

# THE VALIDATION OF A FUZZY EXPERT SYSTEM FOR UMBILICAL CORD ACID-BASE ANALYSIS

J.M. GARIBALDI, J. TILBURY, E.C. IFEACHOR \*

2-4<sup>th</sup> September 1998

*School of Electronic, Communication and Electrical Engineering  
University of Plymouth, Drake Circus, Plymouth, PL4 8AA, UK*

## **Abstract**

Objective assessment of the neonatal outcome of labour is important, but it is a difficult and challenging problem. It is an invaluable source of information which can be used to provide feedback to clinicians, to audit a unit's overall performance, and can guide subsequent neonatal care. Current methods are inadequate as they fail to distinguish damage that occurred during labour from damage that occurred before or after labour. Analysis of the chemical acid-base status of blood taken from the umbilical cord of an infant immediately after delivery provides information on any damage suffered by the infant due to lack of oxygen during labour. However, this process is complex and error prone, and requires expertise which is not always available on labour wards. Previous work had resulted in the development of a crisp expert system for the interpretation of umbilical acid-base status. However, this domain is characterised by uncertainty in both the basic data and the knowledge required for its interpretation. Fuzzy logic provides a technique for representing both these forms of uncertainty in a single framework. Experimental work to establish the imprecision in acid-base parameters, in conjunction with fresh knowledge elicitation sessions, allowed the creation of an expert system to validate and interpret acid-base data using fuzzy logic. The performance of the system was evaluated in a rigorous validation study. This demonstrated excellent agreement with the experts for the numeric outputs, and agreement on a par with the experts for the linguistic outputs. One output of the fuzzy expert system is a novel single dimensional measure that accurately represents the severity of acid-base results.

---

\*jong@cis.plym.ac.uk, julian@cis.plym.ac.uk, e.ifeachor@plymouth.ac.uk

## 1 Introduction

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). Samples of blood may be taken from blood vessels in 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 [1]. 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 status of umbilical cord blood has been recommended by the British Royal College of Obstetricians and Gynaecologists [2].

There are, however, a number of difficulties with the procedure. 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 internal calibration and external quality control checks to ensure continuing accuracy and precision to the manufacturer’s specifications. During a trial on ST-waveform monitoring in Plymouth [3], 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 [4]. This sampling error rate is broadly in line with other studies in which the importance of paired samples was recognised.

A model of clinical expertise required for the accurate interpretation of umbilical acid-base status was developed, and encapsulated in a rule-based expert system [5]. This expert system checks results to ensure their consistency, identifies whether the results come from arterial or venous vessels, and then produces an interpretation of their meaning. This ‘crisp’ expert system was validated, verified and commercially released, and has since been installed at twenty two hospitals all around the United Kingdom [6].

A number of problems were identified in the implementation of conventional crisp rules used in the initial system. The interpretation section of the crisp expert system utilised a number of rules of a form similar to:

IF arterial pH  $< 7.05$  AND arterial  $BD_{ecf} \geq 12 \text{ mmol.l}^{-1}$   
THEN severe arterial metabolic acidemia

Such rules feature sharp boundary cut-offs which are not representative of real decision making processes and do not employ any form of uncertainty representation in the conclusion to imply a less than certain diagnosis.

Fuzzy logic and fuzzy set theory provide a good framework for managing uncertainty and imprecision in medicine [7, 8, 9] and have been successfully applied to a number of areas [10, 8, 11]. It was felt that a fuzzy logic based expert system would offer more realistic and acceptable interpretation. In a fuzzy system, a rule such as above may be replaced by:

IF arterial pH is low AND arterial  $BD_{ef}$  is high  
THEN arterial acidemia is metabolic

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.

A preliminary investigation was performed to convert the crisp expert system directly into a fuzzy expert system (FES) [12] and it was found that, after tuning, this fuzzy system improved the performance of the crisp expert system to a level effectively indistinguishable from the clinical experts [13]. However, although this preliminary fuzzy expert system utilised a set of fuzzy rules to perform the interpretation, it was characterised by a number of restrictions. It functioned with crisp input and output variables with no indication of imprecision, and it *only* interpreted results that had been already validated by the crisp expert system as comprising an error-free arterial-venous pair. An ‘integrated’ fuzzy expert system was then developed, with knowledge gained from the preliminary fuzzy expert system, to validate and interpret *all* acid-base results. When the development of the integrated system was complete, a comprehensive validation process was undertaken to re-evaluate the numeric and linguistic outputs of both the numeric and linguistic interpretations of the system. This paper presents the results of this validation study, and discusses their implications to the further development of this expert system.

## 2 Development of the Fuzzy Expert System

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. Once vessel identification has been carried out, the sample(s) are passed through the interpretation rules. The Mamdani model of infer-

ence was used, with the min operator used for implication. Probabilistic operators were used for *and* and *or* as, in elicitation sessions with experts, the fuzzy output sets produced with the probabilistic operators were favoured as they avoided the ‘plateau’s produced with the standard max, min operators. It was found that the probabilistic family generated smoother transition surfaces for the vessel identification rules and produced higher performance for the interpretation rules. Fuzzy sets were modelled with *sigmoid* membership functions, details of which will be supplied in the full paper. Centre-of-gravity (centroid) defuzzification was performed on the fuzzy output variables to produce numeric outputs, and linguistic approximation was performed to produce linguistic output.

## 2.1 Vessel Identification Rules

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 iff 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.

## 2.2 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 [14]. This ensured that, as far as possible, continuous changes in input parameters resulted in continuous changes in the fuzzy output sets.

## 3 Validation of the Fuzzy Expert System

The cases for each task were selected by the independent engineer from the database of over 10 000 results (approximately 400 abnormal), 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 uniformly 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 [15] 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.

### 3.1 Numeric Interpretation

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

$$condition = acidemia + \frac{component}{20} + \frac{duration}{10} \quad (1)$$

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 asked to rank fifty cases from 'worst' to 'best', in terms of likelihood that the infant may have suffered intrapartum asphyxial damage, on the basis of the acid-base information alone.

### 3.2 Linguistic Interpretation

The experts were given the two sets of pH and  $BD_{ef}$  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.

### 3.3 Statistical Methods

Spearman rank order correlation [16] 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  $\chi^2$  statistic can be used to measure the degree of *association* between two categorical variables, but this statistic makes no distinction between departure from chance association due to *agreement* or *disagreement*. In 1960, Cohen introduced a measure of *agreement* between two categorical variables termed the *kappa* coefficient [17]. This plain kappa statistic measured only *exact* agreement and, to overcome this problem, Cohen later introduced *weighted kappa* to allow for partial agreement [18]. Plain and weighted kappa were used to calculate the degree of agreement between experts and the expert system linguistic outputs.

## 4 Results

### 4.1 Results of Numeric Interpretation

The individual inter-expert and expert-*fuzzy*<sup>2</sup> Spearman rank order correlation coefficients obtained are shown in Table 1. The average inter-expert agreement is calculated by taking the average of each expert against the other *three* experts, and the average *fuzzy*<sup>2</sup> agreement by taking the average of agreement with all *four* experts. 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 4.1, 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).

### 4.2 Results of Linguistic Interpretation

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. This is illustrated in Tables 2 and 3, showing the agreement matrix of plain and weighted kappa for *acidemia* (bold figures

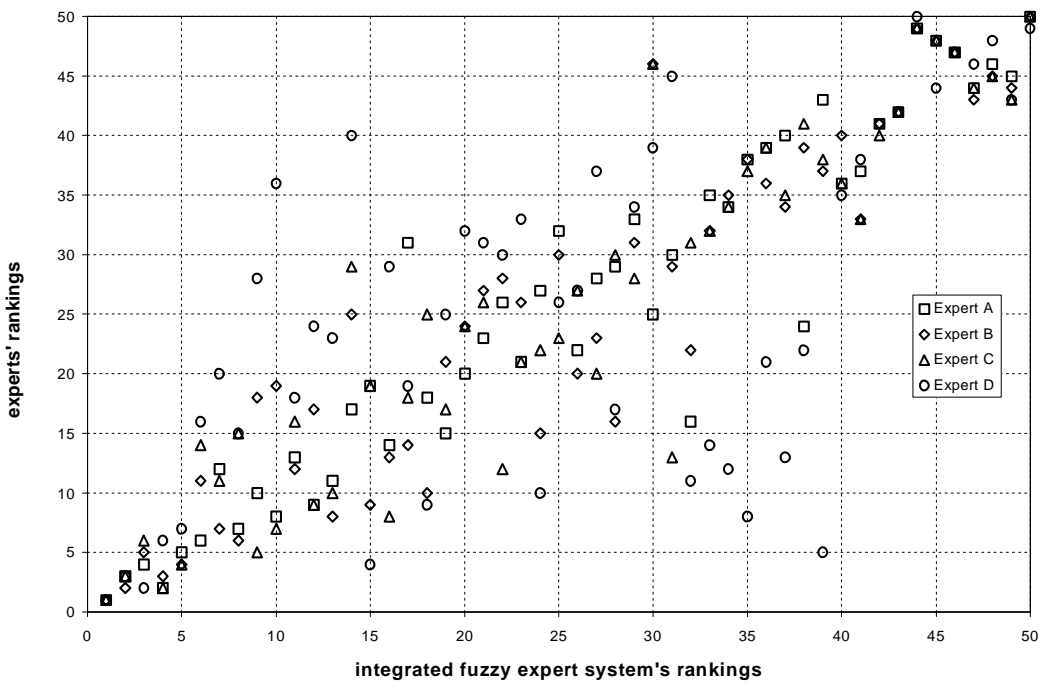


Figure 1: graph of four experts rankings against the integrated fuzzy expert system

indicate kappa above the 0.1% significance level). 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 Discussion, Future Work and Conclusions

The correlations obtained for the numeric interpretation (ranking) task were extremely encouraging. Excellent correlations were obtained against experts *A*, *B* and *C*, and the fuzzy expert system performed better against expert *D* than experts *A* and *C* did. Overall the expert system achieved higher average correlation than any of the experts. Although experts *A*, *B* and *C* were involved in the previous ranking task, it is still an excellent achievement to obtain an average correlation of 0.94 with these three experts. This average correlation achieved on the fresh validation data set was the same as that achieved on the previous data set used in development — a result which imparts a high degree of confidence in the fuzzy model. Whilst the agreement with expert *D* is less good, it must be noted that this expert expressed the opinion that *only* the  $BD_{ecf}$  variable was important in considering the likelihood that the infant may have suffered intrapartum asphyxial damage. This is in marked contrast to the opinions of experts *A*, *B*, and *C*.

However, in contrast, the validation results of the linguistic fuzzy outputs was generally poor. Probably the most positive statement that can be made about this finding was that the fuzzy expert system was on a par with the experts. It is not a simple case that the experts agreed, and the fuzzy expert system didn't — it is more a case that neither the experts nor the system obtained good agreements. This may be due to the 'artificial' nature of the linguistic validation task, in that experts were forced to categorise results by labels, and were unable to express a fuller clinical opinion of the meaning of the data. It may leave scope for enhancing the fuzzy linguistic approximation, but this is unlikely to achieve much unless higher inter-expert agreement could be achieved through a different design of validation task.

The fuzzy expert system presented here, whilst a significant development, still needs to be validated more thoroughly against other clinical data. Unfortunately, suitable *other clinical data* is very hard to obtain. It is planned that a large randomised trial will take place in Scandinavia from 1998 in which a new version of the STAN2 monitor [19] will be evaluated. 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. It is hoped that this data could be used to validate both the fuzzy expert system described here and a fuzzy-logic based expert system for ST waveform analysis [20], The development

Table 1: agreement for numeric interpretation by rank order correlation

<i>Expert</i>	A	B	C	D	<i>fuzzy</i> <sup>2</sup>
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
<i>Average</i>	0.788	0.836	0.777	0.605	0.853

Table 2: inter-expert agreement and expert-*fuzzy*<sup>2</sup> agreement for linguistic *acidemia* interpretation calculated by plain kappa (bold figures indicate kappa beyond 0.1% significance)

<i>Expert</i>	A	B	C	D	<i>fuzzy</i> <sup>2</sup>
A	—	<b>0.642</b>	<b>0.224</b>	0.208	<b>0.671</b>
B	<b>0.642</b>	—	0.175	<b>0.312</b>	<b>0.567</b>
C	<b>0.224</b>	0.175	—	<b>0.253</b>	<b>0.271</b>
D	0.208	<b>0.312</b>	<b>0.253</b>	—	0.124
<i>Average</i>	0.358	0.376	0.217	0.258	0.408

Table 3: inter-expert agreement and expert-*fuzzy*<sup>2</sup> agreement for linguistic *acidemia* interpretation calculated by weighted kappa (bold figures indicate kappa beyond 0.1% significance)

<i>Expert</i>	A	B	C	D	<i>fuzzy</i> <sup>2</sup>
A	—	<b>0.750</b>	<b>0.321</b>	<b>0.340</b>	<b>0.786</b>
B	<b>0.750</b>	—	0.246	<b>0.436</b>	<b>0.633</b>
C	<b>0.321</b>	0.246	—	<b>0.345</b>	<b>0.390</b>
D	<b>0.340</b>	<b>0.436</b>	<b>0.345</b>	—	0.238
<i>Average</i>	0.470	0.477	0.304	0.373	0.512

of the umbilical acid-base fuzzy expert system represents a major achievement and constitutes a significant contribution to the assessment of neonatal outcome.

## Acknowledgements

The authors acknowledge the clinical support of Keith Greene. We are indebted to the experts who took part in the comparison study: Dr Jenny Westgate, Professor Karl Rosén, Mrs Maureen Harris and Professor James Low. Thanks also to all the staff at Derriford Hospital, especially the midwives and auxiliaries, for their cooperation and assistance in the implementation of cord blood sampling. This work was supported by the EPSRC. The crisp expert system software is available as *Expert DataCare* through **Chiron Diagnostics Ltd**, Halstead, Essex, UK.

## References

### References

- [1] O. Siggaard-Andersen. An acid-base chart for arterial blood with normal and pathophysiological reference areas. *Scandinavian Journal Clinical Laboratory Investigation*, 27:239–245, 1971.
- [2] Royal College of Obstetricians and Gynaecologists. 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, 1993.
- [3] J.A. Westgate, M. Harris, J.S.H. Curnow, and K.R. Greene. Plymouth randomised trial of cardiotocogram only versus ST waveform plus cardiotocogram for intrapartum monitoring in 2400 cases. *American Journal Obstetrics Gynecology*, 169:1151–1160, 1993.
- [4] J.A. Westgate, J.M. Garibaldi, and K.R. Greene. Umbilical cord blood gas analysis at delivery: A time for quality data. *British Journal of Obstetrics and Gynaecology*, 101:1054–1063, 1994.
- [5] J.M. Garibaldi, J.A. Westgate, E.C. Ifeachor, and K.R. Greene. 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, 1994.
- [6] J.M. Garibaldi, J.A. Westgate, E.C. Ifeachor, and K.R. Greene. The development and implementation of an expert system for the analysis of umbilical cord blood. *Artificial Intelligence in Medicine*, 10(2):129–144, 1997.

- [7] K.P. Adlassnig. Fuzzy set theory in medical diagnosis. *IEEE Transactions on Systems, Man, and Cybernetics*, 16(2):260–265, 1986.
- [8] M.E. Cohen and D.L. Hudson. The use of fuzzy variables in medical decision making. In M.M. Gupta and T. Yamakawa, editors, *Fuzzy Computing*, pages 263–271. Elsevier Science, North Holland, 1988.
- [9] M. Halim, K.M. Ho, and A. Liu. Fuzzy logic for medical expert systems. *Annals Academy Medicine*, 19(5):672–683, 1990.
- [10] K.P. Adlassnig and G. Kolarz. CADIAC-2: Computer-assisted medical diagnosis using fuzzy subsets. In MM Gupta and E Sanchez, editors, *Approximate Reasoning in Decision Analysis*, pages 219–247. Noth-Holland, New York, 1982.
- [11] H. Watanabe, W.J. Yakowenko, Y.M. Kim, J. Anbe, and T. Tobi. Application of a fuzzy discrimination analysis for diagnosis of valvular heart disease. *IEEE Transactions on Fuzzy Systems*, 2(4):267–276, 1994.
- [12] J.M. Garibaldi and E.C. Ifeachor. 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, 1996.
- [13] J.M. Garibaldi and E.C. Ifeachor. Application of simulated annealing fuzzy model tuning to umbilical cord acid-base interpretation. *IEEE Transactions on Fuzzy Systems*, 7(1):72–84, 1999.
- [14] J.M. Garibaldi. *Intelligent Techniques for Handling Uncertainty in the Assessment of Neonatal Outcome*. PhD thesis, Univeristy of Plymouth, 1997.
- [15] R.D.F. Keith, S.L. Beckley, J.M. Garibaldi, J.A. Westgate, E.C. Ifeachor, and K.R. Greene. A multicentre comparative study of 17 experts and an intelligent computer system for managing labour using the cardiotocogram. *British Journal of Obstetrics and Gynaecology*, 102:688–700, 1995.
- [16] S. Siegel and N.J. Castellan. *Nonparametric Statistics for the Behavioural Sciences (2nd edn)*. McGraw-Hill, New York, 1988.
- [17] J. Cohen. A coefficient of agreement for nominal scales. *Educational Psychological Measurement*, 20:37–46, 1960.
- [18] J. Cohen. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70:213–220, 1968.
- [19] K.G. Rosén and R. Luzietti. The fetal electrocardiogram: ST waveform analysis during labour. *Journal Perinatal Medicine*, 22:501–512, 1994.
- [20] E.C. Ifeachor and N.J. Outram. A fuzzy expert system to assist in the management of labour. In *Proceedings of the International ICSC Symposium on Fuzzy Logic*, pages 97–102. Zurich, Siwtzerland, 1995.