Mesa-Eguiagaray *et al. BMC Medical Research Methodology* (2016) 16:171
DOI 10.1186/s12874-016-0266-4

BMC Medical Research
Methodology

RESEARCH ARTICLE Open Access

Inter-rater reliability of the QuIS as an assessment of the quality of staff-inpatient interactions

Ines Mesa-Eguiagaray[1], Dankmar Böhning[2], Chris McLean[3], Peter Griffiths[3], Jackie Bridges[3] and Ruth M Pickering[1*]

Abstract

Background: Recent studies of the quality of in-hospital care have used the Quality of Interaction Schedule (QuIS) to rate interactions observed between staff and inpatients in a variety of ward conditions. The QuIS was developed and evaluated in nursing and residential care. We set out to develop methodology for summarising information from inter-rater reliability studies of the QuIS in the acute hospital setting.

Methods: Staff-inpatient interactions were rated by trained staff observing care delivered during two-hour observation periods. Anticipating the possibility of the quality of care varying depending on ward conditions, we selected wards and times of day to reflect the variety of daytime care delivered to patients. We estimated inter-rater reliability using weighted kappa, κ_w, combined over observation periods to produce an overall, summary estimate, $\hat{\kappa}_w$. Weighting schemes putting different emphasis on the severity of misclassification between QuIS categories were compared, as were different methods of combining observation period specific estimates.

Results: Estimated $\hat{\kappa}_w$ did not vary greatly depending on the weighting scheme employed, but we found simple averaging of estimates across observation periods to produce a higher value of inter-rater reliability due to over-weighting observation periods with fewest interactions.

Conclusions: We recommend that researchers evaluating the inter-rater reliability of the QuIS by observing staff-inpatient interactions during observation periods representing the variety of ward conditions in which care takes place, should summarise inter-rater reliability by κ_w, weighted according to our scheme A4. Observation period specific estimates should be combined into an overall, single summary statistic $\hat{\kappa}_{w\ random}$, using a random effects approach, with $\hat{\kappa}_{w\ random}$, to be interpreted as the mean of the distribution of κ_w across the variety of ward conditions. We draw attention to issues in the analysis and interpretation of inter-rater reliability studies incorporating distinct phases of data collection that may generalise more widely.

Keywords: Weighted kappa, Random effects meta-analysis, QuIS, Collapsing, Averaging

Background

The Quality of Interactions Schedule (QuIS) has its origin in observational research undertaken in 1989 by Clark & Bowling [1] in which the social content of interactions between patients and staff in nursing homes and long term stay wards for older people was rated to be positive, negative or neutral. The rating specifically relates to the social or conversational aspects of an interaction, such as the degree to which staff acknowledge the patient as a person, not to the adequacy of any care delivered during the interaction. Dean et al. [2] extended the rating by introducing distinctions within the positive and negative ratings, creating a five category scale as set out in Table 1. QuIS is now generally regarded as an ordinal scale ranging from the highest ranking, positive social interactions to the lowest ranking, negative restrictive interactions [3].

Barker et al. [4] in a feasibility study of an intervention designed to improve the compassionate/social aspects of care experienced by older people in acute hospital

* Correspondence: rmp@soton.ac.uk
[1] Medical Statistics Group, Faculty of Medicine, Southampton General Hospital, Mailpoint 805Level B, South Academic Block, Southampton SO16 6YD, UK
Full list of author information is available at the end of the article

Table 1 Definitions of QuIS categories [2]

CATEGORY	Explanation
Positive social (+s)	Interaction principally involving 'good, constructive, beneficial' conversation and companionship.
Positive Care (+c)	Interactions during the appropriate delivery of physical care.
Neutral (N)	Brief, indifferent interactions not meeting the definitions of the other categories.
Negative protective (–p)	Providing care, keeping safe or removing from danger, but in a restrictive manner, without explanation or reassurance: in a way which disregards dignity or fails to demonstrate respect for the individual.
Negative restrictive (–r)	Interactions that oppose or resist peoples' freedom of action without good reason, or which ignore them as a person.

wards, proposed the use of the QuIS as a direct assessment of this aspect of the quality of care received. This is a different context to that for which the QuIS was originally developed and extended, and it may well perform differently: wards may be busier and more crowded, beds may be curtained off, raters may have to position themselves more or less favourably in relation to the patients they are observing. A component of the feasibility work evaluated the suitability of the QuIS in the context of acute wards, and in particular its inter-rater-reliability [5]. Because of the lack of alternative assessments of quality of care it is likely that the QuIS will be used more widely, and any such use should be preceded by studies examining its suitability and its inter-rater reliability.

In this paper we describe the analysis of data from an inter-rater reliability study of the QuIS reported by McLean et al. [5]. Eighteen pairs of observers rated staff-inpatient interactions during two hour long observation periods purposively chosen to reflect the wide variety of conditions in which care is delivered in the hospital setting. The study should thus have captured differences in the quality of care across conditions, for example when staff were more or less busy. It is possible that inter-rater reliability could also vary depending on the same factors, and thus an overall statement of typical inter-rater reliability should reflect variability across observation periods in addition to sampling variability. We aim to establish a protocol for summarising data from inter-rater reliability studies of the QuIS, to facilitate consistency across future evaluations of its measurement properties. We summarise inter-rater reliability using kappa (κ) which quantifies the extent to which two raters agree in their ratings, over and above the agreement expected through chance alone. This is the most frequently used presentation of inter-rater reliability in applied health research, and is thus familiar to researchers in the area. When κ is calculated all differences in ratings are treated equally. Varying severity of disagreement between raters depending on the categories concerned can be accommodated in weighted κ, κ_w, however standard weighting schemes give equal weight to disagreements an equal number of categories apart regardless of their position on the scale, and are thus not ideal for the QuIS. For example, a disagreement between the two adjacent positive categories is not equivalent to a disagreement between the adjacent positive care and neutral categories. Thus we aim to establish a set of weights to be used in κ_w, that reflects the severity of misclassification between each pair of QuIS categories. We propose using meta-analytic techniques to combine the estimates of κ_w from the different observation periods to produce a single overall estimate of κ_w.

Methods

QuIS observation

Following the training described by McLean et al. [5], each of 18 pairs of research staff observed, and QuIS rated all interactions involving either of two selected patients, during a two-hour long observation period. The 18 observation periods were selected with the intention of capturing a wide variety of conditions in which care is delivered to patients in acute wards, as this was the target of the intervention to be evaluated in a subsequent main trial. Observation was restricted to a single, large teaching hospital on the South Coast of England and took place in three wards, on weekdays, and at varying times of day between 8 am to 6 pm, including some periods when staff were expected to be busy (mornings) and others when staff might be less so.

The analysis of inter-rater reliability was restricted to staff-patient interactions rated by both raters, indicated by them reporting an interaction starting at the same time: interactions rated by only one rater were excluded. The percentage of interactions missed by either rater is reported, as is the Intraclass Correlation Coefficient (ICC) of total number of interactions reported by each rater in the observation periods.

κ estimates of inter-rater reliability

Inter-rater agreement was assessed as Cohen's κ [6] calculated from the cross-tabulation of ratings into the $k = 5$ QuIS categories of the interactions observed by both raters:

$$\hat{\kappa} = \frac{p_o - p_e}{1 - p_e}, \tag{1}$$

with p_o being the proportion of interactions with identical QuIS ratings and p_e being the proportion of interactions expected to be identical ($\sum_{i=1}^{k} p_{i.}\ p_{.i}$) calculated from the marginal proportions $p_{i.}$ and $p_{.i}$ of the cross-tabulation.

In the above, raters are only deemed to agree in their rating of an interaction if they record an identical QuIS category, and thus any ratings one point apart (for example ratings of + social and + care) are treated as disagreeing to the same extent as ratings a further distance apart (for example ratings of + social and - restrictive). To better reflect the severity of misclassification between pairs of QuIS categories weighted κ_w can be estimated as follows:

$$\hat{\kappa}_w = \frac{p_{o\ (w)} - p_{e\ (w)}}{1 - p_{e\ (w)}}, \tag{2}$$

where $p_{o\ (w)}$ is the proportion of participants observed to agree according to a set of weights w_{ij}

$$p_{o\ (w)} = \sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} p_{ij}, \tag{3}$$

and $p_{e\ (w)}$ is the proportion of participants expected to agree according to the weights

$$p_{e\ (w)} = \sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} p_{i.}\, p_{.j}. \tag{4}$$

In (3) p_{ij}, for i and $j = 1 \ldots k$, is the proportion of interactions rated as category i by the first rater and category j by the second. A weight w_{ij} is assigned to each combination restricted to lie in the interval $0 \leq w_{ij} \leq 1$. Categories i and j, $i \neq j$ with $w_{ij} = 1$, indicate a pair of ratings deemed to reflect perfect agreement between the two raters. Only if w_{ij} is set at zero, $w_{ij} = 0$, are the ratings deemed to indicate complete disagreement. If $0 < w_{ij} < 1$ for $i \neq j$, ratings of i and j indicate ratings deemed to agree to the extent indicated by w_{ij}. The precision of estimated κ_w from a sample of size n is indicated by the Wald 100(1- α)% confidence interval (CI):

$$\hat{\kappa}_w - z_{\alpha/2} \times SE(\hat{\kappa}_w) \leq \hat{\kappa}_w \leq \hat{\kappa}_w + z_{\alpha/2} \times SE(\hat{\kappa}_w). \tag{5}$$

Fleiss et al. ([6], section 13.1) give an estimate of the standard error of $\hat{\kappa}_w$ as:

$$\widehat{SE}(\hat{\kappa}_w) = \frac{1}{(1 - p_{e(w)})\sqrt{n}} \sqrt{\sum_{i=1}^{k} \sum_{j=1}^{k} p_{i.} p_{.j} [w_{ij} - (\bar{w}_{i.} + \bar{w}_{.j})]^2 - p_{e(w)}^2}, \tag{6}$$

where $\bar{w}_{i.} = \sum_{j=1}^{k} p_{.j} w_{ij}$ and $\bar{w}_{.j} = \sum_{i=1}^{k} p_{i.} w_{ij}$. Unweighted κ is a special case.

We examined the sensitivity of $\hat{\kappa}_w$ to the choice of weighting scheme. Firstly we considered two standard schemes (linear and quadratic) described by Fleiss et al. [6] and implemented in Stata. Linear weighting deems the severity of disagreement between raters by one point to be the same at each point on the scale, and the weighting for disagreement by more than one point is the weight for a one-point disagreement multiplied by the number of categories apart. In quadratic weighting, disagreements two or more points apart are not simple multiples of the one-point weighting, but are still invariant to position on the scale. We believe that the severity of disagreement between two QuIS ratings a given number of categories apart, does depend on their position on the scale. The weighting schemes we devised as better reflections of misclassification between QuIS categories are described in Table 2. In weighting schemes A1 to A6 the severity of disagreements between each positive category and neutral, and each negative category and neutral was weighted to be 0.5; disagreement within the two positive categories was considered to be as severe as that within the two negative categories; and we considered a range of levels of weights (0.5 to 0.9) to reflect this. In schemes B1 to B3 disagreements between each positive category and neutral, and between each negative category and neutral were considered to be equally severe, but were given weight less than 0.5 (0.33, 0.25 and 0.00 respectively); severity of disagreement within the two positive categories was considered to be the same as that within the two negative categories. While in weighting schemes C1-C3, disagreement between the two positive categories (+social and + care) was considered to be less severe than that between the two negative categories (−protective and -restrictive).

Weighting scheme A4 is proposed as a good representation of the severity of disagreements between raters based on the judgement of the clinical authors (CMcL, PG and JB) for the following reasons:

i) There is an order between categories + social > +care > neutral > −protective > −restrictive
ii) Misclassification between any positive and any negative category is absolute and should not be considered to reflect any degree of agreement

Table 2 Weighting schemes

Weighting scheme		+ s	+ c	N	- p	- r	COMMENTS
Unweighted	+ social	1					Ignores the degree of misclassification between categories
	+ care	0	1				
	Neutral	0	0	1			
	- protective	0	0	0	1		
	- restrictive	0	0	0	0	1	
Linear	+ social	1					Standard weights 1 for ordinal variables in Stata. Weights $1-\lvert i-j\rvert/(k-1)$, where i and j index the rows and columns, and k the number of categories
	+ care	1	1				
	Neutral	0.5	0.5	1			
	- protective	0	0	0.5	1		
	- restrictive	0	0	0.5	1	1	
Quadratic	+ social	1					Standard weights 2 for ordinal variables in Stata. Weights $1 - \{(i-j)/(k-1)\}^2$.
	+ care	0.75	1				
	Neutral	0.5	0.75	1			
	- protective	0.25	0.5	0.75	1		
	- restrictive	0	0.25	0.5	0.75	1	
A: Weights given to neutral compared to a positive or negative = 0.5, assuming that misclassification between the positives is equal to misclassification between the negatives.							
Weighted A1	+ social	1					All possibilities from weighting misclassification between the two positives and the two negatives as 1 (will be the same as having only three categories, positive neutral and negative) to weighting it as 0.6. Weighting scheme 4 has a weights of 0.75 (half way between .5 and 1)
	+ care	1	1				
	Neutral	0.5	0.5	1			
	- protective	0	0	0.5	1		
	- restrictive	0	0	0.5	1	1	
Weighted A2	+ social	1					
	+ care	0.9	1				
	Neutral	0.5	0.5	1			
	- protective	0	0	0.5	1		
	- restrictive	0	0	0.5	0.9	1	
Weighted A3	+ social	1					
	+ care	0.8	1				
	Neutral	0.5	0.5	1			
	- protective	0	0	0.5	1		
	- restrictive	0	0	0.5	0.8	1	
Weighted A4	+ social	1					
	+ care	0.75	1				
	Neutral	0.5	0.5	1			
	- protective	0	0	0.5	1		
	- restrictive	0	0	0.5	0.75	1	
Weighted A5	+ social	1					
	+ care	0.7	1				
	Neutral	0.5	0.5	1			
	- protective	0	0	0.5	1		
	- restrictive	0	0	0.5	0.7	1	

Table 2 Weighting schemes *(Continued)*

Weighted A6	+ social	1					
	+ care	0.6	1				
	Neutral	0.5	0.5	1			
	- protective	0	0	0.5	1		
	- restrictive	0	0	0.5	0.6	1	
Weighting scheme		+ s	+ c	N	- p	- r	COMMENTS
B: Weights using less than 0.5 for neutral compared to a positive or negative and assuming that misclassification between the two positives is equal to misclassification between the two negatives							
Weighted B1	+ social	1					
	+ care	0.66	1				
	Neutral	0.33	0.33	1			
	- protective	0	0	0.33	1		
	- restrictive	0	0	0.33	0.66	1	
Weighted B2	+ social	1					
	+ care	0.5	1				
	Neutral	0.25	0.25	1			
	- protective	0	0	0.25	1		
	- restrictive	0	0	0.25	0.5	1	
Weighted B3	+ social	1					
	+ care	0.5	1				
	Neutral	0	0	1			
	- protective	0	0	0	1		
	- restrictive	0	0	0	0.5	1	
C: Weights assuming that misclassification between the two negative categories is less important than misclassification between the two positives and varying the neutral weights							
Weighted C1	+ social	1					
	+ care	0.5	1				
	Neutral	0.25	0.25	1			
	- protective	0	0	0.25	1		
	- restrictive	0	0	0.25	0.75	1	
Weighted C2	+ social	1					
	+ care	0.6	1				
	Neutral	0.4	0.4	1			
	- protective	0	0	0.4	1		
	- restrictive	0	0	0.4	0.8	1	
Weighted C3	+ social	1					
	+ care	0.66	1				
	Neutral	0.5	0.5	1			
	- protective	0	0	0.5	1		
	- restrictive	0	0	0.5	0.83	1	

iii) The most important misclassifications are between positive (combined), neutral and negative (combined) categories

iv) There is a degree of similarity between neutral and the two positive categories, and between neutral and the two negative categories

v) Misclassification *within* positive and negative categories do matter, but to a lesser extent

Variation in $\hat{\kappa}_w$ over observation periods

We examined Spearman's correlation between A4 weighted $\hat{\kappa}_w$ and time of day, interactions/patient hour, mean length of interactions and percentage of interactions less than one minute. ANOVA and two sample t-tests were used to examine differences in A4 weighted $\hat{\kappa}_w$ between wards and between mornings and afternoons.

Overall $\hat{\kappa}_w$ combined over observation periods

To combine g (≥2) independent estimates of κ_w, we firstly considered the naive approach of collapsing over observation periods to form a single cross-tabulation containing all the pairs of QuIS ratings, shown in Table 3a). An estimate, $\hat{\kappa}_{w\ collapsed}$, and its 95% CI, can be obtained from formulae (2) and (6).

We next considered combining the g observation period specific estimates of κ_w using meta-analytic techniques. Firstly, using a fixed effects approach, the estimate $\hat{\kappa}_{wm} = \kappa_w + \varepsilon_m$ in the m^{th} observation period is modelled as comprising the true underlying value of κ_w plus a component, ε_m, reflecting sampling variability dependent on the number of interactions observed within the m^{th} period: where κ_w is the common overall value, and ε_m is normally distributed with zero mean and variance $V_{wm} = SE(\hat{\kappa}_{wm})^2$. The inverse-variance estimate of κ_w, based on the fixed effects model, $\hat{\kappa}_{w\ fixed}$, is a weighted combination of the estimates from each observation period:

$$\hat{\kappa}_{w\,fixed} = \frac{\sum_{m=1}^{g} \omega_m \times \hat{\kappa}_{wm}}{\sum_{m=1}^{g} \omega_m}, \tag{7}$$

with meta-analytic weights, ω_m, given by:

$$\omega_m = \frac{1}{V_{wm}}. \tag{8}$$

Since study specific variances are not known, estimates $\hat{\omega}_m$ with variance estimates $\hat{V}_{wm} = \widehat{SE}(\hat{\kappa}_{wm})^2$ calculated from formula (6) for each of the m periods are used. The standard error of $\hat{\kappa}_{w\,fixed}$ is then:

$$SE\left(\hat{\kappa}_{w\,fixed}\right) = \sqrt{\frac{1}{\sum_{m=1}^{g} \hat{\omega}_m}} \tag{9}$$

from which a 100(1- α)% CI for $\hat{\kappa}_{w\,fixed}$ can be obtained. $\hat{\kappa}_{w\,fixed}$ is the estimate $\hat{\kappa}_{w\,overall}$ combined over strata given by Fleiss et al. [6], here combining weighted $\hat{\kappa}_{wm}$ rather than unweighted $\hat{\kappa}_m$.

Table 3 Cross-tabulation of QuIS ratings collapsed over all observation periods, and for the observation periods with lowest and highest unweighted κ

		Rater 2						Unweighted κ
		+ s	+ c	N	- p	- r	Total	
a) Collapsed table from all observation periods								
Rater 1	+ social	36	23	0	0	0	59 (17%)	0.55
	+ care	22	164	10	4	1	201 (57%)	
	Neutral	3	13	47	2	5	70 (20%)	
	- protective	0	5	2	7	0	14 (4%)	
	- restrictive	3	1	0	0	6	10 (3%)	
	Total	64 (18%)	206 (58%)	59 (17%)	13 (4%)	12 (3%)	354 (100%)	
b) Observation period with lowest unweighted κ								
Rater 1	+ social	2	4	0	0	0	6	0.30
	+ care	1	9	2	0	1	13	
	Neutral	0	2	2	1	0	5	
	- protective	0	0	0	0	0	0	
	- restrictive	0	0	0	0	1	1	
	Total	3	15	4	1	2	25	
c) Observation period with highest unweighted κ								
Rater 1	+ social	1	0	0	0	0	1	0.90
	+ care	0	11	0	0	0	11	
	Neutral	0	0	6	0	0	6	
	- protective	0	0	1	0	0	1	
	- restrictive	0	0	0	0	0	0	
	Total	1	11	7	0	0	19	

Equality of the g underlying, observation period specific values of κ_w, is tested using a χ^2 test for heterogeneity:

$$\chi^2_{heterogeneity} = \sum\nolimits_{m=1}^{g} \omega_m \times \left(\hat{\kappa}_{wm} - \hat{\kappa}_{w\,fixed}\right)^2 \tag{10}$$

to be referred to χ^2 tables with $g - 1$ degrees of freedom. The hypothesis of equality in the g κ_{wm}s is typically rejected if $\chi^2_{heterogeneity}$ lies above the $\chi^2_{g-1}(0.95)$ percentile.

The fixed effects model assumes that all observation periods share a common value, κ_w, with any differences in the observation period specific $\hat{\kappa}_{wm}$ being due to sampling error. Because of our expectation that inter-rater reliability will vary depending on ward characteristics and other aspects of specific periods of observation, our preference is for a more flexible model incorporating underlying variation in true κ_{wm} over the m periods within a random effects meta-analysis. The random effects model has $\hat{\kappa}_{wm} = \kappa_w + \delta_m + \varepsilon_m$, where δ_m is an observation period effect, independent of sampling error (the ε_m terms defined as for the fixed effects model). Variability in observed $\hat{\kappa}_{wm}$ about their underlying mean, κ_w, is thus partitioned into a source of variation due to observation period characteristics captured by the δ_m terms, which are assumed to follow a Normal distribution: $\delta_m \sim N(0, \tau^2)$, with τ^2 the variance in κ_{wm} across observation periods, and sampling variability. The inverse-variance estimate of κ_w for this model is:

$$\hat{\kappa}_{w\,random} = \frac{\sum\nolimits_{m=1}^{g} \Omega_m \times \hat{\kappa}_{wm}}{\sum\nolimits_{m=1}^{g} \Omega_m}, \tag{11}$$

with meta-analytic weights, Ω_m, given by:

$$\Omega_m = \frac{1}{V_{wm} + \tau^2}. \tag{12}$$

Observation period specific variance estimates $\hat{V}_{wm}$ are used, and τ^2 also has to be estimated. A common choice is the Dersimonian-Laird estimator [7] defined as:

$$\hat{\tau}^2 = \frac{\chi^2_{heterogeneity} - (g-1)}{\sum\nolimits_{m=1}^{g} \omega_m - \left(\sum\nolimits_{m=1}^{g} \omega_m^2\right) / \left(\sum\nolimits_{m=1}^{g} \omega_m\right)} \tag{13}$$

usually truncated at 0 if the observed $\chi^2_{heterogeneity} < (g - 1)$. The estimate $\hat{\kappa}_{w\,random}$ is then:

$$\hat{\kappa}_{w\,random} = \frac{\sum\nolimits_{m=1}^{g} \hat{\Omega}_m \times \hat{\kappa}_{wm}}{\sum\nolimits_{m=1}^{g} \hat{\Omega}_m}, \tag{14}$$

with

$$\hat{\Omega}_m = \frac{1}{\hat{V}_{wm} + \hat{\tau}^2}, \tag{15}$$

and an estimate of the standard error of $\hat{\kappa}_{w\,random}$ is:

$$\widehat{SE}(\hat{\kappa}_{w\,random}) = \sqrt{\frac{1}{\sum\nolimits_{m=1}^{g} \hat{\Omega}_m}} \tag{16}$$

leading to 100(1- α)% CIs for $\hat{\kappa}_{w\,random}$.

The role of τ^2 is that of a tuning parameter: When $\tau^2 = 0$ there is no variation in the underlying κ_w, and the fixed effects estimate, $\hat{\kappa}_{w\,fixed}$ is obtained. At the other extreme, as τ^2 becomes larger, the $\hat{\Omega}_m$ become close to constant, so that each observation period is equally weighted and $\hat{\kappa}_{w\,random}$ becomes the simple average of observation period specific estimates:

$$\hat{\kappa}_{w\,averaged} = \frac{\sum\nolimits_{m=1}^{g} \hat{\kappa}_{wm}}{g}. \tag{17}$$

$\hat{\kappa}_{w\,averaged}$ ignores the impact of number of interactions on the precision of the observation period specific estimates. The standard error for $\hat{\kappa}_{w\,averaged}$ is estimated by:

$$\widehat{SE}\left(\hat{\kappa}_{w\,averaged}\right) = \sqrt{\frac{\sum\nolimits_{m=1}^{g} \hat{V}_{wm}}{g^2}}. \tag{18}$$

Obtaining estimates of $\hat{\kappa}_w$ from Stata

The inverse-variance fixed and random effects estimates can be obtained from command metan [8] in Stata by feeding in pre-calculated effect estimates (variable X1) and their standard errors (variable X2). When X1 contains the g estimates of $\hat{\kappa}_{wm}$, X2 their standard errors $\sqrt{\hat{V}_{wm}}$, and variable OPERIOD (labelled "Observation Period") an indicator of observation periods, inverse-variance estimates are obtained from the command:

```
metan X1 X2, second (random) lcols (OPERIOD) xlab(0,
0.2, 0.4, 0.6, 0.8, 1) effect(X1)
```

The "second(random)" option requests the $\hat{\kappa}_{w\,random}$ estimate in addition to $\hat{\kappa}_{w\,fixed}$. The "lcols" and "xlab" options control the appearance of the Forest plot of observation specific estimates, combined estimates, and their 95% CIs.

Results

Across the 18 observation periods 447 interactions were observed, of which 354 (79%) were witnessed by both raters and form the dataset from which inter-rater reliability was estimated. The ICC for the total number of interactions recorded by each rater for the same observation period was high (ICC = 0.97: 95%CI: 0.92 to 0.99, $n = 18$). The occasional absence of patients from ward areas for short periods of time resulted in interactions being recorded for 67 patient hours (compared to the planned 72 h). The mean rate of interactions was 6.7 interactions/patient/hour. More detailed results are given by McLean et al. [5].

In Table 3a) the cross-tabulation of ratings by the two raters can be seen collapsed over the 18 observation periods. Two specific observation periods are also shown: in 3b) the period demonstrating lowest unweighted $\hat{\kappa}$ ($\hat{\kappa}$ = 0.30); and in 3c) the period demonstrating highest unweighted $\hat{\kappa}$ ($\hat{\kappa}$ =0.90). From 3a) it can be seen that the majority of interactions are rated to be positive, between 17% and 20% are rated to be neutral, and 7% as negative (from the margins of the table), and this imbalance in the marginal frequencies would be expected to reduce chance adjusted κ.

Scatterplots of A4 weighted $\hat{\kappa}_{wm}$ against observation period characteristics are shown in Fig. 1. One of the characteristics (interactions/patient/hour) was sufficiently associated with A4 weighted $\hat{\kappa}_{wm}$ to achieve statistical significance ($P = 0.046$).

In Table 4 it can be seen that the various combined estimates of κ_w did not vary greatly depending on the method of meta-analysis or on the choice of weighting scheme. However, there was greater variability in $\chi^2_{heterogeneity}$. For all weighting schemes except unweighted, B2, B3, and C1, there was statistically significant heterogeneity by virtue of $\chi^2_{heterogeneity}$ exceeding the $\chi^2_{17}(0.95)$ cut-point of 27.59.

Figure 2 shows the Forest plot demonstrating the variability in $\hat{\kappa}_{wm}$ over observation periods, $\hat{\kappa}_{w\ fixed}$, and $\hat{\kappa}_{w\ random}$, for the A4 weighting scheme. Estimate $\hat{\kappa}_{w\ fixed}$ and its 95% CI is shown below observation specific estimates to the right of the plot, on the line labelled "I-V Overall". The line below labelled "D+L Overall" presents $\hat{\kappa}_{w\ random}$ and its 95% CI. Both estimates are identical to those shown in Table 4. The final column "% Weight (I-V)" relates to the meta-analytic weights, $\hat{\omega}_m$, not the A4 weighting scheme adopted for κ_w.

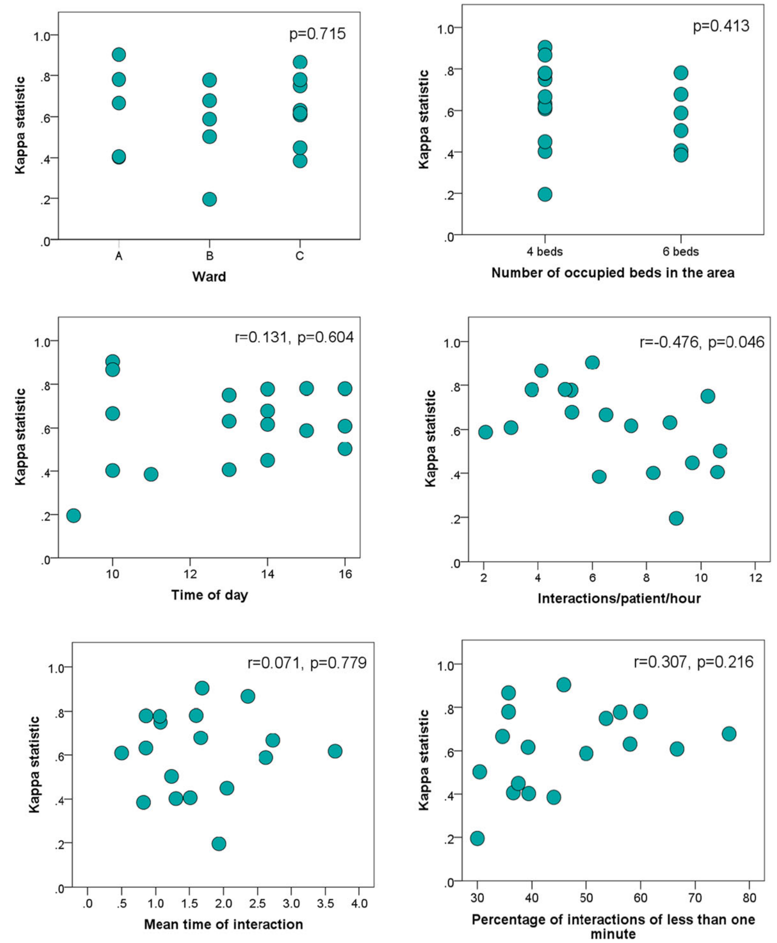

Fig. 1 Variability of A4 weighted $\hat{\kappa}_{wm}$ in relation to observation period characteristics ($n = 18$). P values relate to Spearman's correlation

Table 4 Combined estimates of κ_w with different weighting schemes

Weighting scheme	$\hat{\kappa}_{w\ collapsed}$ (95% CI)	$\hat{\kappa}_{w\ fixed}$ (95% CI)	$\chi^2_{heterogeneity}$	$\hat{\kappa}_{w\ random}$ (95% CI)	$\hat{\kappa}_{w\ averaged}$ (95% CI)
Unweighted	0.55 (0.49, 0.62)	0.52 (0.45, 0.59)	21.20	0.53 (0.45, 0.60)	0.57 (0.48, 0.65)
Linear	0.58 (0.51, 0.65)	0.52 (0.45, 0.59)	35.67	0.56 (0.46, 0.66)	0.59 (0.51, 0.68)
Quadratic	0.61 (0.50, 0.71)	0.53 (0.44, 0.62)	38.71	0.59 (0.45, 0.74)	0.63 (0.52, 0.73)
A1	0.64 (0.56, 0.73)	0.51 (0.43, 0.59)	47.15	0.62 (0.48, 0.77)	0.66 (0.57, 0.75)
A2	0.62 (0.54, 0.70)	0.50 (0.43, 0.57)	45.75	0.60 (0.47, 0.73)	0.64 (0.54, 0.73)
A3	0.60 (0.53, 0.68)	0.51 (0.44, 0.58)	39.28	0.58 (0.47, 0.69)	0.62 (0.53, 0.71)
A4	0.60 (0.53, 0.67)	0.51 (0.44, 0.58)	36.04	0.57 (0.47, 0.68)	0.61 (0.52, 0.70)
A5	0.59 (0.52, 0.66)	0.52 (0.45, 0.59)	33.22	0.56 (0.46, 0.67)	0.60 (0.52, 0.69)
A6	0.58 (0.51, 0.64)	0.52 (0.45, 0.59)	29.10	0.55 (0.46, 0.64)	0.59 (0.51, 0.67)
B1	0.59 (0.53, 0.66)	0.53 (0.46, 0.59)	30.52	0.56 (0.47, 0.66)	0.60 (0.52, 0.69)
B2	0.58 (0.51, 0.65)	0.53 (0.46, 0.59)	26.01	0.55 (0.46, 0.64)	0.59 (0.51, 0.67)
B3	0.59 (0.53, 0.66)	0.53 (0.47, 0.60)	25.11	0.55 (0.47, 0.64)	0.60 (0.51, 0.68)
C1	0.58 (0.51, 0.65)	0.53 (0.46, 0.59)	26.05	0.55 (0.46, 0.64)	0.59 (0.51, 0.67)
C2	0.58 (0.51, 0.65)	0.52 (0.45, 0.59)	28.82	0.55 (0.46, 0.65)	0.60 (0.51, 0.68)
C3	0.58 (0.51, 0.65)	0.52 (0.45, 0.59)	31.26	0.56 (0.46, 0.66)	0.60 (0.51, 0.68)
min-max $\hat{\kappa}_w$ across weighting schemes	0.55–0.64	0.50–0.53	$\chi^2_{17}(0.95)$ =27.59	0.53–0.62	0.57–0.66

Discussion

We consider the most appropriate estimate of inter-rater reliability of the QuIS to be 0.57 (95% CI 0.47 to 0.68) indicative of only moderate inter-rater reliability. The finding was not unexpected, the QuIS categories can be difficult to distinguish and though positioned as closely together as possible, the two raters had different lines of view, potentially impacting on their QuIS ratings. The estimate of inter-rater reliability is based on our A4 weighting scheme with observation specific estimates

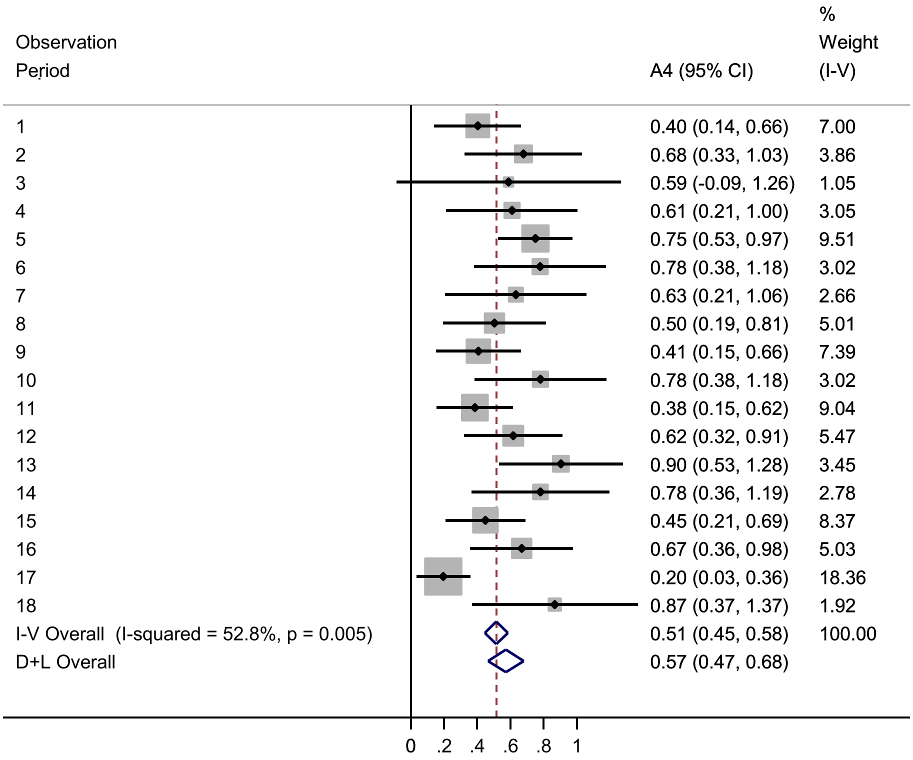

Fig. 2 Forest plot showing observation period specific A4 weighted $\hat{\kappa}_{wm}$, $\hat{\kappa}_{w\ fixed}$, and $\hat{\kappa}_{w\ random}$

combined using random effects meta-analysis. Combined estimates of κ_w were not overly sensitive to the choice of weighting scheme amongst those we considered as plausible representations of the severity of misclassification between QuIS categories. We recommend a random effects approach to combining observation period specific estimates, $\hat{\kappa}_{wm}$, to reflect the inherent variation anticipated over observation periods.

There are undoubtedly other weighting schemes that fulfil all the criteria on which we chose weighting scheme A4, but the evidence from our analyses suggests that it makes relatively little difference to the resultant $\hat{\kappa}_{w\ random}$. In the absence of any other basis for determining weights, our scheme A4 has the virtue of simplicity. A key issue is that researchers should not examine the $\hat{\kappa}_w$ resulting from a variety of weighting schemes, and then choose the scheme giving highest inter-rater reliability. The adoption of a standard set of weights also facilitates comparison of inter-rater reliability across different studies of QuIS.

We compared four approaches to estimating overall κ_w. We do not recommend the simplest of these, $\hat{\kappa}_{w\ collapsed}$, based on estimating κ_w from the cross-tabulation of all ratings collapsed over observation periods: generally collapsing involves a risk of confounding by stratum effects. Comparing the remaining estimates it can be seen that $\hat{\kappa}_{w\ random}$ lies between the fixed effects, $\hat{\kappa}_{w\ fixed}$, and the averaged estimate, $\hat{\kappa}_{w\ averaged}$, for all the weighting schemes we considered. $\hat{\kappa}_{w\ averaged}$ gives equal meta-analytic weight to each observation period, and thus up-weights periods with highest variance compared to $\hat{\kappa}_{w\ fixed}$. The observation periods with highest variance are those with fewest interactions/patient/hour of observation, and it can be seen from Fig. 1 that these periods tend to have highest $\hat{\kappa}_{wm}$. A possible explanation being that with fewer interactions it is easier for observers to see and hear the interactions and thus make their QuIS ratings which would be anticipated to result in more accuracy and agreement. Thus $\hat{\kappa}_{w\ averaged}$ might be expected to over-estimate inter-rater reliability and should be avoided. We recommend a random, rather than fixed effects approach to combining because variation in κ_{wm} across observation periods was anticipated. Observation periods were chosen with the intention of representing the broad range of situations in which staff-inpatient interactions take place. At different times of day staff will be more or less busy, and this more or less guarantees heterogeneity in observation period specific inter-rater reliability.

Böhning et al. [9] identified several practical issues relating to inverse variance estimators in meta-analysis. For example and most importantly, that estimation is no longer unbiased when estimated rather than known variances are used in the meta-analytic weights. This bias is less extreme for larger sample sizes in each constituent study. We included 354 interactions across the 18 observation periods, on average about 20 per period, but it is not clear whether this is sufficient for meaningful bias to be eradicated. A further issue relates to possible misunderstanding of the single combined estimate as applying to all observation periods: a correct interpretation being that the single estimate relates to the mean of the distribution of κ_{wm} over observation periods. An alternative might be to present the range of values that κ_w is anticipated to take over most observation periods. This would be an unfamiliar presentation for most researchers.

Meta-analysis of $\hat{\kappa}$ over studies following a systematic review has been considered by Sun [10] where fixed and random effects approaches are described, but the latter adopting the Hedges [11], rather than the conventional Dersimonian-Laird estimate of τ^2. Alternatives to the DerSimonian-Laird estimator are available including the REML estimate, or the Hartung-Knapp-Sidik-Jonkman method [12]. Friede et al. [13] examine properties of the DerSimonian-Laird estimator when there are only two observation periods and conclude that in such circumstances other estimators are preferable: McLean et al's study [5] was based on sufficient observation periods to make these problems unlikely. Sun addressed the issue of publication bias amongst inter-rater reliability studies found by searching the literature. Here we included data from all observation periods, irrespective of the estimate $\hat{\kappa}_{wm}$. Sun performed subgroup analyses of studies according to the degree of training of the raters involved, and also drew a distinction between inter-rater reliability studies where both raters can be considered to be equivalent and a study [14] comparing ratings from hospital nurses with those from an expert which would more appropriately have been analysed using sensitivity, specificity and related techniques. The QuIS observations were carried out by raters who had all received the training developed by McLean et al: though there was variation in experience of QuIS a further source of inter-rater unreliability relating to the different lines of view from each rater's position was also considered to be important.

In the inter-rater study we describe, in some instances the same rater was involved in more than one observation period, and this potentially violates the assumption of independence across observation periods, which would be anticipated to lead to increased variance in an overall estimate, $\hat{\kappa}_w$. A random effects approach is more suitable in this regard as it catches some of the additional variance, coping with extra-dispersion whether it arises from unobserved heterogeneity or from correlation across observation periods.

Though we have considered analysis choices that need to be made when summarising information on the inter-rater reliability of the QuIS, the issues we address are relevant to inter-rater reliability studies more generally. Firstly, where weighted κ_w rather than unweighted κ is thought to be a better summary of differing degrees of disagreement between raters, it is important that the weighting scheme be decided in advance. Secondly, where a study comprises distinct subsets of data collection, the method of combining information needs to be considered. It is likely that data in larger inter-rater reliability studies would need to be collected in distinct phases, but the lack of attention to combining $\hat{\kappa}_m$ over subsets within a study suggests that researchers often ignore the issue, adopting the easiest approach of collapsing to obtain a single estimate of κ. We would advise taking account of structure in data collection by either a fixed or random effects meta-analysis approach, the latter being appropriate where variation across subsets is anticipated or plausible. Our example dataset illustrates a potential source of bias in the simple average of subset specific estimates, $\hat{\kappa}_m$. Finally, in the context of meta-analysis over studies, Sun considered the issue of bias arising from the selection of studies for publication. In the context of combining over subsets of data collection within a study, it is possible to imagine circumstances where authors might choose to omit selected subsets, but a good reason would have to be given to justify such a step and the omitted data described.

Conclusions

Researchers using the QuIS to evaluate the quality of staff/inpatient interactions should check its suitability in new settings, and (possibly as part of staff training) its inter-rater reliability. In practice such studies are likely to follow a similar protocol to that adopted by McClean et al.: involving the multiple observers to be employed in a subsequent main study, over a variety of wards similar to those planned for the main study; and preferably taking place at different times of day. We recommend inter-rater reliability be estimated using our A4 weighting scheme and a random effects meta-analytic approach to combining estimates over observation periods, $\hat{\kappa}_{w\ random}$, be adopted. The $\hat{\kappa}_{w\ random}$ estimate should be presented with its 95% confidence interval reflecting precision of estimation achieved from the available number and length of observation periods.

Additional file

Additional file 1: Table S1. Cross-classification of ratings for each of the 18 observation periods and period specific covariates[1]. (DOCX 36 kb)

Abbreviations

κ: Unweighted kappa; κ_w: Weighted kappa; ICC: Intra-cluster correlation; QuIS: Quality of Interaction Schedule

Acknowledgements

The authors would like to thank staff and patients at the participating NHS hospitals (staff being observed as well as staff raters).

Funding

The analysis presented here is based on data collected during research funded by the National Institute for Health Research (NIHR) Collaboration for Leadership in Applied Health Research and Care Wessex, the NHS South Central through funding of a clinical lecturer internship undertaken by CMcL and the National Institute for Health Research Health Services and Delivery Research programme. The views and opinions expressed therein are those of the authors and do not necessarily reflect those of the Health Services and Delivery Research programme, NIHR, NHS or the Department of Health.

Availability of data and materials

The inter-rater reliability data analysed in this study is shown in the Additional file 1: Table S1.

Author's contributions

IME carried out statistical analysis. RMP conceived of the paper. DB contributed to the statistical methodology. CMcL, PG and JB carried out the studies from which the dataset was drawn. RMP and IME drafted the manuscript and all authors approved the final version.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Ethical approval for the QuIS inter-rater reliability study was obtained from Oxford 'B' Research Ethics Committee (Reference: 14/SC/1100). Written consent was obtained from patients prior to conducting QuIS observation, and the presence of observers was also explained to non-participating patients and visitors in the vicinity. All patient information was anonymised. Staff were made aware of the study through discussion at team meetings, and through the provision of posters and information sheets sent via email as well as being available in hard copy. Staff present at the time of observations were given opportunity to ask questions and/or decline to participate.

Author details

[1]Medical Statistics Group, Faculty of Medicine, Southampton General Hospital, Mailpoint 805Level B, South Academic Block, Southampton SO16 6YD, UK. [2]Southampton Statistical Sciences Research Institute & Mathematical Sciences, University of Southampton, Southampton, UK. [3]Faculty of Health Sciences, University of Southampton, Southampton, UK.

Received: 1 July 2016 Accepted: 16 November 2016
Published online: 07 December 2016

References

1. Clark P, Bowling A. Observational Study of Quality of Life in NHS Nursing Homes and a Long-stay Ward for the Elderly. Ageing Soc. 1989;9:123–48.
2. Dean R, Proundfoot R, Lindesay J. The quality of interaction schedule (QUIS): development, reliability and use in the evaluation of two domus units. Int J Geriatr Psychiatry. 1993;8(10):819–26.
3. Skea D. SPECIAL PAPER. A Proposed Care Training System: Quality of Interaction Training with Staff and Carers. Int J Caring Sci. 2014;7(3):750–6.
4. Barker HR, Griffiths P, Mesa-Eguiagaray I, Pickering R, Gould L, Bridges J. Quantity and quality of interaction between staff and older patients in UK hospital wards: A descriptive study. Int J Nurs Stud. 2016;62:100–7. doi:10.1016/j.ijnurstu.2016.07.018.
5. McLean C, Griffiths P, Mesa-Eguiagaray I, Pickering RM, Bridges J. Reliability, feasibility, and validity of the quality of interactions schedule (QUIS) in acute

hospital care: an observational study. BMC Health Services Research. (Submitted - January 2016).

6. Fleiss JL, Levin B, Paik MC. Statistical methods for rates and proportions: third edition. John. Hoboken. New Jersey: Wiley & Sons; 2003.
7. Dersimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials. 1986;7:177–1.
8. Harris RJ, Bradburn MJ, Deeks JJ, Harbord RM, Altman DG, Sterne J. metan: fixed- and random-effects meta-analysis. Stata J. 2008;8(1):3–28.
9. Böhning D, Malzahn U, Dietz E, Schlattmann P. Some general points in estimating heterogeneityvariance with the DerSimonian-Laird estimator. Biostatistics. 2002;3:445–57.
10. Sun S. Meta-analysis of Cohen's kappa. Health Serv Outcome Res Methodol. 2011;11:145–63.
11. Hedges LV. A random effects model for effect sizes. Psychol Bull. 1983;93: 388–95.
12. IntHout J, Ionnidis JPA, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Med Res Methodol. 2014;14:2–12. http://www.biomedcentral.com/1471-2288/14/25.
13. Friede T, Röver C, Wandel S, Neuenschwander B. Meta-analysis of two studies in the presence of heterogeneity with applications in rare diseases. Biometrical Journal 2016 (in press).
14. Hart S, Bergquist S, Gajewski B, Dunton N. Reliability testing of the national database of nursing quality indicators pressure ulcer indicator. J Nurs Care Qual. 2006;21:256–65.