - Diverse Approaches. (ed. A. Monk and G. Gilbert), London, Academic Press: pp 145 - 174, 1995.

Shepherd, A. HTA as a framework for task analysis. Ergonomics. 41(11), pp1537 – 1552, 1998.

Shalin, V. L., & Bertram, D. A. (1996). Functions of expertise in a medical intensive care unit. Journal of Experimental and Theoretical Artificial Intelligence, 8(3-4), 209-227.

Acknowledgements

We would like to acknowledge all of the staff of Neonatal Intensive Care Units who have given us their time and demonstrated their expertise.

Biography

Dr Andrew Shepherd, Cognitive Ergonomics Research Group, Department of Human Sciences, Loughborough University, Loughborough Leics, LE11 3TU, UK. Telephone: +44 (0)1509 223028. Fax: +44 (0)1509 223940. Email: A.Shepherd@lboro.ac.uk.

Dr Shepherd is an applied psychologists with particular interests in task analysis, training, interface design, team .

Dr Olga Kostopoulou, Cognitive Ergonomics Research Group, (see above). Telephone: +44 (0)1509 263171 ext. 4271. Fax: +44 (0)1509 223940. Email: O.Kostopoulou@lboro.ac.uk.

Dr Kosopoulou is an applied psychologists with particular interests in diagnosis and faultfinding with a particular interest in medical settings.

Human Error in the Context of Work Activity Systems: A Case Study of a Knowledge-Based Computer System in Routine Use in Obstetrics

M. Harris[1], A.P. Jagodzinski[2] and K.R. Greene[1]
[1] Perinatal Research Group, Department of Obstetrics, Derriford Hospital, Postgraduate Medical School, University of Plymouth.
[2]School of Computing, University of Plymouth.

Keywords: human error, activity theory, case study, context

Abstract

A wider perspective of work activity and organisational context is central to understanding human errors. To help clarify the research approach required to investigate errors in context and inform design, we undertook a comparative case study of two Maternity Units where a knowledge-based computer system was in routine use. An activity theory analysis of the work activity system found the roots of errors were embedded in the wider work system rather than in direct interaction with the computer system. In one Unit, the role of the computer system was undermined as emergent, local adaptations contributed to errors, and the effects went unnoticed at an organisational level. Development of an effective relationship between computer systems and work systems, is an iterative process, which includes re-design of work activity. An appropriate research design and theoretical framework for the design of error tolerant computer systems and work systems needs to accommodate context and the process of change.

Introduction

The user and the user interface, the most visible aspects of human-computer interaction, have received a disproportionate amount of “blame” for human errors. However, it is important to remember that computer systems are embedded in a broader system of activity which is purposive and social. Over the past nine years our group has developed and validated a knowledge-based computer system to assist clinicians in the management of women during labour and delivery (refs.1-3). The Medical Research Council (MRC) have supported the work over the past three years to test the system and develop the user interface prior to a randomised control trial. A wider perspective of human work activity and organisational context is a central concept for understanding errors in human-machine systems (refs. 4-6). However, we found it difficult to identify a research approach which accommodated context and informed design. We undertook a preliminary investigation of an existing knowledge-based computer system in routine use, to clarify the research approach required to investigate errors in context and its potential contribution to the design of our prototype computer system. This paper describes our experience in the design of medical knowledge-based computer systems and:

- Contrasts traditional views of human-computer interaction with more recent perspectives.

- Reports a preliminary case study of an existing traditionally designed knowledge-based computer system in routine clinical use.

Traditional and Recent Perspectives on Human-Computer Interaction

Early studies of human computer interaction focused on cognitive models and quantitative research methods, but the artificial and unrepresentative nature of many studies limited their application (ref. 7). Numerous complex processes occur during a task, any one of which may affect performance. It is therefore difficult to generalise research findings to other tasks or systems (ref. 8).

Growing criticism and dissatisfaction with traditional approaches, combined with the growth of technology to support communication and collaborative activity, have fuelled interest in social and contextual paradigms from sociology, anthropology and philosophy (ref. 7).

Cognitive models: Early studies of cognition were often undertaken in controlled conditions removed from the context of work. Cognitive science attempted to develop theory-based models of information processing and user models to match computer system and user. For example the Goals, Operations, Methods and Selection model (ref. 9) for human information processing attempted to predict human behaviour at the computer interface. While it was a significant attempt to develop theory, most work was in text editing with error free performance by experts. Thus its scope is narrow as learning, skills acquisition and errors are an important part of human-computer interaction. Applying psychological principles to computer design has not been as successful as first hoped. For example recall, recognition and working memory are well described phenomena, but this does indicate that all interactions should be menu-based or that the number of displayed items should be limited to seven (ref. 10).

Distributed Cognition: There is now an increased awareness of the importance of context. More recent studies of work systems have studied how cognition is manifest at a system level rather than at an individual cognitive level. In this view knowledge is dispersed through departments, organisations, people in their roles and in the physical artefacts they use (ref. 11-12). Distributed cognition theory highlights the social and environmental nature of knowledge and views cognition as distributed and collective.

Situated Action: Behavioural and social components of cognition have traditionally been treated as extraneous variables. Situated action offers an alternative view in which cognition, action and situation are bound together in particular social and physical circumstances (ref. 13). In this view, cognition is a social action and context an essential resource. To study only individual internal cognition removes the phenomena of interest. Situated action focuses on individuals in emergent situations where action is flexible in dealing with unexpected events. In these circumstances situated action suggests that plans do not govern human activity but are a post hoc analysis of action. However, situated action excludes the subjective and does not account for learning or enduring structures.

Computer Supported Cooperative Work: Team performance and their shared mental models have been the focus of much research, particularly in industrial, commercial and military sectors. Teams comprise of interdependent individuals with different roles, responsibilities and knowledge (ref. 14). Studies of group dynamics offer some insight into human errors. Time-constrained situations may lead to “recognition primed” decisions, with selection of the first acceptable solution rather than an optimal solution. Social factors may lead to normative errors such as the outward group acceptance of a solution although internally each member disagrees (ref. 15) and lower status members often fail to intervene when higher status members’ performance is poor (ref. 16). Despite a considerable body of research, the factors which contribute to team effectiveness and the role of technological support are not well understood. Support systems which display all available information to team members appear to improve analysis, level of participation, confidence, satisfaction, quality of decisions and reduce dominance of the group by a single member (ref. 17).

Organisation and culture: Computer systems usually do not fit neatly into pre-existing practice. They transform organisations, processes and the way in which people work (ref. 18). These issues were often neglected and consequently problems emerged too late to influence the design process. Organisational theory suggests a multilevel analysis (individuals, task and organisation) before embarking on the design process to assess the appropriate level of intervention for knowledge-based systems (ref. 19). However, many conceptual models are structural and static and fail to accommodate the constant state of flux in organisations.

Ethnography, a qualitative research approach, has been used to study organisations and elicit hidden work and its meaning, to help clarify systems requirements. This approach is gaining acceptance in both engineering and medicine (ref. 20-21) and involves researchers entering the work culture to observe and map social discourse. While qualitative research adds richness and depth of understanding, it is not a complete method and does have some practical problems. The duration of studies is usually prolonged, the amount of data generated can be overwhelming, its role in informing design is not clear and its contribution to cumulative research results is limited.

Human factors: Human activity is extremely complex and influenced by many factors; personal, social and organisational. Unfamiliarity with a task, time shortage and poor signal to noise ratio have been ranked as conditions which increased the risk of error by a factor of ten or more. Surprisingly, factors traditionally considered as major contributors to errors such as sleep disturbances, hostile environments and boredom, appear to have less impact on the risk of errors (ref. 22).
The role of human factors in adverse events has been widely studied in air transport, nuclear power stations and processing plants. Human errors take various forms and have different psychological origins (ref. 6). These include attention failure, memory failure and mistakes through inadequate planning. Mistakes may be rule-based, when a solution is available but is misapplied or omitted, or knowledge-based, where the situation is novel and no solution readily available (ref. 6). In the field of human factors there is also a shift in focus from individuals to the wider activity system. In aviation and medicine, strategies to improve team activity and communication have had considerable impact on human performance (ref. 4).

Activity theory: Activity theory is a conceptual and analytical framework, developed from Soviet psychology, which provides perspectives, tools and a common language for understanding work activity in context. Activity theory avoids dichotomising thought from action, individuals from collectives or social from technical issues. Activity theory proposes that understanding the role of artefacts and their integration into social practice is crucial to understanding human activity. The framework has been applied and developed in the study of human computer interaction to offer an alternative view of human activity by redefining the relationship between individuals, communities, technologies and actions (ref. 23).
The unifying theme of the more recent perspectives is the importance of context and need for investigative strategies which accommodate the complexity of real situations and real systems. However, this poses significant challenge for the investigation of human work activity in context while avoiding mechanistic oversimplification.

A comparative case study of a knowledge-based computer system in routine use in obstetrics

We had planned to visit other hospitals where medical knowledge-based computer systems were thought to be in routine clinical use, to see at first hand the influence of work context and the strategies adopted to study this. However, the systems identified from the literature were either limited prototypes or had been decommissioned.

An existing knowledge-based computer system (Expert DataCare) had been developed by our group for the interpretation of umbilical cord blood acid-base at the time of birth. The development of this system was driven by clinicians and a participatory design approach was adopted to ensure the users' needs, capabilities and preferences informed the design process. The system had been in routine use in the development site for three years and is now also installed in 29 Maternity Units in the U.K. It was therefore decided to undertake a preliminary investigation of this system. While this knowledge-based computer system addressed a different and less complicated problem than our prototype system, it was routinely used in the same domain and by the same target user group.

The aims of the study were to:

Clarify the research approach required to investigate errors in context and its potential contribution to the design of our new prototype computer system.

Describe the human errors which occurred.

Place human errors in the wider context of the work activity system.

Background to the system

Methods of assessing the condition of the infant at delivery have been mainly visual and subjective. Objective information can be obtained at birth from blood from the umbilical cord artery and vein. Blood samples analysed with a blood gas analyser provide information that reflects the oxygenation of the baby during labour. It is now recommended that this is undertaken routinely (ref. 24). However, retrospective analysis of data from a clinical trial in Plymouth (ref. 25) in which cord blood analysis was performed highlighted a number of problems. Approximately 25% of samples contained errors which prevented accurate interpretation. A knowledge-based computer system, interfaced to a blood gas analyser, was developed to check for data errors and provide an interpretation of the results. The aims of the system are to improve the quality of data collected, clarify the importance of the procedure, and ease the introduction of umbilical cord blood analysis into clinical practice. Immediate user feedback affords the opportunity to obtain further samples if required. The data is recorded to a database to provide accessible information for audit and risk management.

Study design

We selected two sites for study: the development site, Unit A, with three years experience of using the system and a comparison site, Unit B, with two years experience using the system. A case study approach was adopted to study the complex work processes in context and explicate interconnected levels and loops of causation (ref. 26).

Data Collection

The study used multiple sources of data to afford investigation of a wider range of issues. Triangulation of data and methods reduced potential problems of construct validity and avoided the weakness inherent in any single approach.

Interviews: The study began with interviews of doctors, midwives and care assistants as the initial data collection method. Nineteen participants, who had experience using the computer system, agreed to take part in the study. The aim of the interviews was to elicit information to uncover the informant's views and perceptions surrounding the use of Expert DataCare.

Unstructured interviews allowed for flexibility, but were more time consuming to analyse. Initial interviews were unstructured, but over time became more focused and structured as particular issues emerged through the process of data collection. A list of topics to cover during the interview was developed to act as a prompt for the researcher. Conversations were tape recorded, transcribed and analysed in detail. Each line of transcript was coded to enable identification of the source and sequence of material.

Observation: Informal observation provided useful additional information and the degree of researcher involvement ranged from complete observer in the comparison site to complete participant in the development site. Data were collected about specific events, social interaction, physical activities and the characteristics of the environment.

Analysis of Artefacts: The computer and artefacts associated with the computer were also included in the analysis. This included the blood gas machine, printed outputs, and the computer database. Examination of these artefacts helped to develop a broader perspective of the role of the computer system and provided insight into the technical operations concerning the application. Documents, including protocols and guidelines, were examined and prompted further lines of inquiry. These were helpful in identifying contradictions which would not have been readily apparent otherwise.

Analytical Framework

We adopted human activity theory as a conceptual framework and analytical tool. Activity theory is considered useful for studying human practice, as a development process, with both individual and social links. Activity theory focuses on *the relationship between people and technology* to uncover how they negotiate with the social and technological environment to solve problems and learn. This theory recognises the contingent and changing nature of activity systems and that human activity is *mediated* and *transformed* by the tools they create.

An activity is a form of *doing* directed at an object (figure 1). The transformation of the object into an outcome motivates the existence of an activity. The relationship between subject and object is mediated by tools; the relationship between subject and community is mediated by rules; the relationship between object and community is mediated by division of labour. This activity forms the minimal meaningful context for individual actions. Engestrom (ref. 28) extended activity theory to analyse work development and focused on contradiction and tensions which are present in every activity system. These tensions, incompatibilities and conflicts (collectively termed contradictions) offer an opportunity for change and development. If the community is receptive and the contradictions made visible, then change can be effected. This may lead to the development of a new activity system or to cessation of the activity. We followed a similar approach in our analysis of work activity and focused on conflicts, tensions and inconsistencies. The views that had been constructed about the activity are presented as emerging perspectives (ref. 29).

Findings

In the exploratory interviews it was found that users viewed the system in different ways between sites. Few participants were clear about the role and the capabilities of the computer system.

Emerging perspectives: Different perspectives were found between Units. In Unit A the output from the system was highly visible, used by individuals to evaluate their care and at department meetings. The output was viewed as a tool and as an important commodity. Respondents in Unit B indicated tensions and conflicts. Excerpts from the interviews illustrate the general perspective of Unit B:

"Its a technological thing. Anyone can do it. I feel I need to be with the mother and baby." *Staff Midwife, Unit B*

"Its a task - something else to do. It could easily be done by others - non professionals." *Staff Midwife, Unit B.*

In Unit B the system inserted itself in a negative way, between midwives and patients, and was perceived as conflicting with their other duties. The use of the output from the system was not visible to the midwives, it was viewed as a task and no value was placed on it.

Contradictions: In Unit A some midwives had initial concerns about the involvement of auxiliary nurses.

"I do feel it is a professional job, it should not be done by the auxiliary nurses." *Midwifery Sister, Unit A.*

"When we first started there were a few midwives who thought that auxiliaries shouldn't do it *[cord blood gas analysis]* and that was a bit upsetting to think they didn't have the confidence in us." *Auxiliary Nurse, Unit A.*

Lack of confidence in the ability of the auxiliary nurses was unjustified. An audit of the computer database found they had a lower error rate than midwives. Figure 2 shows the number of valid paired samples (a sample from both the artery and vein with no measurement errors) and the incidence of inadvertently taking two samples from the same vessel.

Comparison of the computer databases found considerable differences between the two Units. The proportion of significant metabolic acidosis (a measure of oxygen lack in the baby during labour) was found to be 0.7% in Unit A and 2.7% in Unit B. Further analysis of the Units' clinical computer databases found Unit A obtained samples in 98% of deliveries nd the overall sampling error rate was 14%. In Unit B less than 50% of deliveries had cord blood analysis and the overall sampling error rate was 28%.

Placing errors in context of the wider activity system: Each Unit constructed their work in different ways. In Unit A, auxiliary nurses were included as part of the team, analysis of cord blood was performed immediately after delivery, training was provided by a computerised teaching package (CAL) and on-site support for maintenance of the blood gas analyser was provided by the medical physics department. In Unit B, only midwives were involved in the procedure, the analysis of cord blood was delayed for up to 1 hour after delivery, training was "on the job" and provided by midwives who had previously undergone training (cascade training). A midwife was responsible for daily maintenance of the blood gas machine with direct support from the manufacturers.

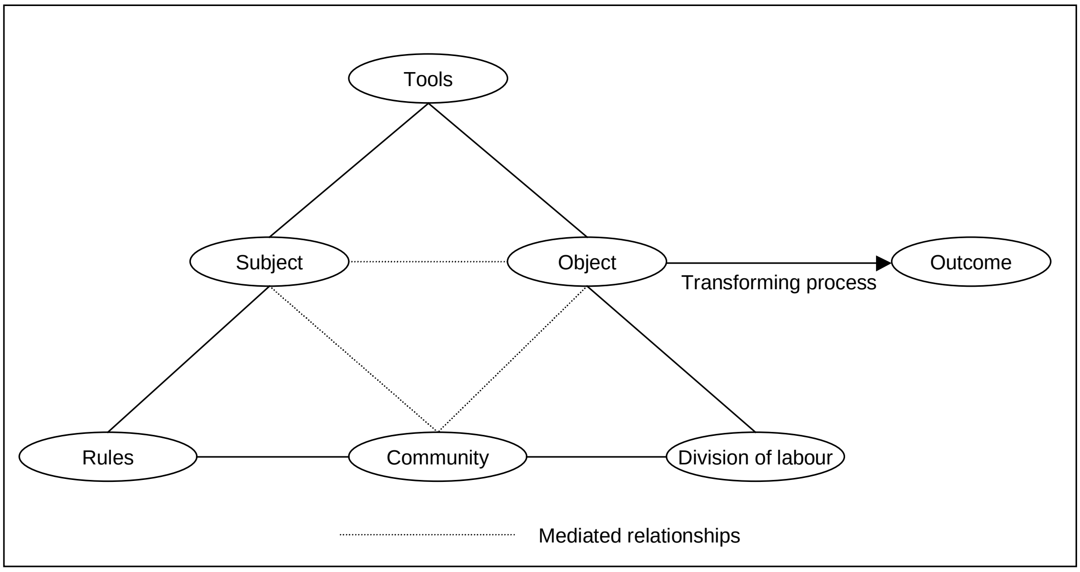


Figure 1 - A Conceptual Diagram of an Activity (ref. 27).

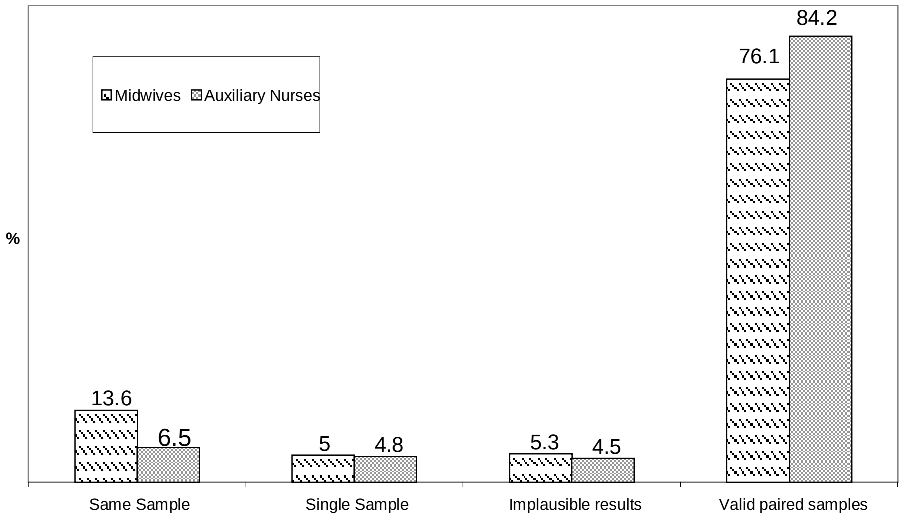


Figure 2 - Comparison of error rates of auxiliary nurses and midwives in Unit A

The implicit rule in Unit B that auxiliary nurses could not undertake cord blood gas analysis resulted in the workload being allocated to midwives. Despite adapting the sampling procedure to allow a delay of up to one hour, the midwives experienced conflict with their other duties. The delay in analysis also weakened the utility of the procedure as it reduced the opportunity to take further blood samples if required. In Unit A, some midwives had initial concerns about the involvement of auxiliary nurses. This implicit rule may be embedded more deeply in hierarchy and tribalism. In Unit B, a clash with the biomedical department's existing rule of purchasing one particular make of blood gas analyser resulted in withdrawal of technical support. A midwife had to take on this role with direct support from the manufacturers. Consequently, this increased training demands and on occasions the system would request a technical operation when no trained staff were available. The main differences between the two Units and the impact of this on the activity are outlined in Figure 3.

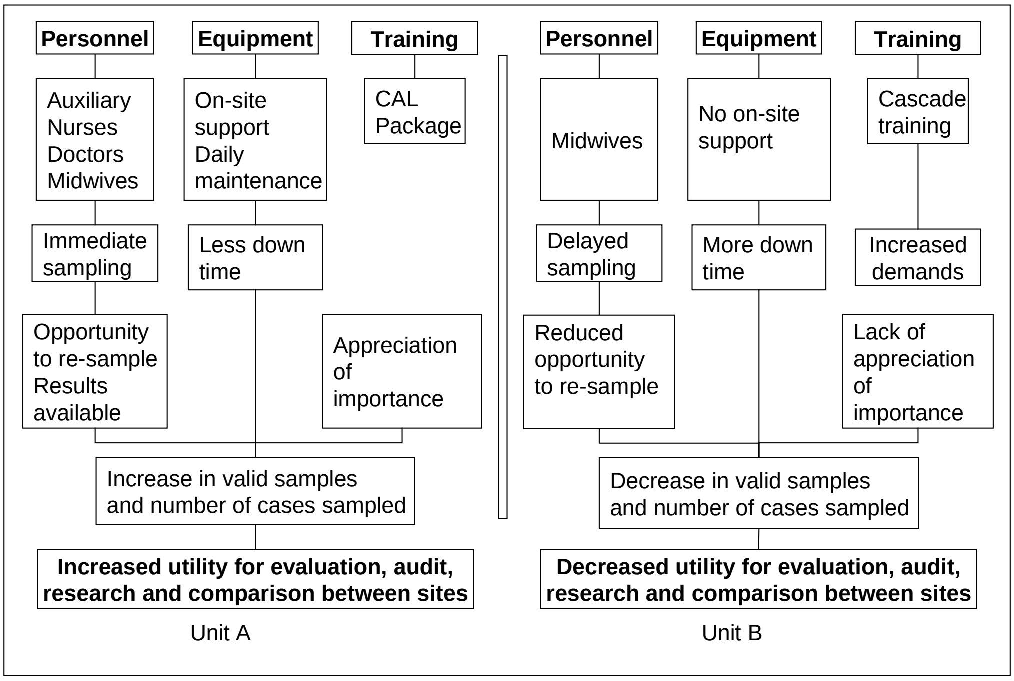


Figure 3 - Differences in work activity system between sites

Discussion

This work combined interviews, observation and analysis of artefacts framed within a case study. Using convergent sources of data strengthened the interpretation of the interview data. Activity theory as an analytical framework offered some insight into the roots of errors, many of which originated from broader activities, attitudes and perspectives surrounding the computer system.

The perspective of Unit A was information orientated and the activity was considered important for research, legal purposes and evaluation of care. The information from the system was highly visible in department audit meetings and at an individual level as a focus for discussion, assessment and improvement of intrapartum care. Unit B's perspective of the work was as "just another task" to be done. There seemed to be a lack of understanding as to how it could help them, individually and professionally, if done properly. The changes to the work system in Unit B did not support attainment of all the objectives of the computer system.

While activity theory suggests that motivated actions within an activity system merge to reach the shared motive, it is not always the case. When this specific type of polymotivation occurs those who benefit are not those who do the work and the technology is likely to fail, or be subverted (ref. 30). The selective sampling in Unit B skewed the data towards a higher incidence of abnormal results. Inappropriate inferences, which could impact on organisational and management decisions could have been made from the data if this discrepancy had not been uncovered.

Ringstaff, Kelly & Dwyer (ref. 31) identified four factors which they considered an important influence on the effectiveness of technology. Psychological factors such as being comfortable with the technology and seeing it as an opportunity rather than a threat; technical support; ease of access; and colleague and institutional support. However, the origins of these factors are embedded in the wider system of activity. People use computers to achieve meaningful goals beyond the actual operation of the system. The physical, social and historical circumstances of work activity give meaning to actions and shape emerging perspectives.

Holland and Reeve (ref. 29) suggest that perspective is a place where contradictions become manifest in consciousness and affect the way work is done. In this study midwives in Unit B did not seem to recognise opportunities for development. The practical difficulties involved in training may have contributed to this lack of awareness. "Cascade" training has a number of limitations. The quality of education becomes diluted over time, particularly in areas with rapid turn over of staff and it is difficult to assess consistency and quality of training. Unit A addressed this practical problem with the use of an interactive computer-based teaching package. The improved morale associated with maintaining a commitment to staff development (ref. 32) may further shape perspectives of an activity and improve performance.

Ease of access and technological support imply institutional support through investment in the technology. However, in Unit B the Biomedical Department's adherence to rigid purchasing policies contributed to errors of omission and was a barrier to effective use of the system. Lack of on-site support impeded operational effectiveness and increased training demands. This type of problem has been highlighted by the Audit Commission (ref. 33) who noted that whether or not a system meets real needs has ceased to be an issue as long as the procurement procedure is followed. In Unit B, the role of the computer system was undermined as emergent, local adaptations contributed to errors. Conflicts and barriers to effective use of computer systems originated from the wider activity system including the rules, tools and division of labour. These effects went unnoticed at an organisational level.

Context has become a touchstone for a better understanding of human-machine errors. However, the study of complex phenomena is a significant challenge for study designs. In this case study we found multiple sources and methods of data collection allowed us to explicate a complex process and uncover important features in the work system. The great strength of the case study is its flexibility. Relevant issues uncovered during the progress of a case study can be explored and guide further data collection to develop a wider perspective. When a study begins with a broad question and where context is considered to influence outcome there is a real danger of attempting to collect data about everything. The case study strategy reduced this risk as it constrained data collection to relevant sources which converged on the phenomena of interest. This approach helped us to place the computer system in context and to understand why errors had occurred. The study also highlighted the need to develop a better understanding of the process of change and its interaction with and effect on context.

While the case study makes no commitment to a particular theoretical stance, it is not atheoretical. The interpretative theory which underpinned this work was considered to be the most salient as it:

Provided an analytical tool which guided the analysis and organisation of data to place it in context of work activity. This framework helped us to make sense of the data and reduced the danger of becoming overwhelmed with large amounts of data.

Provided a conceptual framework which facilitated communication and shared understanding across the various disciplines involved in our research programme.

Conclusions

The roots of human errors are often embedded in underlying organisational issues and the contributing factors are often multiple. The computer technology in this case study was robust. No errors were attributed to failures in the computer system or in the direct interaction between the operator and the computer system (i.e. what is normally called the "interface"). The wider interaction of the human activity system, rather than the human-computer interface, contributed to errors. An unfortunate synergy of individually small factors in the work system can lead to the collapse in the effectiveness of a computer system. This small exploratory study highlighted that:

Formal analytical models are of limited value in the design of error tolerant system.

Computer systems change organisations and the way people work.

The changes that occur are often emergent rather than planned and the process of change may contribute to human error.

The design of technology such as knowledge-based computer systems needs to comprehend these wider issues and undertake multilevel analysis (individuals, task and organisation) to inform both the design process and redesign of the work activity system to reduce the risk of errors.

Analysis of work activity system to inform error tolerant design needs a conceptual framework for understanding context and the process of change to the work system.

This study and the use of activity theory changed our thinking about the design of medical knowledge-based computer systems. Computers are tools which mediate and transform work activity systems. The introduction of new technology changes the work system and changes to the work system change the technology. The design of error tolerant systems cannot make the mistake of separating the technical from the social or the individual from the collective.

The lessons learned in this study are being applied in the design of our prototype computer system to assist clinicians in the management of women in labour. A longitudinal case study of the work activity system in delivery suites is being used to model the work activity system. This will form the basis of models of interaction to inform the design of the computer system and suggest changes to the work activity system to reduce the risk of errors.

A MRC multicentre trial of our knowledge-based computer system, to support clinicians in patient monitoring during childbirth, is planned for the summer of 2000. This will provide an opportunity to undertake longitudinal studies of work activity in 6 UK hospitals.

Acknowledgements

We would like to acknowledge the support of the Medical Research Council, the South & West Regional Health Authority Research and Development Directorate and the Directorate of Obstetrics & Gynaecology, Derriford Hospital, Plymouth.

References

Keith R. D. F., Beckley S., Garibaldi J. M., Westgate J. A., Ifeachor E. C. & Greene K. R. (1995) A multicentre comparative study of 17 experts and an intelligent computer system for managing labour using the cardiotocogram. *British Journal of Obstetrics & Gynaecology* 102: 688 - 700.

Skinner J, Harris M, Greene K.R. (1998) Computerised decision support for managing labour using the cardiotocogram: 500 cases with the range of abnormality. *British Journal of Obstetrics & Gynaecology.* 105 (S 17): 27.

Greene K.R (1996) Intelligent fetal heart rate computer systems in intrapartum surveillance. *Current Opinion in Obstetrics & Gynecology* 8: 123-127.

Helmreich R.L. & Shaefer H.G. (1994) Team performance in the operating room. In: Bogner M. S. (Ed) *Human Errors in Medicine*. New Jersey, Hilsdale.

McCarthy J.C, Healey P.G.T, Wright P.G. & Harrison M.D. (1997) Accountability of work activity in high-consequence work systems: human error in context. *International Journal of Human-Computer Studies* 47: 735-766.

Reason J. (1995) Understanding adverse events: human factors In: Vincent C. (Ed.), *Clinical Risk Management.* London, BMJ Publishing Group.

Carroll J.M. (1997) Human-computer interaction: Psychology as a science of design. *Annual Review Psychology.* 48: 61-83.

Monk A.F. & Wright P.C. (1991) Observations and inventions: new approaches to the study of human-computer interaction. *Interacting with Computers.* 3 (2): 204-216.

Card S.K, Moran T.P, & Newell A. (1983) *The Psychology of Human-Computer Interaction.* Hillsdale, New Jersey, Erlbaum.

Karat J. (1989) The relation of psychological theory to human-computer interaction standards. In: Salvendy G, & Smith J. (Eds.), *Designing and Using Human-Computer Interfaces and Knowledge-Based Systems.* pp.582-588. Amsterdam, Elsevier Science Publications.

Hutchins E. (1990) The technology of team navigation. In: Galagher J. (Ed.), *Intellectual Teamwork.* Hillsdale, New Jersey, Erlbaum.

Rogers Y, & Ellis J. (1994) Distributed cognition: an alternative framework for analysing and explaining collaborative working. *Journal of Information Technology.* 9: 119-128.

Suchman L. (1987) *Plans and Situated Actions: The Problem of Human-Machine Communication.* New York, Cambridge University Press.

Cannon-Bowers J.A, Salas E, & Coverse S. (1993) Shared mental models in expert team decision making. In: Castellan J. (Ed.), *Individual and Group Decision Making. Current Issues.* Hillsdale, New Jersey, Erlbaum

Harvey J.B. (1975) *The Abilene Paradox.* Lexington, Massachusetts, Lexington Books.

Moray N. (1994) Error reduction as a systems problem. In: Bogner M.S. (Ed.), *Human Error in Medicine.* Hillsdale, New Jersey, Erlbaum.

Pinsonneault A. & Kramer K.L (1990) The effects of electronic meetings on group process and outcomes: An assessment of the empirical research. *European Journal of Operations Research.* 46: 143-161.

Clegg C. (1994) Psychology and information technology: the study of cognition in organisations. *British Journal of Psychology.* 85: 449-447.

Huber G.P, Valacich J.S,& Jessup L.M. (1993) A theory of the effects of group support systems on an organization's nature and decisions. In: Jessup L. & Valacich J. (Eds.), *Group Support Systems: New Perspectives.* New York, Macmillan.

Hughes J.A, Sommerville I, Bentley R, & Randall D. (1993) Designing with ethnography: making work visible. *Interacting with Computers.* 5 (2): 239-253.

Mays N. & Pope C. (1995) Rigour and qualitative research. *British Medical Journal.* 311: (6997) 109-112.

Williams J.A. (1988) A data-based method for assessing and reducing human error to improve operational performance. In: Hagen W. (Ed.), *1988 IEEE Fourth Conference on Human Factors and Power Plants.* New York. Institute of Electrical and electronic Engineers. 200-231.

Nardi B. (1996) *Context and Consciousness. Activity Theory & Human Computer Interaction.* London, England, MIT Press.

Royal College of Obstetricians & Gynaecologists (1993) Recommendations arising from the 26th RCOG Study Group: Intrapartum fetal surveillance. In: Spencer J. A. D, & Ward R. H. T. (Eds.), *Intrapartum Fetal Surveillance.* London, RCOG Press.

Westgate J, Garibaldi J.M, & Greene K.R. (1994) Umbilical cord blood gas analysis at delivery: a time for quality data. *British Journal of Obstetrics & Gynaecology*. 101: 1054-1063.

Pettigrew A. (1990) Longitudinal field research on change theory and practice. *Organisational Science.* 1 (3): 267-292.

Engestrom Y. (1993) Developmental work research: reconstructing expertise through expansive learning. In: Nurminen M. & Weir G. (Eds.), *Human Jobs and Computer Interfaces.* Amsterdam, Elsevier.

Engestrom Y. (1995) Objects, contradictions and collaboration in medical cognition: an activity-theoretical perspective. *Artificial Interligence in Medicine*. 7: 395-412.

Holland D. & Reeve J.R. (1996) Activity theory and the view from somewhere: team perspectives on the intellectual work of programming. In: Nardi B. (Ed), *Context and Consciousness. Activity Theory & Human Computer Interaction.* London, England, MIT Press

Norman D. (1993) *Things That Make Us Smart: Defending Human Attributes in the Age of the Machine.* Reading Massachusetts, Addison-Wesley.

Ringstaff C, Kelley L, & Dwyer D. (1993) Breaking the mould of instruction with technology: Formative case studies of the unit of study process. *ACOT Report* (August).

Hannah K.J, & Osis M. (1988) Computers and Staff Development. In: *Nursing Informatics*. Ball M.J, Hannah K.J, Jelger H.G, & Peterson H. (Eds.), New York, Springer-Verlag. 321-328.

Audit Commission (1995). *For your information: A study of information management and systems in the acute hospital.* HMSO. London.

Biography

Maureen Harris, Perinatal Research Group, Department of Obstetrics & Gynaecology, Level 4, Derriford Hospital, Plymouth. PL7 8DH Devon. Telephone 01752 763631. Email:mojo@pnrg.co.uk.

Maureen Harris, a midwife, has worked with the Perinatal Research Group since 1990. In 1995 she was awarded a South & West Regional Research Training Fellowship to develop expertise in health related research. Her current Ph.D work is on integrating work systems and computer systems in association with an MRC funded project to develop medical knowledge-based computers for patient monitoring during labour.

Works in theory but not in practice? Some notes on the precautionary principle.

Kenneth Calman[1] and Denis Smith[2]

1. University of Durham.
2. Centre for Risk and Crisis Management, University of Sheffield.

Introduction

Notions of risk and probability have become dominant constructs within many of our discussions about modern living (See Beck, 1990; Erikson, 1995: Giddens, 1988) and yet they still prove elusive but emotive issues for policy makers to deal with (Smith and Toft, 1998). We should not be surprised therefore at the high level of concern that people have over the range and nature of the hazards that we face. In some respects, different generations have expressed similar concerns about hazard, albeit in different contexts and set against different standards of acceptability. For example, current debates about the risks associated with BSE and genetically modified foods can be seen to provide an echo of earlier debates on both pesticide use (Carsen, 1962) and cancer risks (Steingraber, 1998). The literature on risk provides us with constant warnings of our plight from such hazards as global warming, toxic waste, nuclear radiation, genetically modified organisms and pesticides. What prompts such pessimism within the human condition, on the one hand, and yet allows such dire warnings to go unheeded on the other is a central component of risk debates. The origins of this paradox lie in a number of factors including: the voluntary – involuntary nature of the risk, the sense of helplessness felt by potential victims of the hazards, delayed versus immediate effects of any exposure and the manner in which the hazard (and uncertainty surrounding it) is communicated to those affected. At the heart of risk debates often lies the central role of expertise and the prior knowledge that organisations and expert bodies had of the hazard.

It is clear from the range and ferocity of hazard debates that the various protagonists often take fundamentally different stances regarding the nature of the problems and their potential solutions. Indeed, some have argued that it is difficult to find a mutually accepted solution to a problem when it is clear that the various actors in the debates define the problem in quite different ways (see, Weick, 1988; 1993; 1995; Weick and Roberts, 1993). What invariably frustrates policy makers is the manner in which groups can express concern over one set of hazards whilst, at the same time, exposing themselves to potentially greater hazard through other activities.

The history of policy initiatives for risk management is littered with attempts to “educate” the public or, more recently, to “communicate” the risks more clearly to them in the hope that they will modify their behaviour accordingly. However, this approach might be seen to create a setting in which the views of experts are given greater weight than those of the groups who are exposed to the hazard. More recent research, however, sees the determination of hazard (and, by implication its acceptability) as a negotiated process between protagonists in the debates. This is particularly the case in those situations for which there is little or no a priori evidence for both the probability and consequence associated with the hazard. Inevitably, conditions will arise when it is not possible to find a compromise position between the protagonists, and the state (operating within its regulatory function) will need to intervene in order to prevent an activity from taking place. This use of a “precautionary principle” raises a number of key managerial and research issues. For example, to what extent should the state curtail the use of certain products or prevent certain activities when the hazards themselves are unclear and causal relationships are blurred? Put another way, do or should organisations act with sufficient caution in decisions which have a high implicit hazard potential but where the probabilities of occurrence are low or unknown?

Against a background of high profile public health scares, it is not surprising that the concept of risk has attracted increasing interest over the last few years. As part of this discussion, the term "precautionary principle" has been widely used to justify regulatory actions taken to protect the health of the public. The term, however, may be used in different contexts and with different implications. Despite its current popularity amongst policy makers, the concept has also attracted severe criticism. For example, Furedi (1997) has claimed that the principle "suggests that we are not merely concerned about risks, but are also suspicious of finding solutions to our predicament" (p. 9). Herein lies an apparent paradox. How do we ensure that our 'solutions' to problems do not in themselves create even greater hazards? This has created a sense of insecurity, which Furedi (1997) argues is typified by the number of health scares witnessed in recent years. If we are overly cautions in our approach to risk, then it is conceivable that our inaction may generate more significant hazards than the activity that we are trying to manage. Ultimately, this becomes a debate about the acceptability of a portfolio of risk, which is by definition, more of a socio-political problem rather than a technical-scientific matter. The issue here centers around full disclosure and informed consent and requires that an active attempt is made by the risk generators to communicate in an open and effective manner.

The purpose of this paper is to explore the central dynamic of these relationships by reference to the notion of the precautionary principle and its role within public health policy-making. In particular, it seeks to address the following issues. Firstly, it seeks to clarify the concept of the precautionary principle within health and healthcare policy making. Secondly, it attempts to examine the central role of expertise within risk debates Thirdly, it attempts to establish a framework within which policy making could reflect more the core elements of the precautionary principle. Finally, the paper aims to encourage further debate on the framework proposed, specifically as it applies to health and healthcare.

The Nature of Risk

There is little doubt that "risk" is a much misused and misunderstood construct. At its simplest, we can see it in terms of the probability that a given hazardous event will occur and that this event will have consequences which are deemed to be negative by some, or all, of those who are exposed to it. In such a context, it might be prudent to constrain the activity that gives rise to the hazard, especially if there is a lack of clarity over causality. Where the issue becomes blurred is in terms of those events in which both active intervention and in-action can give rise to a hazard. At what point here do regulators deem an intervention to be justified and what form should this intervention take? At it's crudest, this becomes a 'burden of proof' debate, with those who can muster the greatest evidence likely to win the argument. At its most complex, it becomes a socio-political conflict over the acceptability of risk, even under those conditions where the nature of the risk is not proven. This raises and important question namely, to whom and over what time period is this risk (i.e. probability x consequence) acceptable? The acceptability of any risk will vary according to the derived benefit obtained by those who assume the risk, the range of alternatives available to them, the immediacy of the harm/benefit associated with the risk and the possibility of further corrective action to mitigate the consequences of the action. If we add to this, the mediating role of expertise (in providing an evidential basis for the decision) and the inherent uncertainty in those cases where a precautionary principle is usually applied, then it is clear that there is considerable scope for ambiguity and conflict concerning the development of any precautionary policy.

Part of the problem here centres on the "trans-scientific" nature of certain problems (Weinberg, 1972) in which it is held that the issues go beyond the abilities of science to prove. It is the inability to predict the 'emergent properties' associated with processes, products and systems that creates the problems involved in managing the risks associated with such emergence (see, Fortune and Peters, 1995). Within any scientific analysis, the notion of proof is a central underpinning concept. Without proof, decision-makers invariably factor other issues and considerations into the process. A central element in this process is, therefore, the notion of expert opinion. Further problems are caused by the weak evidential basis for determining emergence, causal relationships and their associated consequences, especially in those cases where there is little a priori data. For delayed-in-effect consequences (associated with the exposure to a substance where the consequences are only negative) one might want to ignore expert <u>opinion</u> and apply the precautionary principle as a matter of course. Clearly in these cases, some form of "harm" is the only certain outcome and policy makers might therefore ignore any derived benefits. However, some would argue that the power of interested parties can be an important dynamic in shaping the manner in which the precautionary principle is applied. This claim is given weight, for example, by the ability of the powerful tobacco lobby to resist stricter controls during the period in which the risks from smoking have been made known. This clearly raises questions of an ethical nature which can be contextualised within the broader literature on medical and business ethics. Put simply, the precautionary principle is akin to Sethi's (1975; 1983) notion of Corporate Social Responsibility in which organisations seek to anticipate the likely social demands placed upon them or, in areas of uncertainty, the possible impact of their actions on current and subsequent generations. To be effective therefore, the precautionary principle requires either an effective anticipatory regime of enforced regulations or fundamental changes made to the core beliefs, values and assumptions of organisations. Both strategies bring with them problems of efficiency and effectiveness. Regulation can only be effective if a strong programme of enforcement and appropriate penalties for violations accompanies it. Changes in the belief and value systems of an organisation are, without doubt, an effective means of ensuring that the precautionary principle is used as a matter of course. However, this is invariably a process that takes time to develop and which may fail if there are changes in the senior staff of the organisation. These are important problems of policy implementation, which demand attention from those who seek to introduce the precautionary principle.

The Precautionary Principle

By invoking the precautionary principle (as it is currently conceptualised) we generally envisage a move towards a situation where legislation is enacted to protect sections of society. This legislation may generate a situation which is:

> restrictive in that choice and freedom of action are limited not only for the individual, but for the good of society as a whole.
>
> punitive, in that those who do not conform for the good of society are subject to some form of punishment such as a fine, or prison sentence.

However, one might also question the long-term effectiveness of such legislation as a means of controlling risk. Invariably, legislation often lags behind both social norms and scientific knowledge. For example, smoking tobacco (as a known hazard) could have been banned within any public spaces many years ago on the basis of our understanding of the risks of secondary smoking. A second failing of legislation is that it invariably exists as a process of negotiation. All too often, the potential victims of the hazard have little, if any, input into that negotiation. One only needs to examine the development of environmental regulation to witness the impact that large corporate concerns can have on the regulatory process, both in terms of delaying legislation or reducing its severity (see, for example, Smith, 1990). Without doubt, the most effective way of dealing with issues of hazard is to ensure that the generators of that hazard behave in an ethical (or socially responsible) manner. Herein lies the paradox. Invariably, it was a failure of organisations to curtail their hazardous activities that resulted in the legislation in the first place. Ensuring that the necessary behavioural shifts take place within organisations and that these are sustained is, perhaps, one of the most difficult and challenging of managerial issues surrounding the use of the precautionary principle.

Given these issues, what then is the role and significance of the precautionary principle? How does it operate in practice? Why in some instances do we legislate and in others only give advice, hoping that organisations and individuals will change their behaviour? If society's core purpose is to protect the health and well-being of the population as a whole from some risk or threat to health, then society also needs to decide on the nature of the precautionary principle framework within which it is to operate. Policy makers need to identify which risks should come under the purview of the principle and which should not. Of particular importance here is the clear articulation of the reasons why certain forms of hazard would be exempt from the principle. The manner in which we communicate concerns about hazard is important (see, Committee on Risk Perception and Communication, 1989) as is the ethical stance that we take with regard to issues of full disclosure and informed consent. It is this aspect of the problem which lies at the heart of the decision making process and which raises the key questions within risk debates. Who should decide? On what evidence? How can potential victims object to the hazard and what happens if they disagree with the mediated response? The following sections try to clarify some of the issues involved and then suggests a way forward.

Making sense of hazard: the role of expertise

"….it is hard to make common sense when each person sees something different or nothing at all" (Weick, 1993, p.636).

It could be argued that one of the primary functions of a manager is to cope with uncertainty. If there were no uncertainty in life, there would be no need to make decisions and therefore, little need for managers. When dealing with issues of hazard, emergence and complexity, this process of managing uncertain outcomes assumes a position of critical importance. Given the role of risk within the management process, one might expect that it would occupy a position of great importance at the strategic levels of the organisation. In reality, however, risk is often relegated in importance in organisations, with managers showing greater interest in short term benefits rather than long term harm. The principal issue invariably becomes one of the centrality of technical expertise, especially when we have little a priori data upon which managers can base their decision-making. For delayed-in-effect hazards the problem becomes further blurred as cause and effect relationships are not always clearly defined or, in some cases, understood. It is here that the precautionary principle should assume a critical role and yet it is often neglected in favour of expert opinion.

There has been considerable discussion relating to the role of experts within policy-making (see, for example, Collingridge and Reeve, 1986; Blowers, 1984) and, in particular, for issues of hazard (Draper, 1991: Irwin and Wynne, 1996; Irwin 1995; Smith, 1990). Collingridge and Reeve (1986) have argued that the use of institutionalised science within scientific conflicts may not lead to a resolution of that conflict. Indeed, they have gone so far as to say that in highly emotive debates the use of science by all parties to the conflict will merely heighten the intensity of that debate. Ultimately, those who can bring the greatest power (and influence) to bear on the debate are likely also to be able to exert control over the scientific arguments as well. This is especially the case where the trans-scientific nature of the problems ensures that there is no clear burden of proof, almost irrespective of the "quality" of the scientific expertise used. In examining debates involving major hazard sites at Canvey Island, Smith (1990) suggested that the success of scientific interventions may well be determined by the economic power exercised by protagonists in the conflict. What emerges from this body of literature is the need to ensure that the regulatory agencies are not unduly influenced by "the national interest" and that they are sufficiently precautionary in the approach to trans-scientific problems of hazard.

Risk, Precaution and the Management of Hazard

A central facet of risk debates is the notion of causality – more specifically, who can we blame for the realisation of the hazard? Once again, the attribution of causality sees the expert occupy a central role. Invariably, the concerns of the public centre around the consequences associated with a hazard whereas the "expert" community may well focus their attention on the probabilistic component of risk. In both cases, attention is focused on causal relationships, although the policy outcomes associated with each perspective may be different. However, clear causal relationships can break down when we move our research out of the laboratory into the "real world" (see Neisser, 1980). Debates concerning risk do not take place in the objectively driven laboratory, but in the value-laden ambiguity of a societal context where notions of categorical proof are elusive (Weinberg, 1972). Consequently, we might not expect science to provide a clear solution to these problems. This, in turn, has an impact upon policy making and requires that greater caution be employed in those circumstances where there is a high level of hazard and uncertainty. The purpose of invoking such a precautionary approach is to prevent a negative event from occurring which might be harmful to health. This intervention may be achieved either by restricting behaviour or by changing patterns of behaviour so as to minimise exposure to the hazard. It is here that the application of the precautionary principle should be straightforward and yet it is also here where it seems to cause the greatest difficulties. In particular, the use of the principle raises a fundamental question. At what point does an organisation or the State assume a moral right to impose such changes on a wider population, when the evidential basis for the decision is weak? Health care provides a number of examples for the discussion of the principle as it is used in practice.

Most surgical operations are associated with risk, yet, except in a few instances (cardiac and liver transplantation) there are few major restrictions upon the actions of those carrying out the procedure. Indeed, the whole practice becomes contextualised within the notion of professional judgement and ethical behaviour. In the absence of strict protocols, clinicians work within the boundaries of what they deem to be their professional judgement and this has raised quite important questions about risk within clinical governance and evidence-based health care (see, Smith, 1998; Treasure, 1998; Walshe and Sheldon, 1998). Pharmaceuticals, on the other hand, are different. Here, the Regulatory Authority has the power to license, or de-license a drug. The latter may be done on a "precautionary basis" because of possible side effects or adverse consequences. The question here is whether the drug should be licensed or removed from the list of prescribable drugs. Thus the use of the drug may be banned completely, or it may remain prescribable. In both cases, it is done in the knowledge that it does have adverse consequences, and therefore when it is used it must be with the patient's informed consent. In those cases where the drug is used, the benefit is seen to outweigh the disadvantages. Much of the debate in such cases is on the nature and strength of the evidence available and around the nature of informed consent. At one level, this distinction may seem bizarre. However, it is clearly a reflection on the nature of an acceptable level of risk. Within non-elective surgical operations there may be a high probability of an adverse consequence, but this has to be set against the likely consequences of no intervention (which may result in a more chronic form of hazard) and the extent of the impact on the wider population (usually limited to a small number of individuals at any one point in time). In contrast the use of pharmaceuticals may adversely affect a considerable percentage of the population and its effects may be delayed in effect, thus adding a spatial and temporal dynamic to the problem. Under these circumstances, one might argue that the use of the precautionary principle is more straightforward and yet this issue is also clouded by problems of informed consent and evidence-based policy making, making it a burden of proof debate.

A brief consideration of some examples shows that the current operation of the principle is complex. For example, invoking the principle does not seem to be related to the level of risk. If this were the case then cigarettes would be banned. Nor does it seem to be related to the numbers of people involved. If this were so then the very small number of people with peanut allergy would not receive such "disproportionate" attention. Nor does it seem to be related to the need of the majority of people. If this were so then the 25% of cigarette smokers in this country would not be allowed, in many instances, to pollute the environment and cause illness in the majority. The freedom of choice of one person inevitably seeks to override the freedom of choice of another. In order to frame the boundaries of this debate it is possible to identify three major issues, which seem to be evident when thinking about the precautionary principle. These can best be expressed in the form of questions.

1. *What is the level of certainty of the risk, and how strong is the evidence?*

This is a key issue within all risk debates. Should action be taken on the basis of an "hypothetical" risk, for which sound evidence is not available, even though the risk is plausible and has been identified by someone? What if two or more reputable groups or individuals dispute the evidence? What if a single-issue group takes up a particular cause and cannot be convinced by the evidence and expert opinion of the level of risk? This is a time when careful and dispassionate consideration of the evidence is required and such consideration should include non-experts in the discussion (see Sheldon and Smith, 1992).

There is a further and crucial aspect of this discussion, and that relates to the level of uncertainty of hazard. This is probably the most difficult situation of all. In this case there seems to be a real hazard, but it is unclear what the probabilities of its occurrence or the consequences associated with it might be. There is no way of knowing, other than by carrying out more research or waiting a period of time. Indeed, many hazards are not quantifiable and the human effects of exposure are just not known. In such an instance, invoking the precautionary principle is a matter of very difficult judgement and may have profound consequences for little or no benefit.

The role of expert scientific committees might be raised here. Their function is to critically examine the evidence available and to determine the nature of any possible risk. Government appointed expert committees are composed of independent scientists and non-scientists and are an important part of the decision-making process. They may wish to suggest action or make recommendations. The Government may accept the advice or reject it. However, it should be emphasised again that is only one part of the decision making process, other implications will also be considered by ministers. The process is ultimately a political one and the problem occurs when it is portrayed only as a scientific (and therefore rational) process.

2. *Does the individual have any choice in whether he or she is subjected to the risk?*

Once again this is a crucial question within a democratic society. It might be argued that if one knows the nature and probability of the hazard, and can make an informed choice about it, then there would be no need to invoke the precautionary principle. In general this is the case. For example, the evidence that exercise improves health is clear but there is no case for legislation to make keep fit classes compulsory. In a similar way, if a pharmaceutical product is associated with adverse effects it might be sufficient to ensure that the professionals concerned have sufficient knowledge to use the drug appropriately, with the patient's informed consent, without regulation being required. But can this generally be said to be the case? For example, in some instances additives are put in foods to preserve them without effectively communicating that fact to consumers, and some foods are withdrawn from all customers because of a known risk, thereby removing choice for the majority. Whilst in most instances, the ability to choose makes the use of the precautionary principle unlikely, this is not universally the case. In particular, it may still be invoked when there is considerable uncertainty about the level of risk. It should be noted that choices made by patients or the public about risks to health cannot always be readily predicted. Much depends on the perception of risk and the manner in which it is communicated. This has a major impact on the nature of informed consent and its effectiveness (see, Treasure, 1998).

3. What is the magnitude and acceptability of the risk being considered?

While this is related to the first question, it has a separate locus. Assessing the magnitude of a hazard always involves an element of judgement and this has three components: the seriousness of the hazard, the immediacy of harm, and the number of people affected. These three factors operate independently but can be synthesised to give an estimate of the magnitude of the consequences. For example, suppose that we know that a chemical does have a known level of risk, let's say one person in one thousand is involved. The effect is serious as it causes a disabling condition, but the number of people exposed is small, only a few hundred. What would the significance of the risk be in this circumstance? This is then complicated by the fact that choice of known pathways of exposure to the chemical may, or may not, exist. However, if those exposed to the hazard are not made fully aware of its dangers then the issue becomes further complicated. Ultimately, a decision has to be made about the acceptability of the risk. The question is who should make that decision?

Making the Decision

Having briefly considered these three questions, the final step is to make a choice between alternative options. Has the evidence, the question of choice and the magnitude of the risk persuaded policy makers that there is sufficient concern to take action? Perhaps it isn't worth making a fuss about something that may have little impact on society for an activity that has enormous benefits? But who is to judge? It is this question, more than any other, which proves to be contentious. What right does an organisation, or the state itself, have to knowingly expose individuals to risk without their fully informed consent? In essence, the critical phrase is "fully informed consent". How can we have informed consent when we do not fully understand the nature of the hazard and its probabilities?

Ultimately, those who legislate have to come to a decision about a particular problem based on all of the information available to them, which in most instances should be more than the scientific evidence. As has been said, this will invariably be a matter of judgement and perhaps what matters most is the transparency of both the approach to the decision and the evidential base upon which the judgement is made. This must include access both to the scientific evidence (for and against) and the nature of any other views expressed. A failure to provide such open access will make the decision process appear one sided.

Some Conclusions

Answering the three questions proposed earlier in this paper should go some way to reaching a more logical and reasoned decision-making process over issues of hazard. However there remain three problems which can arise and which need further discussion and research.

1. *Continuing discontent.* It not infrequently happens that an issue is raised, is fully debated, all the evidence assessed and a decision made which is still not to the liking of those who originally raised concerns about the matter. For example, a chemical is considered to be dangerous, and harmful to human health. It is fully assessed and found not to have the level of harm suggested and is given the "all clear". Yet the criticisms persist. Are they right? Should the debate be opened again? Is a conventional approach to the use of expertise missing something? Is the uncertainty in the judgement too great? There is no simple answer to these questions, and they are likely to be very topic specific, but it is suggested that research needs to be carried out to determine a way of taking this forward.

2. *The timing of the decision*. For a new procedure or chemical, there should normally be time to make decisions about the action to be taken, though this may not always be the case with an infection or acute toxic problem. However, decision taking is even more difficult if the source of the hazard is already in use. In some instances the evidence is clear and an immediate ban is obvious and welcomed. But suppose the evidence is uncertain or even disputed, what should happen then? Do we wait until more evidence is available? In circumstances of choice can we rely on the information being readily available and acted on? How long should we wait before making the decision or taking steps to generate more research? Once again there is a need to consider the three issues concerning the nature of evidence, choice and probability/magnitude of hazard before proceeding to take action to protect the health of the people in a compulsory way.

3. *The role of the scientific expert or committee in policy making*. In this country we are very fortunate in having available scientific expertise of the highest calibre. However, as has been discussed here and elsewhere (see, Collingridge and Reeve, 1986; Smith 1990), decisions on risk are often made on the basis of factors other than scientific ones. In addition scientists may feel uncomfortable in taking the uncertainty presented by complex and difficult science and converting it into policy with confidence and assurance. Yet the public may need a clear statement of the position if the precautionary principle is to be invoked (or not) and legislation enacted to back up the decision.

The role of the scientific committee is thus to set out the evidence, including the uncertainty, and for policy makers to make the final decision. Senior advisers and Government Officers, such as chief scientists and chief medical officers provide the bridge between the two, interpreting and evaluating the science and the policy.

But what happens if there is a disagreement between scientists and policy makers? In many instances this will not matter, after all these are issues of judgement and there is room for differences of view. In some instances however, the differences will be so great (it is considered that public health will be compromised) that it is untenable for the scientist or adviser to remain in post and resignation is the only answer. This is a rare occurrence, but is a necessary part of the procedure.

Finally to return to the question of who should make these decisions. If it is accepted that invoking the precautionary principle leads to legislation, then this cannot be left to scientists or special interest groups. It must be seen to be open and fair and part of the democratic process. It is the responsibility of the legislature, the policy makers and politicians. Those who advise must present the arguments impartially and clearly, mustering the evidence available and indicating where it is weak or uncertain. It should be remembered that the purpose of this whole process is to protect the health of the public.

References

Beck, U. (1992) Risk Society: Towards a new modernity. London: SAGE.

Blowers, A. (1984) Something in the Air. London: Paul Chapman Publishing.

Calman, K.C. (1996) Cancer: Science and Society and the Communication of Risk; British Medical Journal 1996; 313 :pp. 799-802

Carson, R. (1962) Silent Spring.

Collingridge, D. and Reeve, C. (1986) Science Speaks to Power. London : Francis Pinter

Committee on Risk Perception and Communication (1989) Improving Risk Communication. Washington DC: National Academy Press.

Draper, E. (1991) Risky Business. Genetic testing and exclusionary practices in the hazardous workplace. Cambridge: Cambridge University Press.

Erikson, K. (1994) A New Species of Trouble: Explorations in Disaster, Trauma and Community. New York: Norton.

Fortune, J. and Peters, G. (1995) Learning from Failure: The Systems Approach. Chichester: Wiley.

Giddens, A. (1990) The Consequences of Modernity. Cambridge: Polity Press.

Irwin, A. (1995) Citizen Science. London: Routledge.

Irwin, A. and Wynne, B. (Eds.) (1996) Misunderstanding Science? The public reconstruction of science and technology. Cambridge: Cambridge University Press.

Neisser, U. (1980) "On 'social knowing'". Personality and Social Psychology Bulletin, 6, pp. 601-605.

Royal Society (1992) Risk : Analysis, Perception, Management; London : Royal Society

Sethi, S.P. (1975) 'Dimensions of corporate social performance: An analytical framework', California Management Review 17(3) pp. 58-64.

Sethi, S.P. (1983) 'A strategic framework for dealing with schism between business and academe', Public Affairs Review, 1983, pp. 44-59.

Sheldon, T.A. and Smith, D. (1992) Assessing the Health Effects of Waste Disposal Sites: Issues in Risk Analysis and some Bayesian Conclusions in, Clark, M., Smith, D. and Blowers, A. (Eds) (1992) Waste Location: spatial Aspects of Waste Management, Hazards and Disposal. London : Routledge. pp 158-186

Smith, D. (1990) "Corporate Power and the Politics of Uncertainty: Risk Management at the Canvey Island Complex". Industrial Crisis Quarterly, 4 (1) pp.1-26

Smith, D. and McCloskey, J. (1998) "Risk Communication and the Social Amplification of Public Sector Risk". Public Money and Management, 18 (4) pp. 41-50

Smith, D. and Toft, B. (1998) "Issues in Public Sector Risk Management". Public Money and Management, 18 (4) pp. 7-10

Smith, R. (1998) "Regulation of doctors and the Bristol inquiry. Both need to be credible to both the public and doctors", British Medical Journal, 317, pp. 1539-1540.

Steingraber, S. (1998) Living Downstream: an Ecologist looks at Cancer and the Environment. London: Virage

Treasure, T. (1998) "Lessons from the Bristol case. More openness – on risks and on individual surgeons", British Medical Journal, 316, pp. 1685-1686.

Walshe, K. and Sheldon, T. (1998) "Dealing with Clinical Risk: Implications of the Rise of Evidence-Based Health Care" Public Money and Management 19 (4) pp. 15-20

Weick, K.E. (1988) "Enacted sensemaking in crisis situations" Journal of Management Studies, 25, pp. 305-317

Weick, K.E.. (1993) "The collapse of sensemaking in organizations: The Mann Gulch Disaster", Administrative Science Quarterly, 38, pp. 628-652.

Weick, K. (1995) Sensemaking in organizations. Thousand Oaks: Sage Publications.

Weick, K.E. and Roberts, K. H. (1993) 'Collective minds in organizations: Heedful interrelating on flight decks', Administrative Science Quarterly, 38, pp. 357-381.

Weinberg, A.M. (1972) 'Science and Trans-science'. Minerva, 10, pp. 209-222.

Biographies

Professor Sir Kenneth Calman is the Vice Chancellor of the University of Durham. Until September 1998 he was Chief Medical Officer for England and Wales

Professor Denis Smith is Professor of Management at the University of Durham. From the 1st April he will be Professor of Management and Head of the Centre for Risk and Crisis Management at the University of Sheffield