PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Measurement of patient safety: systematic review of the reliability and validity of adverse event detection with record review
AUTHORS	Hanskamp-Sebregts, Mirelle; Zegers, Marieke; Vincent, Charles; van Gurp, Petra J.; de Vet, Riekie; Wollersheim, Hub

VERSION 1 - REVIEW

REVIEWER	A. Rosemary Tate University of Sussex UK
REVIEW RETURNED	11-Mar-2016

GENERAL COMMENTS	An interesting study and clearly a lot of work. However, the statistical methods and results need clearer explanation and justification, and possibly a different approach. For example were three separate Ancovas carried out or one, and if 3 why? What were the response and covariates? How was the influence of the aim of the study and type of instrument assessed? Were the assumptions for Ancova tested and met? What was the justification for categorising number of reviewers etc., when the numbers are (presumably) available. I suggest that the authors consult a statistician for further help with the analysis and also help with the tables. For example Table 4 should report medians and IQRs as some of the variables are obviously highly skewed. A few more detailed comments page 5 2nd section. Please define what you mean by "full text" p. 7 Data extraction - please clarify what is meant by this and how it differs from data abstraction p 11 2nd para. what criteria was used to assess "relatively good" p11. second section. Please justify Kappa classification as fair good etc. The paper referred to is very old and according to some, quite subjective. p. 12 2nd section. How did you get the p-val for prevalence and kappa - was it correlation analysis or one of the ancovas? I suggest that correlation coeff would be best and give the N and coefficient. Either spearman or Pearson as appropriate. Ditto reviewer experience etc. p.13 tables. Please indicate how the p-vals and SD etc were obtained. Table 2 most of the variables are very skew so mean sd not appropriate.

REVIEWER	Alex McConnachie Robertson Centre for Biostatistics, University of Glasgow, Scotland, UK
REVIEW RETURNED	14-Apr-2016

GENERAL COMMENTS	Hanskamp-Sebregts present a systematic review and meta analysis of the reliability of two record review tools for the measurement of adverse events. This review considers the statistical elements of the paper. Overall, the study is a good idea, but from a statistical perspective, I feel that there are several areas where improvements are required. It is not entirely clear from the methods, but I can only assume that no allowance is made for study size / precision when combining the results. We are provided with mean kappa and mean percentage agreement across studies, but these mean values should be weighted in some way, otherwise the smallest and largest studies make equal contributions to the final results. It would be normal to present the results in the form of a forest plot, so the reader can see how the different studies contribute to the final result. Also, it would be usual to consider heterogeneity between studies, and also to use the data to address the issue of publication bias. This latter point is mentioned, but no more. To compare subgroups of studies, the meta analysis should be extended to a meta regression, rather than using ANOVA to simply compare the mean values between groups of studies. This is the main issue with the analysis as it stands. The following are more minor points. Page 11 states that the quality of the studies was "relatively good" - relative to what? Face validity is a vague measure, and unlikely to be reported as being low, so a systematic review would appear of little value. Saying that, it was reported in only one paper. The fact that there are no studies looking at concurrent validity is alarming, given the statements that these tools are the "best" means of measuring AEs. Even if the inter-rater reliability of the these tools is acceptable, there is no evidence that they are measuring anything of value. Surely, with electronic patient records, there must be more efficient ways of measuring AEs than manual record review. Should this be included in the discussion? Should a subgroup analysis be included looking at study quality? Appendix 3 - the rows looking at sensitivity, specificity, PPV and NPV are redundant, since no relevant studies were identified. Also, I feel they are not correct, since acceptable values will always depend on the context. Similarly, I do not think it is correct to use thresholds for p-values. I would remove this section.

VERSION 1 – AUTHOR RESPONSE

Reviewer 1

Were three separate Ancovas carried out or one, and if 3 why? What were the response and covariates?

Author's response

We were interested in the influence of three factors/independent variables (number of reviewers, reviewer experience and reviewer training) on the kappa values, with prevalence as covariate. Putting all these variables into one ANCOVA, there is not enough power to get reliable results with the number of the included studies (n = 20). Therefore, we were forced to carry out three separate ANCOVAs.

The dependent variable ("response") is the kappa value, and prevalence of AE was used as covariate in all analyses.

We made the following edits in text on page 8, lines 147-154:

"We performed three separate ANCOVAs, with prevalence of AE as covariate, to study the influence of the number of reviewers, reviewer experience and reviewer training on the inter-rater reliability (kappa). We adjusted for prevalence of AEs, since a previous study of Lilford et al. (2007) showed correlation between prevalence and kappa.16 It was not possible to incorporate all variables (the number of reviewers, reviewer experience and reviewer training) in one ANCOVA, because the number of studies in our analyses was limited (n = 20)."

How was the influence of the aim of the study and type of instrument assessed?

Author's response

We have analysed the influence of the aim of the study and the type of instrument with an ANCOVA adjusted for prevalence.

We add on page 8, line 159 at the end of the sentence the words 'with an ANCOVA':

"Additionally, we studied the influence of the aim of the study and the type of instrument (GTT versus HMPS) on kappa with two separate ANCOVAs adjusted for prevalence."

Were the assumptions for Ancova tested and met?

Author's response

The assumptions for ANCOVA were tested and met.

The kappa value was normal distributed: skewness: p = 0.536 and kurtosis: p = -0.595.

The significance levels in the three separate ANCOVAs were p > 0.05. Therefore we did not reject the null hypothesis that the error variance of the dependent variable is equal across groups.

Levene's Test of Equality of Error Variances:

- *Design: Intercept + Prevalence rate AE + Group of reviewers - Weighted by No_records_kappa_calculation: F(2,17) = 0.707; p = 0.507*
- *Design: Intercept + Prevalence rate AE + Reviewer experience - Weighted by No_records_kappa_calculation: F(2,17) = 0.395; p = 0.680*

- *Design: Intercept + Prevalence rate AE + Training - Weighted by No_records_kappa_calculation: F(2,10) = 0.769; p = 0.489*

There was no interaction of the covariate 'prevalence of AEs' with the independent variables:

1) Number of reviewers: p = 0.807;

2) Reviewer experience: p = 0.455;

3) Reviewer training: p = 0.404;

4) Instrument: p = 0.342;

Except:

*5) Study aim * Prevalence of AEs: p = 0.034[1]; whereby Study aim: p = 0.056.*

*The interaction 'Study aim * Prevalence' does not raise concern about de main analyses: the influence of number of reviewers, reviewer experience and reviewer training on kappa.*

[1] Note: This finding is not surprising, because the mean prevalence of AEs of the study aim 'measuring the inter rater reliability' is 20.2% (SD 7.0%) and the mean prevalence of AEs of the study aim 'detecting AEs' is 11.5% (SD 4.6%).

We add in text on page 8, line 153:

"The assumptions for ANCOVA were tested and met."

What was the justification for categorising number of reviewers etc., when the numbers are (presumably) available.

Author's response

We were interested in the explanation of the variation of the kappa values: "Which factors influenced the kappa value?" The aim of our analyses was explorative. Looking at the study characteristics, there were also differences in the number of reviewers, the number of reviewed records and training hours before the start of the review process. Classifying these factors/independent variables into three proportional classes we were able to answer our research question and to better interpret the results.

We added in the sentence on page 8 line 154: In order to better interpret the results, we classified ….. > 1 day training respectively.

I suggest that the authors consult a statistician for further help with the analysis and also help with the tables. For example Table 4 should report medians and IQRs as some of the variables are obviously highly skewed.

Author's response

We were assisted by a statistician (see acknowledgments text on page 18) and co-author professor Henrica C.W. de Vet is a biostatistician. She checked the analyses and provided advices of the content of the tables. Both provided advices in the revised manuscript again.

We made edits in Table 4 on page 15 by reporting the medians and IQRs of reviewer experience and training hours. The mean and SD of the prevalence of AE are kept unchanged.

p.13 tables. Please indicate how the p-vals and SD etc were obtained. Table 2 most of the variables are very skew so mean sd not appropriate.

We reported the medians en recalculated p values on pages 14, lines 256 - 260: "This group received the least training (median 6 hours) and assessed the largest number of records (median 213 records). There was no significant difference in the reviewer experience (p = 0.351), the reviewer training (p = 0. 317) and the prevalence of AEs (p = 0.480) between the three groups of reviewers (max 5 reviewers, >5–20 reviewers, >20 reviewers)."

We assume that the reviewer means Table 4, because there is no Table 2 on page 13. Therefore, we made edits in Table 4 on page 15.

Page 5 2nd section. Please define what you mean by "full text"

Author's response

We mean by "full text" that we read and reviewed the complete article.

We added on page 5, line 99: (meaning the complete article)

When the title and abstract did not clearly indicate whether the inclusion criteria were met, the full text (meaning the complete article) was obtained and reviewed by two researchers (MH, MZ).

p. 7 Data extraction - please clarify what is meant by this and how it differs from data abstraction p 11 2nd para.

Author's response

On page 7 we described which data were extracted from each article. The extracted data included the study objective, study population, design and methods and the results of the analysis including statistical parameters about reliability and validity of record review, such as the % inter rater agreement and kappa values. There is no difference between the data extraction on page 7 and the data analysis on pages 12 - 14.

We made no edits in the text.

What criteria was used to assess "relatively good" p11. second section.

Author's response

We removed on page 12, line 214 the word "relatively".

Please justify Kappa classification as fair good etc. The paper referred to is very old and according to some, quite subjective.

Author's response

We refer to the Kappa classification of Landis and Koch (1977), which is the most often used kappa classification.

We made no edits in the text.

p. 12 2nd section. How did you get the p-val for prevalence and kappa - was it correlation analysis or one of the ancovas? I suggest that correlation coeff would be best and give the N and coefficient. Either spearman or Pearson as appropriate. Ditto reviewer experience etc.

Author's response

The p-value (p = .308) on page 13, line 250 was provided with an ANCOVA.

We changed the sentence on page 13 line 249:

" Prevalence was not statistically significant associated with the kappa values (p = 0.069, p = 0.189 and p = 0.726 respectively)."

Reviewer 2

It is not entirely clear from the methods, but I can only assume that no allowance is made for study size / precision when combining the results. We are provided with mean kappa and mean percentage agreement across studies, but these mean values should be weighted in some way, otherwise the smallest and largest studies make equal contributions to the final results. It would be normal to present the results in the form of a forest plot, so the reader can see how the different studies contribute to the final result. Also, it would be usual to consider heterogeneity between studies, and also to use the data to address the issue of publication bias. This latter point is mentioned, but no more.

Author's response

We agree with the reviewer that pooling the kappa values weighted (by 1/var or n) is better than taking the mean value without weighing. We missed information about the 95% confidence interval (CI) of the kappa values in the included studies. Therefore, we could not pooling the kappa values weighted by 1/var, present our results in a forest plot and extend our analysis with a meta regression.

A proxy for accuracy, we used the number of records on which the kappa value is calculated as weighting factor in the ANCOVAs. The conclusions of our analyses do not change.

We added on pages 7 and 8, lines 143-146 the sentence:

"We used in the meta- and subgroup analyses a proxy for accuracy, being the weighting factor consisting of the number of records on which the kappa value is calculated, since we missed information about the 95% confidence intervals (CI) of the kappa values in the included studies."

To compare subgroups of studies, the meta analysis should be extended to a meta regression, rather than using ANOVA to simply compare the mean values between groups of studies.

Author's response

We reported the recalculated p values on pages 13 and 14, lines 252-253: "We found a statistically significant association between the number of reviewers and the pooled kappa values, p = 0.006 (Table 3). There was no association between reviewer experience (p = 0.062) and reviewer training (p = 0.809) versus kappa."

The recalculated p values are reported in Table 3 on page 14.

According to the calculation of the pooled kappa, we reported the pooled kappa of the Global Trigger Tool (0.65, SD 0.19) and the Harvard Medical Practice Study (0.55, SD 0.07) respectively in the abstract line 39 (pooled kappa without SD), on page 12 line 225 and on page 13 lines 241-242.

Page 11 states that the quality of the studies was "relatively good" - relative to what?

Author's response

We removed on page 12, line 214 the word "relatively".

Face validity is a vague measure, and unlikely to be reported as being low, so a systematic review would appear of little value. Saying that, it was reported in only one paper.

Author's response

Indeed, face validity is subjective. We believe, however, that subjective expert opinion is important. However, face validity is only reported in one paper.

We made no edits in the text.

The fact that there are no studies looking at concurrent validity is alarming, given the statements that these tools are the "best" means of measuring AEs. Even if the inter-rater reliability of the these tools is acceptable, there is no evidence that they are measuring anything of value.

Author's response

We agree!! The fact that there are no studies looking at concurrent validity is alarming, given the statements that these tools are the "best" means of measuring AEs is the main message of our manuscript. Therefore, we recommend and make suggestions in the discussion section to test the concurrent validity of record review.

We replaced the sentence: " Although record review is accepted worldwide as the most important and valid method for determining the incidence rates of AE 15 62, concurrent validity has never been evaluated." on page 15, line 281 with the formulation of the reviewer:

"The fact that there are no studies looking at concurrent validity is alarming, given the statements that record review is accepted worldwide as the "best" means of measuring incidence rates of AEs (even called 'the gold standard'). Even if the inter-rater reliability of record review is acceptable, there is no evidence that record review really detects AEs.

Surely, with electronic patient records, there must be more efficient ways of measuring AEs than manual record review. Should this be included in the discussion?

Author's response

We focused in our systematic review on the reliability and validity of record review. The efficiency of methods that are able to measuring AEs is another important research question which should be addressed in a future study.

We made no edits in the text.

Should a subgroup analysis be included looking at study quality?

Author's response

We assessed the quality of the studies using the COSMIN checklist. There was too little differentiation among the studies to justify creating subgroups on study quality.

We made no edits in the text.

Appendix 3 - the rows looking at sensitivity, specificity, PPV and NPV are redundant, since no relevant studies were identified. Also, I feel they are not correct, since acceptable values will always

depend on the context. Similarly, I do not think it is correct to use thresholds for p-values. I would remove this section.

Author's response

We agree and we removed Appendix 3 including the references 23, 24 and 26. We updated the numbers of the appendices, the reference list (see 'Manuscript_revised_clean') and the PRISMA 2009 Checklist. We updated the rating in Appendices 4 and 5 according with the classification as described on page 7 line 136 – 140.

We changed on page 7 line 136 – 140 the sentence in:

"To be able to rate the reliability of record review, we classified the kappa values as "slight" (Ķ: 0.00-0.20), "fair" (Ķ: 0.21-0.40), "moderate" (Ķ: 0.41-0.60), "substantial"(Ķ: 0.61-0.80) and "almost perfect" (Ķ: 0.81-1.00)(Appendices 4 and 5).23"

Based on this classification, we changed line 42 in the abstract: "The reliability of record review is moderate to substantial and improved when a small group of reviewers carried out record review."

VERSION 2 – REVIEW

REVIEWER	Rosemary Tate University of Sussex, UK
REVIEW RETURNED	08-Jun-2016

GENERAL COMMENTS	The statistical methods still have not been clearly explained and, as far as I can comprehend, do not appear to be appropriate for the aims of this study. For example, there is no evidence of a meta-analysis being carried out, and nowhere is it properly explained how the weights are employed. If it is not possible to produce a forest plot, then there is no meta analysis. Another example: the revised table 4 (with medians as requested) clearly shows a large difference in experience between the n<=5 group and the other two groups, suggesting that the wrong (maybe parametric?) test has been used to obtain the p-values; Correlation analysis should be used to investigate relationships; tables should be much clearly labelled and details of how the p-vals are obtained should be detailed in the caption. Tables 3 and 4 appear to contradict each other. Table 3 shows that the best agreement (kappa=0.8) is when N<6 and when experience is least (<100) (kappa = 0.71). However, Table 4 (and the text) shows that the reviewers with n<6 have the most experience (with median more than double that of the next largest group). This is difficult to comprehend. I do have quite a few more comments, but in my opinion at present there are too many flaws in the analysis and reporting (and also probably erroneous conclusions) for this to merit publication.

REVIEWER	Alex McConnachie Robertson Centre for Biostatistics University of Glasgow Scotland
REVIEW RETURNED	12-Jun-2016

GENERAL COMMENTS	I would like to thank Hanskamp-Sebregts et al for their responses to my original comments, and I am satisfied with the changes that have been made. One final amendment could be to add a sentence to the study limitations, along the lines that due to limitations in the reporting of reliability statistics in the source papers, namely the lack of confidence intervals, it was not possible to formally estimate the pooled kappa statistics for the two measurement tools, to assess between-study heterogeneity, or to carry out analyses of the likelihood of publication bias. I think it is worth making the point that when reporting studies (of any type) it is important to include confidence intervals for estimated quantities, both to allow proper interpretation of individual study results, but also to allow others to use those results in future research.

VERSION 2 – AUTHOR RESPONSE

Comments of the reviewers	**Author's response**
Reviewer 1	
The statistical methods still have not been clearly explained and, as far as I can comprehend, do not appear to be appropriate for the aims of this study.	We hope that the modifications made below will do justice to the comments made by the reviewers to the first revision of the manuscript.
For example, there is no evidence of a meta-analysis being carried out, and nowhere is it properly explained how the weights are employed. If it is not possible to produce a forest plot, then there is no meta analysis.	We agree with the reviewer that if you carry out a meta-analysis, a forest plot is expected. Only half of the studies showed a 95% confidence intervals (CI) around the kappa values (Appendices 4 and 5). Therefore, it was impossible to draw a forest plot with lines indicating the confidence intervals and only point estimates would remain. As weights for the statistical pooling, we used the number of records on which the kappa value is calculated as a proxy for accuracy, since we missed information about the 95% CI of the kappa values in the included studies. We changed the sentence on page 7, line 141: "We used the number of records on which the kappa value is calculated as weighing factor in the statistical pooling as a proxy for accuracy, since we missed information about the 95% confidence intervals (CI) of the kappa values in the included studies."
Another example: the revised table 4 (with medians as requested) clearly shows a large difference in experience between the n<=5 group and the other two groups, suggesting that the wrong (maybe parametric?) test has been used to obtain the p-values.	For the analysis of variables for which medians are presented in the revised Table 4, we used the non parametric Kruskal-Wallis test to obtain p values. The p value of the prevalence rate was obtained by an ANOVA. We changed the sequence of the sentences in the 'Data Synthesis and Analysis' on page 8 from line 146 to line 163 on page 9 and added the text in red:

	"To examine differences in kappa values depending on the number of reviewers, reviewer experience and reviewer training, we present descriptive statics per subgroup (mean with standard deviation (SD) or median with interquartile range (IQR) for non normal distributions, minimum and maximum). In order to better interpret the results, we classified the number of reviewers per study, reviewer experience and reviewer training into three proportional classes: max 5 reviewers, >5–20 reviewers, >20 reviewers; <100 records per reviewer, 100–300 records per reviewer, >300 records per reviewer, and < 1 day training, 1 day training, > 1 day training respectively. We used the non parametric Kruskal-Wallis test for the group characteristics which are not normally distributed and an ANOVA for the group characteristics with a normal distribution. We checked whether the assumptions for ANCOVA were met. It was not possible to incorporate all variables (the number of reviewers, reviewer experience and reviewer training) in one ANCOVA, because the number of studies in our analyses was limited (n = 20). Therefore, we performed three separate ANCOVAs, with prevalence of AE as covariate. We adjusted for prevalence of AEs, since a previous study of Lilford et al. (2007) showed correlation between prevalence and kappa.[16]"
Correlation analysis should be used to investigate relationships	Our study aim was explorative and therefore we were interested in differences in kappa values between subgroups of studies according to number of reviewers, reviewer experience, level of training and prevalence rate. We changed the term 'association' by 'differences between subgroups' throughout the manuscript. We added in the 'Introduction' on page 4, line 75: "We assumed that the inter-rater reliability of record review was higher for studies with a small number of reviewers, more reviewer experience and a higher training level."
Tables should be much clearly labelled and details of how the p-vals are obtained should be detailed in the caption	We thank the reviewer for this suggestion. We changed the title of Table 3 on page 14, line 272: "Differences in pooled kappa values (n = 20) among subgroups according to number of reviewers, reviewer experience, and reviewer training" We added the caption to Table 3 on page 14, lines 275 and 276: [1] "Pooled kappa weighted for the number of records on which the kappa value is calculated " "* P values are obtained with an ANCOVA with prevalence rate as covariate"

	We added the caption to Table 4 on page 15, lines 281-283: [2] "Unweighted statistics for reviewer experience, training and prevalence rate" * P values are obtained by the non parametric Kruskal-Wallis test ** P value is obtained with an ANOVA"
Tables 3 and 4 appear to contradict each other. Table 3 shows that the best agreement (kappa=0.8) is when N<6 and when experience is least (<100) (kappa = 0.71). However, Table 4 (and the text) shows that the reviewers with n<6 have the most experience (with median more than double that of the next largest group). This is difficult to comprehend.	Table 3 and Table 4 have different classifications of the included studies. In Table 3, we presented the weighted kappas of reviewer experience classified into three proportional classes. In Table 4, we presented reviewer experience by the (median) number of reviewed records that are actually reviewed per reviewer to explain the statistically significant difference between the three group of reviewers (max 5 reviewers, > 5 – 20 reviewers and > 20 reviewers). We changed and added sentences in the 'Data Synthesis and Analysis' on page 8. We also changed the title and added a caption to Tables 3 and 4.
Reviewer 2	
One final amendment could be to add a sentence to the study limitations, along the lines that due to limitations in the reporting of reliability statistics in the source papers, namely the lack of confidence intervals, it was not possible to formally estimate the pooled kappa statistics for the two measurement tools, to assess between-study heterogeneity, or to carry out analyses of the likelihood of publication bias. I think it is worth making the point that when reporting studies (of any type) it is important to include confidence intervals for estimated quantities, both to allow proper interpretation of individual study results, but also to allow others to use those results in future research.	We totally agree with the reviewer and we incorporated his good suggestion in the study limitations of the discussion section. We added on page 17 line 332 the sentence: "Second, it was not possible to formally estimate the pooled kappa statistics for the GTT and MRR, to assess between-study heterogeneity, or to carry out analyses of the likelihood of publication bias, because confidence intervals were lacking in approximately half of the reliability studies."