Pergamon

Expert Systems With Applications, Vol. 11, No. 4, pp. 537–541, 1996
Copyright © 1996 Elsevier Science Ltd
Printed in Great Britain. All rights reserved
0957-4174/96 $15.00+0.00

PII: S0957-4174(96)00069-3

Multicentre Validation of an Intelligent System for Managing Labour

R. D. F. Keith, S. Beckley, J. M. Garibaldi†, J. A. Westgate, E. C. Ifeachor and K. R. Greene

Plymouth Perinatal Research Group, Postgraduate Medical School, University of Plymouth, Plymouth, UK

Abstract—*Fetal condition during labour is inferred from a continuous visual recording of the fetal heart rate and uterine contractions, called the cardiotocogram (CTG). Difficulties in CTG interpretation is a major problem that can lead to both unnecessary Caesarean sections and damage to the infant. Our group has developed an intelligent system which applies expert knowledge to support clinical decision making during labour. The system was previously found to obtain a performance comparable to local experts in two internal evaluations. This study presents a validation of the system which compared its management with 17 experts from 16 leading centres within the U.K. in 50 cases selected from a database of 2400 high risk labours. This study found that the majority of experts agreed well and were largely consistent in their management of the cases. The system obtained a performance that was indistinguishable from the experts. Copyright © 1996 Elsevier Science Ltd*

INTRODUCTION

An assessment of fetal condition during labour is inferred from a continuous visual recording of the fetal heart rate and uterine contractions, called the cardiotocogram (CTG). Considerable expertise is required to interpret the complex changes seen on the recording to accurately identify the fetus coping appropriately with the stress of labour from the truly compromised fetus and to decide when to seek further information or intervene. Difficulties in CTG interpretation lead to unnecessary intervention (Haverkamp et al., 1979) as well as a failure to intervene when necessary (Vincent et al., 1991). The financial burden of this imprecision is considerable; evidence obtained from a leading teaching hospital suggests that almost one third of Caesarean section (CS) deliveries were unnecessary (Barrett et al., 1990) each of which cost £1500 (Clark et al., 1991). On the other hand, "birth asphyxia" U.K. litigation settlements approach a total average cost of £1 million each (compensation+legal costs) (Leigh, 1993) and it has been estimated that there are probably 350 cases annually.

Consequently, many question the value of continuous fetal heart rate monitoring (Neilson, 1993). However, it is important to remember that the machines used to obtain the CTG are recorders and not monitors. The monitor is the clinician who interprets the recording. This is a very important distinction because the effectiveness of fetal monitoring is limited not only by the ability of the variables to reflect fetal condition but also by the ability of the clinical staff to interpret them.

Over the past 15–20 years, computer systems have been developed to attempt to provide quantitative heart rate analysis to overcome the inconsistencies of visual interpretation (Dawes et al., 1991; Krause, 1990; Maeda, 1990), but these systems have not proved successful. This is because the correct assessment of fetal condition and appropriate management depends not only on heart rate changes but also requires physiological knowledge, fetal blood sampling, knowledge of the specific patient and the dynamics of that individual labour (Recommendations, 1993) as well as expert judgement which is based on knowledge and experience.

Since 1989, our group has been developing a prototype intelligent system which applies captured expert knowledge to manage labour. The system classifies the same features from the CTG as experienced clinicians using numerical algorithms and a small neural network. This approach has been shown to obtain a comparable performance with experts (Keith et al., 1994a,b). The CTG information, together with the patient information (obstetric history, risk factors etc) and labour events (administration of drugs, epidural insertion, etc.) are collectively passed to an expert system for processing. The expert system interprets this combined data using a database of over 400 rules which are used to recommend action. Importantly, as the knowledge is rule based, it

† Also at: School of Electronics, Communication and Electrical Engineering, University of Plymouth, Plymouth, UK.

allows the system to explain the reasoning which led it to recommend a certain action. In this way, the clinician is not expected to blindly follow the system's recommendations but can reach an informed judgement in the same way they might by discussing the case with an experienced informed colleague. This approach will assist with knowledge transfer (teaching) and also anticipates the legal implications of the system because the final responsibility will remain, where it belongs, with the clinician (midwife or doctor).

The system has been evaluated in two internal studies which found that it was able to attain a performance comparable to local clinicians for a small number of cases (Ifeachor et al., 1991; Keith et al., 1994a,b). The next stage was to assess the system's performance more rigorously by comparing its management with nominated experts from multicentres in a larger number of cases.

METHOD

The Head of Department in 16 U.K. centres was contacted by letter and invited to nominate an individual within their unit regularly involved in the interpretation of the CTG and the management of labour who was regarded as expert.

The case notes and CTGs were obtained for 50 cases selected from a database of 2400 high risk labours in which the condition of the infant at birth had been measured by cord blood gas analysis and Apgar scores. None of the cases had previously been reviewed by the system. The number of cases chosen was considered to be the maximum that an expert could review in a single day. The cases included a wide range of outcomes, from the birth asphyxiated to the very normal.

A synopsis of each patient's obstetric history was compiled from the case notes. The CTGs (annotations removed) were photocopied and reformed to obtain a continuous recording. Every 15 minutes of recording was marked and numbered. All the relevant clinical details (e.g. administration of drugs, presence of meconium liquor and results of vaginal examinations) were written on the trace within the appropriate time markers.

Each expert reviewed the 50 cases twice, at least one month apart under the supervision of a referee sent from our unit. A total of four referees were used. All were extensively briefed on the protocol to ensure that each review took place under similar conditions. The reviews were independent and blind to outcome. The reviewers were presented with the written history of each case and then the referee pulled the CTG from a sealed opaque box in 15 minute segments. This prevented the reviewer from seeing future segments before the current segment had been commented upon and also prevented them from estimating the length of the CTG which would have provided a clue to the length of the labour. The reviewers scored each segment according to the concern they had for the fetus and the management they considered most appropriate at that point in time by considering all past events (Table 1). The aim for the experts, as in the clinical situation, was to achieve minimal intervention without jeopardizing the safety of the fetus.

The review of each case was concluded when either the CTG ended or a score was reached which permitted delivery (five in the first stage and five or three in the second stage). If an expert requested an FBS (score 3) then the referee provided a result secretly read from an FBS graph. Similarly, a measure of cervical dilatation was given if a reviewer wished to check on progress. Scoring sheets were provided which included space for the experts to note any comments they wished to make, for example why they had recommended a particular score or what remedial action they would have taken.

The System's Review

A method was developed to extract the fetal heart rate and uterine contraction data from the paper CTGs using a Hewlett–Packard scanner (Scanjet IIc) in a format useable by the system. Prior to the study, each node in the system's knowledge tree was coded with the appropriate protocol score and was displayed for each 15 minute segment of recording. This allowed the system's recommendations to be directly compared to those of the experts.

The system reviewed each case twice, with a different operator for each review. The role of the operator was to provide the additional case information and estimates of FBS results when requested by the system. In the first

TABLE 1
The Protocol

Score	Comment
1	I am not concerned for this fetus
2	I have concerns for this fetus, but they are not sufficient to request an FBS. I may take some remedial action
3	I am sufficiently concerned to request an FBS or, if possible, a simple vaginal delivery
4	The information I have leads me to be seriously concerned for this fetus. I am not going to recommend immediate delivery although I am thinking of delivery and will do so if things deteriorate further
5	I am so concerned for this fetus that I want immediate delivery.

TABLE 2
Analysis of Agreement

Reviewer	Agreement (%)	Kappa
A	71.34	0.39
B	62.93	0.22
C	72.33	0.42
D	73.37	0.44
E	73.56	0.44
F	68.90	0.34
G	72.03	0.41
H	67.00	0.30
I	74.27	0.46
J	70.63	0.38
K	72.82	0.43
L	73.65	0.44
M	68.04	0.32
N	71.21	0.39
O	71.71	0.40
P	71.31	0.39
Q	58.17	0.12
System	67.33	0.31

review the operator was an engineer who understood obstetric terminology but had no labour ward experience. In the second review, the operator was an obstetrician not previously involved in the development of the system. The reviews were independent and the operators had no knowledge of each other's results.

The scores recorded by the experts and the system for each case were entered in a spreadsheet. All entries were double checked by two workers. The experts were coded, A to Q inclusive and the system was coded, S, with each of the two reviews identified as 1 and 2.

A method was derived to measure the agreement between any two reviewer's recorded sequences of scores for a given case. This measure obtained a value of 1 for perfect agreement and 0 when there was no similarity (Keith, 1993). The agreement was calculated for all pairs of reviews and were formed in an agreement table for each case.

RESULTS

Analysis of Scores

Each reviewers' average agreement (inter-agreement) with the other experts was calculated and expressed as kappa values (Landis & Koch, 1977) (Table 2). These were used to test the null hypothesis that neither the system nor the experts achieved an agreement significantly better than that expected by chance.

The experts' and system's results shown in Table 2 all reached significance which confirms that they agreed better than chance.

Analysis of Decisions to Obtain Delivery by Caesarean Section

There were 31 cases in which at least one expert or the system recommended immediate operative intervention by Caesarean section (CS).

Agreement in the Cases Recommended for Caesarean Sections. The measure of agreement for a reviewer's decision to deliver by CS was taken as the sum of other expert reviews also making the same recommendation. This number was averaged over all the cases in which the particular reviewer recommended a CS in either review. The maximum score an expert could score was 32 which would be obtained if all other 16 experts had all recommended delivery by CS in both their two reviews (32). The maximum score the system could obtain was 34, that is if all 17 experts recommended delivery by CS in both reviews (34). These figures are given in Table 3.

Agreement in the Timing of Caesarean Sections. In the 11 cases where the system recommended CS, on average 18/34 (52.9%) of the expert reviews also recommended CS within 15 minutes of the system. An average of 23/34 (67.6%) did so within 30 minutes and 28/34 (82.4%) recommended CS within 60 minutes of the system.

For expert L who obtained the best timing agreement, on average 18/32 (56.3%) of all other expert reviews also recommended CS within 15 minutes of L's decisions. An average of 24/32 (75.0%) did so within 30 minutes and 28/32 (87.5%) did so within 60 minutes.

For expert H who obtained the least timing agreement, on average 12/32 (37.5%) of all other expert reviews also recommended CS within 15 minutes of H's decisions. An average of 17/32 (53.1%) did so within 30 minutes and 24/32 (75.0%) did so within 60 minutes.

Intervention in Cases with Poor Perinatal Outcome

Three categories of poor outcome were considered.

Birth Asphyxia. This group was characterized by cord arterial pH<7.05 and BDecf≥12 and Apgar score at 5 minutes≤7 with neonatal morbidity.

Severe Metabolic Acidosis. This group was characterized by cord arterial pH<7.05 and BDecf≥12 and Apgar score at 5 minutes>7 and no neonatal morbidity.

Acidosis. Characterized by cord pH<7.05 but BDecf<12 with no neonatal morbidity. These cases have acidosis but without a significant metabolic component.

The experts' and the system's management was considered for each group (Table 4).

Intervention in Cases with Good Clinical Outcome

There were 11 cases which obtained good outcome (cord artery pH>7.15, vein pH7.20, 5 minute Apgar score≥9 and no resuscitation) after a normal vaginal delivery. Operative intervention in these cases was unnecessary. Experts P and M and the system obtained the best performance in these cases by recommending no unnecessary intervention. Experts B and Q obtained the worst performance by recommending two unnecessary Caesarean sections.

TABLE 3
Agreement in Decisions to Deliver by CS

Reviewer	Total No. CS	Agreement
		Average No. expert reviews in agreement / max score possible
A	21	24/32 (75.0%)
B	27	20/32 (62.5%)
C	20	26/32 (81.3%)
D	20	26/32 (81.3%)
E	22	25/32 (78.1%)
F	19	26/32 (81.3%)
G	19	27/32 (84.4%)
H	25	21/32 (65.6%)
I	17	29/32 (90.6%)
J	21	25/32 (78.1%)
K	19	27/32 (84.4%)
L	19	27/32 (84.4%)
M	13	30/32 (93.8%)
N	17	28/32 (87.5%)
O	20	25/32 (78.1%)
P	18	29/32 (90.6%)
Q	20	22/32 (68.8%)
System	11	31/34 (91.2%)

Summary of the System's Performance

This study found that the system:

(1) Recommended *no unnecessary intervention* in cases born in good condition (spontaneous vaginal delivery and umbilical artery pH>7.20) which was better than all but two of the experts.

(2) Recommended delivery by CS in 11 cases. At least 15 of the 17 experts also recommended CS delivery in these cases. The majority did so within 15 minutes of the system and two thirds did so within 30 minutes.

(3) Identified as many of the birth asphyxiated cases as the majority of experts and one more than acted upon clinically.

(4) Agreed with experts well and significantly better than chance (67.33%, kappa=0.31, $P \ll 0.001$).

TABLE 4
Operative Intervention in the Cases with Poor Outcome

	Operative interventions in cases with **birth asphyxia** (max 3)		Operative intervention in cases with **severe metabolic acidosis** (max 4)		Operative intervention in cases with **acidosis** but without a significant metabolic component (max 5)	
Experts	Review 1	Review 2	Review 1	Review 2	Review 1	Review 2
A	2	1	3	4	1	2
B	3	3	3	4	3	3
C	2	2	4	4	4	5
D	2	2	4	4	5	4
E	2	2	4	4	5	4
F	2	2	2	4	4	4
G	2	2	4	4	5	4
H	3	3	4	4	5	4
I	2	2	3	4	3	4
J	2	2	3	4	5	3
K	2	1	4	3	3	2
L	3	2	4	4	5	5
M	2	2	3	3	2	4
N	2	2	4	3	4	4
O	2	2	4	4	5	5
P	2	2	3	2	3	4
Q	2	2	3	3	2	4
System	2	2	2	2	2	2
Clinical	1		2		2	

DISCUSSION AND CONCLUSIONS

The system's performance was indistinguishable from the experts' but it was more consistent even when used by an engineer with no previous labour ward experience. This clearly demonstrates, on these 50 cases, the potential for the system to improve patient care. However, a review of the system's performance found that in two cases (X,Y) the system did not recommend intervention as swiftly as the experts with signs of fetal compromise during the second stage of labour. In these cases the system assigned a validation score of 4 to indicate it was considering delivery whereas the experts tended to recommended a forceps delivery. In a further two cases (A,B) the system did not interpret a rapid fall in baseline heart rate from 160–180 bpm to 90–110 bpm as a severely abnormal event which led most experts to recommend delivery.

No action was recommended in these cases in clinical practice and all had a spontaneous vaginal delivery. However, two cases were delivered with a respiratory acidosis (X,B) and two cases had a significant metabolic acidosis (Y,A). It would be necessary to modify two rules in the system for it to obtain the management of the majority of the experts in these four cases without changing its management in any other case. These modifications would be simple to implement and would enable the system to obtain an overall performance as good as the best of the experts.

The challenge remains to formulate a method to effectively transfer expertise to the labour ward and thereby address the real and practical problems which face fetal monitoring today. This study demonstrates that intelligent systems may provide the vehicle to achieve this and thereby transform the cardiotocograph from a difficult to use, ineffective recorder of fetal heart rate, to an interactive and effective decision support tool capable of raising the skills of staff.

FUTURE RESEARCH

We are currently adapting the system to form part of a teaching package which has the potential to quickly lead to a method of staff training and accreditation. We are also continuing the development of the system for clinical decision support with Medical Research Council funding. The next stage is to rigorously evaluate it on many more cases and to develop a sophisticated user interface for use on the labour ward. A blinded study on 900 cases will then establish that the system will at least do no harm if fully implemented. If successful, a large multicentre randomized trial will be undertaken to fully assess the impact of the system on clinical practice. This is scheduled to take place in 1997.

Acknowledgements—We gratefully acknowledge the support of the Northcott Devon Medical Foundation, Science and Engineering Research Council, South Western Regional Health Authority and the Polytechnic Central Funding Council.

REFERENCES

Barrett, J. F. R., Jarvis, G. J., MacDonald, H. N., Buchan, P. C., Tyrell, S. N. & Lilford, R. J. (1990). Inconsistencies in clinical decisions in obstetrics. *Lancet*, **336**, 549–551.

Clark, L. (1991). Mugford, M. & Paterson, C. (1991). How does the mode of delivery affect the cost of maternity care?. *British Journal of Obstetrics and Gynaecology*, **98**, 519–523.

Dawes, G. S., Moulden, M. & Redman, C. W. G. (1991). System 8000: Computerized antenatal FHR analysis. *Journal of Perinatal Medicine*, **19**, 47–51.

Haverkamp, A. D., Orleans, M. & Langendoerfer, S. (1979). A controlled trial of the differential effects of intrapartum fetal monitoring. *American Journal of Obstetrics and Gynecology*, **134**, 399–408.

Ifeachor, E. C., Keith, R. D. F., Westgate, J. & Greene, K. R. (1991) An expert system to assist in the management of labour. In Liebowitz, J. (Ed.), *Expert Systems World Congress Proceedings*, 16–19 December, Orlando, FL, Vol. 4, pp. 2615–2622. Oxford: Pergamon Press.

Keith, R. D. F. (1993). Intelligent fetal monitoring and decision support in the management of labour. Ph.D. thesis, University of Plymouth, U.K.

Keith, R. D. F., Westgate, J., Ifeachor, E. C. & Greene, K. R. (1994). Suitability of artificial neural networks for feature extraction from the cardiotocogram during labour. *Medical, Biological, Engineering and Computing—Electrocardiography, Myocardial Contraction and Blood Flow Supplement*, **32**, S51–S57.

Keith, R. D. F., Westgate, J., Hughes, G. W., Ifeachor, E. C. & Greene, K. R. (1994). Preliminary evaluation of an intelligent system for the management of labour. *Journal of Perinatal Medicine*, **22**, 345–350.

Krause, W. (1990). Natalie by Niess. A computer-aided monitoring system for supervision of labour. In Maeda, K. (Ed.), *Computers in perinatal medicine*, pp. 103–111. Amsterdam: Elsevier Science.

Landis, J. R. & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, **33**, 159–174.

Leigh, M. A. (1993). The Lawyer's view of fetal monitoring. *Proceedings of the Department of Health Conference on Fetal Monitoring*, 28–29 January, Cumberland Lodge, Windsor, pp. 33–40.

Maeda, K. (1990). Computerised analysis of cardiotocograms and fetal movements. In Lilford, R. (Ed.), *Balliere's clinical obstetrics and gynaecology*, Vol. 4, pp. 797–781.

Neilson, J. P. (1993). Cardiotocography during labour; An unsatisfactory technique but nothing better yet. *British Medical Journal*, **306**, 347–348.

Recommendations arising from the 26th RCOG Study Group: Intrapartum fetal surveillance. In Spencer, J. A. D. & Ward, R. H. T. (Eds), *Intrapartum fetal surveillance*, pp. 387–395. London: Royal College Obstetricians Gynaecologists Press.

Vincent, C. A., Martin, T. & Ennis, M. (1991). Obstetric accidents: The patient's perspective. *British Journal of Obstetrics and Gynaecology*, **98**, 390–395.