ORIGINAL ARTICLE

Inter- and intra-rater reliability for classification of medication related events in paediatric inpatients

D L Kunac, D M Reith, J Kennedy, N C Austin, S M Williams

Qual Saf Health Care 2006;**15**:196–201. doi: 10.1136/qshc.2005.014407

See end of article for authors' affiliations

Correspondence to: Mrs D L Kunac, Research Fellow, School of Pharmacy, University of Otago, P O Box 913, Dunedin, New Zealand; desiree.kunac@stonebow.otago.ac.nz

Accepted for publication 12 February 2006

Background: In medication safety research studies medication related events are often classified by type, seriousness, and degree of preventability, but there is currently no universally reliable ''gold standard'' approach. The reliability (reproducibility) of this process is important as the targeting of prevention strategies is often based on specific categories of event. The aim of this study was to determine the reliability of reviewer judgements regarding classification of paediatric inpatient medication related events.

Methods: Three health professionals independently reviewed suspected medication related events and classified them by type (adverse drug event (ADE), potential ADE, medication error, rule violation, or other event). ADEs and potential ADEs were then rated according to seriousness of patient injury using a seven point scale and preventability using a decision algorithm and a six point scale. Inter- and intra-rater reliabilities were calculated using the kappa (κ) statistic.

Results: Agreement between all three reviewers regarding event type ranged from ''slight'' for potential ADEs (κ = 0.20, 95% CI 0.00 to 0.40) to ''substantial'' agreement for the presence of an ADE (κ = 0.73, 95% CI 0.69 to 0.77). Agreement ranged from ''slight'' (κ = 0.06, 95% CI 0.02 to 0.10) to ''fair'' (κ = 0.34, 95% CI 0.30 to 0.38) for seriousness classifications but, by collapsing the seven categories into serious versus not serious, ''moderate'' agreement was found (κ = 0.50, 95% CI 0.46 to 0.54). For preventability decision, overall agreement was ''fair'' (κ = 0.37, 95% CI 0.33 to 0.41) but ''moderate'' for not preventable events (κ = 0.47, 95% CI 0.43 to 0.51).

Conclusion: Trained reviewers can reliably assess paediatric inpatient medication related events for the presence of an ADE and for its seriousness. Assessments of preventability appeared to be a more difficult judgement in children and approaches that improve reliability would be useful.

Medication related patient injury—so-called adverse drug events (ADEs)—and errors in the use of medications are commonly associated with the pharmacological treatment of patients in hospital.[1–3] In order to analyse these events with the aim of developing prevention strategies, data on the frequency, type, seriousness, and degree of preventability[4] of the event is required. If calculation of rates of events and the targeting of prevention strategies are based on specific categories of event, then the concept of reliability (or reproducibility) of the classification process is important.

Such classifications require some form of professional review and the general approach has previously been to have two independent physicians make these judgements.[4] Judgements require not only an up to date clinical knowledge, but also consideration of standards of care and the recognition of distinction between those injuries caused by disease or patient condition and those due to a medication.[5] Variations in judgements made by reviewers are an important source of measurement error.[6]

Where two or more reviews have been undertaken independently, it is possible to conduct reliability studies to determine the level of reviewer agreement in the measurement process. Reliability refers to the consistency of ratings or to the ability of two or more reviewers to reach the same conclusions about a specific case.[7]

In epidemiological studies of adverse events (including ADEs), the statistic most often used to measure agreement between two reviewers is the kappa statistic (κ). Kappa is a chance corrected index of agreement and is calculated by the equation $((O - E)/(1 - E)$, where O = observed agreement and E = expected agreement by chance.[6] Using kappa, reliability of 0.00 is considered poor agreement, 0.01–0.20 considered slight agreement, 0.21–0.40 fair agreement, 0.41–0.60 moderate agreement, 0.61–0.80 substantial agreement, and 0.81–1.00 almost perfect agreement.[6 8] The percentage agreement is sometimes reported and is calculated by dividing the number of agreed cases by the total number of cases.

Although a standardised approach for classification of events has recently been proposed,[4] there is no ''gold standard'' currently available. Not only does this mean that a variety of classification scales are in use, but the rigor with which these classifications are undertaken varies considerably between studies. Few paediatric ADE studies have published estimates of reliability. Where estimates of reliability are provided, there is often too little information available to allow comparison between studies or to allow an understanding of what factors may influence reliability judgements.

This study was undertaken to determine the reliability of reviewer judgements regarding classification of paediatric inpatient medication related events by type, seriousness, and degree of preventability.

METHODS

Study design and setting

A prospective observational cohort study was conducted over a 12 week period from 18 March to 9 June 2002 at a university affiliated urban general hospital in Dunedin, New Zealand. All admissions to the neonatal intensive care unit (NICU), postnatal ward, and paediatric ward during the study period were eligible for inclusion. Patients were excluded if the hospital admission was for less than 24 hours,

Table 1 Medication related event types: definitions and examples (adapted from Kaushal *et al*[3])

Event type	Definition	Example
ADE	Actual injuries resulting from medical interventions related to a medicine	Troublesome drug rash requiring intervention
Preventable ADE	Actual injuries resulting from the use of medication in error	The development of a rash after administration of flucloxacillin in a patient known to be allergic to penicillins
Non-preventable ADE	Actual injuries resulting from the use of a medication not associated with error, also termed adverse drug reactions	The development of a rash after administration of flucloxacillin in a patient with no known drug allergies
Potential ADE	Events that have a significant potential for injuring a patient but do not actually cause harm. This may be because they are intercepted before reaching the patient or, due to particular circumstances or chance, the patient is able to tolerate the error	A prescription order written for a 10-fold overdose of digoxin that is intercepted and corrected by the pharmacist before reaching the patient. A non-intercepted potential ADE would be the administration of a non-steroidal anti-inflammatory agent to a patient with asthma who does not experience any adverse effects
Medication error	Harmless errors associated with the use of a medication	Administration of one regular dose of non-critical medication given more than 2 hours later than scheduled
Rule violation	Faulty medication orders with little potential for harm or extra work because they are typically interpreted correctly by pharmacy and nursing staff without additional clarification	Prescription written for regular medication but not dated
Other events	Any reported events not classified as one of the other four event types	Mild side effects that are tolerated without need of intervention or general practice related issues

ADE, adverse drug event.

if medical staff deemed it inappropriate for a patient to be involved, or if the admission was due to an intentional overdose. This resulted in 495 eligible study patients who had a total of 520 admissions (84 paediatric medical, 61 paediatric surgical, 57 NICU, 318 postnatal).

Medication related events were identified by the principal investigator (DK) using a multipronged approach which involved:

- chart review for all admissions;[9]
- attendance at multidisciplinary ward meetings;
- interview of parents/carers (and children) when further information or clarification of information was required (a total of 106 of the 110 parents approached (96.4%) gave consent and were interviewed);
- voluntary and verbally solicited reports from staff;[4] all paediatric ward staff were educated about the study and were invited to take part by submitting voluntary reports of any actual events or potentially unsafe medication systems that they noted during their daily activities. This was either via the hard copy medEVENT form designed for the study or communicated verbally direct to the investigator during daily ward visits or via telephone. In addition, when the investigator visited ward areas, reports were solicited from staff on duty at the time.

All suspected medication related incidents (N = 701) were reviewed by a panel of three health professionals who independently categorised the events in various ways. The panel included a paediatric clinical pharmacologist (DR, reviewer 1), a neonatologist (NA, reviewer 2), and a clinical pharmacist (JK, reviewer 3). Prior to this process, the reviewers underwent a calibration exercise using simulated test cases and the reviewer form. As a result of discussions regarding these test cases, a clear set of guidelines were agreed; this included explanatory notes about the review process and contained definitions and examples for the different event categories, as shown in table 1. An anonymised computer generated summary was created for each event. Assessments were performed individually by the reviewers using a standardised form.

The review panel was required to judge event type (ADE, potential ADE, medication error, rule violation or other event), seriousness, and preventability. The reviewers rated ADEs and potential ADEs for seriousness based on International Committee on Harmonisation (US) guidelines.[10] The reviewers assessed preventability on the basis of the practitioners' presumed knowledge at the time the medication was prescribed. A preventable versus not preventable decision was made using a set of questions developed by Schumock and Thornton.[11] Confidence about the preventability classification of events was rated on a six point scale, based on the four point score devised by Dubois and Brook.[12] The preventability scores were collapsed into preventable (score 1–3) and not preventable (4–6) events. Medication errors, by definition, were automatically deemed ''not serious'' and ''preventable'' events. Rule violations, being very trivial events, were separated out as ''not applicable'' in the classification of preventability of events.

Statistical analysis

Inter-rater and intra-rater reliabilities for key judgements were calculated using the percentage of agreement and the kappa statistic (κ) using STATA for Windows Version 8.0 (Stata Corporation, College Station, TX, 2003). Three-way kappa was used to evaluate reliability between all three reviewers and two-way kappa analysis performed for evaluation of reviewer pairs. Because the marginal totals for some outcomes for some pairs of reviewers were very different, the maximum possible value of kappa was also calculated. Kappa max (max κ) was calculated using the equation: 1 − (minimum disagreement/expected disagreement).[8]

Table 2 Inter-rater reliability for all three reviewers for event type

Outcome	Frequency reported by each reviewer, N (%)			
	Reviewer 1	Reviewer 2	Reviewer 3	κ (95% CI)
Event type				
ADE	72 (10.3)	62 (8.8)	44 (6.3)	0.73 (0.69 to 0.77)
Potential ADE	40 (5.7)	156 (22.3)	101 (14.4)	0.20 (0.00 to 0.40)
Medication error	256 (36.5)	300 (42.8)	254 (36.2)	0.38 (0.34 to 0.42)
Rule violation	283 (40.4)	104 (14.8)	221 (31.5)	0.42 (0.38 to 0.46)
Other	50 (7.1)	79 (11.3)	81 (11.6)	0.37 (0.33 to 0.41)
Overall agreement				0.40 (0.38 to 0.42)

ADE, adverse drug event.

RESULTS

Level of agreement between all three reviewers

Agreement between all three reviewers regarding event type ranged from ''slight'' for potential ADEs to ''substantial'' for the presence of an ADE. Overall, using all five categories of event, ''fair'' agreement was found between reviewers (table 2).

The level of agreement between the three reviewers for seriousness is shown in table 3. Agreement was not much better than chance for seriousness categories of ''potential death'' (D) (there were no fatalities documented during the study period) and ''intervention to prevent permanent impairment'' (O). The strength of agreement was ''moderate'' for not serious events and also ''moderate'' when the seriousness categories were collapsed into serious versus not serious events.

The level of agreement between the three reviewers for preventability is shown in table 4. For preventability decision (yes/no), overall agreement was ''fair'' but ''moderate'' for not preventable events. For the preventability score and when scores were collapsed into three categories, overall agreement was again ''fair'' for preventability of events.

Level of agreement between reviewer pairs

The levels of agreement between reviewer pairs for event type, seriousness, and preventability are shown in table 5. For event type, the best agreement occurred between reviewers 1 and 3 where the level of agreement was found to be ''moderate''. Only ''fair'' agreement occurred between the other two reviewer pairs. The intra-rater reliability for each reviewer for a repeat categorisation of event type (12 months apart) of 100 randomly selected events is shown in table 5. Each of the three reviewers was found to be consistent.

The classification of events into the seven categories of seriousness demonstrated only ''fair'' agreement between each of the reviewer pairs. The maximum value of κ for reviewers 1 and 2 was 0.63 because the reviewers judged the seriousness in very different ways. For example, the second reviewer described 55 (7.9%) events as seriousness category O, whereas reviewer 1 described three (0.43%) events as seriousness category O. In addition, it appears that reviewer 3 was more likely to classify potential ADEs as more serious events than the other reviewers (table 3). By collapsing the categories down into two (serious versus not serious events), agreement between the reviewer pairs 2 and 3 and between 1 and 2 improved to ''moderate'' agreement. The κ/κmax ratio also improved for these reviewer pairs demonstrating ''substantial'' agreement for seriousness of events. There was only ''fair'' agreement between reviewers 1 and 3 when considering κ values and the κ/κmax ratio.

For the judgements made by reviewer pairs regarding the yes or no decision as to whether or not an event was preventable, the best agreement was found between reviewers 1 and 3 with $\kappa = 0.50$ and $\kappa/\kappa max = 0.51$, which is regarded as ''moderate'' agreement. Only ''fair'' agreement was found for the other two reviewer pairs. Similar findings were found for the preventability scores and when preventability was collapsed into three categories.

DISCUSSION

The level of agreement between all three reviewers was found to be ''substantial'' for judgments regarding whether or not an event was an ADE (patient injury related to a medication). However, for classification into the other event types, the level of agreement was lower, especially for potential ADEs where agreement was only ''slight''. Moderate agreement was achieved when the seriousness categories were collapsed into serious *v* not serious events. The degree of preventability appeared a more difficult judgement, with only ''fair'' agreement found between the three reviewers. Despite classification guidelines and prior discussion between the reviewers, there appeared to be some marked differences in interpretation between them. The judgements of each reviewer regarding event categorisation were, however, found to be consistent over time.

Table 3 Inter-rater reliability for all three reviewers for seriousness

Outcome	Frequency reported by each reviewer, N (%)			
	Reviewer 1	Reviewer 2	Reviewer 3	κ (95% CI)
Seriousness category				
D, potential death	0 (0)	1 (0.1)	14 (2.0)	0.06 (0.02 to 0.10)
L, life threatening	10 (1.4)	11 (1.6)	15 (2.1)	0.30 (0.26 to 0.34)
H, hospitalisation	27 (3.9)	11 (1.6)	12 (1.7)	0.34 (0.30 to 0.38)
P, persistent disability	20 (2.9)	5 (0.7)	14 (2.0)	0.16 (0.12 to 0.20)
O, intervention to prevent permanent impairment	3 (0.4)	55 (7.9)	3 (0.4)	0.07 (0.03 to 0.11)
N, not serious	641 (91.4)	618 (88.2)	643 (91.7)	0.50 (0.46 to 0.54)
Overall agreement				0.34 (0.32 to 0.36)
Seriousness category grouping				
Serious* *v* not serious				0.50 (0.46 to 0.54)

*Seriousness categories (D, L, H, P, O) combined.

Table 4 Inter-rater reliability for all three reviewers for preventability

Outcome	Frequency reported by each reviewer, N (%)			
	Reviewer 1	Reviewer 2	Reviewer 3	κ (95% CI)
Preventability decision (yes or no)				
Not preventable	23 (3.28)	18 (2.57)	21 (3.00)	0.47 (0.43 to 0.51)
Preventable	329 (46.93)	398 (56.78)	326 (46.50)	0.35 (0.31 to 0.39)
Not applicable	349 (49.79)	285 (40.66)	354 (50.50)	0.37 (0.33 to 0.41)
Overall agreement				0.37 (0.33 to 0.41)
Preventability score				
1 (definitely preventable)	257 (36.67)	307 (43.80)	298 (42.51)	0.33 (0.29 to 0.37)
2 (strong evidence for preventability)	33 (4.71)	44 (6.28)	13 (1.85)	0.18 (0.14 to 0.22)
3 (preventability more likely than not)	25 (3.57)	42 (6.00)	14 (2.00)	0.31 (0.27 to 0.35)
4 (preventability not quite likely)	11 (1.57)	17 (2.43)	5 (0.71)	0.17 (0.13 to 0.21)
5 (slight to modest evidence not preventable)	7 (1.00)	2 (0.29)	7 (1.00)	0.06 (0.02 to 0.10)
6 (definitely not preventable)	19 (2.71)	6 (0.86)	11 (1.57)	0.24 (0.20 to 0.28)
Not scored	349 (49.79)	283 (40.37)	353 (50.36)	0.37 (0.33 to 0.41)
Overall agreement				0.33 (0.29 to 0.37)
Preventability category*				
Preventable	315 (44.94)	393 (56.06)	325 (46.36)	0.37 (0.33 to 0.41)
Not applicable	349 (49.79)	283 (40.37)	353 (50.36)	0.37 (0.33 to 0.41)
Not preventable	37 (5.28)	25 (3.57)	23 (3.28)	0.53 (0.49 to 0.57)
Overall agreement				0.38 (0.34 to 0.42)

*Based on preventability scores (score 1–3 = preventable, score 4–6 = not preventable).

There are a limited number of paediatric studies of ADEs and medication errors that report reliability data for event classification by type of event, seriousness, or degree of preventability.[3 13–15] Kaushal *et al*[3] reported 87–100% agreement, κ = 0.65–1.0, but actual values specific to event type, seriousness, and preventability were not stated. In each of the studies by King *et al*[14] and Potts *et al*,[15] inter-rater reliability is reported for event type but not for seriousness or preventability. In the remaining study, Kozer and colleagues[13] found substantial agreement between two paediatric emergency physicians for whether an error occurred (κ = 0.79) and for a three category ranking of severity of events (κ = 0.70).

Event type

In the present study, although we found ''substantial'' agreement between all three reviewers for the presence of an ADE (κ = 0.73), there was only ''fair'' to ''moderate'' agreement for classification of the other event types. It is evident that the reviewers classified event types very differently (table 2); in particular, reviewer 1 classified very few events as potential ADEs compared with reviewers 2 and 3. This led to only a ''fair'' level of agreement for event type overall between the three reviewers (κ = 0.40).

For reviewer pairs, the present study showed best agreement between reviewers 1 and 3 (κ = 0.51) for event type overall. Previous paediatric studies reported higher levels of agreement between two reviewers for event type. King *et al*[14] found ''substantial'' agreement (κ = 0.64, 95% CI 0.45 to 0.82) for 20 randomly selected incident reports from paediatric inpatients at a tertiary care paediatric hospital when independently rated by two physicians. Potts *et al*[15] reported a κ value of 0.96, indicating ''almost perfect'' agreement between a clinical pharmacist and physician when a 10% random sample of patients from a paediatric critical care unit was reviewed. Unfortunately, these reports do not provide a breakdown of levels of agreement for the different event types, so it is difficult to compare our study findings any further with other paediatric inpatient reliability data.

However, the finding for the presence of an ADE (κ = 0.73) in the present study is consistent with the adult literature. For ADE *v* potential ADE or problem order, Bates and colleagues reported ''almost perfect'' agreement in two

Table 5 Level of agreement between reviewer pairs

Inter-rater	Reviewer 1 versus 2			Reviewer 2 versus 3			Reviewer 1 versus 3		
	κ (95% CI)	κ max	κ/κ max	κ (95% CI)	κ max	κ/κ max	κ (95% CI)	κ max	κ/κ max
Event type	0.36 (0.32 to 0.40)	0.98	0.37	0.35 (0.33 to 0.39)	0.77	0.45	0.51 (0.47 to 0.55)	0.88	0.57
Seriousness*	0.36 (0.32 to 0.40)	0.63	0.59	0.37 (0.33 to 0.41)	0.59	0.63	0.29 (0.25 to 0.33)	0.78	0.38
Serious *v* not serious	0.53 (0.45 to 0.61)	0.82	0.64	0.58 (0.50 to 0.66)	0.80	0.73	0.37 (0.29 to 0.45)	0.98	0.37
Preventability decision (yes/no)	0.31 (0.25 to 0.37)	0.81	0.38	0.30 (0.24 to 0.36)	0.81	0.37	0.50 (0.42 to 0.58)	0.99	0.51
Preventability score	0.30 (0.24 to 0.36)	0.81	0.37	0.24 (0.18 to 0.30)	0.83	0.29	0.45 (0.39 to 0.51)	0.90	0.50
Preventability category†	0.33 (0.27 to 0.39)	0.80	0.42	0.31 (0.25 to 0.37)	0.81	0.38	0.51 (0.45 to 0.57)	0.96	0.53
Intra-rater	Reviewer 1			Reviewer 2			Reviewer 3		
Event type	0.64 (0.52 to 0.76)	0.74	0.87	0.55 (0.45 to 0.65)	0.87	0.63	0.69 (0.59 to 0.79)	0.88	0.78

*Based on seven categories of seriousness;
†Based on preventability scores (score 1–3 preventable, score 4–6 not preventable).

studies ($\kappa = 0.83$ and $\kappa = 0.98$),[2 16] and ''substantial'' agreement ($\kappa = 0.68$) for classification as medication error, rule violation or neither.[17] For adult inpatients at community based nursing homes in the United States, Gurwitz[18] reported ''substantial' agreement'' ($\kappa = 0.80$) between two independent physicians for the presence of an ADE. It appears that judgements regarding classification of events as ADEs are more reliable than classification of other event types. This would seem reasonable as ADE classification is based on objective evidence of actual patient injury, whereas classification for other event types is subjective and based on reviewer opinion regarding potential for patient harm (potential ADEs) and whether the cause of the event was due to error (medication error) or violation of a rule or guideline (rule violation).

Seriousness

In the present study, when reviewers classified events for seriousness into one of the seven categories, the level of agreement was only ''fair'' between all three reviewers and for the reviewer pairs. When collapsed into two seriousness categories (serious and not serious), ''moderate'' agreement between the three reviewers was achieved. Many of the published paediatric studies of ADEs have included some assessment of seriousness of events using a variety of different rating scales. However, only two studies appear to have evaluated and published inter-rater reliability data regarding the severity or seriousness scale being used. Both studies report ''substantial'' agreement between two independent physician reviewers.[3 13] The lower level of agreement in the present study may in part be due to differences in the rating scales used (seven categories in the present study compared with 3–4 point scales in the previous paediatric reports), but may also be attributed to bias among reviewers in the present study. The very different frequencies (table 3) show that the reviewers classified seriousness of events in very different ways.

In adult inpatient studies, three to four category scales of seriousness have been evaluated for reliability, producing mixed results. Using a three point scale and classification by two independent physicians, Bates[16] reported a κ value of 0.89 for life threatening *v* serious or significant and a κ value of 0.63 for significant *v* serious or life threatening. In a later study by Bates and colleagues,[2] using a four point scale adapted from Folli *et al*,[19] found (as we had) low κ values despite a high percentage agreement. Actual findings were life threatening *v* serious or significant $\kappa = 0.37$ (85% agreement) and significant *v* serious or life threatening $\kappa = 0.32$ (66% agreement). Again using the same four point Folli scale[19] but subsequently collapsed into severe *v* not severe events, Gurwitz *et al*[18] reported ''substantial'' agreement ($\kappa = 0.62$).

Preventability

In the present study the low levels of agreement regarding preventability indicate that the reviewers had difficulty determining whether an error was associated with an event. It may be that such judgements are difficult in the paediatric setting due to unlicensed use of medicines in children[20–25] and the resulting lack of standardised paediatric clinical practice guidelines. It is believed that judgements regarding appropriateness of care are strongly influenced by perceived outcomes and that practice guidelines aid reviewers to make assessments by clarifying the accepted standard of care.[26] The appropriate standard of care may have therefore been unclear to our reviewers, making preventability judgements difficult.

Few paediatric inpatient studies have reviewed events for degree of preventability and, unfortunately, those that have also used the Schumock and Thornton[11] assessment criteria[27–29] have not reported inter-rater reliability data for preventability judgements. Kaushal *et al*,[3] using a five point scale collapsed into preventable *v* not preventable events, reported the level of agreement for preventability to be within the range $\kappa = 0.65$–1.0. In the present study, rule violations—being very trivial events—were considered separately within the ''not applicable'' category. It is not clear whether rule violations were considered as part of the ''preventable'' event group by Kaushal *et al*[3] but, if so, this may account for a higher level of agreement than the present study findings.

In adult inpatient studies, using a four point scale proposed by Dubois[12] and collapsed into preventable *v* not preventable events, ''substantial'' to ''almost perfect'' agreement has been found between two independent physician reviewers. In two separate studies of hospitalised adults, Bates and colleagues have reported κ values for preventability of 0.71[16] and 0.92.[2] For adult inpatients in community based nursing homes, Gurwitz[18] also found ''substantial'' agreement ($\kappa = 0.73$) for preventability.

The lower levels of agreement in the present study probably reflect bias among reviewers, but may also be attributed to different event types being included in the ''preventable'' grouping.

Limitations

The present study is limited because the data come from hospital records of paediatric admissions at one academic institution and represent the agreements from three reviewers, so they may not be generalisable to other geographical locations or other reviewers. Also, our study only investigated one implicit review instrument. Reordering, rewording, or restructuring the subcategories of our review form could produce better degrees of reliability.

Implications for future research

Our findings have several implications for the design of future research studies involving medication related event classification. Firstly, in research studies involving classification of events, independent review should be undertaken by at least two reviewers so that reliability of judgements may be determined. Although in the present study there seemed to be some differences in interpretation between reviewers despite classification guidelines, structured review criteria, and early joint review of ''test'' cases, such strategies to identify any differences in interpretation before the start of the study are essential. Secondly, for assessment of seriousness, reviewer judgements could be streamlined by direct classification of events as serious *v* not serious events, but this may not be as useful clinically. Thirdly, in research studies where event classification is undertaken, inter- and intra-rater reliability data should be reported in sufficient detail to allow the reader to assess the reproducibility of the

Key messages

- The reliability of reviewer ratings for medication related event classification was ''substantial'' for the presence of an ADE, ''moderate'' for seriousness of the event, but only ''fair'' for the more complex judgement regarding preventability of events.
- Trained reviewers can reliably assess paediatric inpatient medication related events for the presence of an ADE and for seriousness.
- Assessments of preventability appeared to be a more difficult judgement in children and approaches that improve reliability would be useful.

classification method used. Finally, where marginal totals are markedly different, inclusion of the κ:κmax ratio is useful as this may account for lower than expected levels of agreement.

Authors' affiliations

D L Kunac, J Kennedy, School of Pharmacy, University of Otago, Dunedin, New Zealand
D M Reith, S M Williams, Department of Women's and Children's Health, Dunedin School of Medicine, University of Otago, Dunedin, New Zealand
N C Austin, Christchurch Women's Hospital, Christchurch, New Zealand

This research was supported by a Fellowship awarded to Desireé Kunac by the Child Health Research Foundation of New Zealand.

Competing interests: none.

Ethical approval for this study was granted by the Otago Ethics Committee.

REFERENCES

1 **Davis P**, Lay-Yee R, Briant R, *et al*. *Adverse events in New Zealand public hospitals. Principal findings from a national survey. Occasional Paper No 3*. Wellington: Ministry of Health. Available from http://www.moh.govt.nz/moh.nsf (accessed 9 February 2006).
2 **Bates DW**, Cullen DJ, Laird N, *et al*. Incidence of adverse drug events and potential adverse drug events: Implications for prevention. *JAMA* 1995;**274**:29–34.
3 **Kaushal R**, Bates DW, Landrigan C, *et al*. Medication errors and adverse drug events in pediatric inpatients. *JAMA* 2001;**285**:2114–20.
4 **Morimoto T**, Gandhi TK, Seger AC, *et al*. Adverse drug events and medication errors: detection and classification methods. *Qual Saf Health Care* 2004;**13**:306–14.
5 **Brennan TA**, Localio RJ, Laird NL. Reliability and validity of judgments concerning adverse events suffered by hospitalized patients. *Med Care* 1989;**27**:1148–58.
6 **Landis JR**, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;**33**:159–74.
7 **Dawson B**, Trapp R. *Basic and clinical biostatistics*, 4th ed. New York: McGraw-Hill, 2004.
8 **Dunn G**. What is good agreement? In: *Design and analysis of reliability studies: the statistical evaluation of measurement errors*. New York: Oxford University Press, 1989:37–8.
9 **Kaushal R**. Using chart review to screen for medication errors and adverse drug events. *Am J Health Syst Pharm* 2002;**59**:2323–5.
10 **ICH Steering Committee**. *ICH harmonised tripartite guideline. Clinical safety data management: definitions and standards for expedited reporting*. Available from http://www.ich.org (accessed 9 February 2006).
11 **Schumock GT**, Thornton JP. Focusing on the preventability of adverse drug reactions. *Hosp Pharm* 1992;**27**:538.
12 **Dubois RW**, Brook RH. Preventable deaths: who, how often, and why? *Ann Intern Med* 1988;**109**:582–9.
13 **Kozer E**, Scolnik D, Macpherson A, *et al*. Variables associated with medication errors in pediatric emergency medicine. *Pediatrics* 2002;**110**:737–42.
14 **King WJ**, Paice N, Rangrej J, *et al*. The effect of computerized physician order entry on medication errors and adverse drug events in pediatric inpatients. *Pediatrics* 2003;**112**:506–9.
15 **Potts AL**, Barr FE, Gregory DF, *et al*. Computerized physician order entry and medication errors in a pediatric critical care unit. *Pediatrics* 2004;**113**:59–63.
16 **Bates DW**, Leape LL, Petrycki S. Incidence and preventability of adverse drug events in hospitalized adults. *J Gen Intern Med* 1993;**8**:289–94.
17 **Bates DW**, Boyle DL, Vander Vliet MB, *et al*. Relationship between medication errors and adverse drug events. *J Gen Intern Med* 1995;**10**:199–205.
18 **Gurwitz JH**, Field TS, Avorn J, *et al*. Incidence and preventability of adverse drug events in nursing homes. *Am J Med* 2000;**109**:87–94.
19 **Folli HL**, Poole RL, Benitz WE, *et al*. Medication error prevention by clinical pharmacists in two children's hospitals. *Pediatrics* 1987;**79**:718–22.
20 **Collier J**. Paediatric prescribing: using unlicensed drugs and medicines outside their licensed indications. *Br J Clin Pharmacol* 1999;**48**:5–8.
21 **Conroy S**. Unlicensed and off label drug use for paediatric patients. Optimal dosing schedules with gentamicin are needed for premature neonates. *BMJ* 1998;**317**:204–5.
22 **Conroy S**, McIntyre J, Choonara I. Unlicensed and off label drug use in neonates. *Arch Dis Child Fetal Neonatal Ed* 1999;**80**:F142–4.
23 **Conroy S**, Choonara I, Impicciatore P, *et al*. Survey of unlicensed and off label drug use in paediatric wards in European countries. European Network for Drug Investigation in Children. *BMJ* 2000;**320**:79–82.
24 **Turner S**, Gill A, Nunn T, *et al*. Use of ''off-label'' and unlicensed drugs in paediatric intensive care unit. *Lancet* 1996;**347**:549–50.
25 **Turner S**, Longworth A, Nunn AJ, *et al*. Unlicensed and off label drug use in paediatric wards: prospective study. *BMJ* 1998;**316**:343–5.
26 **Goldman RL**. The reliability of peer assessments of quality of care. *JAMA* 1992;**267**:958–60.
27 **Easton-Carter KL**, Chapman CB, Brien JE. Emergency department attendances associated with drug-related problems in paediatrics. *J Paediatr Child Health* 2003;**39**:124–9.
28 **Easton KL**, Chapman CB, Brien JA. Frequency and characteristics of hospital admissions associated with drug-related problems in paediatrics. *Br J Clin Pharmacol* 2004;**57**:611–5.
29 **Easton KL**, Parsons BJ, Starr M, *et al*. The incidence of drug-related problems as a cause of hospital admissions in children. *Med J Aust* 1998;**169**:356–9.