The Development of a Fuzzy Expert System for the Analysis of Umbilical Cord Blood

Jonathan M. Garibaldi and Emmanuel C. Ifeachor

School of Electronic, Communication and Electrical Engineering,
University of Plymouth, Drake Circus, Plymouth, PL4 8AA, U.K.

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 newborn infant and 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 results. The development and validation of an expert system to overcome these difficulties is described. The initial development utilised conventional 'crisp' logic within the rule base and this system was evaluated to commercial release. 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. Subsequent development went on to incorporate fuzzy logic into part of the expert system knowledge base, but tests of this preliminary fuzzy system showed that it performed worse than the original crisp expert system. A tuning algorithm was then employed to modify the fuzzy model and this process resulted in improved performance to a level comparable to clinicians and superior to the crisp system. Finally, the entire knowledge base was converted to utilise fuzzy logic and this 'integrated' fuzzy expert system was validated against international expert opinion.

Keywords, Expert systems, fuzzy logic, validation, umbilical cord acid-base balance, neonatal outcome

1 Introduction

Childbirth is a stressful experience for both mother and infant. Even during normal labour every infant is being regularly deprived of oxygen as maternal contractions, which increase in frequency and duration throughout labour until delivery, restrict blood supply to the placenta. This oxygen deprivation can lead to fetal 'distress', permanent brain damage and, in the extreme, fetal death. Once an infant has been delivered the attending clinicians must make an immediate assessment of the need for neonatal resuscitation.

In the 1950's Virginia Apgar introduced a scoring system [1] specifically to determine any requirement for neonatal resuscitation. The system, originally intended to be assessed objectively by an independent observer, sums five clinical factors (heart rate, respiratory effort, reflex irritability, muscle tone and skin colour) scored 0, 1 or 2 to give a total between 0 and 10, at 1 minute and 5 minutes after birth. This system has been adopted almost universally in the developed world. However, it is now widely assigned by the attending clinician, is therefore subjective and is often assigned retrospectively. The Apgar score is also affected by factors that existed prior to the onset of labour such as congenital abnormalities of the fetus and events immediately after delivery. Since its introduction, the Apgar score has frequently been misused to evaluate obstetric care or to predict long term neurological outcome.

The need remains to establish an objective, physiologically based, immediate assessment of the health of the newborn infant which can be used to accurately evaluate obstetric care. Such an assessment could be used to guide neonatal care, provide individual feedback to clinicians, audit overall hospital performance, teach inexperienced clinicians and assess the impact of new technologies.

1.1 Umbilical Cord Acid-Base Analysis and The Need for an Expert System

An assessment of neonatal outcome may be obtained from analysis of blood in the umbilical cord of an infant immediately after delivery [19]. The umbilical cord vein carries blood from the placenta to the fetus and the two smaller cord arteries return blood from the fetus. The blood from the placenta has been freshly oxygenated, and has a relatively high partial pressure of oxygen (pO_2) and low partial pressure of carbon dioxide (pCO_2). Oxygen in the blood fuels *aerobic* cell metabolism, with carbon dioxide produced as 'waste'. Thus the blood returning from the fetus has relatively low oxygen and high carbon dioxide content. Some carbon dioxide dissociates to form carbonic acid in the blood, which increases the acidity (lowers the pH). If oxygen supplies are too low, *anaerobic* (without oxygen) metabolism can supplement aerobic metabolism to maintain essential cell function, but this produces lactic acid as 'waste'. This further acidifies the blood, and can indicate serious problems for the fetus.

A sample of blood is taken from each of the blood vessels in the clamped umbilical cord and a blood gas analysis machine measures the pH, pO_2 and pCO_2. A parameter termed *base deficit of extracellular fluid* (BD_{ecf}) can be derived from the pH and pCO_2 parameters [21]. 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. An interpretation is then made based on the pH and BD_{ecf} parameters from both arterial and venous blood.

There are, however, a number of difficulties with such umbilical acid-base analysis. Difficulties in obtaining the samples can result in two samples from the same vessel or mixed samples, whilst blood in the syringe can alter due to exposure to air. Blood gas analysis machines require regular calibration and quality control checks to ensure continuing performance to the manufacturer's specifications, Careful retrospective analysis of the acid-base results obtained during a trial on electronic fetal

monitoring [24] highlighted a 25% failure rate to obtain arterial and venous paired samples with all parameters [23]. This sampling error rate is broadly in line with other studies in which the importance of paired samples was recognised. The study also highlighted the fact that considerable expertise was required to reliably recognise these errors and accurately interpret the results.

To overcome these difficulties an expert system has been developed for the analysis of umbilical cord acid-base data, encapsulating the knowledge of leading obstetricians, neonatologists and physiologists gained over years of acid-base interpretation. The expert system combines knowledge of the errors likely to occur in acid-base measurement, physiological knowledge of plausible results and statistical knowledge of a large database of results. It automatically checks for errors in input parameters, identifies the vessel origin (artery or vein) of the results and provides an interpretation in an objective, consistent and intelligent manner.

The expert system was developed in three main incremental stages. Initially, a crisp expert system was developed incorporating conventional forward-chaining logic [11, 13]. The crisp system underwent an extensive software verification and validation process and has been installed at over twenty five UK hospitals [10]. Next, the existing crisp system was extended by deriving a 'preliminary' fuzzy expert system in which the crisp rules for interpretation of error-free results were converted directly into a fuzzy rule set [12, 6]. This preliminary model was automatically tuned to match expert opinion using an algorithm based on simulated annealing [7]. Finally, the limitations of the preliminary fuzzy expert system were overcome through the creation of an 'integrated' fuzzy expert system in which fresh knowledge elicitation resulted in new fuzzy rule sets for the tasks of identification of vessel origin and interpretation of results [5]. The performance of both aspects of this integrated system was validated in a further comparison with expert opinion [9, 8].

2 The Crisp Expert System

2.1 Development of the Crisp Expert System

The expert system module has two main purposes;

- to validate the results and identify the vessel origin, and
- to interpret the results.

The development of the expert system took place in close collaboration with several clinicians experienced in the interpretation of umbilical cord acid-base data. A database of over 2 000 cord samples had already been collected, which was used to formulate and verify the rules in conjunction with the experts' knowledge of fetal physiology. Frequency distributions of the pH, pCO_2 and pO_2 values were plotted to establish the median values and lower 2.5^{th} centile ranges of each; means and standard deviations cannot be used on the data as all the distributions are skewed and not Normal. Frequency distributions of the differences between vessels were also plotted and were used to establish the minimum allowable differences for arterial-venous paired samples. The populations were checked against other published data to ensure that they were not specific to the local data.

Initially a set of rules was generated by two of the clinicians after a knowledge elicitation session. These rules were then encoded and applied to the database in a variety of ways. Firstly the full results were passed to the expert system and the interpretations recorded. Next each input parameter was marked as containing an error in all combinations and these results were also passed to the expert system. This generated the interpretations of the expert system with successively less information and enabled the internal consistency of the rules to be checked. A number of other techniques, such as passing plausible random numbers to the expert system, were used to examine the behaviour of the expert system rules. The output generated was examined by the clinicians and the rules modified to eliminate inconsistencies and refine interpretations. This process continued iteratively until the rules were deemed acceptable. The expert system was tailor written in the 'C' language and featured a forward-chaining algorithm, suitable for the classification rules for both validation and interpretation. The knowledge representation was organised as a set of frames, with attributes such as pH, pCO_2, BD_{ecf}, validation flags and originating vessel (artery or vein) for each sample.

Each sample's results are passed to the expert system module for error checking. Two classes of errors are detected; *analyser errors* and *physiological errors*. Analyser errors are reported at the time of sample measurement, such as when the electrodes fail to reach a stable reading. Physiological errors are an additional class of error detected specifically by the expert system by examining whether the results are consistent with the range of possibilities for cord blood. For example, there is a strong relationship between the pH and the pCO_2, as shown in Figure 1, where the 99.9% confidence of prediction intervals have been calculated by regression analysis. Analysis of the residuals has shown that variance across the pCO_2 axis is not uniform for all pH, indicating that the prediction intervals are not strictly valid. Hence, the exclusion limits were constructed beyond these intervals, widening with the increased variance in pCO_2 as pH decreases. Results that fall outside these limits — caused, for example, by the presence of non-blood fluid in the sample — are reported as errors. Once validated the vessel origin or the results is determined from a simple rule base governing the minimum differences expected physiologically.

Paired results undergo a further, more sophisticated, stage of validation to ensure that they make 'physiological sense' when viewed together; if this is found not to be the case, an error will be marked against the suspect results. If the pH and pCO_2 values for a sample are accepted as valid, the base deficit of the extracellular fluid (BD_{ecf}) is calculated by equations from Siggaard-Andersen [22]. The pH and BD_{ecf} of both vessels are examined to categorise the results into one of 54 interpretations, ranging from 'normal' to 'severe metabolic acidemia'. An interpretation is performed on single samples as well as paired samples, although the information is very much more limited and the user is advised to retry with a paired sample.

2.2 Evaluation of the Crisp Expert System

The crisp system underwent an extensive software verification and validation process to ensure that it was safe for transfer to clinical use. This expert system does not fall naturally into the traditional classification of either a *decision making system*

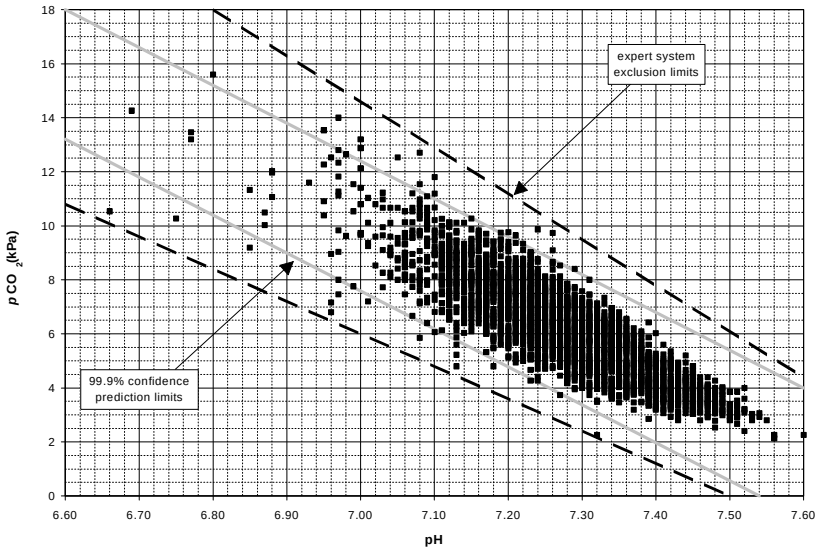

Fig. 1. Scatter diagram of cord blood pH against pCO_2 with 99.9% prediction limits and expert system exclusion limits

or a *decision support system*. 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 — it effectively transforms the four-dimensional numerical input data into a single textual interpretation. Although the system's textual interpretation could be used to guide the provision of neonatal resuscitation or intensive care, it does not recomend a decision for direct clinical intervention. Thus, the requirements of evaluating the performance of the expert system for clinical release were reduced to three objectives:

1. to ensure that the system was safe,
2. to ensure that the interpretations agreed with respected experts, and
3. to demonstrate the potential for economic benefit.

These three objectives were accomplished by carrying out the following tasks:

- *subsystem validation* — subsystem validation 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 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. The few minor problems that this process highlighted were corrected.

- *face validation* — during face validation the expert system's performance was subjectively compared against human expert performance [17]. 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 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' [25].
- *hazard analysis* — during *hazard analysis* a 'black-box' approach is used, in which the behaviour of the system is observed in response to all conceivable external events; this contrasts to the 'white-box' approach of subsystem validation, in which the code itself is examined to anticipate failures. 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. A suite of automatic test procedures was created to simulate communications functions and user inputs. The communication functions of the blood gas analyser were encapsulated in a simulation program, and a second program was created to run the expert system and to simulate user input. A set of bespoke databases were created to drive simultaneously both test programs with sets of key strokes to simulate, sample results to transmit and the target states of the system. The target states comprised the system screen that should be on view, the expected expert system interpretation and the output databases. After all these tests had been completed, and satisfactorily passed, the system was deemed to be functionally correct according to specification.
- *sensitivity analysis* — sensitivity analysis is defined as "systematically changing expert system input variable values and parameters over some range of interest and observing the effect upon system performance" [17]. The test suite described above was utilised to perform a comprehensive sensitivity analysis on the expert system categorisations. 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. Altogether almost 1 000 samples were created to test across the entire range 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 by the experts 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.

- *economic assessment* — 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 [16]. Umbilical cord acid-base assessment has the potential to be of large economic benefit through the reduction in unwarranted negligence litigation. This is a complex issue which cannot be fully addressed here, but it has been shown [5, 10] that the improvement in reliability of results obtained after the introduction of the expert system was enough to warrant its extra running costs.
- *clinical assessment* — the expert system was placed at the local hospital and at a nearby hospital for extensive field trials before release. During these 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. This resulted in a small number of changes to the user interface.

3 The Preliminary Fuzzy Expert System

3.1 Development of the Preliminary Fuzzy Expert System

A number of problems were identified in the implementation of conventional crisp rules used in the initial system. The interpretation rules featured sharp boundary cut-offs which were not representative of real decision making processes and did not employ any form of uncertainty representation in the conclusion to imply a less than certain diagnosis. It was felt that a fuzzy logic based expert system would offer more realistic and acceptable interpretation. The use of fuzzy logic allows for more gradual changes between categories and allows for a representation of certainty in the rule consequence through the ability to fire rules with varying strength dependent on the antecedents. Additionally, fuzzy logic can allow the results to be presented to clinicians in a more natural form. An investigation was performed to convert the crisp expert system directly into a fuzzy expert system. The purposes of this study were:

- to determine the feasibility of converting the existing crisp rules into a set of fuzzy rules, without the necessity of additional expert knowledge elicitation sessions, and
- to investigate whether the fuzzy system would offer any improvement in performance over the crisp system in its interpretation of results.

It was decided to restrict the initial fuzzy expert system only to the interpretation of true paired samples (samples verified as being an arterial and venous pair with error free pH and BD_{ecf} parameters) as these rules represented a self-contained subset of the crisp system. There were 21 such crisp rules operating on four input parameters (pH_A, BD_A, pH_V, BD_V) which needed conversion directly into equivalent fuzzy rules. Examination of the crisp rules showed that each of these four input parameters could be naturally divided into three fuzzy terms, corresponding

to meanings of *low*, *medium* and *high*. These four fuzzy input variables had the position and width of their terms determined by the cut-offs encoded into the crisp rules and were modelled with sigmoid membership functions. Thus, for example, arterial pH fuzzy variable had its transition from low to medium at 7.05 as this value is used throughout the crisp rules. The term-sets for each fuzzy input variable are shown in Figures 2 and 3.

Two output fuzzy variables were used, severity of acidemia (*acidemia*) and duration of acidemia (*duration*). From the crisp rules it was determined that the acidemia variable had five terms in its term-set: *severe*, *moderate*, *significant*, *mild* (*non-significant*) and *none*; and that the duration variable had three terms: *chronic*, *intermediate* and *acute*. These were also modelled with sigmoid membership functions, with the base variable and cross over of each term determined arbitrarily. The term-sets for the fuzzy output variables are shown in Figure 4.

3.2 Evaluation of the Preliminary Fuzzy Expert System

The fuzzy expert system re-analysed some 8 000 true paired samples and the output was compared to the crisp system. The crisp system produced a category in the range 80 to 120 which had been designed to correspond to an ordered scale, such that 80 was the worst outcome (severe metabolic acidaemia) and 120 was the best outcome (normal). It was immediately apparent that the ordering of classifications produced by the initial fuzzy expert system differed from that of the crisp expert system. A test was designed to determine which order of results was the most appropriate by comparing the expert systems to practising clinicians. The clinicians were asked to rank a set of 50 difficult to classify results in order from 'worst' to 'best', in terms of likelihood of the infant having suffered damage during labour. Two clinicians involved in the creation of the rules and four clinicians experienced in the interpretation of cord acid-base results took part in the comparison study. They consisted of one Professor of Physiology, one Consultant, one Senior Registrar, two Clinical Research Fellows and a Senior Midwife. Additionally, the relationship of each system's ordering to the results ordered by Apgar score at 5 minutes and 1 minute was examined. The Spearman Rank Correlation statistic was used to compare the order of results.

The initial results of the agreement with clinicians are shown in Table 1. It can be seen that the average inter-clinician correlation was very high (0.91), indicating that the clinicians agreed with each other very well on the order of results. The crisp expert system correlated reasonably well with the clinicians (0.80), but the performance of the fuzzy expert system ($fuzzy^0$) was significantly worse. The agreement with the Apgar score is shown in Table 2. Given the fact that other clinical factors affect the Apgar score, the precise level of clinicians' agreement was not particularly important, but the fact that there was significant correlation beyond chance indicates that the clinicians' ordering did reflect actual clinical outcome. The important point is that the crisp expert system performed with a level close to the clinicians, but again the fuzzy system performed significantly worse.

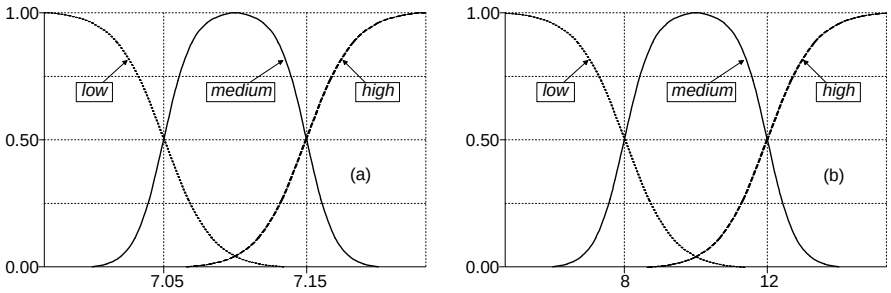

Fig. 2. Term-sets of (a) the arterial pH (pH_A) and (b) the arterial BD_{ecf} (BD_A) fuzzy input variables

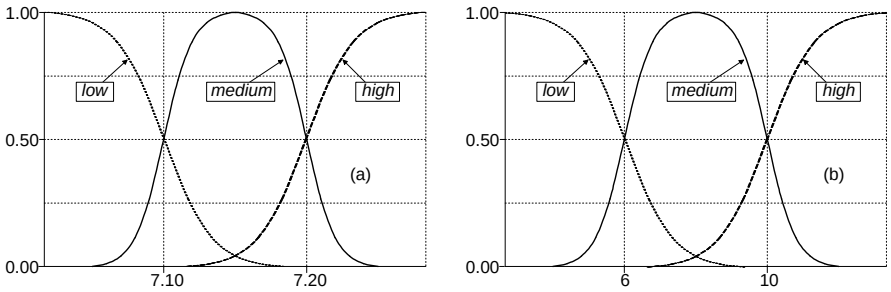

Fig. 3. Term-sets of (a) the venous pH (pH_V) and (b) the venous BD_{ecf} (BD_V) fuzzy input variables

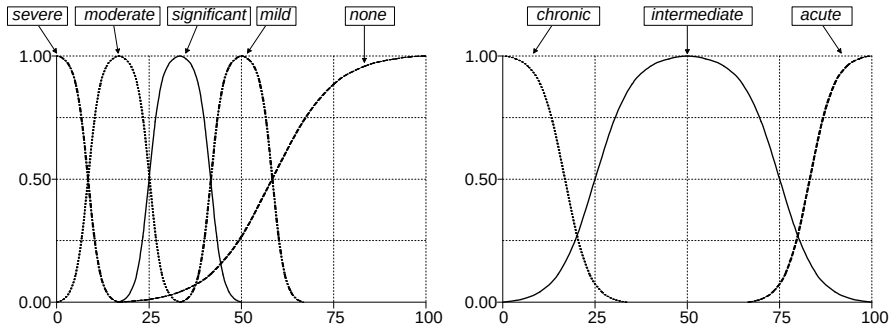

Fig. 4. Term-sets of (a) the *acidemia* and (b) the *duration* fuzzy output variable

Table 1. Results of clinicians', crisp and initial fuzzy expert systems' agreement with clinicians

Agreement	*Corr.* (ρ_s)	*Sig.* (p)
clinicians $\Leftrightarrow$ clinicians	0.91	$\ll 0.001$
crisp system $\Leftrightarrow$ clinicians	0.80	$\ll 0.001$
fuzzy^0 system $\Leftrightarrow$ clinicians	0.67	$\ll 0.001$

Table 2. Results of clinicians', crisp and initial fuzzy expert systems' agreement with outcome (Apgar^5, Apgar^1)

Agreement	*Corr.* (ρ_s)	*Sig.* (p)
clinicians $\Leftrightarrow$ outcome	0.47	$\ll 0.001$
crisp system $\Leftrightarrow$ outcome	0.39	$\ll 0.001$
fuzzy^0 system $\Leftrightarrow$ outcome	0.17	≈ 0.001

3.3 Tuning the Preliminary Fuzzy Expert System

Clearly, the initial performance of the fuzzy expert system was a disappointment. It performed worse that the crisp system both in comparison with expert opinion and in comparison with the Apgar score. The design of any fuzzy model is a complex multi-step process, in which the designer is faced with a large number of parameters such as the type of inference methodology, the number of linguistic variables and their fuzzy terms, the fuzzy rule set, which fuzzy operators to use, the shape and location of membership functions, and the method of defuzzification. Unfortunately, there is currently very little theoretical guidance as to which of the design choices are appropriate for a particular domain. Given the poor performance of the fuzzy system and the large number of alterations that could have been made to improve its performance, an automatic method of tuning these design parameters was required.

In general, the formulation of an optimal fuzzy model in terms of its performance at a given task in the particular domain is a problem of N-dimensional non-linear optimization, in which N is very large even for the most trivial of fuzzy systems. The *simulated annealing* algorithm is a general purpose algorithm for performing approximate optimization in large dimensional problems [15]. It is generally useful for combinatorial optimization problems, and/or problems where derivatives of the cost function being optimized are not available. An adaptation of the *Simulated Annealing* algorithm for continuous minimization by the Simplex method [18] was applied to automatically tune the performance of the preliminary fuzzy model to the clinical expert opinion already obtained [7].

The results obtained for the tuned fuzzy expert system (*fuzzy**) are shown in Table 3. It can be seen that the tuned system achieved an excellent agreement with the clinicians and matched the clinicians in its agreement with outcome. Its modified

performance was better than the crisp expert system and effectively indistinguishable from the clinicians. This result was validated by examining the relative correlation of the expert systems against Apgar scores from the entire database of cases of *true paired* samples with abnormal acid-base status ($n = 383$, defined by the crisp cut-offs $pH_A < 7.05$, $BD_A \geq 12\text{mmol.l}^{-1}$, $pH_V < 7.10$, and $BD_V \geq 10\text{mmol.l}^{-1}$). The results in Table 4 were obtained, in which it can be seen that for these novel abnormal cases, the *fuzzy** system again achieved the highest correlation of all three systems, at a level similar to that for the tuning set (Table 3).

Table 3. Results of the tuned fuzzy expert systems' agreement with clinicians and outcome (Apgar^5, Apgar^1) —compare these results to the pre-tuning results shown in Tables 1 and 2

Agreement	*Corr.* (ρ_s)	*Sig.* (p)
fuzzy* system $\Leftrightarrow$ clinicians	0.93	$\ll 0.001$
fuzzy* system $\Leftrightarrow$ outcome	0.51	$\ll 0.001$

Table 4. Validation of expert systems with outcome (Apgar^5, Apgar^1) for all abnormal cases ($n = 383$)

Agreement	*Corr.* (ρ_s)	*Sig.* (p)
crisp system $\Leftrightarrow$ outcome	0.44	$\ll 0.001$
fuzzy^0 system $\Leftrightarrow$ outcome	0.26	< 0.001
***fuzzy** system $\Leftrightarrow$ outcome**	0.52	$\ll 0.001$

4 The Integrated Fuzzy Expert System

4.1 Development of the Integrated Fuzzy Expert System

Although the preliminary fuzzy expert system was tuned to achieve a high level of agreement with clinicians and to associate with Apgar scores, it was characterised by a number of restrictions:

- *crisp input variables* — the variables to the fuzzy expert system were represented as fuzzy singletons, equivalent to the crisp value obtained from the blood gas analyser;
- *adapted crisp expert system rules* — the rule set was adapted directly from the crisp rule set, rather than from a rule set specifically designed for a fuzzy expert system;
- *restricted rule set* — the rule set was restricted to the 21 crisp rules that dealt with the interpretation of **full paired** results only, i.e. samples which the crisp expert system had previously validated as comprising an error-free pH and BD_{ecf} from both artery and vein; and

- *crisp output variables* — the two output variables *acidemia* and *duration* elicited from the crisp rule set, although represented internally as fuzzy sets, had been centre-of-gravity defuzzified and combined to give a single crisp output.

The integrated fuzzy expert system was enhanced from the preliminary system in several ways. Each of the input variables was fuzzified to have a width which explicitly represented that this input value was an estimate of the 'true' parameter value. A new set of fuzzy rules was developed for both the vessel identification and the interpretation capabilities. Fresh knowledge elicitation sessions were undertaken with the same experts that had developed the crisp rules. Two sets of fuzzy rules were employed; the *vessel identification* rules and the *interpretation* rules. The sample(s) parameters are passed through the vessel identification rules to determine whether they represent an arterial-venous pair. As two samples may both be accidentally obtained from the vein, both from the arteries, one may be mixed arterial-venous, or both may be mixed, a 'safe' vessel identification rule may be that if all parameters differ by more than a specified uncertainty, then the samples can definitely be taken as a true arterial-venous pair. The expected imprecision in each parameter was established through a number of clinical experiments. A fuzzy rule-base was designed to produce the behaviour that if (and only if) all parameters differed by more than these values then the results were labelled as an arterial-venous pair, with smooth transitions between each of the categories.

Once vessel identification has been carried out, the sample(s) are passed through the interpretation rules. The basic principles of acid-base analysis elicited from the experts were that: (i) *acidemia* is based on the absolute value of arterial pH (lower arterial pH implies worse *acidemia*), refined by the value of the venous pH; (ii) *component* is based on arterial BD_{ecf} (high BD_{ecf} implies *metabolic* component, low BD_{ecf} implies *respiratory* component), refined by venous BD_{ecf}; and (iii) *duration* is based on pH and BD_{ecf} differences (smaller differences imply *chronic* duration, larger differences imply *acute* duration), refined by absolute arterial values. These basic principles were encapsulated in the fuzzy rules such that there was smooth transition over all input and output sets. This ensured that, as far as possible, continuous changes in input parameters resulted in continuous changes in the fuzzy output sets. Unknown values were explicitly represented such that results with invalid or missing parameters could be processed by the fuzzy expert system.

Three fuzzy output variables (*acidemia*, *component*, and *duration*) were utilised in rule consequences, with the availability of graphical output of the consequence fuzzy sets. In addition to several alternative numerical representations of uncertainty in the interpretation, linguistic approximation of the fuzzy output variables was also introduced to allow textual output from the fuzzy expert system.

4.2 Validation of the Integrated Fuzzy Expert System

The cases for each task were selected by the independent engineer from the database of over 10 000 results (approximately 400 abnormals), but this provided serious problems. Cases could not be selected from the entire database on a uniform random basis, as this would have resulted in approximately 75% paired arterial-venous

samples, and approximately 98% *normal* interpretations. In essence it was desired to uniformally span the *target* outputs, so that a roughly even spread across the various output sets would have been obtained from the combined experts (and expert system). However, this pre-supposed that the output was known — which it obviously wasn't for the validation study. Other studies [14] have used an in-house expert to select difficult and/or representative cases, but due to the restricted number of experts available this was not feasible. The problem was solved by using the crisp expert system categorisation already obtained on the data to guide the selection of cases. Two sets of fifty cases were randomly selected to roughly span the crisp expert system categorisations. This ensured that a few cases were obtained from a variety of conditions, including results that had parameter errors, results from a single vessel, and results ranging from metabolic acidemia to normal.

The centroids of the integrated fuzzy expert system were combined into a single index by:

$$condition = acidemia + \frac{component}{20} + \frac{duration}{10} \tag{1}$$

where the relative weighting of the three terms was determined empirically. Given that the three output variables are arranged in such a way that low scores indicate a worsening condition for the infant, to the extreme *severe, metabolic, chronic acidemia*, this index can be thought of as indicating the *health* of the infant as represented by its acid-base balance at birth. The experts were again asked to rank fifty cases from 'worst' to 'best', in terms of likelihood that the infant may have suffered damage during labour, on the basis of the acid-base information alone.

The experts were given the two sets of pH and BD_{ecf} parameters from each of fifty cases, and were asked to indicate their opinion of the closest linguistic interpretation for three linguistic variables; *acidemia, component*, and *duration*. For each variable they were instructed to mark *zero*, *one* or *two* terms to indicate the closest match. This was specifically designed to allow the expert to mark two adjacent labels if they felt a result fell in-between two labels, or to mark no label if there was insufficient information, or no label was appropriate.

Spearman rank order correlation [20] can be used to determine the degree of association between two sets of rank-ordered data. This was used to calculate the difference between the expert system's ranking of cases, specified by the index described above, and the experts' ordering. Note that this is effectively the same as minimising the mean square error between the desired rankings and the obtained rankings. To measure the agreement between two expert's linguistic categorisation a measure of (nominal) categorical agreement was required. The kappa statistic [2] was used to measure *exact* agreement between experts and the expert system linguistic outputs and weighted kappa [3] was used for partial agreement.

4.3 Results

The individual inter-expert and expert-*fuzzy*^{2} Spearman rank order correlation coefficients obtained are shown in Table 5. The average inter-expert agreement is calculated by taking the average of each expert against the other *three* experts, and the average *fuzzy*^{2} agreement by taking the average of agreement with all *four* ex-

perts. As can be seen, the fuzzy expert system performed exceptionally well against experts *A*, *B*, and *C*. These three experts had taken place in the previous study, and the average expert system agreement with these three is 0.94 — slightly lower correlation was obtained against expert *D*, although the fuzzy expert system was no worse than the other experts. These results are illustrated in Figure 5, in which each of the expert's rankings are plotted against the fuzzy expert system rankings — perfect agreement would result in a diagonal line from (1,1) to (50,50).

Table 5. Agreement for numeric interpretation by rank order correlation

Expert	A	B	C	D	*fuzzy*^{2}
A	—	0.899	0.888	0.577	0.950
B	0.899	—	0.908	0.701	0.931
C	0.888	0.908	—	0.537	0.925
D	0.577	0.701	0.537	—	0.606
Average	0.788	0.836	0.777	0.605	0.853

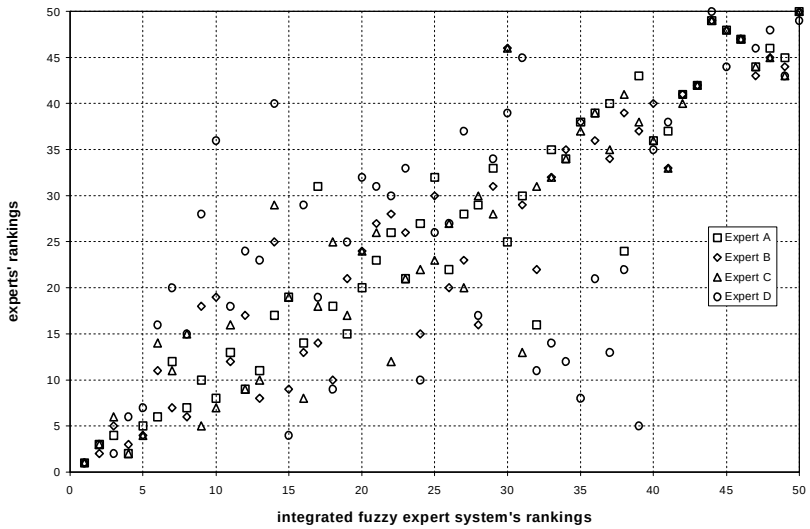

Fig. 5. Graph of four experts rankings against the integrated fuzzy expert system

The results of the linguistic interpretation were investigated by means of comparison of the linguistic output of the *acidemia*, *component* and *duration* variables

with the categorisations of the experts. In all cases, both the inter-expert agreement and the agreements between the fuzzy expert system and the experts were generally found to be relatively low, even for weighted kappa. An attempt was made to investigate the effect of different pH and BD_{ecf} weights on these linguistic agreements, but in general it was found that performance was not significantly increased above the results achieved with default weights.

5 Summary

The successful development of a crisp expert system for umbilical acid-base assessment was the first achievement of this work. The junior, or inexperienced clinician does not, in general, have the knowledge to accurately assess umbilical acid-base information. The crisp expert system assists such a clinician by producing a consistent and reliable interpretation based on rules that embody expert knowledge. The crisp expert system technology has been licensed to a commercial company, and has been placed at over twenty hospitals in the United Kingdom. Expert systems reaching routine clinical use have been rare, largely as a result of difficulties in clinical validation. An expert system that offers advice for clinical intervention, albeit as advice intended for decision support such that the final responsibility for decision remains with the clinician, is expected to be infallible [4]. The design of the crisp expert system as an interpretation support system is believed to have greatly contributed to its swift transfer to clinical use, through the reduction in validation requirements.

Although the crisp expert system represents a major advance to the current clinical assessment of umbilical acid-base, its lack of explicit uncertainty handling is a limitation. The preliminary fuzzy expert system was developed to overcome this limitation through the introduction of explicit uncertainty in the knowledge base, but it was initially found to perform worse than the crisp expert system when compared to human experts. This observation led to the development of the fuzzy model tuning algorithm based on the method of simulated annealing to perform large dimensional function optimisation. This was used to automatically tune the performance of the preliminary fuzzy expert system to match the clinicians.

The final ‘integrated fuzzy expert system’ explicitly represented both imprecision in the input data and uncertainty in the interpretive knowledge base. Modification of the linguistic variables, fuzzy membership functions, and the fuzzy rule base in response to fresh knowledge elicitation sessions resulted in a fuzzy model that performed as well as the previously tuned system, but without the need for any subsequent tuning. This integrated fuzzy expert system also incorporated a fuzzy rule base for performing vessel identification, in addition to the rule base for performing interpretation. The integrated fuzzy expert system was tested in a validation study and was found to perform favourably compared to the human experts.

The achievements of this work can be summarised as having successfully modelled the clinical expert knowledge necessary for the assessment of umbilical acid-base information. The need for basic data validation of acid-base parameters prior to interpretation has been largely accepted clinically. The requirement to interpret the

pH **and** BD_{ecf} from *both* arterial and venous samples in order to reach an accurate assessment has also been accepted, although it is often still not done in practice. The introduction of the crisp expert system is widely regarded as a significant clinical achievement. The subsequent development of the fuzzy expert system, incorporating explicit uncertainty handling, has increased the embedded intelligence within the expert system to a level which is indistinguishable from the best clinical experts. This expert system represents a major step towards the establishment of an objective measure of obstetric care.

References

1. V. Apgar (1953). A proposal for a new method of evaluation of the newborn infant. *Current Researches in Anesthesia and Analgesia*, 32:260–267.
2. J. Cohen (1960). A coefficient of agreement for nominal scales. *Educational Psychological Measurement*, 20:37–46.
3. J. Cohen (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70:213–220.
4. J. Fox (1993). On the soundness and safety of expert systems. *Artificial Intelligence Medicine*, 5:159–179.
5. J.M. Garibaldi (1997). *Intelligent Techniques for Handling Uncertainty in the Assessment of Neonatal Outcome*. PhD thesis, Univeristy of Plymouth.
6. J.M. Garibaldi and E.C. Ifeachor (1996). The comparison of a crisp and fuzzy expert system with practising and expert clinicians. In *Proceedings of the 2nd International Conference on Neural Networks and Expert Systems in Medicine and Healthcare*, pages 229–237, Plymouth, UK.
7. J.M. Garibaldi and E.C. Ifeachor (1999). Application of simulated annealing fuzzy model tuning to umbilical cord acid-base interpretation. *IEEE Transactions Fuzzy Systems*, Accepted for publication.
8. J.M. Garibaldi, J. Tilbury, and E.C. Ifeachor (1999). The design and validation of a fuzzy expert system for umbilical cord acid-base analysis. *In preparation for IEEE Transactions Biomedical Engineering*.
9. J.M. Garibaldi, J. Tilbury, and E.C. Ifeachor (1998). The validation of a fuzzy expert system for umbilical cord acid-base analysis. In *Proceedings of the 3rd International Conference on Neural Networks and Expert Systems in Medicine and Healthcare*, pages 230–240, World Scientific, Singapore.
10. J.M. Garibaldi, J.A. Westgate, and E.C. Ifeachor (1999). The evaluation of an expert system for the analysis of umbilical cord blood. *Artificial Intelligence Medicine*, Accepted for publication.
11. J.M. Garibaldi, J.A. Westgate, E.C. Ifeachor, and K.R. Greene (1994). The development of an expert system for the analysis of umbilical cord blood at delivery. In *Proceedings of the International Conference on Neural Networks and Expert Systems in Medicine and Healthcare*, pages 394–402, Plymouth, UK.

12. J.M. Garibaldi, J.A. Westgate, E.C. Ifeachor, and K.R. Greene (1996). The development of a crisp and fuzzy expert system for the analysis of umbilical cord blood at delivery. In *Proceedings of the 3rd World Congress on Expert Systems*, pages 202–209, Seoul, Korea.
13. J.M. Garibaldi, J.A. Westgate, E.C. Ifeachor, and K.R. Greene (1997). The development and implementation of an expert system for the analysis of umbilical cord blood. *Artificial Intelligence Medicine*, 10:129–144.
14. R.D.F. Keith, S.L. Beckley, J.M. Garibaldi, J.A. Westgate, E.C. Ifeachor, and K.R. Greene (1995). A multicentre comparative study of 17 experts and an intelligent computer system for managing labour using the cardiotocogram. *British Journal Obstetrics Gynaecology*, 102:688–700.
15. S. Kirkpatrick, C.D. Gelatt, and M.P. Vecchi (1983). Optimization by simulated annealing. *Science*, 220:671–680.
16. P.L. Miller and D.F. Sittig (1990). The evaluation of clinical decision support systems: What is necessary versus what is interesting. *Medical Informatics*, 15(3):185–190.
17. R.M. O'Keefe, O. Balci, and E.P. Smith (1987). Validating expert system performance. *IEEE Expert*, 2(4):81–90.
18. W.H. Press, S.A. Teukolsky, W.T. Vettering, and B.P. Flannery (1992). *Numerical Recipes in C: The Art of Scientific Computing (2nd edn)*. Cambridge University Press, Cambridge.
19. Royal College of Obstetricians and Gynaecologists (1993). Recommendations Arising from the 26th RCOG Study Group. In J.A.D. Spencer and R.H.T. Ward, editors, *Intrapartum Fetal Surveillance*, page 392. RCOG Press, London.
20. S. Siegel and N.J. Castellan (1988). *Nonparametric Statistics for the Behavioural Sciences (2nd edn)*. McGraw-Hill, New York, 1988.
21. O. Siggaard-Andersen (1971). An acid-base chart for arterial blood with normal and pathophysiological reference areas. *Scandanavian Journal Clinical Laboratory Investigation*, 27:239–245.
22. O. Siggaard-Andersen (1976). *The Acid-Base Status of the Blood (4th edn)*. Munksgaard, Copenhagen.
23. J.A. Westgate, J.M. Garibaldi, and K.R. Greene (1994). Umbilical cord blood gas analysis at delivery: A time for quality data. *British Journal Obstetrics Gynaecology*, 101:1054–1063.
24. J.A. Westgate, M. Harris, J.S.H. Curnow, and K.R. Greene (1993). Plymouth randomised trial of cardiotocogram only versus ST waveform plus cardiotocogram for intrapartum monitoring in 2400 cases. *American Journal Obstetrics Gynecology*, 169:1151–1160.
25. J. Wyatt and D. Spiegelhalter (1990). Evaluating medical expert systems: What to test and how? *Medical Informatics*, 15(3):205–217.