ORIGINAL ARTICLE

Do panels vary when assessing intrapartum adverse events? The reproducibility of assessments by hospital risk management groups

D Kernaghan, G C Penney

Qual Saf Health Care 2006;15:359–362. doi: 10.1136/qshc.2006.018572

See end of article for authors' affiliations

Correspondence to:
D Kernaghan, Scottish
Programme for Clinical
Effectiveness in
Reproductive Health
(SPCERH), Room S7130,
Department of Obstetrics
and Gynaecology,
Simpson Centre for
Reproductive Health, 51
Little France Crescent,
Edinburgh EH16 4SA, UK;
dawn.kernaghan@ed.ac.
uk

Accepted 21 July 2006

Introduction: A national audit project, Scotland-wide Learning from Intrapartum Critical Events (SLICE), included local assessment of quality of care in cases of perinatal death and neonatal encephalopathy due to intrapartum events. Concerns had been raised about interobserver variation in case assessment by different panels. We therefore studied the extent of agreement and disagreement between assessment panels, and examined the areas in which agreement and disagreement tended to occur.

Methods: 8 cases were randomly selected from all 42 cases identified during a 6-month period (1 January–1 July 2005). Each case was independently reviewed by three panels: the local hospital clinical risk-management group and two specially convened external panels. Panels assessed quality of care in three areas: admission assessment, recognition of incident, and method and timing of delivery. Predefined standards of care were provided for these three areas. Panels were also asked to assess the overall quality of care.

Results: For each area of care, agreement between the two external panels was lowest. The lowest levels of agreement between panels were seen in assessment of overall care (50% crude agreement between external panel 1 and the hospital ($\kappa=0.24$, $AC_1=0.36$); 29% crude agreement between external panels 1 and 2 ($\kappa=-0.11$, $AC_1=0.1$); 47% crude agreement between external panel 2 and the hospital ($\kappa=0.36$, $AC_1=0.46$). The lowest level of agreement among all three panels was also in the assessment of overall care (crude agreement 48%; $\kappa=0.16$, $AC_1=0.34$).

Conclusion: Moderate to substantial agreement among the three panels was achieved for the three areas in which explicit standards were provided. Therefore, a systematic approach to analysis of adverse events in perinatal care improves reproducibility.

The quality of antenatal and intrapartum care has an important role in determining the outcome for an infant experiencing a hypoxic event in labour. It is one of the factors that interacts with the timing and severity of the insult and the constitutional susceptibility of the infant to determine whether an infant experiences no sequelae, neonatal encephalopathy, stillbirth or neonatal death.¹

A recently completed national audit project, Scotland-wide Learning from Intrapartum Critical Events (SLICE), included local assessment of quality of care in cases of perinatal death and neonatal encephalopathy due to intrapartum events. Assessment was undertaken by hospital clinical risk-management groups. Completed assessment forms were anonymised by the hospital and collated centrally so that generalisable clinical lessons could be elicited.

Preliminary work undertaken in the context of the Scottish Confidential Audit of Severe Maternal Morbidity² raised concerns regarding interobserver variation in case assessment by different panels. Because of these concerns, we studied in the context of the SLICE project, the extent of agreement and disagreement between assessment panels, and examined the areas in which agreement and disagreement tend to occur.

METHODS

During 6 months of national data collection (1 January–1 July 2005), 42 cases were identified that fitted the SLICE inclusion criteria (appendix). These cases had been assessed by the local hospital risk management group where the event occurred, using a structured case assessment form. The local risk management groups had access to information

containing patient details. Eight of these cases were randomly selected for the study of interobserver variation. Of these, five infants were alive at the end of the perinatal period (first week of life), of these two intrapartum stillbirths and one early neonatal death.

For each of the eight selected cases, the hospital of birth was contacted and the clinical case notes and other relevant information were photocopied and anonymised. Two independent panels were convened to review the cases. Each panel comprised a neonatologist, a neonatal nurse, an obstetrician and a midwife.

For each case, an identical case assessment proforma was completed by each of three assessment panels (the local hospital risk management group and the two independent panels). The case assessment proforma was developed from current evidence-based guidelines and research articles and provided guidance in several areas—for example, fetal monitoring in labour, indications for fetal blood sampling and decision-to-delivery interval in caesarean section.^{3–5}

Each panel was asked to assess the quality of care in three areas: admission assessment, recognition of incident, and method and timing of delivery. Panels were also asked to assess the overall quality of care.

The grading system used was based on the system first developed and used in the UK-wide confidential enquiry into

Abbreviations: CESDI, Confidential Enquiry into Stillbirths and Deaths in Infancy; SLICE, Scotland-wide Learning from Intrapartum Critical Events; SSBIDS, Scottish Stillbirth and Infant Death Survey

Table 1 Multirater agreement (agreement among the three panels)

	Crude agreement	% Crude agreement	AC ₁ statistic	κ
Admission assessment	13/18*	72	0.67	0.23
Recognition of incident	13/21	62	0.52	0.36
Method and timing of delivery	12/21	57	0.44	0.47
Overall assessment of care	10/21	48	0.34	0.16

*Explanation of numerators and denominators: potentially, 24 assessments were available for analysis (three panels assessing eight cases). For areas of care where no assessment was made, the case was excluded from generation of sums of agreement. Two panels failed to assign a grade of care for admission assessment in two cases; therefore, all three-panel assessments of these two cases were excluded. The numerator is the number of assessments in which panels agreed.

maternal deaths.⁶ For each area of care, four grading options were available:

1. Good: care is appropriate and well managed
2. Incidental suboptimal: lessons can be learnt, although the final outcome was not affected
3. Minor suboptimal: different management may have resulted in a different outcome
4. Major suboptimal: different management might have been expected to result in a more favourable outcome. The management of this case contributed considerably to the adverse outcome.

Pairwise agreement was measured for three pairs of assessments (panel 1 v panel 2, panel 1 v hospital and panel 2 v hospital). Agreement among all three panels (multirater agreement) was measured by calculating the percentage of cases for which there was agreement. For areas of care where no assessment was made, the case was excluded from generation of sums of agreement.

For further analyses, grades were dichotomised by aggregating grades 1 and 2 versus grades 3 and 4.

Generalised κ scores⁷ and AC₁ statistics⁸ were calculated to summarise the extent of agreement among the three panels. A κ value of 1 represents perfect agreement, whereas a κ value of 0 represents agreement equal to that of chance alone. A negative value implies that agreement is worse than chance alone. Landis and Koch⁹ have provided the criteria by which to evaluate κ statistics. The AC₁ statistic can be interpreted similarly.¹⁰

Table 2 Multirater agreement (agreement among the three panels); aggregation of grades 1 and 2 versus aggregation of grades 3 and 4

	Crude agreement	Crude agreement (%)	Aggregation of grades (AC ₁ statistic)	Aggregation of grades (κ)
Admission assessment	14/18	77	0.65	0.2
Recognition of incident	17/21	81	0.63	0.61
Method and timing of delivery	19/21	90	0.81	0.80
Overall assessment of care	15/21	71	0.49	0.36

Table 3 Admission assessment: pairwise agreement

	Admission assessment			
	Agreement in cases assessed	% Crude agreement	AC ₁	κ
Panel 1 v hospital	7/8	83	0.73	0.61
Panel 1 v panel 2	4/6	66	0.49	-0.14
Panel 2 v hospital	5/6	83	0.82	-0.09

Ethics committee approval for this study was granted by the Multi-Centre Research Ethics Committee for Scotland (B).

RESULTS

Three pairwise measures of agreement were made for each area of care (admission assessment, recognition of incident, method and timing of delivery, and overall care). The agreement between the two independent panels was lowest for each area of care. Crude agreement was <50% for all areas of care other than admission assessment, where it reached 66%.

The lowest levels of agreement between panels were seen in assessment of overall care. There was 50% crude agreement between external panel 1 and the hospital ($\kappa = 0.24$, AC₁ = 0.36), 29% crude agreement between external panels 1 and 2 ($\kappa = -0.11$, AC₁ = 0.1), and 47% crude agreement between external panel 2 and the hospital ($\kappa = 0.36$, AC₁ = 0.46).

Agreement among all three panels (multirater agreement) was also measured for each area of care. Again, agreement was lowest in the assessment of overall care (crude agreement 48%; $\kappa = 0.16$, AC₁ = 0.34; table 1). When grades 1/2 and 3/4 were aggregated, all measures of agreement increased. Again, agreement was lowest in the assessment of overall care (crude agreement 71%; $\kappa = 0.36$, AC₁ = 0.49; table 2).

DISCUSSION

Currently in Scotland, the Scottish Stillbirth and Infant Death Survey (SSBIDS) counts and classifies all perinatal deaths, allowing the calculation of mortality rates. However, in contrast with the confidential enquiry system for maternal deaths, there is no routine assessment of the circumstances of perinatal losses. SLICE was developed to assess the circumstances and quality of care in cases of perinatal mortality and morbidity in Scotland. Assessments of care were carried out by the clinical risk management group in the hospital where the event occurred. In an attempt to make assessments more objective and reproducible, predefined standards of care and a structured case assessment proforma were distributed to each hospital at the start of the data collection period.

There are several different ways to summarise the extent of agreement between pairs and groups of raters.¹⁰ κ is the most

Table 4 Recognition of incident: pairwise agreement

	Recognition of incident			
	Agreement in cases assessed	% Crude agreement	AC ₁	κ
Panel 1 v hospital	6/8	75	0.46	0.62
Panel 1 v panel 2	3/7	43	0.26	0.12
Panel 2 v hospital	4/7	57	0.46	0.55

widely used statistic in the medical literature; however, it has limitations. The κ statistic corrects for the agreement, which is expected to occur by chance; but it is this adjustment that can lead to the apparent paradox: high percentage agreement but low κ score.^{11, 12} This occurs because κ is sensitive to the prevalence of the assessment grades and is also affected by observer bias.¹³ (This problem is illustrated by the results generated by the multirater admission assessment. Agreement was reached in 13 of 18 cases (72%). Despite this apparently good level of agreement, the κ statistic was 0.23, which is unexpectedly low. In 12 of the 13 assessments, agreement was reached because raters graded the care as good. As assessments were not evenly spread across the four possible grades, the κ score was affected by the high prevalence of "grade 1".) Feinstein and Cicchetti¹² suggested that one way of dealing with this problem is simply to use the percentage agreement and not impose the corrections for chance that are associated with the κ statistic. An alternative solution suggested by Gwet⁸ is to calculate an AC₁ statistic that provides a chance-corrected result without the problems associated with κ .

In the fifth annual report of the Confidential Enquiry into Stillbirths and Deaths in Infancy (CESDI),¹⁴ a similar study was described where every fifth case of intrapartum-related death underwent a repeat review. Agreement in assessment of the grade of overall care was present in 52 of the 111 cases (47%; $\kappa = 0.18$).

One of the recommendations of this report was that standards should be predefined to provide guidance to the assessing panel. Within our own setting, standards were predefined for those areas of care relating to the admission assessment, recognition of incident, and method and timing of delivery. These standards allowed panels to assess care in a systematic manner and also permitted analysis of the areas in which agreement and disagreement occurred. In contrast with the CESDI study, each case in this study underwent two, rather than one, further panel assessments.

In the three areas where standards were provided, there appeared to be moderate to substantial agreement among the three panels. However, this level of agreement was not sustained when each panel was asked to assess the overall care. All three measures of agreement for this category show moderate agreement at best, with levels of agreement similar to those seen in the CESDI review.

Case assessment by panels of clinicians has been used in the Confidential Enquiry into Maternal Deaths since its inception in 1952. This approach has been modified and extended to other areas of healthcare,¹⁵ but no formal appraisal of its validity has taken place. It is assumed that the assessment of care made by the inquiry panel represents the reference measure or gold standard; however, we have shown in this study that the greatest disagreement was in fact between the two independent panels.

There are few areas in healthcare where robust evidence exists or may ever exist. It would be impossible to have predefined standards for every potential clinical scenario. Evidence also suggests that if many of the scenarios rarely occur in practice, the increased burden on the respondents may not be justified. Indeed, it may alienate them from the task.¹⁶

For each case, the final assessment of care is based on the consensus opinion of the panel. It was hoped that when grading the quality of care, the panels could show a departure from or adherence to the accepted standard, but where no standard exists this assessment would be based on the collective experience and opinion of the panel. A wider range of experience and knowledge can be accumulated by using a group consensus approach, and interaction between the group members hopefully stimulates a reflection of a wider range of possible options.

Table 5 Method and timing of delivery: pairwise agreement

	Method and timing of delivery			
	Agreement in cases assessed	% Crude agreement	AC1	κ
Panel 1 v hospital	6/8	75	0.68	0.63
Panel 1 v panel 2	3/7	43	0.26	0.18
Panel 2 v hospital	4/7	57	0.45	0.36

Table 6 Overall assessment of care: pairwise agreement

	Overall assessment of care			
	Agreement in cases assessed	% Crude agreement	AC1	κ
Panel 1 v hospital	4/8	50	0.36	0.24
Panel 1 v panel 2	2/7	29	0.1	-0.11
Panel 2 v hospital	4/7	47	0.46	0.36

Despite the theoretical benefits of this approach, it does not immediately appear to have translated into reproducible assessments among panels, where no standards exist. However, although perfect agreement between panels does not exist, the grading of overall care assigned is acceptable, as shown by the improvement in agreement scores when similar grades of care are aggregated (tables 2 and 3).

Although the counting and classification of perinatal deaths is vital for managing healthcare resources, a more thorough assessment of the circumstances surrounding these events may lead to the identification of areas of suboptimal care, which can be modified or prevented in the future. Local review benefits from the fact that more immediate results may be seen than in delayed retrospective evaluations. Some evidence also suggests that feedback by peers is more effective in prompting change than feedback by a neutral observer.¹⁷ There are concerns, however, that local review may be compromised by a lack of objectivity, owing to close involvement in individual cases. External reviews allow a potentially more objective assessment to be carried out.

In 2002, the stillbirth rate in the UK rose for the first time in 50 years. If we are to provide the best standards of perinatal care for women and their babies to reverse this trend, local and national reviews of adverse events in perinatal care must continue to be carried out. A systematic approach should be adopted to reap maximum rewards from such reviews.

ACKNOWLEDGEMENTS

We thank the panel members: Dr John Tierney, Sr Audrey Taggart, Dr David Lloyd, Sr Mary Gilchrist, Sr Elaine Moore, Sr Cindy Sykes, Dr Paul Holmes and Dr Julie Clare Becher; Mr Robert Lee for statistical advice; and the hospital clinical risk managers and risk management groups who identified, assessed, photocopied and anonymised cases.

Authors' affiliations

D Kernaghan, Scottish Programme for Clinical Effectiveness in Reproductive Health (SPCERH), Department of Obstetrics and Gynaecology, Simpson Centre for Reproductive Health, Edinburgh, UK
G C Penney, Scottish Programme for Clinical Effectiveness in Reproductive Health, Aberdeen Maternity Hospital, Aberdeen, UK

Funding: This study was funded by NHS Quality Improvement Scotland. However, the views expressed are those of the authors.

Competing interests: None.

REFERENCES

- 1 **Low JA.** Determining the contribution of asphyxia to brain damage in the neonate. *J Obstet Gynaecol Res* 2004;**30**:276–86.
- 2 **Penney G,** Adamson L, Kernaghan D. *Scottish Confidential Audit of Severe Maternal Morbidity.* Aberdeen: SPCEH, 2005.
- 3 **National Institute for Clinical Effectiveness (NICE).** *The use of electronic fetal monitoring: the use and interpretation of cardiotocography in intrapartum fetal surveillance.* London: NICE, 2001.
- 4 **RCOG Clinical Effectiveness Support Unit.** *The use of electronic fetal monitoring. Evidence-based clinical guideline number 8.* London: RCOG, 2001.
- 5 **Thomas J,** Paranjothy S. *National Sentinel Caesarean Section Audit Report.* London: Royal College of Obstetricians and Gynaecologists, Clinical Effectiveness Support Unit, 2001.
- 6 **Lewis G,** Drife J. *Why mothers die 2000–2, The Confidential Enquiries into Maternal Deaths in the United Kingdom, 2004.*
- 7 **Fleiss JL.** Measuring nominal scale agreement among many raters. *Psychol Bull* 1971;**76**:378–82.
- 8 **Gwet K.** Kappa statistic is not satisfactory for assessing the extent of agreement between raters. *Series: Statistical methods for inter-rater reliability assessment*, No1. 2002.
- 9 **Landis JR,** Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;**33**:159–74.
- 10 **Gwet K.** *Handbook of inter-rater reliability.* Gaithersburg MD: STATAXIS Publishing Company, 2001.
- 11 **Cicchetti DV,** Feinstein AR. High agreement but low kappa: II. Resolving the paradoxes. *J Clin Epidemiol* 1990;**43**:551–8.
- 12 **Feinstein AR,** Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. *J Clin Epidemiol* 1990;**43**:543–9.
- 13 **Costa SC,** Costa PA, Bernardes J. Agreement studies in obstetrics and gynaecology: inappropriateness, controversies and consequences. *BJOG* 2005;**112**:667–9.
- 14 **Confidential Enquiry into Stillbirths and Deaths in Infancy (CESDI) 5th Annual Report.** Maternal and Child Health Research Consortium, 1998.
- 15 **Grimley EJ.** *Review of confidential enquiries for the National Institute for Clinical Effectiveness.* London: NICE, 2001.
- 16 **Murphy MK,** Black NA, Lamping DL, *et al.* Consensus development methods, and their use in clinical guideline development. *Health Technol Assess* 1998;**2**:1–88.
- 17 **van den HP,** Grol R, van den Hoogen HJ, *et al.* Practice visits as a tool in quality improvement: mutual visits and feedback by peers compared with visits and feedback by non-physician observers. *Qual Health Care* 1999;**8**:161–6.

APPENDIX

SLICE Inclusion criteria

- All intrapartum stillbirths ≥ 34 weeks

OR

- Admission to the neonatal unit

PLUS

- ≥ 34 weeks gestation, no major congenital anomaly, infection or inborn error of metabolism

PLUS

- Presentation within 72 hours of birth with evidence of abnormality of neurological exam to include either:

1. Abnormal state of consciousness (lethargy, stupor, coma)
2. Seizures
3. Abnormal reflexes or hyporeflexia (including oculomotor or papillary abnormalities)
4. Abnormal tone