The evaluation of an expert system for the analysis of umbilical cord blood

Jonathan M. Garibaldi [a,*], Jennifer A. Westgate [b],
Emmanuel C. Ifeachor [a]

[a] *School of Electronic, Communication and Electrical Engineering, University of Plymouth, Drake Circus, Plymouth PL4 8AA, UK*
[b] *Department of Obstetrics and Gynaecology, Middlemore Hospital, Private Bag 93311, Otahuhu, Aukland, New Zealand*

Abstract

An assessment of neonatal outcome may be obtained from analysis of blood in the umbilical cord of an infant immediately after delivery. This can provide information on the health of the new-born infant, guide requirements for neonatal care, but there are problems with the technique. Samples frequently contain errors in one or more of the important parameters, preventing accurate interpretation and many clinical staff lack the expert knowledge required to interpret error-free results. The development and implementation of an expert system to overcome these difficulties has previously been described. This expert system validates the raw data, provides an interpretation of the results for clinicians and archives all the results, including the quality control and calibration data, for permanent storage. Issues regarding the clinical evaluation of this system are now detailed further, along with some clinical results illustrating the potential of such a system. © 1999 Elsevier Science B.V. All rights reserved.

* Corresponding author. Tel: + 44-1752-233514; fax: + 44-1752-232583.
E-mail address: jong@cis.plym.ac.uk (J.M. Garibaldi)

1. Introduction

An assessment of neonatal outcome may be obtained from analysis of blood in the umbilical cord of an infant immediately after delivery. Samples of blood may be taken from the umbilical cord of the neonate immediately on delivery and a blood gas analysis machine measures the pH, partial pressure of carbon dioxide (pCO_2) and partial pressure of oxygen (pO_2). A parameter termed base deficit of extracellular fluid (BD_{ecf}) can be derived from the pH and pCO_2 parameters [15]. This can distinguish the cause of a low pH between the distinct physiological conditions of respiratory acidosis, due to a short-term accumulation of CO_2 and a metabolic acidosis, due to lactic acid from a longer-term oxygen deficiency. Analysis of the acid–base balance of arterial and venous blood from a clamped umbilical cord provides objective information on the severity and duration of any lack of oxygen during labour. Such assessment of the acid-base status of umbilical cord blood has recently been recommended by the British Royal College of Obstetricians and Gynaecologists [14].

There are, however, a number of difficulties with the procedure. During a clinical trial in Plymouth [20], routine cord blood sampling on every delivery was initiated. Careful retrospective analysis of the cord blood gas results highlighted a 25% failure rate to obtain arterial and venous paired samples with all parameters [19]. This sampling error rate is broadly in line with other studies in which the importance of paired samples was recognised and this is despite the fact that the sampling took place within a research study which featured regular staff training sessions. The study also highlighted the facts that clinical staff were not very good at identifying sampling errors, did not recognise the occurrence of two samples from the same vessel and were poor at interpreting the results.

The development and implementation of an on-line expert system for the validation and interpretation of acid–base data for blood taken from the umbilical cord of the neonate immediately after delivery has previously been described [3]. In this paper, more emphasis is placed on the evaluation of the expert system. First the general requirements for the evaluation of medical expert systems and how these requirements impact on our expert system, are discussed. Next the actual steps taken in the verification and validation of the expert system prior to its commercial release are detailed. Two small studies based on the clinical data collected by the expert system are used to illustrate the potential of the expert system. Finally, the evaluation of the expert system so far and the further steps required for formal clinical evaluation are discussed.

2. Evaluation of medical expert systems

Many authors have used the terms verification, validation, assessment and evaluation in a differing and inconsistent manner in the literature [10,11]. In this paper the following terminology, designed specifically for the European AIM project [2], is adopted:

- Verification: is the process of ensuring that the expert system is functioning according to its specification,
- Validation: is the process of ensuring that the knowledge embedded within the expert system is an accurate representation of the domain
- Assessment: is the process of determining the effect that the expert system has in the clinical setting—this can be further split into two further sub-tasks
  1. Human factors assessment: determining whether the system is useful and usable to its clinical users and
  2. Clinical assessment: determining whether the system makes a measurable difference (improvement) to clinical care
- Evaluation: is a global term that refers to the collective processes of verification, validation and assessment.

Although most authors assert that thorough evaluation of medical expert systems is an essential pre-requisite to their routine use in the clinical situation, it is widely acknowledged that this is very difficult in practice [8,21]. A formal clinical evaluation should either demonstrate that the new treatment, technique or technology improves patient care, or show that it maintains patient care whilst decreasing cost. The usual method of evaluating novel medical treatment, the double-blinded randomised control trial (RCT), in which neither the administering clinicians or treated patients are aware of which arm of the trial the patients are in, would be extremely difficult to implement for an expert system for a number of reasons:

- The lack of external criteria against which to measure the expert system
- To 'blind' the clinician an independent third-party would have to interact with the expert system, thus adding an additional level of interpretation and indirection
- The effect of the expert system will depend on the initial skill levels of the clinicians involved in the trial
- The transfer of knowledge from the expert system to clinicians through interaction with the system over time may influence results

An alternative to the RCT is the less demanding 'test of no harm' or 'safety-test', in which the safety of the expert system is considered and the establishment that the system cannot harm a patient is sufficient—full clinical assessment is not necessary from a safety point of view [6].

Traditionally expert systems have been characterised as:

- Decision making systems—the expert system reaches decisions on patient care and presents the decisions as the correct patient management, for example an intelligent anaesthetic control system, or
- Decision support systems—the expert systems reaches decisions on patient care and presents its recommendation to the human clinicians, who then reach their treatment decision based on the expert system's recommendations, their own judgements and other clinical factors.

The acid–base expert system presented here does not fall naturally into either of these categories. The expert system takes a set of data and performs validation and interpretation of the data, but does not offer (even a suggestion of) a decision for clinical action. Thus it effectively transforms the four-dimensional numerical input

data into a single textual interpretation. This puts it into a category that is less interventionist than even a decision support system. It is an expert system by the definition of Jackson [5]:

> An expert system is a computer program that represents and reasons with knowledge of some specialist subject with a view to solving problems or giving advice

in that it represents and reasons with knowledge of the specialist subject of umbilical cord acid–base analysis with a view to solving the problem of validating and interpreting the raw data. However, as it neither makes nor suggests a decision, such an expert system might be termed an interpretation support system. Hence, this new category can be characterised as:

● Interpretation support systems—the expert system performs an intelligent analysis of raw data and presents processed data to the clinician in a form which is more natural, but does not recommend any specific clinical action.

Many other (non expert system) technologies have been introduced into clinical use in the last 30–40 years, usually without the stringent evaluation requirements that have been advocated for expert systems. Many of these technologies are microprocessor based, for example the cardiotocograph (CTG), yet had little or no formal evaluation before their introduction. Although it might be argued that they have suffered as a result, their introduction has been implemented and is often widespread. As many of these technologies are very specialised and also involve significant amounts of data-preprocessing, the differences between these systems and an interpretation support expert system are small and blurred. Thus, it is argued, the evaluation requirements for clinical assessment of each of the expert system types is more accurately represented by Table 1.

The final aspect of evaluation that must be addressed is the medico-legal consideration. It is still not clear whether an expert system will be viewed as a 'product' or a 'service' by the courts, if it is subject to litigation [1]. If it is viewed as a product, then it would be subject to product liability laws, which are particularly strict in the USA. However, if the expert system is considered to be a service, then it must reach the standard expected of an 'informed and sensible body of opinion' [21].

Table 1
Expert system types and evaluation requirements for clinical assessment

Expert system category	RCT	Safety test
Decision making	Required	Required
Decision support	Desirable	Required
Interpretation support	Optional	Required

Table 2
How each task relates to components of evaluation

Task	Verification	Validation	Assessment
Subsystem validation	●		
Face validation	●	●	
Hazard analysis	●		
Sensitivity analysis	●	●	
Economic assessment			●
Field tests in Plymouth		●	●
Field tests in Exeter		●	●

Umbilical cord blood acid–base analysis is a classic example of a domain where no gold-standard exists. It is currently difficult to establish the degree of brain damage, even in extreme circumstances and totally impossible (with current technology) to identify absolutely whether any diagnosed damage actually occurred during labour. Consequently, it is only possible to validate the performance of the expert system against the opinions of respected clinical experts.

3. Evaluation for commercial release

The industrial partner collaborating in this project was a British company certified to conform to the requirements of the BS5750 quality assurance standard. In essence, this standard simply states that within an organisation all procedures should be specified and that adherence to the specifications should be provable. In practice, this implies that each procedure should have a specification document which identifies what tasks should be carried out and the documentation that should be produced as a result. As the collaborative partner was a blood gas analyser manufacturer whose products feature complex electronics and embedded software destined for critical clinical use, the procedures for software testing were already established. These procedures concentrated on ensuring that the software was clinically safe. In addition, their BS5750 requirements implied that any third-party software developer had to comply with these existent testing procedures.

Thus, a number of specific tasks were carried out in compliance with the company's BS5750 requirements and to allow the release of the expert system. These tasks were to:
1. Ensure that the system was safe,
2. Ensure that the interpretations agreed with respected experts and
3. Demonstrate the potential for economic benefit.

The tasks carried out in the evaluation of the expert system will each be described. Table 2 shows how each of these tasks relates to the evaluation terminology presented at the beginning of this section.

3.1. Subsystem validation

Subsystem validation was carried out to ensure that the software development cycle complied with BS5750 quality standards. This involved extensive 'destruction testing' of the software in which, as far as possible, every aspect of the software was tested. Specifically, each line of code was examined to ensure that its behaviour was well determined. The code was structured such that the use of `goto` was eliminated and the use of break to prematurely exit from control loops was avoided. The code was compiled under the highest level of warning messages and was refined until it produced no warnings. In addition, each time any function (including library functions) was called, all the possible return values for the function were anticipated and the calling code was amended to take appropriate action in each case.

The principle is that each subsystem (function) should not be able to exhibit any behaviour other than anticipated. Any non-anticipated behaviour is catered for through the use of a software exception routine, such that a message is displayed to the user screen with a description of the exception condition and an instruction for the user to call the technical support department. If this occurs the expert system is immediately halted. Conceptually, it is better for the system to halt with an exception condition, rather than to continue to run in a state that could result in an ill-founded expert system interpretation. The few minor problems that this process highlighted were corrected.

3.2. Face validation

During face validation project team members, potential expert system users and people knowledgeable about the application domain, subjectively compare the performance of the expert system against human expert performance [10]. Face validation was partially integrated into the development phase, during the process of rule elicitation. Once the rules had been established, the complete rule set was given to a number of other experienced clinicians. Each clinician was asked to highlight any interpretation rules that they would disagree with. Additionally, all 'non-normal' results that occurred during the initial field trials in Plymouth and Exeter (see below) were regularly reviewed by the resident experts.

The result of this face validation was the minor modification of one rule, with it being split into two sub-rules. At the end of this process, no cases of non-trivial disagreement between clinicians and expert system had been discovered. This was then taken to be sufficient for an adequate demonstration of the legal criterion of reaching the standard expected of an 'informed and sensible body of opinion' described above.

3.3. Hazard analysis

The process of hazard analysis was prescribed as part of the BS5750 requirements of the industrial partner. In contrast to the 'white-box' approach of subsystem validation, in which the code itself is examined to anticipate failures, in hazard

analysis a 'black-box' approach is used, in which the behaviour of the system is observed in response to all conceivable external events. Each potential hazard is identified and documented and the appropriate behaviour of the system is specified. The hazard is then instigated or simulated (as far as possible) and the actual behaviour of the system is recorded. The specific behaviour of the system is not particularly important—the specified behaviour to an unlikely hazard could be an immediate system crash with unintelligible error message—although commercial considerations usually impose the requirement of graceful and predictable behaviour to all hazards.

As an example, consider the expert system function to allow a download of all data to floppy disk. Table 3 shows the hazards that were identified for this procedure and the anticipated program behaviour in each case.

Such manual hazard analysis tested all user-selectable functions and commands, but did not address either the screen sequences which result from normal user input or the functioning of the expert system module itself. Both of these aspects were considered too time consuming and too complex to test manually, due to the large number of combinations of each. Therefore, a suite of automatic test procedures was created to attempt to ensure the correct functioning of these software components. As an additional advantage this test suite was designed to be utilised both in the development of the system and when the software had to be updated, for example, to communicate with a new blood gas analyser developed by the industrial partner.

The automatic test program has the ability to simulate communications functions and user inputs. The communication functions of the blood gas analyser were encapsulated in a simulation program, which could be connected via a loop-back to the test machine and run in one instance of the test program. A second program was created to run the expert system and to simulate user input, which was run through a second instance of the test program. A set of bespoke databases were created to drive simultaneously both test programs with a set of key strokes to simulate, a set of sample results to transmit and a set of target output states of the system. The target output states could comprise not only the system screen that should be on view, but also the expected expert system interpretation. A set of

Table 3
An example of hazard analysis for the process of database download onto floppy disk

Action	Behaviour
Select database download to A:	Databases are copied to floppy disk
Select database download to C:	Databases are copied to hard disk
User selects 'Cancel' button	No databases are copied
Floppy disk is write-protected	Warn user and prompt for new disk
Floppy disk is unformatted	Warn user and prompt for new disk
Floppy disk is nearly full	Prompt for new disk when disk is full
Floppy disk contains previous download	Warn user and abandon download immediately
Floppy disk removed during download	Prompt user to re-enter disk in drive

target databases that should result from a complete run of the test suite was also created.

The complete test suite attempted to verify every 'path' through the system, from the initial 'Main Menu' to the final results screen and back to the 'Main Menu'. As the control structure allowed looped paths, theoretically there are an infinite number of paths through the system. For example, if a bad sample is detected, the operator is prompted to select Retry, Ignore or Abandon. If Retry is selected, the user returns to the same sample screen and inputs another sample—which could be bad—etc. Rather than following these potentially infinite loops, the test program tried each looping branch once and then twice only and each optional looping branch zero times, once and twice only. After all these tests had been completed and satisfactorily passed, the system was deemed to be functionally correct according to specification.

3.4. Sensitivity analysis

O'Keefe [10] defines sensitivity analysis as "systematically changing expert system input variable values and parameters over some range of interest and observing the effect upon system performance". The test suite described above was utilised to perform a comprehensive sensitivity analysis on the expert system categorisations. The same principle as described above was adopted, with one test program simulating a blood gas analyser by reading pre-defined results from a database and transmitting them as if they were actual sample results, while a second instance of the test program simulated a user running the expert system and verifying that the obtained results matched those expected.

As there were too many possibilities of input data to test exhaustively, a method was devised to pick a selection of important results that lay in the middle and at the edges of all rule boundaries. The interpretation is essentially a four-dimensional classification task. Initially a two-dimensional partition was fixed in pH_A and BD_A and a two-dimensional partition graph of pH_V against BD_V was drawn by the experts to construct the classifications for this sub-space. As an example, the first partition was for $pH_A < 7.05$ and $BD_A \geq 12$ mmol l^{-1}. This process was then repeated by varying the BD_A partition first, for example $BD_A \geq 10$ and $BD_A < 12$ mmol l^{-1} and so on until the entire four-dimensional space was classified; each expert system categorisation was assigned a category number (ESN). These two-dimensional partition graphs were utilised to select data samples that lay on the corners, borders and middle of each partition. Fig. 1 shows a partition graph with the data points that lie on corners, borders and middles highlighted. Appropriate values of the pH and pCO_2 for each vessel that would produce closest possible values of pH and BD_{ecf} to the required point were calculated independently and entered into a database. As can be seen from Fig. 1, there are 26 sample points for this partition, which corresponds to $pH_A < 7.05$ and $BD_A \geq 12$ mmol l^{-1}. Hence, with a pH_A value on the border (7.04) and far away (6.80) and a BD_A value on the border (12 mmol l^{-1}) and far away (15 mmol l^{-1}), there are 104 ($26 \times 2 \times 2$) sample points.
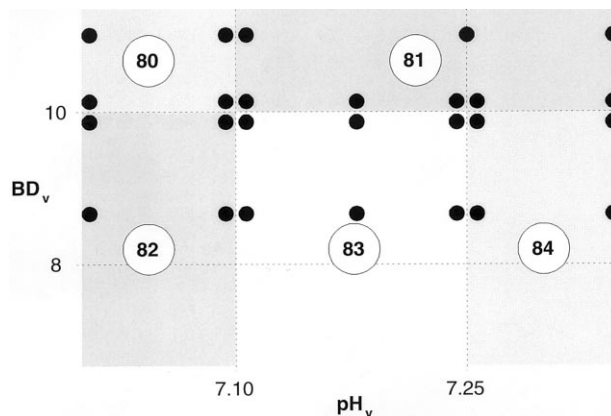
Fig. 1. Partition graph showing data points chosen for simulated sampling in sensitivity analysis of expert system categorisation, with ESN for each region.

Altogether almost 1000 samples were created to systematically test the expert system categorisation across the entire ranges of each parameter. In each case the expected expert system category was forecast and the test program verified that the specified result was obtained. The process did highlight a small number of cases (six) in which the interaction of validation and categorisation rules produced a different output to that expected. These cases were closely examined and it was judged that the actual output was more 'reasonable' than the anticipated output. Hence the anticipated output was adjusted and the test continued. These tests are now run at each software modification/update to re-verify and re-validate the functioning of the expert system.

3.5. Economic assessment

If a full randomised control trial of an expert system is not possible, or not required, then it has not necessarily been shown to be of any benefit. Given that the expert system can be shown to be safe ('do no harm'), an economic assessment of the benefits of the expert system may be enough to justify its use [8]. Umbilical cord acid–base (UAB) assessment has the potential to be of large economic benefit. To address the economic factors, it is necessary to discuss the problems of litigation in more detail. It is estimated that only around 10–20% of neurological handicap can be attributed to intrapartum events [9,17]—either intrapartum asphyxia or traumatic damage, caused for example, by instrumental (forceps) delivery.

In the current litigious climate, the parents of a brain-damaged infant are often encouraged to sue and the defendant is likely to be the obstetric clinician. For the plaintiff to be successful, the prosecution must prove three things:
1. The presence of brain damage,
2. Causality and
3. Negligence.

Although this would apparently favour the defendant, it is not uncommon for the judge (or jury) to find in favour of the unfortunately brain-damaged child simply because there were some problems with the birth and the clinician has no adequate defence—despite the normal legal requirement to prove guilt. As it is relatively straight forward to find problems with almost any birth, a clinician without defence is always susceptible.

Normal UAB results will greatly increase the credibility of a defendants case and are often sufficient in themselves to defend a case successfully [18]. In Plymouth, several cases have been deflected from litigation (the case dropped by the prosecution before going to court) on the basis of normal acid–base results alone. Each lost case currently costs an estimated £2 million in settlement and legal fees (split fairly evenly). In the UK, this money comes out of the general health-care budget rather than any specific legal fund.

Table 4 shows a breakdown of deliveries according to whether the infant has cerebral palsy (CP) and the possible influence that UAB results might have on any litigation. It may be assumed that cases in the left-hand column, in which the infant does have CP as a result of intrapartum events, are made no more costly to settle through the presence of adverse UAB results. Even if more cases are settled as a consequence (which is doubtful) this would most probably be offset by the reduction in legal fees due to early settlements. For cases in the middle column, in which the infant has CP as a result of other non-intrapartum events, the vast majority (around 98% or greater) will have normal UAB results and hence will stand a much larger chance of being successfully defended. The 2% which may have abnormal UAB results, even though intrapartum events did not cause the CP, might be thought to make defence harder. In fact, these would probably make the situation little worse than if no UAB results are present and are hugely outnumbered by those with normal UAB results.

The right-hand column represents the cases that clinicians often worry about. There is no cerebral palsy, but abnormal UAB results were found. Although this group is large (2000 per 100 000), in fact they do not cause a problem because litigation will not be undertaken (and certainly would not be successful) due to the failure to satisfy criterion 1 above through the absence of CP.

Table 4
The likely effect that umbilical acid–base assessment would have on litigation for CP

	CP 200/100 000		No CP 99 800/100 000
	Intrapartum: 20% 40/100 000	Other: 80% 160/100 000	
Normal results	Trauma: 10%? 4/100 000 saved/settled?	98% 157/100 000 saved	98% 97 800/100 000 no litigation
Abnormal results	Asphyxia: 90%? 36/100 000 settled?	2% 3/100 000 false litigation?	2% 2000/100 000 no litigation

Given a conservative CP rate of around 2 per 1000, there will be around ten cases each year in a hospital such as Plymouth, with 5000 deliveries per year. As eight of these will probably have an antenatal or postnatal cause and $\approx 157/160$ of these will have normal UAB results, they are all likely to be defended. UAB analysis costs roughly £20 000 per year in equipment and maintenance costs. Therefore, it is only necessary to save one additional CP litigation case every 100 years (£2 000 000/£20 000) out of the 800 potential cases to justify the cost. This of course is only possible if the UAB results were obtained and were reliable. As the expert system costs a mere £200 per year at most (£2500 spread over at least 12.5 years), the expert system must only improve reliability rates of UAB analysis by 1% (£200/£20 000) to ensure that its additional cost over 'manual' UAB analysis is justified! Although the reliability rates have not been formally investigated in isolation, an improvement of 4% reliability has been shown in Plymouth in the year since the introduction of the expert system [3]. Note that the costs incurred in the research development have not been included in this cost-benefit analysis. Thus, the analysis demonstrates the potential benefit to the obstetric unit that purchases the expert system, but does not demonstrate the financial benefit of the project as a whole in developing the system.

It should be remembered that the reduction of unwarranted litigation is desirable not only to save money, but also to avoid the increase in 'defensive' medical practice which is ultimately harmful to the patient: the obstetric clinician should not be held responsible for events that were outside their control. As a final comment, there is a good argument for considering 'no-fault' compensation for any individual unfortunate enough to have cerebral palsy, but this is (and should be) distinct from the argument over current trends in litigation.

4. Clinical assessment

The expert system was placed at the local hospital and at a nearby hospital for extensive field trials before release. The local hospital at Plymouth had performed UAB analysis manually since March 1992 and the expert system was introduced in July 1993. In contrast, the nearby hospital in Exeter did not perform UAB analysis at all and the introduction of the expert system was used to initiate UAB analysis. The users at Plymouth therefore represented a set of users familiar with a manual system who had to change to the expert system, whereas the users at Exeter represented novice users to both the clinical practice and the expert system. During the field trials the output of the expert system was regularly reviewed for all abnormal cases by the resident clinical experts and feedback was obtained from the users on the usability of the system.

In this section two clinical studies based on data collected by the expert system are used to illustrate the potential of the expert system in clinical audit and training. Firstly, a comparison is made between data obtained from the two hospitals at Plymouth and Exeter which demonstrates statistically significant differences. Secondly, a comparison is made between neonatal umbilical acid–base results collected

at Plymouth and the generally accepted normal ranges for adult acid–base interpretation. This demonstrates the need for specific interpretation rules for neonatal UAB as encapsulated within the expert system.

4.1. A comparison of two centres

4.1.1. Methods

Routine cord blood sampling for every delivery began in Exeter on the 1 October 1993, after a short period of familiarisation with the system. Data collection took place in the middle of July 1994, so data for the nine complete months from 1 October 1993 to 30 June 1994 were taken from both centres for comparison. Median and centile ranges were used to describe the populations as all have markedly skewed distributions. Comparisons of location were made using Student's *t*-test, as the high numbers in both groups ensured that the test was reliable and the proportions of expert system categorisation were tested with χ^2, all at 5% significance level.

Every blood gas analyser will have its own particular calibration, which will depend on internal calibration parameters and individual electrodes. Consequently each machine will have minor performance differences. The quality control results from both centres were examined to standardise the pH and pCO_2 results so that the pH and BD_{ecf} results could be properly compared, without the minor machine differences influencing the statistics. Three sets of quality control material with preset levels of pH and pCO_2 parameters were measured regularly to ensure the correct functioning of the analysers, as part of the routine clinical maintenance of the machines. Comparison of these quality control results at each level allowed for compensation of differences in machine performance (cross-calibration). Regression analysis of the monthly means for each parameter at each level showed that there was no overall trend in machine calibration.

Examination of mean differences by Student's *t*-test showed that there was no difference in mean pH at each of the three levels, but that there was a significant mean difference in pCO_2 readings. Regression analysis on the means of each of the three levels at the two sites showed that the Exeter pCO_2 results could be corrected by:

$$pCO_2 \text{ (corrected)} = 1.07 \, pCO_2 \text{ (measured)} - 0.81 \tag{1}$$

which represents a correction of between approximately 0.0 and 0.5 kPa over the range of pCO_2 ($R^2 \approx 1.0$). Once this was applied, the mean differences in pCO_2 quality control results at each of the three levels were eliminated. This correction was then applied to all pCO_2 readings from Exeter and the results used to re-calculate the BD_{ecf} parameters, before further statistical description and comparison presented below.

4.1.2. Results

In Plymouth, 3318 samples were taken from 3544 deliveries (93.6%). Of these samples, 95.7% were intended to be from both artery and vein and 4.3% were single

Table 5
Median and 25th to 97.5th centile range for acid–base parameters at Plymouth and Exeter

Vessel	Plymouth ($n = 2684$)			
	pH	pCO_2 (kPa)	pO_2 (kPa)	BD_{ecf} (mmol 1^{-1})
Artery	7.27 (7.06–7.40)	6.7 (4.5–9.7)	2.1 (0.8–3.7)	3.6 (-0.9–11.0)
Vein	7.36 (7.16–7.49)	5.1 (3.3–7.6)	3.6 (1.9–5.3)	3.6 (-0.5–9.4)
Vessel	Exeter ($n = 1222$)			
	pH	pCO_2 (kPa)	pO_2 (kPa)	BD_{ecf} (mmol 1^{-1})
Artery	7.23 (7.03–7.36)	7.1 (4.6–10.3)	2.4 (0.9–4.3)	4.7 (-0.4–12.2)
Vein	7.34 (7.15–7.46)	5.3 (3.4–7.9)	4.1 (2.1–6.1)	4.1 (-0.6–10.3)

samples. Of the intended artery and vein samples, the expert system classified 84.5% (2684) as validated arterial–venous pairs. In Exeter, 1848 samples were taken from 2116 deliveries (87.3%). Of these samples, 88.7% were intended to be from both artery and vein and 11.3% were single samples. Of the intended artery and vein samples, the expert system classified 74.5% (1222) as validated arterial–venous pairs. Thus only 75.7% of deliveries in Plymouth and only 57.8% in Exeter resulted in validated paired samples. The median and 2.5th to 97.5th centile range for each parameter are given in Table 5.

Examination of the distributions of pH in the artery and vein (Fig. 2) showed a significant shift to lower pH's in both vessels in Exeter compared to Plymouth (pH_A: $t = 13.9$, $p \approx 10^{-42}$; pH_V: $t = 6.6$, $p \approx 10^{-11}$). Similarly, the distributions of BD_{ecf} in the artery and vein (Fig. 3) both showed a significant shift to higher BD_{ecf}'s in Exeter compared to Plymouth (BD_A: $t = -10.8$, $p \approx 10^{-26}$; BD_V: $t = -3.6$, $p \approx 0.0003$).

As the differences between vessels are clinically significant [13], these too were examined by Student's t-test. The mean pH difference of 0.11 at Exeter was found to be significantly greater than that of 0.09 at Plymouth ($t = -11.7$, $p \approx 10^{-30}$) and the mean BD_{ecf} difference of 0.6 mmol 1^{-1} at Exeter was also found to be significantly greater than that of 0.0 mmol 1^{-1} at Plymouth ($t = -12.0$, $p \approx 10^{-31}$).

The expert system categorisations (ESN) were examined and the results are shown in Table 6. As can be seen, the observed frequencies of the 'worst' ESN groups (80s and 90s) are higher at Exeter than at Plymouth, although the results did not reach significance ($\chi_4^2 = 6.43$, $df = 4$, $P = 0.169$). When cast in a 2×2 contingency table of ESN groups 80s and 90s (arterial pH < 7.05) compared to the rest, the 41 (3.4%) cases at Exeter were significantly more than the 57 (2.1%) cases at Plymouth (Yates corrected $\chi_1^2 = 4.71$, $P = 0.030$). Although the same trend was apparent when grouped by ESN 80s ($pH_A < 7.05$ and $BD_A \geq 12$ mmol 1^{-1}) compared to the rest, the difference (17 cases at Exeter and 22 cases at Plymouth) did not reach statistical significance (Yates corrected $\chi_1^2 = 2.23$, $P = 0.136$).
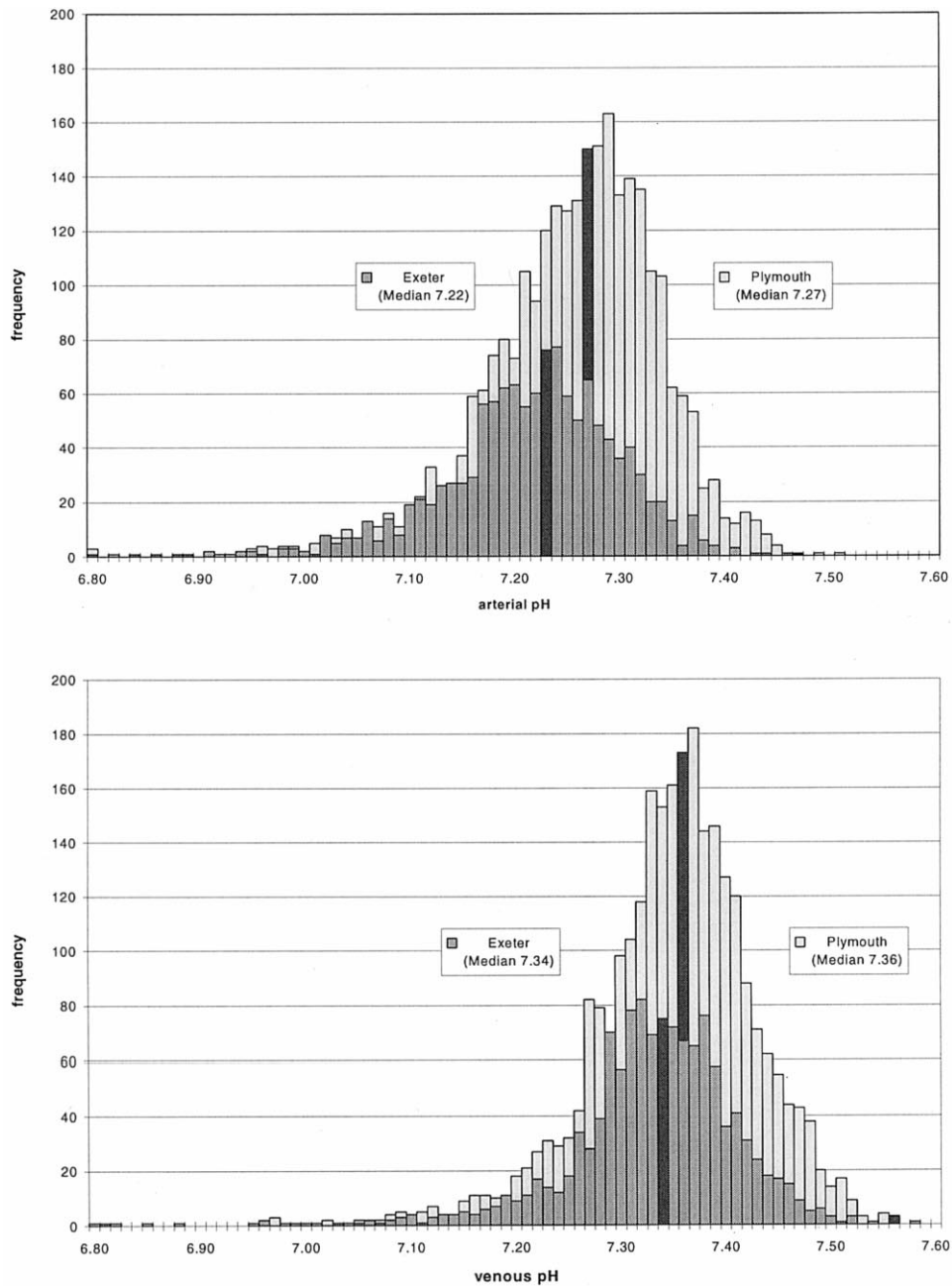
Fig. 2. Distributions of arterial and venous pH at Plymouth and Exeter.
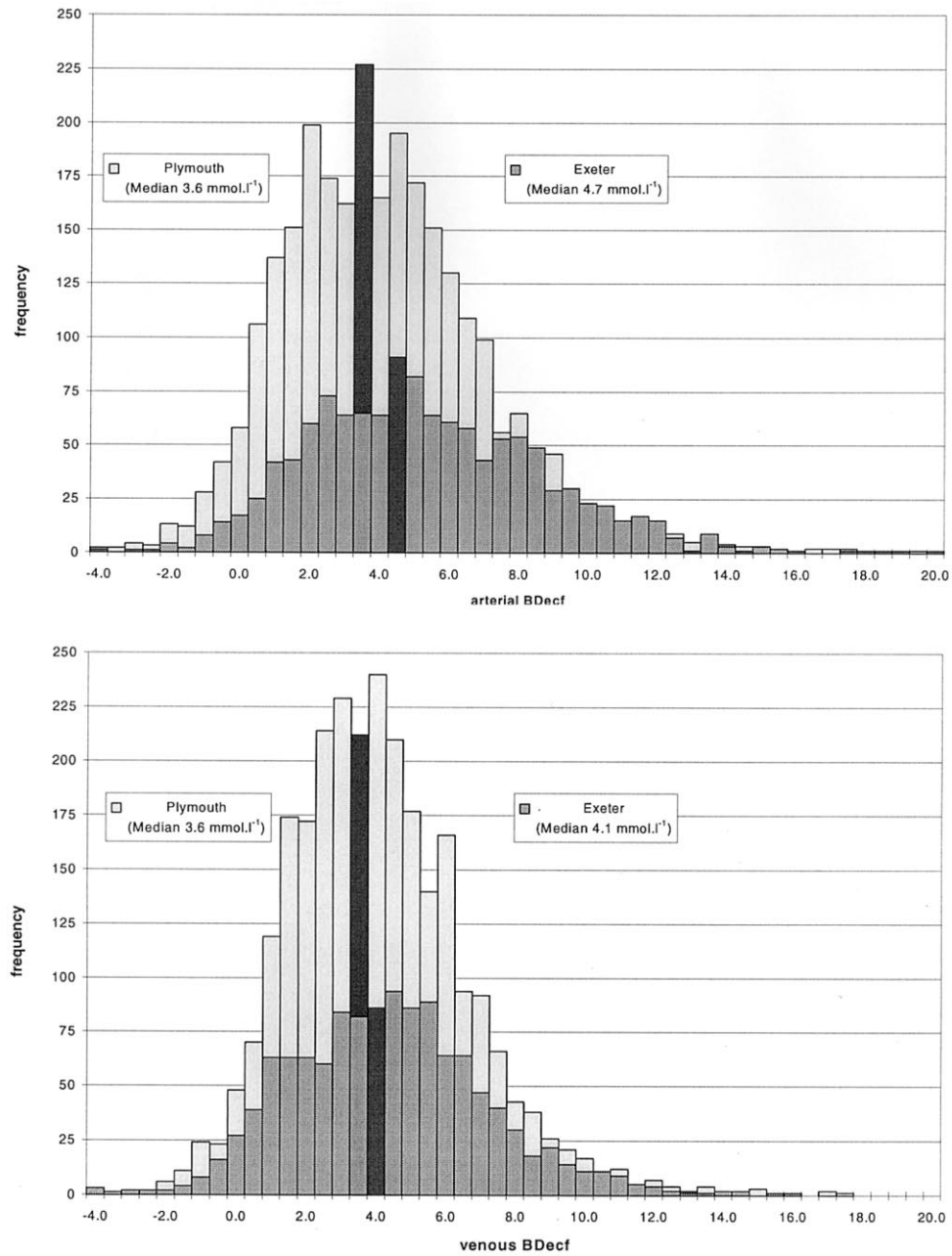
Fig. 3. Distributions of arterial and venous BD$_{ecf}$ at Plymouth and Exeter.

4.2. A comparison of umbilical and adult acid–base

Siggaard-Andersen published an acid–base chart for adult and infant blood, in which the pH is plotted against the logarithm of pCO_2 such that lines of constant BD_{ecf} would appear as straight lines [16]. The chart also indicated ranges for various types of abnormal acid–base status, including acute and chronic forms of base deficit, in which the term chronic was defined to be over a period of 6–12 h. Note that adult is used in this context to mean non-fetal or not in the immediate neonatal period.

The normal range for adult arterial acid–base was centred around a pH of 7.40, pCO_2 of 5.3 kPa (40 mmHg) and a BD_{ecf} of 0 mmol l^{-1}. This is clearly not appropriate for the fetus at the end of labour, where the normal range for arterial pH is around 7.27 or below and the normal range for arterial pCO_2 is around 6.7 kPa (50 mmHg) or above. This corresponds to a BD_{ecf} of around 3.6 mmol l^{-1}. Venous values for umbilical cord blood are nearer to the adult values, but still clearly different. In addition, there is probably no fetal equivalent of the chronic base deficit defined for adults in which the acids are not the result of anaerobic metabolism. These differences are illustrated in Figs. 4 and 5, in which the values of the approximately 10 000 arterial and venous results from paired vessels collected in Plymouth are superimposed onto the standard Siggaard-Andersen Acid–Base Chart.

It can be seen from these charts that there are no results in the positive base excess portion and that the vast majority of results are outside the normal range defined for adults. Most lie in the range between acute hypercapnia (short term accumulation of pCO_2) equivalent to an acute respiratory acidosis and acute base deficit (short term accumulation of non-carbonic acid) equivalent to an acute metabolic acidosis. Again the term acute is defined for adults and may not correspond to the clinical usage of the term in the perinatal period. Additionally, ten arterial points and four venous points lie off the top left-hand side of the charts. It is clear from these charts that the adult term normal is inappropriate for umbilical acid–base assessment: this reinforces the point that specific knowledge of fetal physiology is required for accurate interpretation of umbilical acid–base.

Table 6
Observed and expected frequencies of expert system categories at Plymouth and Exeter $\chi^2_4 = 6.43$

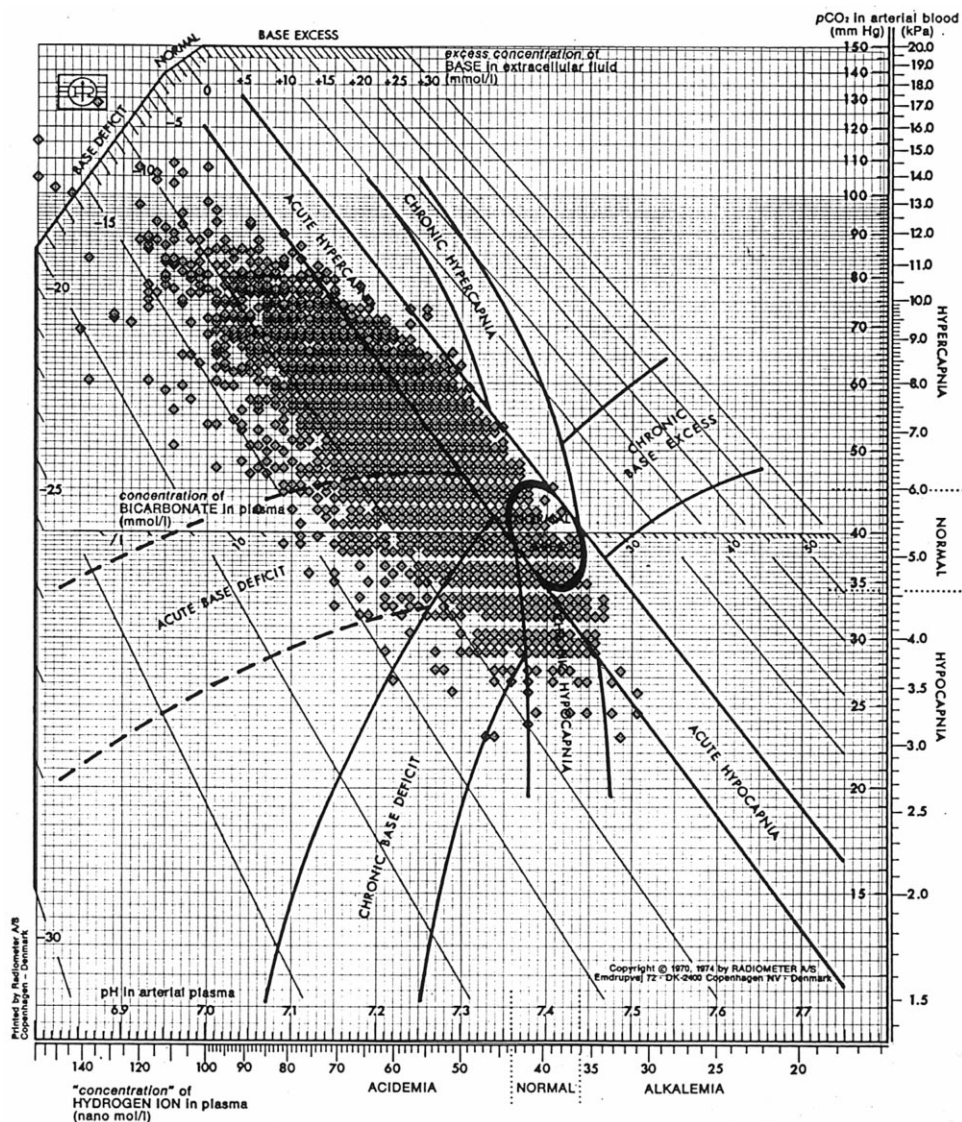| Centre	ESN						Total
	80s	90s	100s	110s	120		
Plymouth	22	35	18	12	2597		2684
		26.8		40.5		19.2		10.3		2587.1	
Exeter	17	24	10	3	1168		1222
		12.2		18.5		8.8		4.7		1177.9	
Total	39	59	28	15	3765		3906

Fig. 4. Neonatal umbilical arterial acid–base results superimposed onto the Siggaard-Andersen Acid–Base Chart for adults.

5. Discussion and conclusions

This paper has detailed the various aspects of the evaluation of the umbilical acid–base expert system. The evaluation process can be divided into two main sections:

1. Evaluation required for commercial release and

2. Subsequent clinical assessment.

Although the evaluation process has been presented separately from the design and development process described in the previous paper [3], it is important to emphasise that the two processes took place concurrently to a large extent.

Early on in the design phase, it was decided that the system would produce an interpretation of data only. This, it is believed, contributed to the relatively rapid
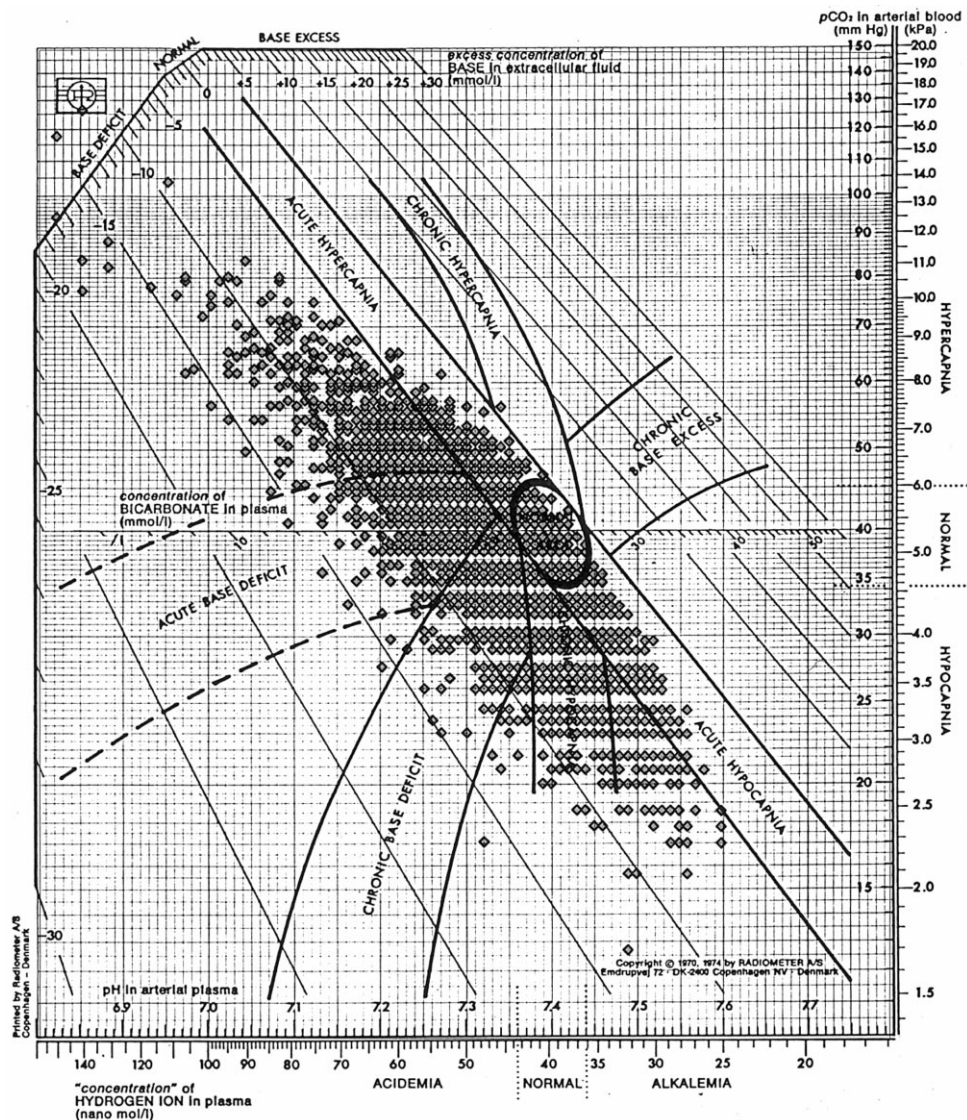

Fig. 5. Neonatal umbilical venous acid–base results superimposed onto the Siggaard-Andersen Acid–Base Chart for adults.

transfer of the system from the research environment to commercial release. This was beneficial for two reasons:

1. It reduced the requirement for clinical assessment of the system, as discussed in Section 2, to little more than proof of clinical safety and
2. It avoided the presentation of a threat to the clinicians, that a decision support system might.

Although clinicians continue to debate the necessity and utility of umbilical acid–base assessment, there has been wide recognition that stringent data validation is valuable if the procedure is to be undertaken. As data validation is viewed as a tedious task, the expert system is seen as a benefit. Questions have been raised on details of the interpretations offered, but as they are not presented as decisions, the clinicians feel happy to place their own interpretation on the data, if they disagree.

There were significant differences in the population statistics between Plymouth and Exeter found in Section 4. This is in the context of two sets of identically validated results, collected on cross-calibrated machines. These results do not seem to be due to selective sampling and may be caused by a difference in practice in the two centres. Neither centre has sufficient other data to examine whether these acid–base differences are reflected in other outcome measures, such as long-term neurological development. Indeed, it would also be necessary to have far more cases in each centre before such a question could be adequately addressed. There was no obvious difference in possibly the two most important factors reflecting management of labour; caesarean section or epidural rates. The BD_{ecf} difference could be a result of longer second stages in Exeter as large arterial–venous BD_{ecf} differences are commonly the result of reduced umbilical blood flow in the second stage [13]. It is believed that the second stage has become shorter in Plymouth as a result of routine cord sampling, which commenced in March 1992. Unfortunately, data on duration of second stage was not available to examine this. The fact that the system has shown a significant difference between the two centres demonstrates its potential as an audit tool. It provides an exciting opportunity to explore the impact of change in clinical practice, both over time and between centres.

It is important to stress that these clinical studies have not fully assessed the expert system in the accepted sense. They simply demonstrate the potential of the system in terms of clinical comparisons of umbilical acid–base data. In order to fully assess the impact of the expert system, it is necessary to define how the information that the expert system provides is to be used and then to design a clinical trial to measure the effect of this information. There are, in fact, four distinct main uses of the information:

1. Individual feedback—the clinician(s) responsible for delivery use the acid–base information to assess their own management of a particular case retrospectively
2. Group audit statistics—population data are used to compare changes over time or between centres as illustrated in the comparison of Plymouth and Exeter
3. Neonatal guidance—paediatricians responsible for care of an infant in poor condition use the umbilical acid–base data as 'baseline' information to assess changes in condition or to guide feeding and resuscitation regimes

4. Medicolegal protection—umbilical acid–base information is used to protect obstetricians from undeserved and unwarranted litigation (see Section 3.5)

Each of these uses implies different assessment criteria:

1. Individual feedback—interviews with obstetric clinicians, performance data
2. Group audit statistics—long term outcome variables, other perinatal data, large multi-centre trials
3. Neonatal guidance—interviews with paediatric clinicians, neonatal data
4. Medicolegal protection—records of legal cases or deflections

In conclusion, the expert system was successfully evaluated to the requirements for commercial release. Full clinical assessment of the expert system, as with any other medical expert system, is a much harder problem, which would require the availability of appropriate data collected within a long term multi-centre randomised control trial. There is currently probably not sufficient political impetus or the necessary financial commitment for such a study to take place in this country.

6. Future work

More work needs to be carried out on the clinical evaluation of the expert system, in order to try to quantify its effect on clinical practice. It is hoped that work can be undertaken to follow up the current clinical installations, some of which have been in place for over 2 years now. This may lead to the availability of more maternal and neonatal data, if any of these clinical sites have computerised database systems and may enable the opinions of the clinical users to be ascertained to provide data on the usage and subjective impact of the system.

A new version of the monitor for intrapartum care featuring analysis of the ST segments of the ECG waveform of the fetus during labour has been developed, known as the STAN2 monitor [12]. It is planned that a large randomised trial will take place in Scandinavia from 1998 in which the STAN2 monitor is compared against the conventional CTG monitor. This will feature the collection of antepartum, intrapartum and postpartum data in a controlled and uniform manner, for up to around 30 000 deliveries. There would be additional follow up of any damaged infants delivered during the trial, which might continue for a number of years subsequently. The data will be used to validate a fuzzy-logic based expert system for ST waveform analysis [4] and it is hoped that this expert system could be used in the trial for the collection and validation of the umbilical cord acid–base data.

There are also plans for a similar sized multi-centre trial to take place in Britain, co-ordinated from Plymouth, to test a CTG interpretation expert system [7]. If either of these trials takes place, the data collected could prove an invaluable source for evaluating more accurately the existing crisp expert system and to gauge whether a fuzzy-logic based system would provide any additional clinical benefit.

Acknowledgements

References

[1] Brahams D, Wyatt J. Decision-aids and the law. Lancet 1989;2:632–4.
[2] Engelbrecht R, Rector A, Moser W. Verification and validation. In: van Gennip EMSJ, Talmon JL, editors. Assessment and Evaluation of Information Technologies. Amsterdam: IOS Press, 1995:51–66.
[3] Garibaldi JM, Westgate JA, Ifeachor EC, Greene KR. The development and implementation of an expert system for the analysis of umbilical cord blood. Artif Intell Med 1997;10:129–44.
[4] Ifeachor EC, Outram NJ. A fuzzy expert system to assist in the management of labour. In: Proceedings of the International ICSC Symposium on Fuzzy Logic. Zurich, Switzerland, 1995:97–102.
[5] Jackson P. Introduction to Expert Systems, 2. Reading, MA: Addison-Wesley, 1990.
[6] Jennett B. Assessment of clinical technologies. Int J Technol Health Care 1988;4:435–45.
[7] Keith RDF, Beckley SL, Garibaldi JM, Westgate JA, Ifeachor EC, Greene KR. A multicentre comparative study of 17 experts and an intelligent computer system for managing labour using the cardiotocogram. Br J Obstet Gynaecol 1995;102:688–700.
[8] Miller PL, Sittig DF. The evaluation of clinical decision support systems: What is necessary versus what is interesting. Med Inform 1990;15(3):185–90.
[9] Nelson KB, Ellenberg JH. Antecedents of cerebral palsy. New Engl J Med 1986;315:81–6.
[10] O'Keefe RM, Balci O, Smith EP. Validating expert system performance. IEEE Expert 1987;2(4):81–90.
[11] O'Leary TJ, Goul M, Moffitt KE, Radwan AE. Validating expert systems. IEEE Expert 1990;5(3):51–8.
[12] Rosén KG, Luzietti R. The fetal electrocardiogram: ST waveform analysis during labour. J Perinatal Med 1994;22:501–12.
[13] Rosén KG, Murphy KW. How to assess fetal metabolic acidosis from cord samples. J Perinatal Med 1991;19:221–6.
[14] Royal College of Obstetricians and Gynaecologists. Recommendations Arising from the 26th RCOG Study Group. In: Spencer JAD, Ward RHT, editors. Intrapartum Fetal Surveillance. London: RCOG Press, 1993:392.
[15] Siggaard-Andersen O. An acid–base chart for arterial blood with normal and pathophysiological reference areas. Scand J Clin Laboratory Invest 1971;27:239–45.
[16] Siggaard-Andersen O. The Acid–Base Status of the Blood, 4. Copenhagen: Munksgaard, 1976.
[17] Stanley FJ, Blair E. Why have we failed to reduce the frequency of cerebral palsy? Med J Aust 1991;154:623–6.
[18] Thorp JA, Dildy GA, Yeomans ER, Meyer BA, Parisi VM. Umbilical cord blood gas analysis at delivery. Am J Obstet Gynecol 1996;175:517–22.
[19] Westgate JA, Garibaldi JM, Greene KR. Umbilical cord blood gas analysis at delivery: a time for quality data. Br J Obstet Gynaecol 1994;101:1054–63.

[20] Westgate JA, Harris M, Curnow JSH, Greene KR. Plymouth randomised trial of cardiotocogram only versus ST waveform plus cardiotocogram for intrapartum monitoring in 2400 cases. Am J Obstet Gynecol 1993;169:1151–60.

[21] Wyatt J, Spiegelhalter D. Evaluating medical expert systems: what to test and how? Med Inform 1990;15(3):205–17.