Inter-rater reliability (IRR) report for NRAD

Method

As part of NRAD's Inter-Rater Reliability (IRR) process, cases were independently reviewed in order to assess the reliability of NRAD's methods, as follows:

Sample 1 (for assessing the reliability of the screening phase): A random 10% sample (selected by a computer programme) of the first 500 cases screened was selected (50 cases) for repeat screening for suitability for inclusion by the expert clinical screening panel.

Sample 2 (for assessing the reliability of the panel assessment): It was initially planned to repeat the panel assessment in a 10% sample or 40 cases, whichever was highest. However, owing to resources, only 23 cases were selected at random from the first 23 panels (see note on this method at the end of this IRR report).

Sample 3 (for assessing the reliability of reviewing the post-mortem report): Where post-mortems were undertaken, an independent pathologist with expertise in the coronial process repeated the review for alternate cases, a sample of 68 cases.

Results in the first and repeat assessments were compared by calculating the overall percentage of cases where there was agreement in the assessment. The kappa statistic was also used to measure agreement. The kappa statistic quantifies the degree to which the assessors agree over and above what could be expected by chance and is a more meaningful measure than overall agreement. For questions where the great majority of answers are in one category, we would expect a high percentage agreement purely by chance; in these circumstances, the kappa statistic will be more stringent and distinguish how much agreement there is beyond mere chance.

The kappa statistic ranges from 1 (perfect agreement) to –1 (complete disagreement). A kappa statistic of 0 implies a level of agreement that would be expected by chance alone. Statistics of 0.41–0.60 are usually regarded as reflecting moderate agreement and 0.61–0.80 as reflecting good agreement.

Results

1. Assessing the reliability of the screening phase (using sample 1)

		Rater 2		
		Included	Excluded	Total
Rater 1	Included	18	4	22
	Excluded	1	27	28
	Total	19	31	50

90% (45/50) agreement, kappa statistic=0.79

2. Assessing the reliability of the panel assessment (using sample 2)

2.1. Patient had asthma

		Rater 2		
		Definitely/ probably/ possibly	Unlikely/ no	Total
Rater 1	Definitely/probably/possibly	19	1	20
	Unlikely/no/insufficient information	2	1	3
	Total	21	2	23

87% (20/23) agreement, kappa statistic=0.33

2.2 Asthma caused or contributed to death

		Rater 2		
		Yes	No	Total
Rater 1	Yes	12	2	14
	No	5	4	9
	Total	17	6	23

70% (16/23) agreement, kappa statistic=0.32

2.3. Patient had at least one major factor that contributed towards death

		Rater 2		
		Yes	No	Total
Rater 1	Yes	9	2	11
	No	2	10	12
	Total	11	12	23

83% (19/23) agreement, kappa statistic=0.65

2.4. Overall assessment

		Rater 2				
		Good practice	Room for improvement	Less than satisfactory	Insufficient information	Total
Rater 1	Good practice	1	4	0	0	5
	Room for improvement	2	7	1	0	10
	Less than satisfactory	1	4	1	0	6
	Insufficient information	0	0	0	2	2
	Total	4	15	2	2	23

48% (11/23) agreement, kappa statistic=0.20

3. Assessing the reliability of reviewing the post-mortem report (using sample 3)

3.1 Was the post-mortem useful?

		Rater 2		
		Yes	No	Total
Rater 1	Yes	60	0	60
	No	6	2	8
	Total	66	2	68

91% (62/68) agreement, kappa statistic=0.37

3.2 Was the report of a sufficient standard on which to base a final conclusion?

		Rater 2		
		Yes	No	Total
Rater 1	Yes	47	9	56
	No	8	4	12
	Total	55	13	68

75% (51/68) agreement, kappa statistic=0.17

3.3 Was asthma the cause of death?

		Rater 2			
		Yes	No	Unable to conclude	Total
Rater 1	Yes	33	1	3	37
	No	3	10	5	18
	Unable to conclude	7	1	5	13
	Total	43	12	13	68

71% (48/68) agreement, kappa statistic=0.49

Conclusion

Statistics of 0.01–0.20 are usually regarded as slight agreement, 0.21–0.40 as fair agreement, 0.41–0.60 as moderate agreement and 0.61–0.80 as good agreement.

Screening phase: There was good agreement at the screening phase, with a high kappa statistic of 0.79. This is very encouraging in terms of the utility of the method to include/exclude cases suitable for panel review in the future.

Panel assessment: There was good agreement between first and repeat panel assessors in regard to whether patients had at least one major factor contributing to death, with a kappa statistic of 0.65. However, there was only fair agreement in regard to whether patients had asthma (0.33) and to whether patients died from asthma (0.32). These values are much lower than anticipated. Both assessors were consistent in judging that the vast majority of these patients had asthma, but they had difficulty in agreeing when this was not so – between them, they identified four such cases, but could agree on only one. The panel assessors also had difficulty in agreeing on when patients did not die from asthma, since between them they identified 11 deaths not due to asthma and agreed on only four of these.

Note that there was another sample of 27 cases that were purposely selected to go to the panel for reassessment (rather than a random selection) – these results are not described in detail here other than to note that the four respective kappa statistics of 0.17, 0.25, 0.32 and 0.31 were lower than those reported for the random sample.

Post-mortem report assessment: There was moderate agreement, kappa statistic of 0.49, between post-mortem assessors as to whether asthma was the cause of death. In regard to whether the post-mortem was useful, there was less agreement (kappa statistic of 0.37), and this is reflected by the two assessors between them having identified eight cases in which the post-mortem was not useful but could agree on only two on these. It is clear from the response overall that the vast majority of the post-mortems were useful – the difficulty that the assessors had was in agreeing when they were not useful. Finally, there was minimal agreement, kappa statistic of 0.17, as to whether the report was of a sufficient standard on which to base a final conclusion. The two assessors identified 21 cases in which the report was not of a sufficient standard, but could agree on only four of these.

Note on method for selecting the sample to test for the reliability of panel assessment

Panel assessors will ring a member of the NRAD team to book a place for panel assessment. The panel assessor will then fill one of the 8–15 places that have been allocated for that panel. Each case will then be allocated to one main assessor. The only criterion for a case not to be allocated to a panel assessor is that the place of death cannot be from the same region as where the panel assessor works. For this reason, cases were selected using a randomly generated rank system approach. Each place was given a random number between 0 and 1 (eg 0.12, 0.34) and then ordered in ascending order and ranked 1–15. Cases that were reviewed in place ranked as 1 would be selected for second review. If this case had already gone to panel twice, then cases that were placed in rank 2 would be selected for second review, and so on.