Cochrane

Comparison of the Cochrane risk of bias tool 1 (RoB 1) with the updated Cochrane risk of bias tool 2 (RoB 2)

Richter B, Hemmingsen B

Cochrane Metabolic and Endocrine Disorders Group,
Institute of General Practice, Medical Faculty of the
Heinrich-Heine University Düsseldorf, Germany

richterb@uni-duesseldorf.de

This study was commissioned and funded by Cochrane. The report was prepared to describe the background, methods and results for the Cochrane community. It has not been formally peer-reviewed and may be subject to change if submitted for publication.

Contents

Summary

Objective

This pilot compared the Cochrane risk of bias tools 1 and 2 (RoB 1 vs RoB 2), using data from a Cochrane review on insulin analogues for type 1 diabetes mellitus. Both study publications and clinical study reports (CSRs) served as resources for data extraction and risk of bias assessments.

Methods

Two raters with content and methodological expertise applied RoB 1 on publications without using CSRs investigating outcomes on a study and endpoint level. Thereafter, RoB 2 was used on the same set of publications employing the Excel® RoB 2 assessment spreadsheet. We compared results of risk of bias assessments between review authors and calculated measures of agreement for our outcome measures. Subsequently, RoB 1 assessments were compared to their RoB 2 counterpart for domains where mapping was possible, i.e 'selective reporting' in RoB 1 was not formally mapped to 'selection of the reported result' in RoB 2. For available CSRs we repeated the same procedure, this time without using regular publications. We planned to compare the impact of both risk of bias tools on the results of meta-analyses by dichotomizing results of risk of bias analyses into overall high versus low risk of bias. We recorded time needed to complete evaluation of risk of bias domains for both tools and noted specific problems in the use of both tools.

Results

The risk of bias comparison sample was based on a Cochrane index review including 26 studies, evaluating therapies with different types of insulin. Additional information could be obtained from 24 CSRs, clinical study synopses or both. A total of 24 studies were parallel randomised controlled trials, two studies were cross-over trials. Twenty studies had a non-inferiority design, six studies were performed as superiority trials. All studies except one cross-over trial were open-label studies. For each of the 11 (13 for CSRs) individual outcomes per included study we applied one specific result. Eleven outcome measures were analysed as dichotomous outcomes and two outcome measures (health-related quality of life, HbA1c levels) were analysed as continuous outcomes.

When using publications as data source consensus was necessary for RoB 1 domains 'performance bias' as well as 'detection bias' with reference to participant-reported outcome measures. There were numerous definitions of hypoglycaemic events resulting from insulin therapy with different associated risks of bias. Clinical expertise had to be exercised to establish subjective and (semi)objective labels for certain types of hypoglycaemic episodes. Use of RoB 2 for publications resulted in few differences in judgements between raters. There were no substantial differences for most risk of bias domains of RoB 1 compared to RoB 2, disregarding the fact that judgements in RoB 1 on subjective outcomes investigated in open-label studies caused high risk of bias judgements, whereas in RoB 2 use of the risk of bias algorithm resulted in 'some concerns'.

Compared to publications CSRs had much more detailed information regarding all aspects of risk of bias evaluations. For both RoB 1 and RoB 2 there were no difficulties in achieving agreement between raters and differences between both tools were small. RoB 1 tended to aggravate risk of bias judgements due to limited choice of answer (high/low/unclear) and no guidance from signalling questions embedded in RoB 2.

Because of small differences in overall risk of bias for RoB 1 (low risk of bias if there was low risk of bias for all domains) compared to RoB 2 (low risk of overall bias) we could not fully investigate impacts on meta-analyses. However, we demonstrated an example where effect size based on RoB 2 differed from RoB 1.

Despite the learning curve for RoB 2 mean time to assess risk of bias for all outcomes was comparable to RoB 1 assessment times (approx. 30 min). Time needed to complete risk of bias evaluation for all

outcomes using either RoB 1 or RoB 2 with CSRs as data source did not differ and ranged between 45 min to 1 ½ hr per CSR.

There were no major problems with the new RoB 2 tool. Consensus was necessary when assessing publications for domains 'deviations from intended interventions' and 'selection of the reported result', mainly because of missing trial protocols. Consensus was partly necessary for domain 'missing outcome data' with reference to potential relations of missing data to participants' health status, and interpretation demanded clinical expertise.

From a user point of view the RoB 2 Excel® tool should follow the way review authors usually extract data from publications and automatically reconstruct study-based data entry into the outcome-based structure of the spreadsheet.

Conclusions

If applied correctly, i.e. risk of bias analysis on a study *and* outcome level, RoB 1 and RoB 2 judgements are broadly similar. However, due to the way the RoB 2 Excel® tool is designed review authors are automatically encouraged to think in terms of endpoints/results which hopefully will lead to better quality in risk of bias assessments in Cochrane reviews. RoB 2 has a more granular approach by means of signalling questions and a broader range of possible answers guiding review authors to better address complexity and context of clinical trials. Some signalling questions would profit from additional elaborations, more specific wording, or both. In particular, data source is important for both RoB 1 and RoB 2 with CSRs being an excellent source for appropriate risk of bias appraisal.

Background

A necessity has arisen to revise the Cochrane RoB 1 tool due to moderate reliability and substantial variability in the use of various risk of bias domains and associated judgements (1). However, it is unclear how this will impact on the amount of time and effort review authors need to invest to adequately use the new RoB 2 tool, not only for establishing new Cochrane reviews but also for updating published Cochrane reviews. Therefore, guidance to help authors understand what the new tools implies and how risk-of-bias assessments should be integrated in Cochrane systematic reviews is needed.

This pilot used an index Cochrane review based on both publications and clinical study reports (CSRs). Within this review both the RoB 1 tool and RoB 2 tool were applied to compare key features, difficulties in use and potential consequences for the results of the systematic review process.

Aims

This pilot had two distinct aims:

a) to compare inter-rater reliability between comparable domains for RoB 1 and RoB 2 tools on a range of outcomes evaluated in the index Cochrane review using either publications or CSRs as data sources;
b) to compare the usability/applicability of RoB 1 and RoB 2, including median time taken to perform assessments and specific problems using the tools, as well as potential effects on the results and interpretation of the index Cochrane review.

Methods

For **(a)** we applied the following procedures:

The pilot was conducted using the index Cochrane review "(Ultra-)long-acting insulin analogues for people with type 1 diabetes mellitus" (2).

A pilot of five randomly selected studies was used to stabilize the learning curve employing the new RoB 2 instrument.

The RoB 1 tool offers three author judgements ('low risk', 'unclear risk', 'high risk') per risk of bias domain. The RoB 2 tool also provides three author judgements ('low', 'high', 'some concerns') per risk of bias domain. In order to contrast tools, we set 'unclear risk' in RoB 1 in any domain equal to 'some concerns' in the associated RoB 2 domain.

We compared randomisation sequence generation and allocation concealment ('selection bias') in RoB 1 to 'randomisation process' in RoB 2. If both random sequence generation and allocation concealment in RoB 1 were judged as low risk of bias we had a direct comparison to low risk of bias in the domain 'randomisation process' in RoB 2. If either random sequence generation or allocation concealment or both in RoB 1 were judged as 'unclear' we set 'selection bias' to 'some concerns' in order to compare it to 'some concerns' of RoB 2. If one of the items of 'selection bias' in RoB 1 was judged as high risk of bias 'selection bias' obtained a high risk of bias judgement.

Blinding of participants and personnel ('performance bias') in RoB 1 was compared to 'deviations from intended interventions' in RoB 2. Blinding of outcome assessment ('detection bias') in RoB 1 was compared to 'measurement of the outcome' in RoB 2. Finally, we compared incomplete outcome data ('attrition bias') in RoB 1 with 'missing outcome data' in RoB 2.

Selective reporting ('reporting bias') in RoB 1 is not directly comparable to 'selection of the reported results' in RoB 2. However, we provided some information on both risk of bias domains, as well as data on 'overall bias' in RoB 2 which shows the algorithm result of the five aggregated standard risk of bias domains of RoB 2.

The following 16 outcome measures were prespecified in the index review protocol:

- primary outcomes: all-cause mortality, health-related quality of life, serious/severe hypoglycaemia;

- secondary outcomes: cardiovascular mortality, non-fatal myocardial infarction, non-fatal stroke, end-stage renal disease, blindness, serious adverse events, non-serious adverse events, severe nocturnal hypoglycaemia, mild/moderate hypoglycaemia, socioeconomic effects, glycosylated haemoglobin A1c (HbA1c) levels, HbA1c < 7% without severe hypoglycaemia.

We defined health-related quality of life, non-serious adverse events and mild/moderate hypoglycaemia as subjective outcomes because these measures were participant-reported.

We used information from all publications (duplicate publications, companion documents, or multiple reports of a primary trial and trials registers) describing a single study, thereby imitating what is expected from a regular Cochrane review author.

Information on RoB 1 domains was entered in the 'characteristics of included studies' table in RevMan 5 (4).

We applied the RoB 2 tool on the same set of publications by using the Excel® RoB 2 assessment spreadsheet (1, 3) and transferred results to RevMan Web 2020 (5).

Thereafter, we established a data matrix for both RoB 1 and RoB 2 transferring results to an Excel® spreadsheet.

We planned to engage at least three review authors including a person experienced in using RoB 2. Two UK contributors had considerable methodological expertise, two contributors from the editorial base of the Cochrane Metabolic and Endocrine Disorders Group (CMED) had both considerable clinical and methodological expertise. Due to the COVID-19 crisis RoB 2 evaluation could be completed by two raters only. For a subgroup of included studies two UK contributors also evaluated RoB 1. Agreements and differences in judgements were generally comparable to the evaluations of the other two raters (details not shown).

For comparison of inter-rater agreement between RoB 1 and RoB 2 we applied the following procedure: a) using publications as data source for included studies of the index review we compared raters using RoB 1, b) thereafter, for the same set of studies we compared raters using RoB 2 (because the instrument was new to raters only descriptive elements of the calibration exercise were reported), and c) for corresponding domains we contrasted the results of using RoB 1 compared with RoB 2 using publications only.

Finally, we compared RoB 1 and RoB 2 using clinical study reports/synopses as data source of included studies. CSRs are important documents for regulatory approval, and we expected more detailed information on all aspects of study design, data on endpoints and information for evaluation of risk of bias. With regard to our pilot we wanted to find out whether amount and quality of information in data sources had an effect on risk of bias judgements. For the index review in case of conflicting data between publications and CSRs, information in CSRs obtained priority due to their official status for the approval process likely introducing more accurate data handling by applicants.

For **(b)** we applied the following procedures:

Regarding usability/applicability of RoB 1 and RoB 2 we noted the time it took to perform assessments across outcomes for a particular study and calculated the median assessment time across all studies.

We recorded specific problems using the tools by checking raters' comments written as footnotes or by means of personal communication.

To investigate potential effects on the results and interpretation of the index Cochrane review we evaluated the impact of RoB 1 versus RoB 2 on key results of meta-analyses performed in the index review.

Data analysis

We measured inter-rater agreement for two unique raters by means of the kappa-statistic measure of agreement scaled 0 for what would be expected to be observed by chance and 1 for perfect agreement. We used the following interpretations suggested by Landis and Koch (6): values less than 0 imply poor agreement, values 0 to 0.20 slight, values 0.21 to 0.40 fair, values 0.41 to 0.60 moderate, values 0.61 to 0.80 substantial and values 0.81 to 1.00 almost perfect agreement, respectively.

Measurement of agreement and Cohen's kappa-statistic with 95% confidence intervals (CIs) were done by Stata 17 software (StataCorp. 2021. Stata Statistical Software: Release 17. College Station, TX: StataCorp LLC).

We also calculated Gwet's AC1 as a measure of interrater agreement for two raters' categorical assessments (7). According to Gwet "AC stands for agreement coefficient and digit 1 indicates the first-order chance correction, which accounts for full agreement only as opposed to full and partial agreement (second-order chance correction)". Gwet's agreement coefficient can be used in more contexts than kappa because it does not depend upon the assumption of independence between raters. Because two raters from the editorial base of the CMED have collaborated on several Cochrane review projects this statistic was also used. Gwet's AC1 was calculated by Statsdirect Vers. 3.3.5 (8). Due to the so-called "kappa paradox" potentially appearing in situations where two raters achieve high agreement leading to low kappa-values (9) we reported the percentage of observed agreement and the Gwet's statistic in the results section.

Results

The risk of bias comparison sample comprised 11 (13 for CSRs) individual outcome assessments and was performed for 26 included studies containing two unpublished studies. For each outcome and analysis of an included study we applied one specific result. Due to scarce data we combined non-fatal myocardial infarction and non-fatal stroke into non-fatal-myocardial infarction/stroke and end-stage renal disease and blindness into end-stage renal disease/blindness. No adequate information for risk of bias evaluation was available for the outcome socioeconomic effects. Only CSRs provided sufficient details on the outcome measures severe nocturnal hypoglycaemia and the HbA1c < 7% without severe hypoglycaemia. Eleven outcome measures were analysed as dichotomous outcomes and two outcome measures (health-related quality of life, HbA1c levels) were analysed as continuous outcomes.

For 24 of the 26 included studies we were able to obtain clinical study reports, clinical study synopses or both. A total of 24 studies were parallel RCTs, two studies were cross-over trials. Twenty studies had a non-inferiority design, six studies were performed as superiority trials. All studies except one cross-over trial were open-label studies.

a) Inter-rater reliability between comparable domains for RoB 1 and RoB 2

(a1) Comparison of raters using RoB 1 and publications

Agreement on randomisation sequence generation and allocation concealment were 70% and 87%, respectively. Some differences between raters for low vs unclear risk of bias judgements were observed (Table 1, for details see Appendix a1). Agreement on the risk of bias domain 'Blinding of participants & personnel' for objective outcomes ranged between 82% and 100% and for subjective outcomes (health-related quality of life, non-serious adverse events, mild/moderate hypoglycaemia) between 71% and 100%. Agreement on the risk of bias domain 'Blinding of outcome assessment' for objective outcomes ranged between 67% and 100%, with the exception of disagreement for the endpoint end-stage renal disease/blindness which was reported in one study only. For subjective outcomes agreement ranged between 25% and 57%. We achieved consensus on patient-reported outcomes in open-label trials being associated with high risk of bias, even if validated questionnaires were analysed by blinded personnel. Agreement on the risk of bias domain 'Incomplete outcome data' for objective outcomes ranged between 33% and 88% (with disagreement for the single study reporting the endpoint end-stage renal disease/blindness) and for subjective outcomes between 50% and 62%. Consensus had to be achieved for selective reporting (judgement set to low risk of bias if protocol or ClinicalTrials.gov data were available, otherwise unclear risk of bias) and other bias (judgement set to no risk of bias if other statements could not be backed up by published evidence as source of bias).

Table 1: level of agreement rater A vs rater B

Randomisation sequence	70% [0.43]		
Allocation concealment	87% [0.85]		
Outcome (no. of studies with outcome)	**A**	**B**	**C**
All-cause mortality (7)	100% [1]	100% [1]	86% [0.84]
Health-related quality of life (4)	100% [1]	25% [-0.09]	50% [0.32]
Severe hypoglycaemia (22)	82% [0.80]	73% [0.69]	64% [0.57]
Cardiovascular mortality (7)	100% [1]	86% [0.84]	86% [0.84]
Non-fatal myocardial infarction/stroke (3)	67% [0.54]	67% [0.53]	33% [-0.20]
End-stage renal disease/blindness (1)	100% [1]	0% [-1]	0% [-1]
Serious adverse events (19)	84% [0.82]	68% [0.53]	63% [0.56]
Diabetic ketoacidosis (8)	100% [1]	100% [1]	88% [0.82]
Non-serious adverse events (17)	71% [0.61]	53% [0.34]	59% [0.50]
Mild/moderate hypoglycaemia (21)	71% [0.62]	57% [0.44]	62% [0.55]
HbA1c (23)	87% [0.85]	74% [0.66]	65% [0.59]
Selective reporting	18% [-0.23]		
Other risk of bias	0% [-1]		

[ ] = Gwet's AC1; A: Blinding of participants & personnel; B: Blinding of outcome assessment; C: Incomplete outcome data

(a2) Comparison of raters using RoB 2 and publications

The calibration exercise on RoB 2 described the initial problems raters encountered when using the until then unknown tool.

Domain 1 (‘randomisation process’) achieved almost complete agreement.

Domain 2 (‘deviations from intended interventions’) needed some consensus especially because of signalling question 2.3 (“Were there deviations from the intended intervention that arose because of the experimental context?”): since in almost all cases no trial protocol was available answers to this question were either ‘No information‘ or ‘Probably no‘ which either resulted in the algorithm result ‘Some concerns‘ or ‘Low‘.

Domain 3 (‘missing outcome data’) achieved almost complete agreement. Some consensus was necessary for signalling question 3.3 (“Could missingness in the outcome depend on its true value?”): answering the question if losses to follow-up or withdrawals from the study were related to participants' health status appeared rather judgemental, especially with regard to whether missing data could have a clear impact on outcomes in case attrition rates were comparable and explained. Moreover, we rarely discovered information on missing data per outcome in publications. Usually, trial flow diagrams just showed the numbers and reasons for drop-outs. However, imputation data and missing data for specific outcomes were usually not reported in tables and the text of publications.

Domain 4 (‘measurement of the outcome’) achieved almost complete agreement. Signalling question 4.5 (“Is it likely that assessment of the outcome was influenced by knowledge of intervention received?”) needed significant content expertise because different judgements on whether there were strong beliefs in either beneficial or harmful effects of the interventions resulted in the algorithm result ‘Some concerns‘ or ‘High‘.

Domain 5 (‘selection of the reported result’) achieved almost complete agreement. Signalling question 5.1 (“Were the data that produced this result analysed in accordance with a pre-specified analysis plan that was finalized before unblinded outcome data were available for analysis?”) was difficult to answer because most often no protocol was available. Some information was available when comparing trial registry information with the publication.

(a3) Comparison RoB 1 versus RoB 2 using publications

Following the (a1) and (a2) scheme we used consented judgements for both tools to compare RoB 1 with RoB 2 (Table 2, for details see Appendix a3).

Agreement on ‘randomisation sequence generation and allocation concealment’ (RoB 1) compared with ‘randomisation process’ (RoB 2) was 91%. Agreement on ‘blinding participants and personnel’ (RoB 1) compared with ‘deviations from intended interventions’ (RoB 2) ranged between 82% and 100% for objective outcomes. There was disagreement for all subjective outcomes because these were judged as high risk of bias in RoB 1 but mostly as low risk of bias in RoB 2 resulting from seven signalling questions providing more detailed appraisal. Agreement on ‘incomplete outcome data (RoB 1) compared with ‘missing outcome data’ (RoB 2) across all outcomes ranged between 86% and 100%. Agreement on ‘blinding outcome assessment’ (RoB 1) compared with ‘measurement of the outcome’ (RoB 2) ranged between 91% and 100% for objective outcomes (with the exception of end-stage renal disease/blindness reported in one study only). There was disagreement for all subjective outcomes because these were judged as high risk of bias in RoB 1 but mostly as some concerns in RoB 2 resulting from five signalling questions providing more detailed appraisal.

Although ‘selective reporting’ (RoB 1) cannot directly compared with ‘selection of the reported result’ (RoB) we were interested whether any kind of ‘selection process’ resulted in different judgements if evaluated for subjective or objective outcomes: for both outcomes agreement was comparable (non-serious adverse events 53% (Gwet’s AC1 0.06); HbA1c: 57% (Gwet’s AC1 0.14)).

Overall risk of bias (RoB 2) for objective outcomes was either judged as low (45% to 100%) or some concerns (14% to 55%). All subjective outcomes were judged as some concerns.

Table 2: level of agreement RoB 1 vs RoB2 using publications

Randomisation sequence + allocation concealment vs randomisation process	91% [0.85]			
Outcome (no. of studies with outcome)	**A**	**B**	**C**	**D**
All-cause mortality (7)	86% [0.84]	100% [1]	100% [1]	86% l; 14% s
Health-related quality of life (3)	0% [-0.44]	100% [1]	0% [-1]	100% s
Severe hypoglycaemia (22)	82% [0.78]	91% [0.90]	91% [0.90]	45% l; 55% s
Cardiovascular mortality (7)	86% [0.84]	86% [0.84]	100% [1]	86% l; 14% s
Non-fatal myocardial infarction/stroke (3)	100% [1]	100% [1]	100% [1]	67 l; 33% s
End-stage renal disease/blindness (1)	100% [1]	100% [1]	0% [-1]	100% l
Serious adverse events (19)	84% [0.82]	89% [0.88]	95% [0.94]	53% l; 47% s
Diabetic ketoacidosis (8)	100% [1]	100% [1]	100% [1]	63% l; 37% s
Non-serious adverse events (17)	0% [-0.38]	88% [0.87]	0% [-1]	100% s
Mild/moderate hypoglycaemia (21)	0% [-0.39]	90% [0.90]	0% [-1]	100% s
HbA1c (23)	83% [0.79]	91% [0.91]	100% [1]	48% l; 52% s

Note: ‘unclear’ in RoB 1 was set to ‘some concerns’ for RoB 2 comparison
[ ] = Gwet’s AC1
A: Blinding participants and personnel (RoB 1) vs deviations from intended interventions (RoB 2)
B: Incomplete outcome data (RoB 1) vs missing outcome data (RoB 2)
C: Blinding outcome assessment (RoB 1) vs measurement of the outcome (RoB 2)
D: Overall bias (RoB 2): l = low; s = some concerns

(a4) Comparison RoB 1 versus RoB 2 using clinical study reports

We used consented judgements for both tools to compare RoB 1 with RoB 2 (Table 3, for details see Appendix a4).

Agreement on ‘randomisation sequence generation and allocation concealment’ (RoB 1) compared with ‘randomisation process’ (RoB 2) was 100%. Agreement on ‘blinding participants and personnel’ (RoB 1) compared with ‘deviations from intended interventions’ (RoB 2) ranged between 86% and 100% for objective outcomes. For subjective outcomes agreement ranged between 0% and 20%, usually set as high risk of bias in RoB 1 but mostly as low risk of bias in RoB 2. Agreement on ‘incomplete outcome data (RoB 1) compared with ‘missing outcome data’ (RoB 2) across all outcomes was 100%. Agreement on ‘blinding outcome assessment’ (RoB 1) compared with ‘measurement of the outcome’ (RoB 2) was 100% for objective outcomes and ranged between 0% and 20% for subjective outcomes because these were mostly judged as high risk of bias in RoB 1 but as some concerns in RoB 2.

Although ‘selective reporting’ (RoB 1) cannot directly compared with ‘selection of the reported result’ (RoB) we were interested whether any kind of ‘selection process’ resulted in different judgements if evaluated for subjective or objective outcomes: for both outcomes agreement was comparable (non-serious adverse events 92% (Gwet’s AC1 0.91); HbA1c: 92% (Gwet’s AC1 0.92)).

Overall risk of bias (RoB 2) for objective outcomes was either judged as low (86% to 100%) or some concerns (8% to 14%). Subjective outcomes were judged as low in 0% to 20% of cases and as some concerns in 80% to 100% of cases.

Table 3: level of agreement RoB 1 vs RoB2 using clinical study reports

Randomisation sequence + allocation concealment vs randomisation process	100% [1]			
Outcome (no. of studies with outcome)	**A**	**B**	**C**	**D**
All-cause mortality (24)	96% [0.96]	100% [1]	100% [1]	92% l; 8% s
Health-related quality of life (5)	20% [-0.54]	100% [1]	20% [-0.18]	20% l; 80% s
Severe hypoglycaemia (24)	100% [1]	100% [1]	100% [1]	87.5% l; 12.5% s
Cardiovascular mortality (24)	96% [0.96]	100% [1]	100% [1]	92% l; 8% s
Non-fatal myocardial infarction/stroke (7)	86% [0.84]	100% [1]	100% [1]	86% l; 14% s
End-stage renal disease/blindness (1)	100% [1]	100% [1]	100% [1]	100% l
Serious adverse events (24)	96% [0.96]	100% [1]	100% [1]	92% l; 8% s
Diabetic ketoacidosis (19)	95% [0.94]	100% [1]	100% [1]	95% l; 5% s
Non-serious adverse events (24)	0% [-0.35]	100% [1]	0% [-1]	100% s
Severe nocturnal hypoglycaemia (20)	95% [0.95]	100% [1]	100% [1]	95% l; 5% s
Mild/moderate hypoglycaemia (22)	0% [-0.35]	100% [1]	0% [-1]	100% s
HbA1c (25)	96% [0.96]	100% [1]	100% [1]	88% l; 12% s
HbA1c <7% without severe hypoglycaemia (4)	100% [1]	100% [1]	100% [1]	100% l

Note: ‘unclear’ in RoB 1 was set to ‘some concerns’ for RoB 2 comparison.
[ ] = Gwet’s AC1; A: Blinding participants and personnel (RoB 1) vs deviations from intended interventions (RoB 2)
B: Incomplete outcome data (RoB 1) vs missing outcome data (RoB 2); C: Blinding outcome assessment (RoB 1) vs measurement of the outcome (RoB 2); D: Overall bias (RoB 2): l = low; s = some concerns

b) Usability/applicability of RoB 1 and RoB 2 tools

(b1) Time taken to assess risk of bias for RoB 1 and RoB2

Mean time to assess risk of bias using RoB 1 for all outcomes using publications ranged between 19 min and 270 min (Table 4). Assessment times were highly comparable for reviewers with both content and methodological expertise (ranging between 19 and 27 min). Despite the learning curve for RoB 2 mean assessment time was highly comparable between reviewers and similar to RoB 1 assessment time, ranging from 27 to 28 min.

Table 4: median, mean, minimum and maximum assessment time in minutes for RoB 1 vs RoB 2 (publications)

RoB 2	Reviewer 1	Reviewer 2	RoB1	Reviewer 1	Reviewer 2	Reviewer 3	Reviewer 4
Median	30	22	Median	25	15		
Mean	28	27	Mean	27	19	270	60
Min	20	15	Min	20	10		
Max	35	55	Max	35	35		

Though CSRs sometimes contain thousands of pages, assessment time for risk of bias appraisal could be significantly reduced due to easy navigation in pdf-files. The mean time taken to assess risk of bias for all outcomes using CSRs for both RoB 1 or RoB 2 was approximately 60 min (ranging between 45 min to 90 min).

(b2) Specific problems in the use of both RoB tools

Because of the long experience with RoB 1 major problems did not arise due to the fact that within CMED review authors have to analyse risk of bias on study level and outcome level. Outcome measures were not grouped because every endpoint had to be interpreted with clinical expertise (e.g. distinguishing severe hypoglycaemia as an objective outcome measure from mild/moderate hypoglycaemia as a subjective outcome measure).

Regarding “selection bias” (random sequence generation & allocation concealment) agreement on risk of bias judgement was good, including inspection of baseline imbalances potentially indicating that randomisation did not work correctly.

For “performance bias” (blinding of participants and personnel) and “detection bias” (blinding of outcome assessment) there was good agreement for objective outcomes. Consensus was necessary for subjective outcomes, even amongst experienced review authors. This was due to the nature of the outcome measures, e.g. mild/moderate hypoglycaemia might appear objective due to measurement of blood glucose by technical means. However, all these measures were performed and reported by participants in the open-label trials which results in some risk of bias. The same is true for health-related quality of life using validated questionnaires where the outcome assessor is the study participant as is the case with non-serious adverse events.

There was good agreement on “attrition bias” (incomplete outcome data). However, interpretation was limited because of scarce data on missing data per outcome and the necessity to focus on information from the trial flow diagram specifying in various detail drop-outs, missed follow-up and withdrawals.

Consensus was necessary for “reporting bias” (selective reporting). In RoB 1 this is a study level bias making it difficult to distinguish per outcome measure. Usually, in CMED reporting bias is evaluated by integrating the results of the table 'Matrix of study endpoints (publications and trial documents)' with another table 'High risk of outcome reporting bias according to the “Outcome Reporting Bias In Trials” (ORBIT) classification' (10). In the above-mentioned matrix endpoints quoted in trial documents (e.g. ClinicalTrials.gov, FDA/EMA documents, manufacturer's website, published design papers) are compared to endpoints quoted in the publication and the abstract of the publication. Since most review authors do not use such a procedure, we abstained from employing our regular approach. Instead, we just achieved consensus whether a protocol was available in a trial register and if so whether primary and secondary outcomes were comparable between information in the trial register and the publication.

We agreed not to use the “other risk of bias” domain because of lacking evidence on most postulated biases.

Using RoB 2, there was good agreement for domain 1 (“randomisation process”), no signalling question was difficult to answer.

We had to achieve consensus for signalling question 2.3 of domain2 (“deviations from intended interventions”), mainly because usually no trial protocol was available and answers ranged between ‘No information’ and ‘Probably no’ resulting in either ‘Some concerns’ or ‘Low’ risk of bias judgement according to the algorithm result.

We had to achieve consensus for signalling question 3.3 of domain 3 (“missing outcome data”), mainly because opinions sometimes differed whether missing outcome data, though occurring for documented reason with comparable attrition rates, could have been related to participants’ health status. To answer ‘Yes’ to signalling question 3.4 would be appropriate if reported reasons for missing data differed between the intervention groups. However, the extent and consequence of a difference is highly judgemental demanding content expertise making it necessary to obtain consensus.

There was good agreement for domain 4 (“measurement of the outcome”).

We had to achieve consensus for signalling question 5.1 of domain 5 (“selection of the reported result”). Usually, in publications there is no sufficient detail of researchers’ pre-specified intentions to compare planned outcome measures with the those presented in the publication. Without a protocol answers were either ‘Probably yes’ or ‘No information’ resulting in ‘Low’ or ‘Some concerns’ of risk of bias judgement according to the algorithm result.

There was good agreement regarding ‘Overall bias’.

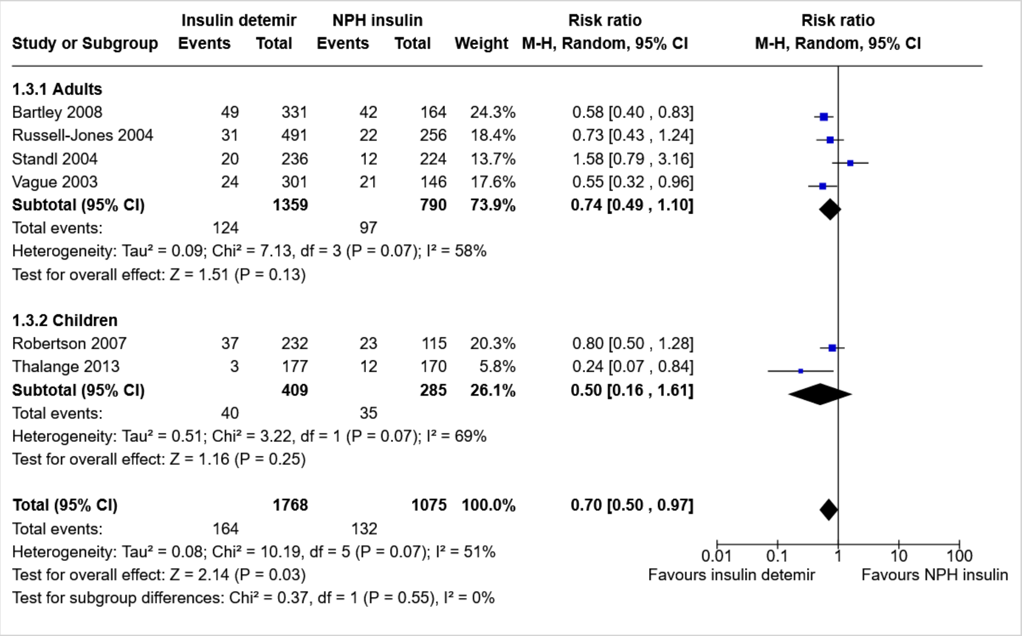
(b3) Impact of RoB1/RoB2 on the results of meta-analyses

There were no major differences between RoB 1 and RoB 2 when comparing individual risk of bias domains with each other (randomisation/allocation concealment vs randomisation process; blinding of participants & personnel vs deviations from intended interventions; incomplete outcome data vs missing outcome data; blinding of outcome assessment vs measurement of the outcome). However, subjective nature of outcome assessment resulted in 'high' risk of bias in RoB 1 compared to 'some concerns' in RoB 2. The signalling questions in RoB 2 and a wider choice of possibilities to answer (yes, probably yes, probably no, no, no information) permitted a more granular approach to appraise risk of bias domains.

To investigate the possible impact of RoB 1 vs RoB2 judgements the following two figures illustrate the result on the effect estimates of severe hypoglycaemia (defined as a (semi)objective outcome measure requiring third party help in case of a hypoglycaemic episode) for the comparison insulin detemir vs NPH insulin.

Figure 1 includes all studies reporting this outcome, separated into the subgroups adults and children, with low risk of bias for all domains in RoB 1.

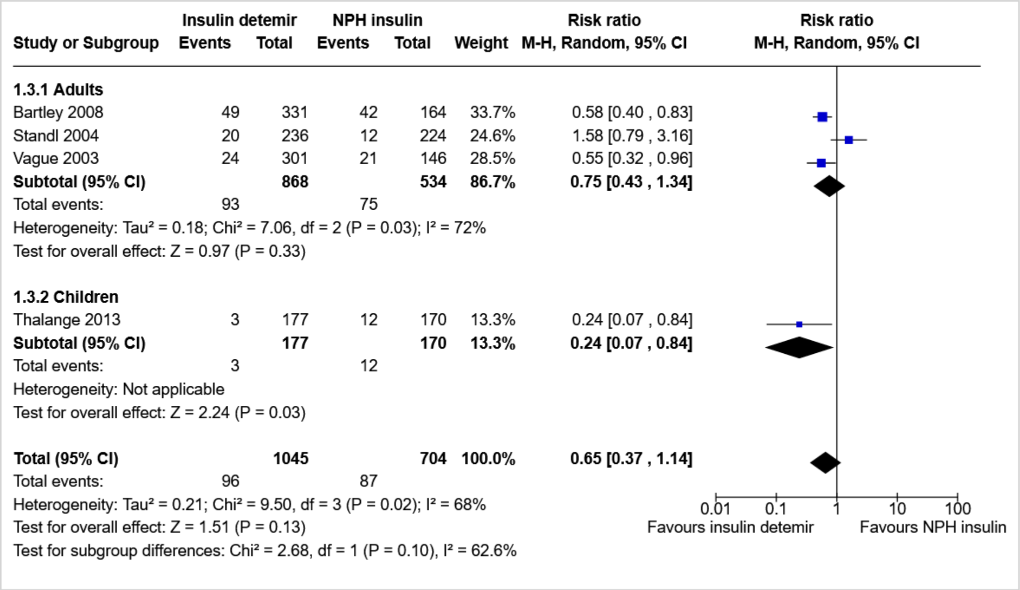
Figure 1: risk ratio for severe hypoglycaemia (RoB 1)



	Insulin detemir		NPH insulin			Risk ratio
Study or Subgroup	**Events**	**Total**	**Events**	**Total**	**Weight**	**M-H, Random, 95% CI**
1.3.1 Adults						
Bartley 2008	49	331	42	164	24.3%	0.58 [0.40 , 0.83]
Russell-Jones 2004	31	491	22	256	18.4%	0.73 [0.43 , 1.24]
Standl 2004	20	236	12	224	13.7%	1.58 [0.79 , 3.16]
Vague 2003	24	301	21	146	17.6%	0.55 [0.32 , 0.96]
Subtotal (95% CI)		**1359**		**790**	**73.9%**	**0.74 [0.49 , 1.10]**
Total events:	124		97			
Heterogeneity: Tau² = 0.09; Chi² = 7.13, df = 3 (P = 0.07); I² = 58%						
Test for overall effect: Z = 1.51 (P = 0.13)						
1.3.2 Children						
Robertson 2007	37	232	23	115	20.3%	0.80 [0.50 , 1.28]
Thalange 2013	3	177	12	170	5.8%	0.24 [0.07 , 0.84]
Subtotal (95% CI)		**409**		**285**	**26.1%**	**0.50 [0.16 , 1.61]**
Total events:	40		35			
Heterogeneity: Tau² = 0.51; Chi² = 3.22, df = 1 (P = 0.07); I² = 69%						
Test for overall effect: Z = 1.16 (P = 0.25)						
Total (95% CI)		**1768**		**1075**	**100.0%**	**0.70 [0.50 , 0.97]**
Total events:	164		132			
Heterogeneity: Tau² = 0.08; Chi² = 10.19, df = 5 (P = 0.07); I² = 51%						
Test for overall effect: Z = 2.14 (P = 0.03)						
Test for subgroup differences: Chi² = 0.37, df = 1 (P = 0.55), I² = 0%						

Figure 2 includes the same set of studies with low overall risk of bias for RoB 2. Two studies (Robertson 2007; Russel-Jones 2004) were excluded because they were associated with some concerns in the overall risk of bias assessment of RoB 2.

Figure 2: risk ratio for severe hypoglycaemia (RoB 2)



Study or Subgroup	Insulin detemir Events	Insulin detemir Total	NPH insulin Events	NPH insulin Total	Weight	Risk ratio M-H, Random, 95% CI
1.3.1 Adults						
Bartley 2008	49	331	42	164	33.7%	0.58 [0.40 , 0.83]
Standl 2004	20	236	12	224	24.6%	1.58 [0.79 , 3.16]
Vague 2003	24	301	21	146	28.5%	0.55 [0.32 , 0.96]
Subtotal (95% CI)		**868**		**534**	**86.7%**	**0.75 [0.43 , 1.34]**
Total events:	93		75			
Heterogeneity: Tau² = 0.18; Chi² = 7.06, df = 2 (P = 0.03); I² = 72%						
Test for overall effect: Z = 0.97 (P = 0.33)						
1.3.2 Children						
Thalange 2013	3	177	12	170	13.3%	0.24 [0.07 , 0.84]
Subtotal (95% CI)		**177**		**170**	**13.3%**	**0.24 [0.07 , 0.84]**
Total events:	3		12			
Heterogeneity: Not applicable						
Test for overall effect: Z = 2.24 (P = 0.03)						
Total (95% CI)		**1045**		**704**	**100.0%**	**0.65 [0.37 , 1.14]**
Total events:	96		87			
Heterogeneity: Tau² = 0.21; Chi² = 9.50, df = 3 (P = 0.02); I² = 68%						
Test for overall effect: Z = 1.51 (P = 0.13)						
Test for subgroup differences: Chi² = 2.68, df = 1 (P = 0.10), I² = 62.6%						

For adults, the effect estimate did not change substantially (RoB 1: risk ratio (RR) 0.74, 95% CI 0.49 to 1.10 vs RoB 2: RR 0.75, 95% CI 0.43 to 1.34)).

For children, the effect estimates differed (RoB 1: RR 0.50, 95% CI 0.16 to 1.61 vs RoB 2: RR 0.24, 95% CI 0.07 to 0.84)). However, only one study in children for RoB 2 evaluation was available for this comparison.

For all participants, the effect estimates slightly differed (RoB 1: RR 0.70, 95% CI 0.50 to 0.97 vs RoB 2: RR 0.65, 95% CI 0.37 to 1.14)).

This example demonstrates the possibility that risk of bias evaluation on the basis of RoB 1 or RoB 2 might have an effect on the results of meta-analyses. Interestingly, this could be shown with the rather unusual case where RoB 2 was more rigorous than RoB 1. In most cases in our pilot RoB 1, due to the restricted choice of answers, resulted in stronger judgements (e.g. high risk of performance/detection bias for subjective outcomes compared to some concerns using RoB 2).

Due to the small differences in overall risk of bias for RoB 1 (low risk of bias if there was low risk of bias for all domains) compared to RoB 2 (low risk of overall bias) we did not further investigate impacts on meta-analyses.

Discussion

Our two major aims, i.e. to compare inter-rater reliability between comparable domains for RoB 1 and RoB 2 tools using either publications or CSRs as data sources and to compare usability/applicability of the tools including potential effects on the results and interpretation of the index Cochrane review could be achieved due to the distinct database of 26 studies including 24 CSRs of these trials.

Overall, appraisal of risk of bias using RoB 1 or RoB 2 did not show substantial differences in comparable domains. However, it was evident that for both instruments content and methodological expertise is important to achieve high agreement and reliability between different raters. Moreover, data source is relevant as could be demonstrated by improved information from CSRs compared to regular publications.

Differences in judgements for RoB 1 using publications, if any, appeared for low risk of bias vs unclear risk of bias assessments. Consensus was necessary for participant-reported outcome measures in open-label trials which resulted in high risk of bias for performance bias and detection bias. For selective reporting bias we attributed low risk of bias if there was trial registry information and definitions of primary and secondary outcomes did not differ substantially between trial registries and publications. If no information from trial registries existed, we attributed unclear risk of bias. For ‘other risk of bias’ we consented on unclear risk because no firm empirical evidence existed to reliably attribute other risk of bias statements. For a subgroup of included studies two UK contributors also evaluated risk of bias. Agreements and differences in judgements were generally comparable to the evaluations of the other two raters. Due to the open-label design of studies one rater often judged performance bias as high for most outcomes without distinguishing between subjective and objective outcome measures. Generally, there were difficulties in evaluation of hypoglycaemia because publications reported numerous definitions of this outcome. As a result, for hypoglycaemic episodes we focussed on severe hypoglycaemia and severe nocturnal hypoglycaemia as (semi)objective outcome measures and on mild/moderate hypoglycaemia as a subjective outcome measure. The same applied to detection bias. Differences were less distinct for incomplete outcome data (some differences in judgement for low vs unclear risk of bias judgements).

When contrasting RoB 1 with RoB 2 there were no major differences for most comparable risk of bias domains. Subjective outcome measures needed consensus as well as content expertise because the way endpoints were measured and reported had a strong effect on risk of bias judgements. Within CMED RoB 1 evaluation is always done one a study and endpoint level. Apparently, this is not the case for all Cochrane Review Groups creating problems when subjective and objective outcome measures are lumped as 'all outcomes' for performance bias, detection bias and attrition bias. Due to the way the RoB 2 Excel tool is structured review authors are automatically directed to think in terms of endpoints/results which consequently will lead to better quality in risk of bias assessments in Cochrane reviews. In addition, signalling questions in RoB 2 are a major progress, ensuring that review authors address more aspects within each risk of bias domain. In contrast, RoB 1 tended to aggravate risk of bias judgements due to limited choice of answers and no guidance from signalling questions

Our pilot also showed that data sources are important for both RoB 1 and RoB 2. CSRs provided much better information for all risk of bias domains than publications, even if some data from trials registries were available.

Limitations: due to the small differences in overall risk of bias for RoB 1 (low risk of bias if there was low risk of bias for all domains) compared to RoB 2 (low risk of overall bias) we could not thoroughly investigate impacts on meta-analyses of the index review. However, we provided an example where potential differences in the use of the two tools resulted in different effect sizes which should be investigated in follow-up studies. Another limitation of our pilot was that we could not completely involve all four raters as planned. The worldwide COVID-19 crisis had major impacts on scarce resources amongst Cochrane contributors which resulted in partial involvement for evaluation of RoB 1 and no possibility to replicate our approach for all raters using RoB 2. However, use of experienced review authors from the CMED editorial base had the advantage to involve both clinical and methodological experts reflecting the necessity to arrange a team of review authors with content and methodological expertise. A recent publication investigating RoB 2 showed low inter-rater reliability and challenges in its application (11). Authors reported that insufficient knowledge of the subject matter was one of their problems and demanded clinicians and methodologists being involved in the reviewer team. They also recommended a pilot run of the evaluation which we applied in our approach. Moreover, authors pointed to potential problems due to negative formulations in some of the signalling questions (i.e. 'yes' sometimes indicates high risk of bias) which could be especially difficult for non-native English speakers. The authors did not use the Excel® spreadsheet and only investigated the primary outcome of each study resulting in a mean assessment time to apply RoB 2 of 28 min per outcome. In our pilot the Excel® RoB 2 spreadsheet appeared as the best approach to evaluate risk of bias. However, from a user point of view the tool should follow the way review authors extract data from publications, i.e. authors usually extract an entire dataset from one or more publications. Currently, the Excel® RoB 2 tool is outcome-based which is cumbersome because for several outcomes many Excel files need to be established with significant time delay to check a publication over and over again for risk of bias information per outcome, especially if there is a time lag between data extraction. Alternatively, one could extract all outcomes for one study in a single Excel spreadsheet but later on one needs to reconstruct the original outcome-based Excel structure which is error-prone. Therefore, automatic reconstruction from a study-based risk of bias assessment into an outcome-based Excel spreadsheet would be very useful.

In conclusion, for high quality risk of bias assessments RoB 2 is superior to RoB 1. The more granular approach by means of signalling questions leads review authors quasi semi-automatically to a better examination of risk of bias. Review authors should ensure content and methodological expertise within their team to adequately address the complexity of outcome measures' definition. For a broad usage within the systematic reviewer community the RoB 2 Excel® tool should be improved as regards to more specific wording and the regular flow of data extraction. In particular, data source is important for both RoB 1 and RoB 2 with CSRs being an excellent source for appropriate risk of bias appraisal.

References

(1) Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ. 2019;366:l4898)

(2) Hemmingsen B, Metzendorf Maria-Inti, Richter B. (Ultra-)long-acting insulin analogues for people with type 1 diabetes mellitus. Cochrane Library 2021; https://doi.org/10.1002/14651858.CD013498.pub2

(3) RoB 2 tool: version 22, August 2019. https://www.riskofbias.info/welcome/rob-2-0-tool (last accessed 15 May 2021)

(4) Review Manager 5 (RevMan 5) [Computer program]. Version 5.3. Copenhagen: Nordic Cochrane Centre, The Cochrane Collaboration, 2014

(5) Review Manager Web (RevMan Web). Computer program. Version 1.22.0. The Cochrane Collaboration, 2020. Available at revman.cochrane.org

(6) Landis, J. R., and G. G. Koch. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–174. http://doi.org/10.2307/2529310.

(7) Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. British Journal of Mathematical and Statistical Psychology 2008;61:29-48

(8) StatsDirect Ltd. StatsDirect statistical software. http://www.statsdirect.com. England: StatsDirect Ltd. 2013

(9) Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. J Clin Epidemiol 1990;43(6):543-9. doi: 10.1016/0895-4356(90)90158-l

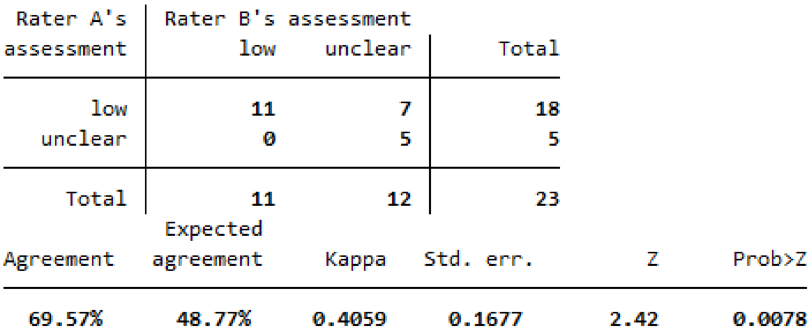
(10) Kirkham JJ, Dwan KM, Altman DG, Gamble C, Dodd S, Smyth R, et al. The impact of outcome reporting bias in randomised controlled trials on a cohort of systematic reviews. BMJ 2010;340:c365

(11)Minozzi S, Cinquini M, Gianola S, Gonzalez-Lorenzo M, Banzi R. The revised Cochrane risk of bias tool for randomized trials (RoB 2) showed low interrater reliability and challenges in its application. J Clin Epidemiol. 2020 Oct;126:37-44. doi: 10.1016/j.jclinepi.2020.06.015

Appendices

(a1) RoB 1: Rater A versus Rater B using publications

Randomisation sequence (for all outcomes)

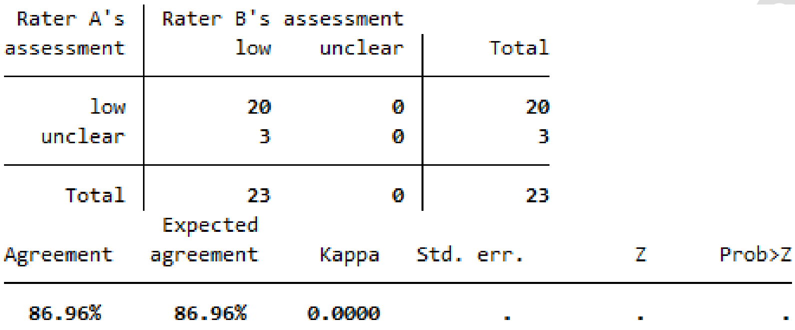


```
 Rater A's | Rater B's assessment
assessment |      low    unclear |     Total
-----------+---------------------+----------
       low |       11          7 |        18
   unclear |        0          5 |         5
-----------+---------------------+----------
     Total |       11         12 |        23
             Expected
Agreement   agreement     Kappa   Std. err.          Z      Prob>Z
-----------------------------------------------------------------
  69.57%      48.77%     0.4059     0.1677        2.42      0.0078
```

95% CI for kappa: 0.11 to 0.70

Gwet's AC1 with 95% CI: 0.43 (0.05 to 0.81)

Allocation concealment (for all outcomes)



```
 Rater A's | Rater B's assessment
assessment |      low    unclear |     Total
-----------+---------------------+----------
       low |       20          0 |        20
   unclear |        3          0 |         3
-----------+---------------------+----------
     Total |       23          0 |        23
             Expected
Agreement   agreement     Kappa   Std. err.          Z      Prob>Z
-----------------------------------------------------------------
  86.96%      86.96%     0.0000          .           .           .
```

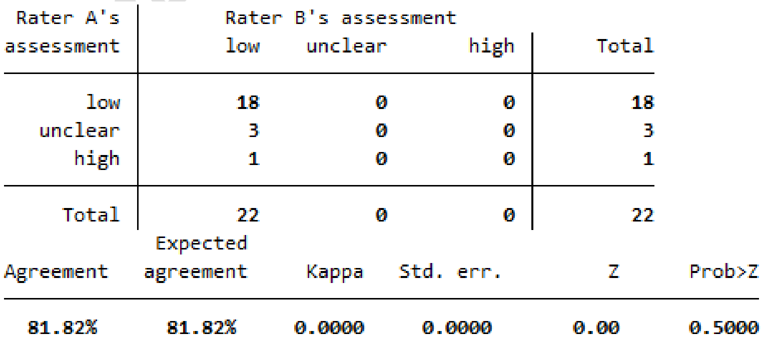
Gwet's AC1 with 95% CI: 0.85 (0.67 to 1.03)

Blinding participants and personnel

ALL-CAUSE MORTALITY: perfect agreement

HEALTH-RELATED QUALITY OF LIFE: perfect agreement

SEVERE HYPOGLYCAEMIA

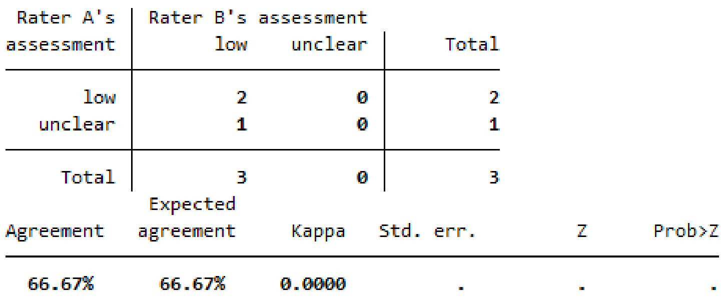


```
 Rater A's |       Rater B's assessment
assessment |      low    unclear       high |     Total
-----------+--------------------------------+----------
       low |       18          0          0 |        18
   unclear |        3          0          0 |         3
      high |        1          0          0 |         1
-----------+--------------------------------+----------
     Total |       22          0          0 |        22
            Expected
Agreement  agreement     Kappa   Std. err.         Z      Prob>Z
---------------------------------------------------------------
  81.82%     81.82%     0.0000     0.0000       0.00      0.5000
```

Gwet's AC1 with 95% CI: 0.80 (0.61 to 0.99)

CARDIOVASCULAR MORTALITY: perfect agreement

NON-FATAL MYOCARDIAL INFARCTION/STROKE

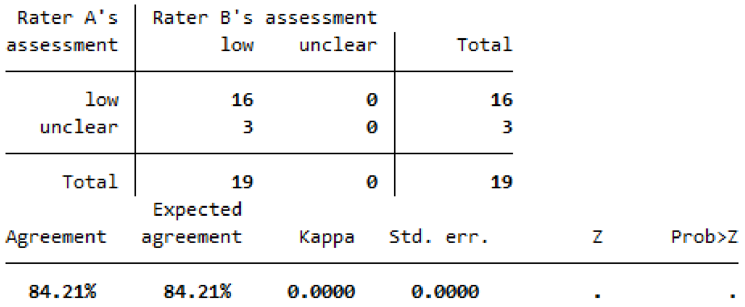


```
 Rater A's | Rater B's assessment
assessment |       low    unclear |     Total
-----------+----------------------+----------
       low |         2          0 |         2
   unclear |         1          0 |         1
-----------+----------------------+----------
     Total |         3          0 |         3
             Expected
Agreement   agreement     Kappa   Std. err.         Z      Prob>Z
-----------------------------------------------------------------
  66.67%      66.67%     0.0000          .          .          .
```

Gwet's AC1 with 95% CI: 0.54 (-0.43 to 1.50)

END-STAGE RENAL DISEASE/BLINDNESS: perfect agreement

SERIOUS ADVERSE EVENTS

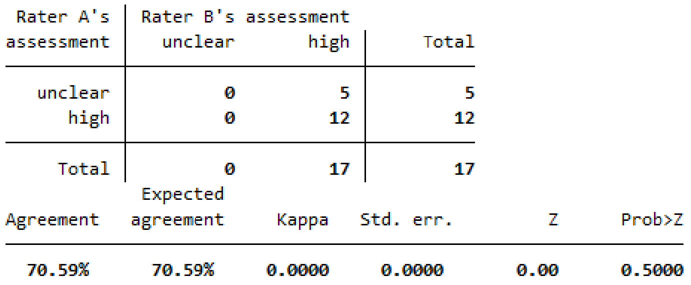


```
 Rater A's | Rater B's assessment
assessment |       low    unclear |     Total
-----------+----------------------+----------
       low |        16          0 |        16
   unclear |         3          0 |         3
-----------+----------------------+----------
     Total |        19          0 |        19
             Expected
Agreement   agreement     Kappa   Std. err.         Z      Prob>Z
-----------------------------------------------------------------
  84.21%      84.21%     0.0000     0.0000          .          .
```

Gwet's AC1 with 95% CI: 0.82 (0.59 to 1.04)

DIABETIC KETOACIDOSIS: perfect agreement

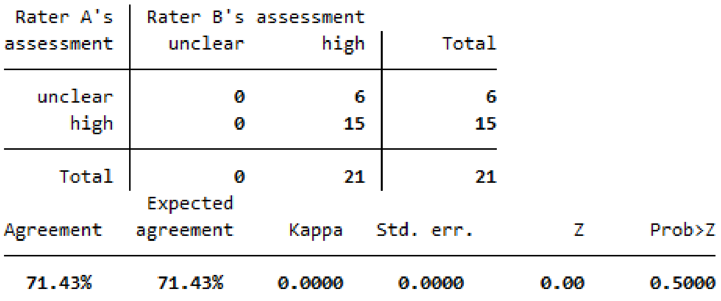
NON-SERIOUS ADVERSE EVENTS



```
 Rater A's | Rater B's assessment
assessment |   unclear       high |     Total
-----------+----------------------+----------
   unclear |         0          5 |         5
      high |         0         12 |        12
-----------+----------------------+----------
     Total |         0         17 |        17
             Expected
Agreement   agreement     Kappa   Std. err.         Z      Prob>Z
-----------------------------------------------------------------
  70.59%      70.59%     0.0000     0.0000       0.00      0.5000
```

Gwet's AC1 with 95% CI: 0.61 (0.24 to 0.98)

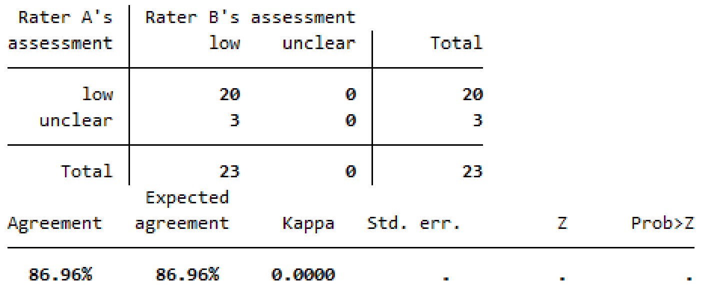
MILD/MODERATE HYPOGLYCAEMIA



```
 Rater A's | Rater B's assessment
assessment |   unclear       high |     Total
-----------+----------------------+----------
   unclear |         0          6 |         6
      high |         0         15 |        15
-----------+----------------------+----------
     Total |         0         21 |        21
             Expected
Agreement   agreement     Kappa   Std. err.         Z      Prob>Z
-----------------------------------------------------------------
  71.43%      71.43%     0.0000     0.0000       0.00      0.5000
```

Gwet's AC1 with 95% CI: 0.62 (0.30 to 0.95)

HBA1C



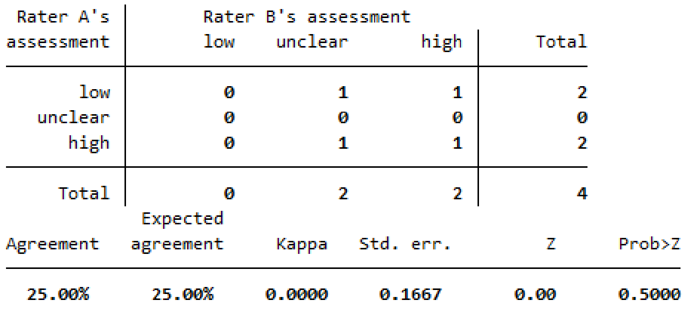
```
 Rater A's │ Rater B's assessment
assessment │       low   unclear │     Total
───────────┼───────────────────┼──────────
       low │        20         0 │        20
   unclear │         3         0 │         3
───────────┼───────────────────┼──────────
     Total │        23         0 │        23
             Expected
Agreement   agreement     Kappa   Std. err.          Z      Prob>Z
──────────────────────────────────────────────────────────────────
  86.96%      86.96%     0.0000          .           .          .
```

Gwet's AC1 with 95% CI: 0.85 (0.67 to 1.03)

Blinding outcome assessment

ALL-CAUSE MORTALITY: perfect agreement

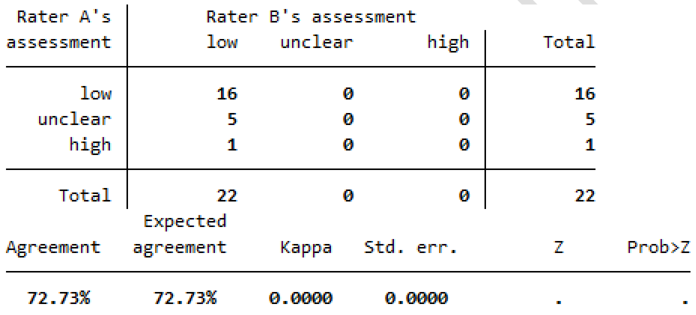
HEALTH-RELATED QUALITY OF LIFE



```
 Rater A's │      Rater B's assessment
assessment │       low   unclear      high │     Total
───────────┼─────────────────────────────┼──────────
       low │         0         1         1 │         2
   unclear │         0         0         0 │         0
      high │         0         1         1 │         2
───────────┼─────────────────────────────┼──────────
     Total │         0         2         2 │         4
             Expected
Agreement   agreement     Kappa   Std. err.          Z      Prob>Z
──────────────────────────────────────────────────────────────────
  25.00%      25.00%     0.0000     0.1667        0.00      0.5000
```

Gwet's AC1 with 95% CI: -0.09 (-0.82 to 0.64)

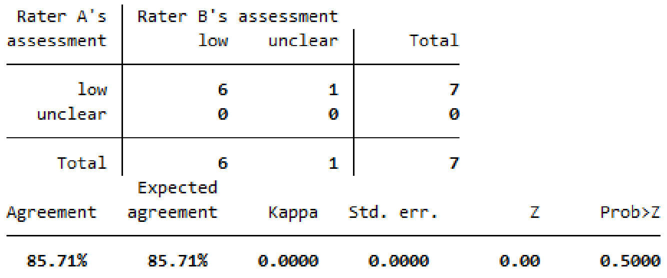
SEVERE HYPOGLYCAEMIA



```
 Rater A's │      Rater B's assessment
assessment │       low   unclear      high │     Total
───────────┼─────────────────────────────┼──────────
       low │        16         0         0 │        16
   unclear │         5         0         0 │         5
      high │         1         0         0 │         1
───────────┼─────────────────────────────┼──────────
     Total │        22         0         0 │        22
             Expected
Agreement   agreement     Kappa   Std. err.          Z      Prob>Z
──────────────────────────────────────────────────────────────────
  72.73%      72.73%     0.0000     0.0000           .          .
```

Gwet's AC1 with 95% CI: 0.69 (0.45 to 0.93)

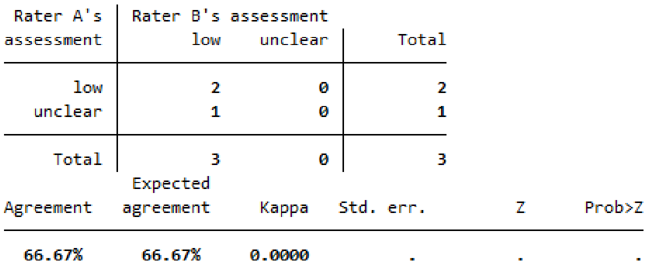
CARDIOVASCULAR MORTALITY



```
 Rater A's │ Rater B's assessment
assessment │       low   unclear │     Total
───────────┼───────────────────┼──────────
       low │         6         1 │         7
   unclear │         0         0 │         0
───────────┼───────────────────┼──────────
     Total │         6         1 │         7
             Expected
Agreement   agreement     Kappa   Std. err.          Z      Prob>Z
──────────────────────────────────────────────────────────────────
  85.71%      85.71%     0.0000     0.0000        0.00      0.5000
```

Gwet's AC1 with 95% CI: 0.84 (0.49 to 1.18)

NON-FATAL MYOCARDIAL INFARCTION/STROKE

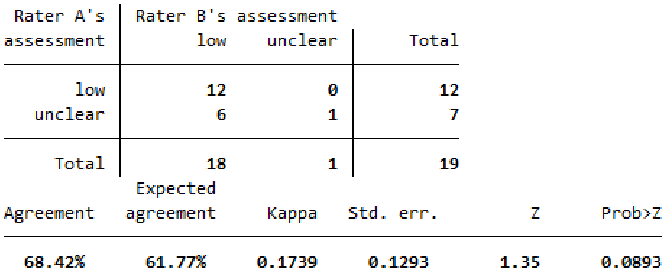


```
 Rater A's │ Rater B's assessment
assessment │      low    unclear │     Total
───────────┼─────────────────────┼──────────
       low │        2          0 │         2
   unclear │        1          0 │         1
───────────┼─────────────────────┼──────────
     Total │        3          0 │         3
             Expected
Agreement   agreement     Kappa   Std. err.         Z      Prob>Z
─────────────────────────────────────────────────────────────────
  66.67%      66.67%     0.0000          .          .          .
```

Gwet's AC1 with 95% CI: 0.53 (-0.43 to 1.50)

END-STAGE RENAL DISEASE/BLINDNESS: 1 low vs unclear

SERIOUS ADVERSE EVENTS

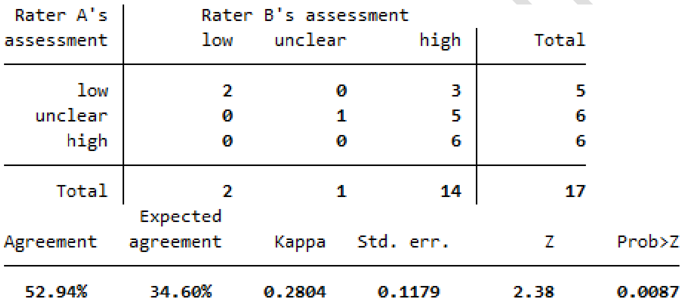


```
 Rater A's │ Rater B's assessment
assessment │      low    unclear │     Total
───────────┼─────────────────────┼──────────
       low │       12          0 │        12
   unclear │        6          1 │         7
───────────┼─────────────────────┼──────────
     Total │       18          1 │        19
             Expected
Agreement   agreement     Kappa   Std. err.         Z      Prob>Z
─────────────────────────────────────────────────────────────────
  68.42%      61.77%     0.1739     0.1293       1.35      0.0893
```

95% CI for kappa: -0.13 to 0.48
Gwet's AC1 with 95% CI: 0.53 (0.13 to 0.92)

DIABETIC KETOACIDOSIS: perfect agreement

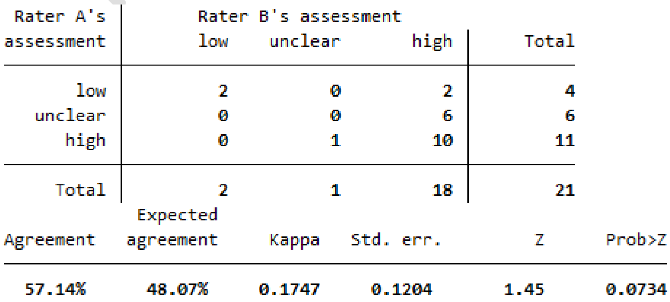
NON-SERIOUS ADVERSE EVENTS



```
 Rater A's │        Rater B's assessment
assessment │      low    unclear       high │     Total
───────────┼────────────────────────────────┼──────────
       low │        2          0          3 │         5
   unclear │        0          1          5 │         6
      high │        0          0          6 │         6
───────────┼────────────────────────────────┼──────────
     Total │        2          1         14 │        17
             Expected
Agreement   agreement     Kappa   Std. err.         Z      Prob>Z
─────────────────────────────────────────────────────────────────
  52.94%      34.60%     0.2804     0.1179       2.38      0.0087
```

95% CI for kappa: 0.01 to 0.55
Gwet's AC1 with 95% CI: 0.34 (-0.01 to 0.69)

MILD/MODERATE HYPOGLYCAEMIA

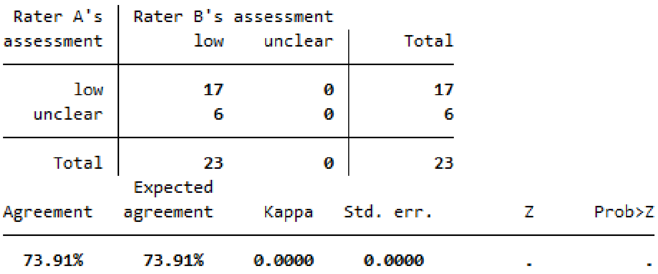


```
 Rater A's │        Rater B's assessment
assessment │      low    unclear       high │     Total
───────────┼────────────────────────────────┼──────────
       low │        2          0          2 │         4
   unclear │        0          0          6 │         6
      high │        0          1         10 │        11
───────────┼────────────────────────────────┼──────────
     Total │        2          1         18 │        21
             Expected
Agreement   agreement     Kappa   Std. err.         Z      Prob>Z
─────────────────────────────────────────────────────────────────
  57.14%      48.07%     0.1747     0.1204       1.45      0.0734
```

95% CI for kappa: -014 to 0.49

Gwet’s AC1 with 95% CI: 0.44 (0.13 to 0.75)

HBA1C

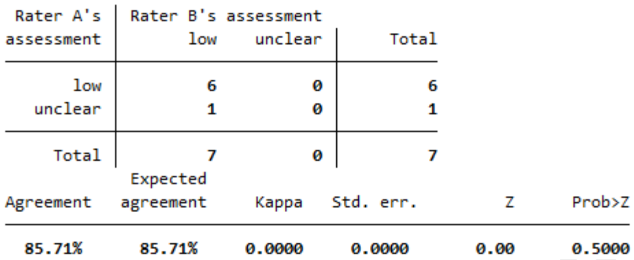


```
  Rater A's | Rater B's assessment
 assessment |       low    unclear |     Total
------------+----------------------+----------
        low |        17          0 |        17
    unclear |         6          0 |         6
------------+----------------------+----------
      Total |        23          0 |        23
            Expected
Agreement   agreement     Kappa   Std. err.         Z      Prob>Z
-----------------------------------------------------------------
  73.91%      73.91%     0.0000     0.0000          .          .
```

Gwet’s AC1 with 95% CI: 0.66 (0.37 to 0.95)

Incomplete outcome data

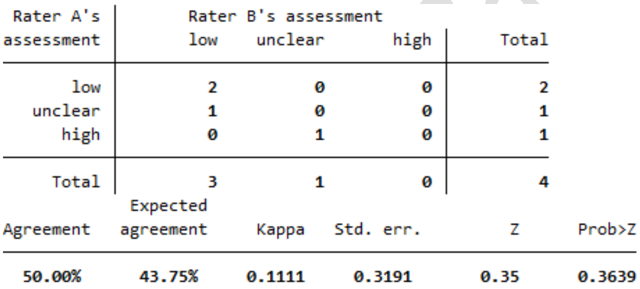
ALL-CAUSE MORTALITY



```
  Rater A's | Rater B's assessment
 assessment |       low    unclear |     Total
------------+----------------------+----------
        low |         6          0 |         6
    unclear |         1          0 |         1
------------+----------------------+----------
      Total |         7          0 |         7
            Expected
Agreement   agreement     Kappa   Std. err.         Z      Prob>Z
-----------------------------------------------------------------
  85.71%      85.71%     0.0000     0.0000       0.00      0.5000
```

Gwet’s AC1 with 95% CI: 0.84 (0.49 to 1.18)

HEALTH-RELATED QUALITY OF LIFE

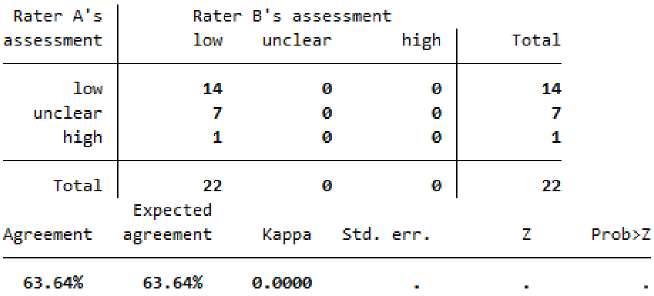


```
  Rater A's |        Rater B's assessment
 assessment |       low    unclear       high |     Total
------------+---------------------------------+----------
        low |         2          0          0 |         2
    unclear |         1          0          0 |         1
       high |         0          1          0 |         1
------------+---------------------------------+----------
      Total |         3          1          0 |         4
            Expected
Agreement   agreement     Kappa   Std. err.         Z      Prob>Z
-----------------------------------------------------------------
  50.00%      43.75%     0.1111     0.3191       0.35      0.3639
```

95% CI for kappa: -0.29 to 0.51

Gwet’s AC1 with 95% CI: 0.32 (-0.49 to 1.13)

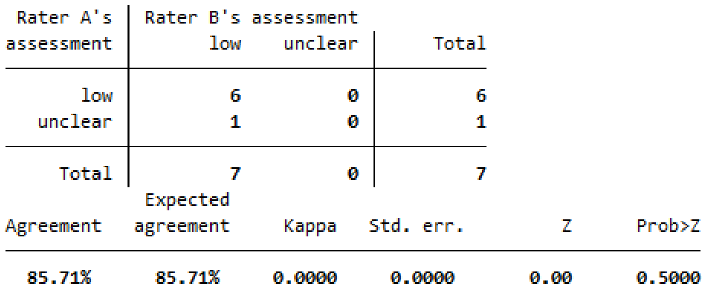
SEVERE HYPOGLYCAEMIA



```
  Rater A's |        Rater B's assessment
 assessment |       low    unclear       high |     Total
------------+---------------------------------+----------
        low |        14          0          0 |        14
    unclear |         7          0          0 |         7
       high |         1          0          0 |         1
------------+---------------------------------+----------
      Total |        22          0          0 |        22
            Expected
Agreement   agreement     Kappa   Std. err.         Z      Prob>Z
-----------------------------------------------------------------
  63.64%      63.64%     0.0000          .          .          .
```

Gwet’s AC1 with 95% CI: 0.57 (0.30 to 0.84)

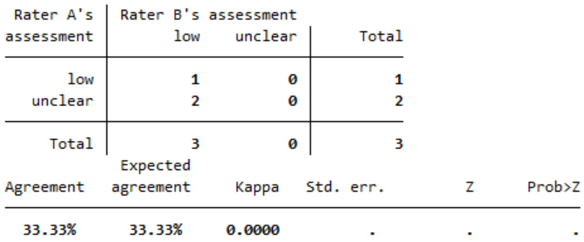
CARDIOVASCULAR MORTALITY



```
 Rater A's | Rater B's assessment
assessment |       low    unclear |     Total
-----------+----------------------+----------
       low |         6          0 |         6
   unclear |         1          0 |         1
-----------+----------------------+----------
     Total |         7          0 |         7
             Expected
Agreement   agreement     Kappa   Std. err.         Z      Prob>Z
-----------------------------------------------------------------
  85.71%      85.71%     0.0000     0.0000       0.00      0.5000
```

Gwet’s AC1 with 95% CI: 0.84 (0.49 to 1.18)

NON-FATAL MYOCARDIAL INFARCTION/STROKE

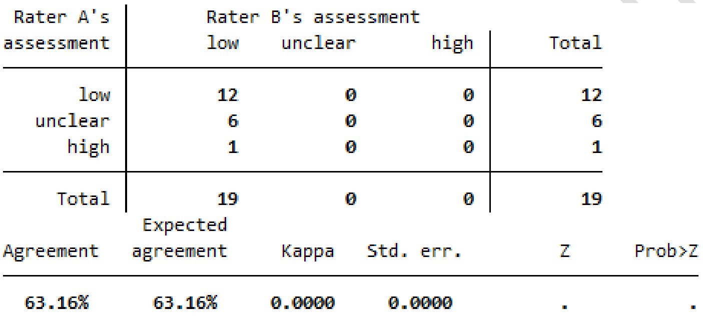


```
 Rater A's | Rater B's assessment
assessment |       low    unclear |     Total
-----------+----------------------+----------
       low |         1          0 |         1
   unclear |         2          0 |         2
-----------+----------------------+----------
     Total |         3          0 |         3
             Expected
Agreement   agreement     Kappa   Std. err.         Z      Prob>Z
-----------------------------------------------------------------
  33.33%      33.33%     0.0000          .          .           .
```

Gwet’s AC1 with 95% CI: -0.20 (-1.54 to 1.14)

END-STAGE RENAL DISEASE/BLINDNESS: 1x unclear versus low

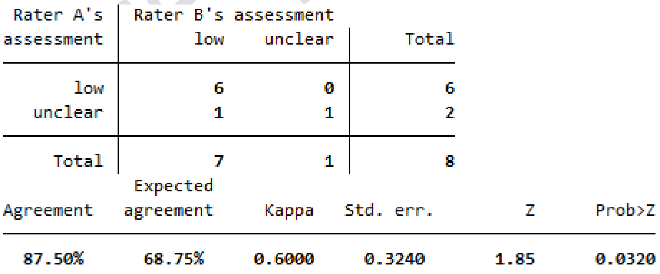
SERIOUS ADVERSE EVENTS



```
 Rater A's |        Rater B's assessment
assessment |       low    unclear       high |     Total
-----------+---------------------------------+----------
       low |        12          0          0 |        12
   unclear |         6          0          0 |         6
      high |         1          0          0 |         1
-----------+---------------------------------+----------
     Total |        19          0          0 |        19
             Expected
Agreement   agreement     Kappa   Std. err.         Z      Prob>Z
-----------------------------------------------------------------
  63.16%      63.16%     0.0000     0.0000          .           .
```

Gwet’s AC1 with 95% CI: 0.56 (0.27 to 0.86)

DIABETIC KETOACIDOSIS

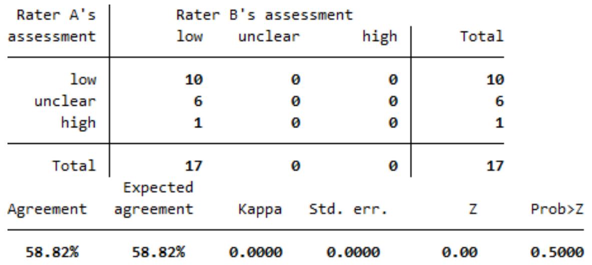


```
 Rater A's | Rater B's assessment
assessment |       low    unclear |     Total
-----------+----------------------+----------
       low |         6          0 |         6
   unclear |         1          1 |         2
-----------+----------------------+----------
     Total |         7          1 |         8
             Expected
Agreement   agreement     Kappa   Std. err.         Z      Prob>Z
-----------------------------------------------------------------
  87.50%      68.75%     0.6000     0.3240       1.85      0.0320
```

95% CI for kappa: -0.07 to 1

Gwet’s AC1 with 95% CI: 0.82 (0.46 to 1.18)

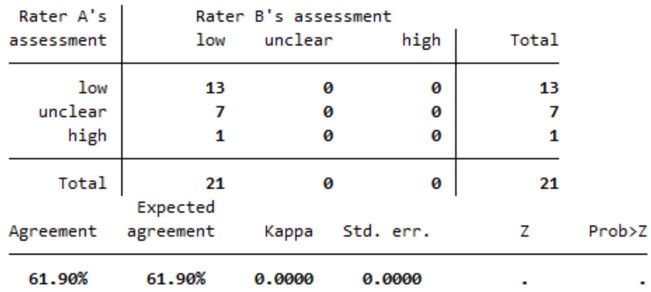
NON-SERIOUS ADVERSE EVENTS



```
 Rater A's |        Rater B's assessment
assessment |       low    unclear       high |     Total
-----------+---------------------------------+----------
       low |        10          0          0 |        10
   unclear |         6          0          0 |         6
      high |         1          0          0 |         1
-----------+---------------------------------+----------
     Total |        17          0          0 |        17
             Expected
Agreement   agreement     Kappa   Std. err.         Z      Prob>Z
-----------------------------------------------------------------
  58.82%      58.82%     0.0000     0.0000       0.00      0.5000
```

Gwet’s AC1 with 95% CI: 0.50 (0.18 to 0.83)

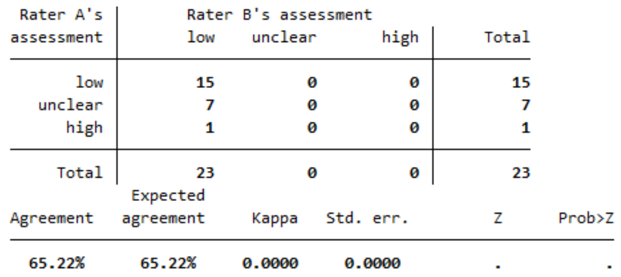
Mild/moderate hypoglycaemia



```
 Rater A's |        Rater B's assessment
assessment |       low    unclear       high |     Total
-----------+---------------------------------+----------
       low |        13          0          0 |        13
   unclear |         7          0          0 |         7
      high |         1          0          0 |         1
-----------+---------------------------------+----------
     Total |        21          0          0 |        21
             Expected
Agreement   agreement     Kappa   Std. err.         Z      Prob>Z
-----------------------------------------------------------------
  61.90%      61.90%     0.0000     0.0000          .           .
```

Gwet’s AC1 with 95% CI: 0.55 (0.26 to 0.83)

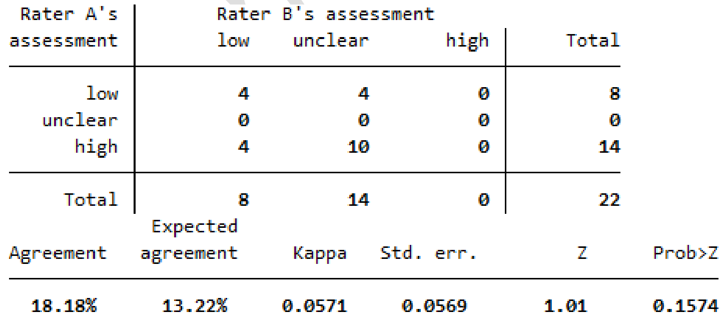
HbA1c



```
 Rater A's |        Rater B's assessment
assessment |       low    unclear       high |     Total
-----------+---------------------------------+----------
       low |        15          0          0 |        15
   unclear |         7          0          0 |         7
      high |         1          0          0 |         1
-----------+---------------------------------+----------
     Total |        23          0          0 |        23
             Expected
Agreement   agreement     Kappa   Std. err.         Z      Prob>Z
-----------------------------------------------------------------
  65.22%      65.22%     0.0000     0.0000          .           .
```

Gwet’s AC1 with 95% CI: 0.59 (0.33 to 0.85)

Selective reporting



```
 Rater A's |        Rater B's assessment
assessment |       low    unclear       high |     Total
-----------+---------------------------------+----------
       low |         4          4          0 |         8
   unclear |         0          0          0 |         0
      high |         4         10          0 |        14
-----------+---------------------------------+----------
     Total |         8         14          0 |        22
             Expected
Agreement   agreement     Kappa   Std. err.         Z      Prob>Z
-----------------------------------------------------------------
  18.18%      13.22%     0.0571     0.0569       1.01      0.1574
```

95% CI for kappa: -0.06 to 0.17
Gwet’s AC1 with 95% CI: -0.23 (-0.48 to 0.03)

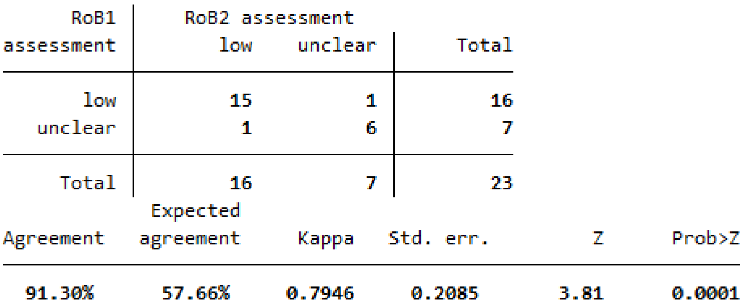
Other risk of bias

Disagreement because of ‘sponsor bias’ (unclear) Rater A versus ‘none’ Rater B

(a2) RoB 1 versus RoB 2 (using publications)

[Unclear in RoB 1 was set to some concerns for RoB 2 comparison]

Randomisation sequence + allocation concealment vs randomisation process (for all outcomes)
Note: if rando + allo were low > low; if either rando or allo were unclear > some concerns

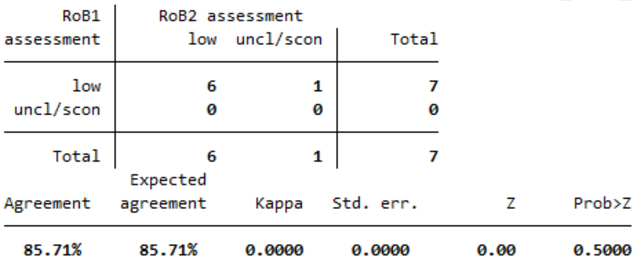


```
      RoB1 |   RoB2 assessment
assessment |       low    unclear |     Total
-----------+----------------------+----------
       low |        15          1 |        16 
   unclear |         1          6 |         7 
-----------+----------------------+----------
     Total |        16          7 |        23 
             Expected
Agreement   agreement     Kappa   Std. err.         Z      Prob>Z
-----------------------------------------------------------------
  91.30%      57.66%     0.7946     0.2085       3.81      0.0001
```

95% CI for kappa: 0.52 to 1
Gwet’s AC1 with 95% CI: 0.85 (0.64 to 1.06)

Blinding participants and personnel vs deviations from intended interventions

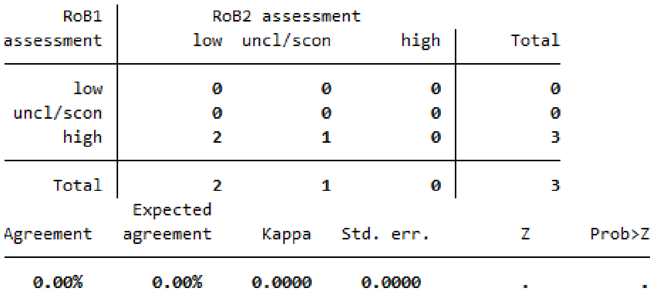
ALL-CAUSE MORTALITY



```
      RoB1 |   RoB2 assessment
assessment |       low  uncl/scon |     Total
-----------+----------------------+----------
       low |         6          1 |         7 
 uncl/scon |         0          0 |         0 
-----------+----------------------+----------
     Total |         6          1 |         7 
             Expected
Agreement   agreement     Kappa   Std. err.         Z      Prob>Z
-----------------------------------------------------------------
  85.71%      85.71%     0.0000     0.0000       0.00      0.5000
```

Gwet’s AC1 with 95% CI: 0.84 (0.49 to 1.17)

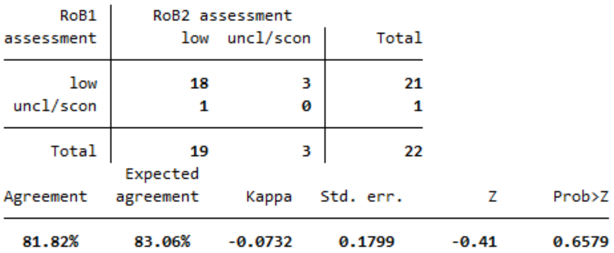
HEALTH-RELATED QUALITY OF LIFE



```
      RoB1 |          RoB2 assessment
assessment |       low  uncl/scon       high |     Total
-----------+---------------------------------+----------
       low |         0          0          0 |         0 
 uncl/scon |         0          0          0 |         0 
      high |         2          1          0 |         3 
-----------+---------------------------------+----------
     Total |         2          1          0 |         3 
             Expected
Agreement   agreement     Kappa   Std. err.         Z      Prob>Z
-----------------------------------------------------------------
   0.00%       0.00%     0.0000     0.0000          .           .
```

Gwet’s AC1 with 95% CI: -0.44 (-0.53 to -0.35)

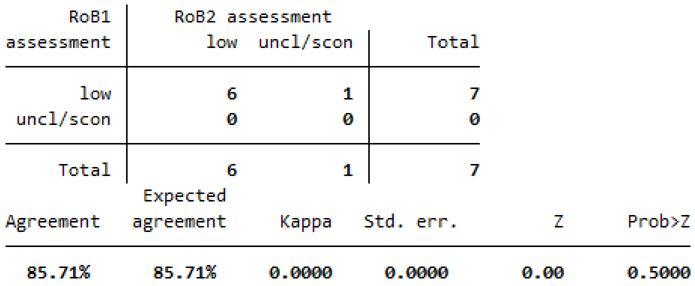
SEVERE HYPOGLYCAEMIA



```
     RoB1 |   RoB2 assessment
assessment|      low  uncl/scon |     Total
----------+---------------------+----------
       low|       18          3 |        21
 uncl/scon|        1          0 |         1
----------+---------------------+----------
     Total|       19          3 |        22
             Expected
Agreement   agreement     Kappa   Std. err.         Z      Prob>Z
-----------------------------------------------------------------
  81.82%      83.06%    -0.0732    0.1799       -0.41      0.6579
```

Gwet's AC1 with 95% CI: 0.78 (0.55 to 1.01)

CARDIOVASCULAR MORTALITY



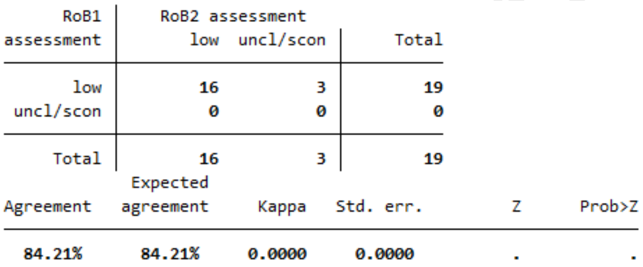
```
     RoB1 |   RoB2 assessment
assessment|      low  uncl/scon |     Total
----------+---------------------+----------
       low|        6          1 |         7
 uncl/scon|        0          0 |         0
----------+---------------------+----------
     Total|        6          1 |         7
             Expected
Agreement   agreement     Kappa   Std. err.         Z      Prob>Z
-----------------------------------------------------------------
  85.71%      85.71%     0.0000    0.0000        0.00      0.5000
```

Gwet's AC1 with 95% CI: 0.84 (0.49 to 1.18)

NON-FATAL MYOCARDIAL INFARCTION/STROKE: perfect agreement

END-STAGE RENAL DISEASE/BLINDNESS: perfect agreement

SERIOUS ADVERSE EVENTS

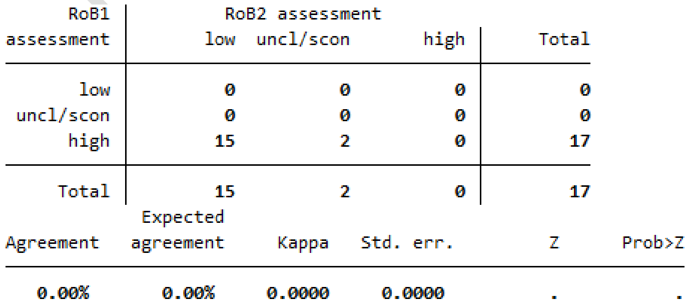


```
     RoB1 |   RoB2 assessment
assessment|      low  uncl/scon |     Total
----------+---------------------+----------
       low|       16          3 |        19
 uncl/scon|        0          0 |         0
----------+---------------------+----------
     Total|       16          3 |        19
             Expected
Agreement   agreement     Kappa   Std. err.         Z      Prob>Z
-----------------------------------------------------------------
  84.21%      84.21%     0.0000    0.0000           .           .
```

Gwet's AC1 with 95% CI: 0.82 (0.59 to 1.04)

DIABETIC KETOACIDOSIS: perfect agreement

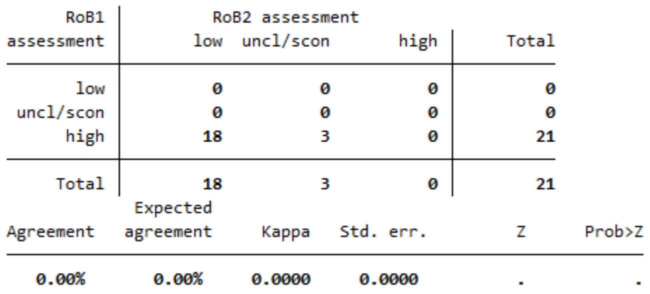
NON-SERIOUS ADVERSE EVENTS: disagreement



```
     RoB1 |          RoB2 assessment
assessment|      low  uncl/scon       high |     Total
----------+--------------------------------+----------
       low|        0          0          0 |         0
 uncl/scon|        0          0          0 |         0
      high|       15          2          0 |        17
----------+--------------------------------+----------
     Total|       15          2          0 |        17
             Expected
Agreement   agreement     Kappa   Std. err.         Z      Prob>Z
-----------------------------------------------------------------
   0.00%       0.00%     0.0000    0.0000           .           .
```

Gwet's AC1 with 95% CI: -0.38 (-0.44 to -0.33)

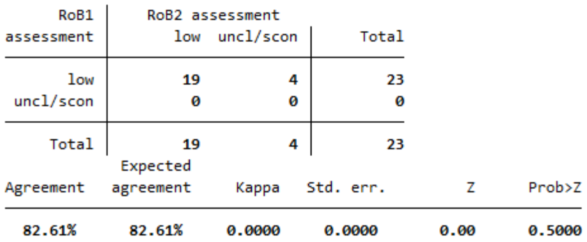
MILD/MODERATE HYPOGLYCAEMIA



```
      RoB1 |        RoB2 assessment
assessment |       low  uncl/scon       high |     Total
-----------+---------------------------------+----------
       low |         0          0          0 |         0
 uncl/scon |         0          0          0 |         0
      high |        18          3          0 |        21
-----------+---------------------------------+----------
     Total |        18          3          0 |        21
             Expected
Agreement   agreement     Kappa   Std. err.         Z      Prob>Z
-----------------------------------------------------------------
  0.00%       0.00%     0.0000     0.0000          .          .
```

Gwet's AC1 with 95% CI: -0.39 (-0.44 to -0.33)

HBA1C



```
      RoB1 |    RoB2 assessment
assessment |       low  uncl/scon |     Total
-----------+----------------------+----------
       low |        19          4 |        23
 uncl/scon |         0          0 |         0
-----------+----------------------+----------
     Total |        19          4 |        23
             Expected
Agreement   agreement     Kappa   Std. err.         Z      Prob>Z
-----------------------------------------------------------------
  82.61%     82.61%     0.0000     0.0000       0.00      0.5000
```

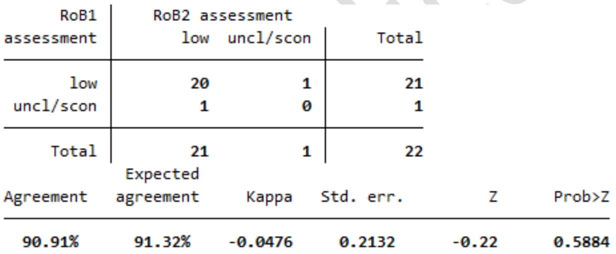
Gwet's AC1 with 95% CI: 0.79 (0.58 to 1.01)

Incomplete outcome data vs missing outcome data

ALL-CAUSE MORTALITY: perfect agreement

HEALTH-RELATED QUALITY OF LIFE: perfect agreement

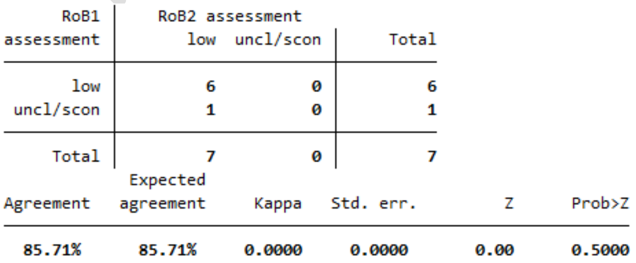
SEVERE HYPOGLYCAEMIA



```
      RoB1 |    RoB2 assessment
assessment |       low  uncl/scon |     Total
-----------+----------------------+----------
       low |        20          1 |        21
 uncl/scon |         1          0 |         1
-----------+----------------------+----------
     Total |        21          1 |        22
             Expected
Agreement   agreement     Kappa   Std. err.         Z      Prob>Z
-----------------------------------------------------------------
  90.91%     91.32%    -0.0476     0.2132      -0.22      0.5884
```

Gwet's AC1 with 95% CI: 0.90 (0.76 to 1.04)

CARDIOVASCULAR MORTALITY



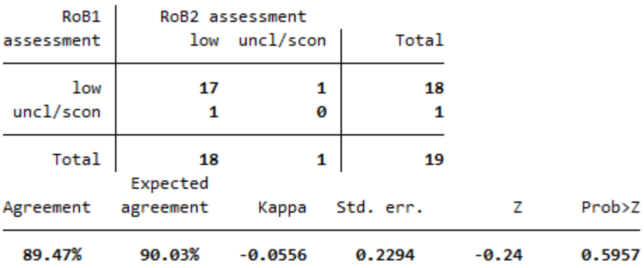
```
      RoB1 |    RoB2 assessment
assessment |       low  uncl/scon |     Total
-----------+----------------------+----------
       low |         6          0 |         6
 uncl/scon |         1          0 |         1
-----------+----------------------+----------
     Total |         7          0 |         7
             Expected
Agreement   agreement     Kappa   Std. err.         Z      Prob>Z
-----------------------------------------------------------------
  85.71%     85.71%     0.0000     0.0000       0.00      0.5000
```

Gwet's AC1 with 95% CI: 0.84 (0.49 to 1.18)

NON-FATAL MYOCARDIAL INFARCTION/STROKE: perfect agreement

END-STAGE RENAL DISEASE/BLINDNESS: perfect agreement

SERIOUS ADVERSE EVENTS

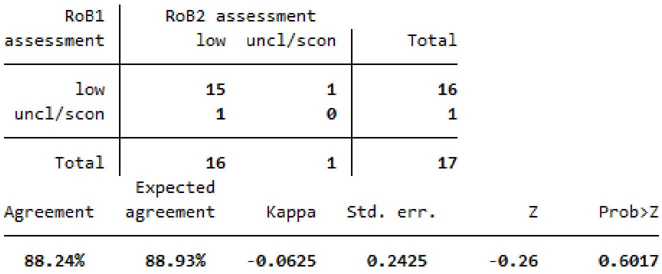


```
      RoB1 |   RoB2 assessment
assessment |       low  uncl/scon |     Total
-----------+----------------------+----------
       low |        17          1 |        18
 uncl/scon |         1          0 |         1
-----------+----------------------+----------
     Total |        18          1 |        19
              Expected
Agreement    agreement     Kappa   Std. err.         Z      Prob>Z
-----------------------------------------------------------------
  89.47%      90.03%     -0.0556     0.2294      -0.24      0.5957
```

Gwet’s AC1 with 95% CI: 0.88 (0.71 to 1.05)

DIABETIC KETOACIDOSIS: perfect agreement

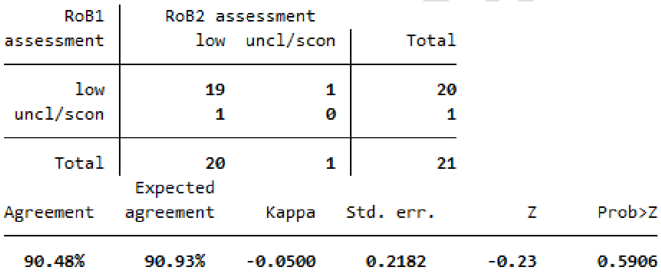
NON-SERIOUS ADVERSE EVENTS



```
      RoB1 |   RoB2 assessment
assessment |       low  uncl/scon |     Total
-----------+----------------------+----------
       low |        15          1 |        16
 uncl/scon |         1          0 |         1
-----------+----------------------+----------
     Total |        16          1 |        17
              Expected
Agreement    agreement     Kappa   Std. err.         Z      Prob>Z
-----------------------------------------------------------------
  88.24%      88.93%     -0.0625     0.2425      -0.26      0.6017
```

Gwet’s AC1 with 95% CI: 0.87 (0.68 to 1.06)

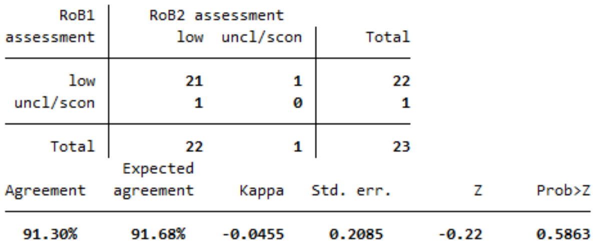
MILD/MODERATE HYPOGLYCAEMIA



```
      RoB1 |   RoB2 assessment
assessment |       low  uncl/scon |     Total
-----------+----------------------+----------
       low |        19          1 |        20
 uncl/scon |         1          0 |         1
-----------+----------------------+----------
     Total |        20          1 |        21
              Expected
Agreement    agreement     Kappa   Std. err.         Z      Prob>Z
-----------------------------------------------------------------
  90.48%      90.93%     -0.0500     0.2182      -0.23      0.5906
```

Gwet’s AC1 with 95% CI: 0.90 (0.74 to 1.05)

HBA1C



```
      RoB1 |   RoB2 assessment
assessment |       low  uncl/scon |     Total
-----------+----------------------+----------
       low |        21          1 |        22
 uncl/scon |         1          0 |         1
-----------+----------------------+----------
     Total |        22          1 |        23
              Expected
Agreement    agreement     Kappa   Std. err.         Z      Prob>Z
-----------------------------------------------------------------
  91.30%      91.68%     -0.0455     0.2085      -0.22      0.5863
```

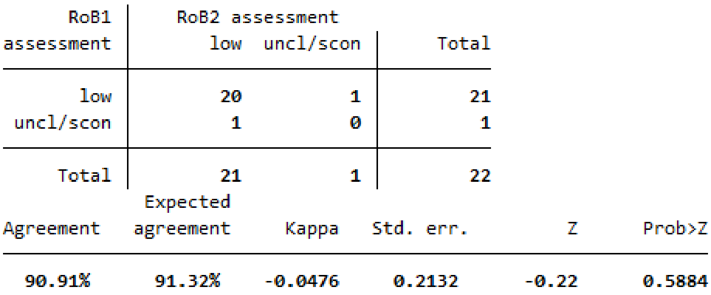
Gwet’s AC1 with 95% CI: 0.91 (0.77 to 1.04)

Blinding outcome assessment vs measurement of the outcome

ALL-CAUSE MORTALITY: perfect agreement

HEALTH-RELATED QUALITY OF LIFE: disagreement: RoB1 all high vs RoB 2 all some concerns

SEVERE HYPOGLYCAEMIA



```
      RoB1 |     RoB2 assessment
assessment |       low  uncl/scon |     Total
-----------+----------------------+----------
       low |        20          1 |        21 
 uncl/scon |         1          0 |         1 
-----------+----------------------+----------
     Total |        21          1 |        22 
             Expected
Agreement   agreement     Kappa   Std. err.         Z      Prob>Z
-----------------------------------------------------------------
  90.91%      91.32%    -0.0476     0.2132      -0.22      0.5884
```

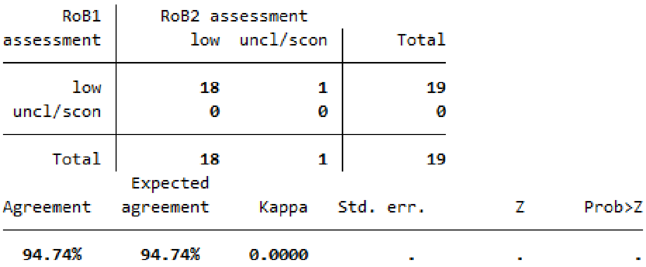
Gwet’s AC1 with 95% CI: 0.90 (0.76 to 1.04)

CARDIOVASCULAR MORTALITY: perfect agreement

NON-FATAL MYOCARDIAL INFARCTION/STROKE: perfect agreement

END-STAGE RENAL DISEASE/BLINDNESS: 1 unclear vs low

SERIOUS ADVERSE EVENTS



```
      RoB1 |     RoB2 assessment
assessment |       low  uncl/scon |     Total
-----------+----------------------+----------
       low |        18          1 |        19 
 uncl/scon |         0          0 |         0 
-----------+----------------------+----------
     Total |        18          1 |        19 
             Expected
Agreement   agreement     Kappa   Std. err.         Z      Prob>Z
-----------------------------------------------------------------
  94.74%      94.74%     0.0000         .           .          .
```

Gwet’s AC1 with 95% CI: 0.94 (0.83 to 1.06)

DIABETIC KETOACIDOSIS: perfect agreement

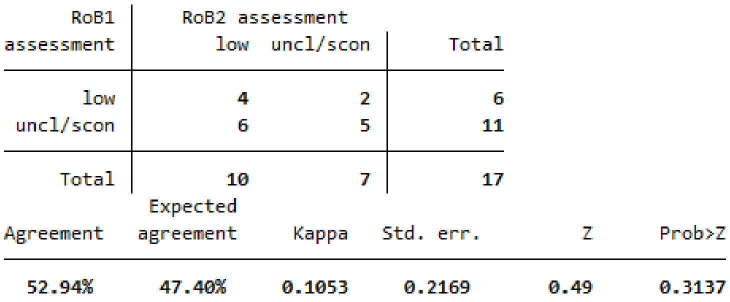
NON-SERIOUS ADVERSE EVENTS: disagreement; RoB1 all high vs RoB 2 all some concerns

MILD/MODERATE HYPOGLYCAEMIA: disagreement; RoB1 all high vs RoB 2 all some concerns

HBA1C: perfect agreement

Selective reporting vs selection of the reported result (exercise only)

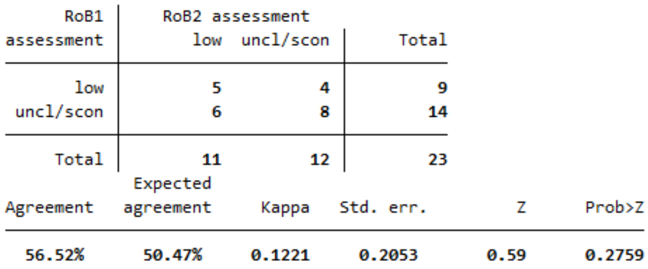
Example for NON-SERIOUS ADVERSE EVENTS



```
      RoB1 |  RoB2 assessment
assessment |      low  uncl/scon |     Total
-----------+---------------------+----------
       low |        4          2 |         6
 uncl/scon |        6          5 |        11
-----------+---------------------+----------
     Total |       10          7 |        17
             Expected
Agreement   agreement     Kappa   Std. err.         Z      Prob>Z
-----------------------------------------------------------------
  52.94%      47.40%     0.1053     0.2169       0.49      0.3137
```

95% CI for kappa: -0.31 to 0.52
Gwet's AC1 with 95% CI: 0.06 (-0.42 to 0.54)

Example for HBA1C



```
      RoB1 |  RoB2 assessment
assessment |      low  uncl/scon |     Total
-----------+---------------------+----------
       low |        5          4 |         9
 uncl/scon |        6          8 |        14
-----------+---------------------+----------
     Total |       11         12 |        23
             Expected
Agreement   agreement     Kappa   Std. err.         Z      Prob>Z
-----------------------------------------------------------------
  56.52%      50.47%     0.1221     0.2053       0.59      0.2759
```

95% CI for kappa: -0.28 to 0.52
Gwet's AC1 with 95% CI: 0.14 (-0.27 to 0.56)

Overall bias (RoB2)

( ) = number of studies

All-cause mortality (n = 7): 86% low; 14% some concerns
Health-related quality of life (n = 4): 100% some concerns
Severe hypoglycaemia (n = 22): 45% low; 55% some concerns
Cardiovascular mortality (n = 7): 86% low; 14% some concerns
Non-fatal myocardial infarction/stoke (n = 3): 67 low; 33% some concerns
End-stage renal disease/blindness (n = 1): 100% low
Serious adverse events (n = 19): 53% low; 47% some concerns
Diabetic ketoacidosis (n = 8): 63% low; 37% some concerns
Non-serious adverse events (n = 17): 100% some concerns
Mild/moderate hypoglycaemia (n = 21): 100% some concerns
HbA1c (n = 23): 48% low; 52% some concerns

(a4) RoB 1 versus RoB 2 (using clinical study reports)

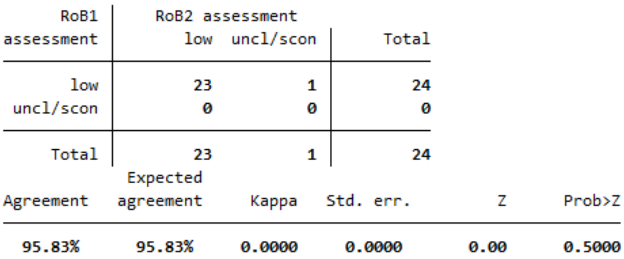
[Unclear in RoB 1 was set to some concerns for RoB 2 comparison]

Randomisation sequence + allocation concealment vs randomisation process (for all outcomes)
Note: if rando + allo were low > low; if either rando or allo were unclear > some concerns

Perfect agreement

Blinding participants and personnel vs deviations from intended interventions

ALL-CAUSE MORTALITY

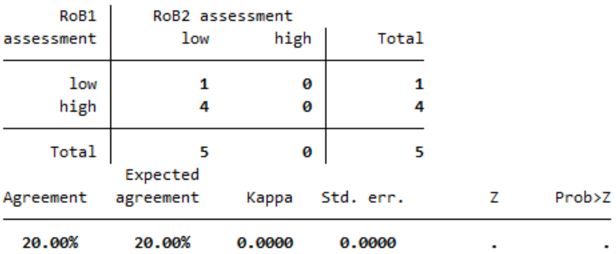


```
     RoB1 |   RoB2 assessment
assessment|      low  uncl/scon |     Total
----------+---------------------+----------
      low |       23          1 |        24
uncl/scon |        0          0 |         0
----------+---------------------+----------
    Total |       23          1 |        24

             Expected
Agreement   agreement     Kappa   Std. err.         Z      Prob>Z
-----------------------------------------------------------------
  95.83%      95.83%     0.0000     0.0000       0.00      0.5000
```

Gwet's AC1 with 95% CI: 0.96 (0.87 to 1.04)

HEALTH-RELATED QUALITY OF LIFE



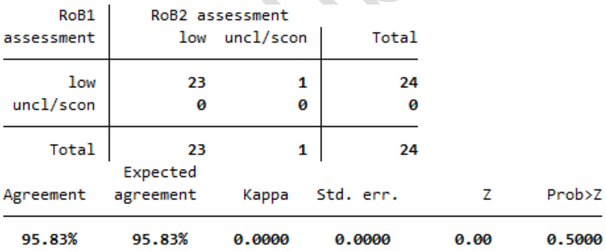
```
     RoB1 |   RoB2 assessment
assessment|      low       high |     Total
----------+---------------------+----------
      low |        1          0 |         1
     high |        4          0 |         4
----------+---------------------+----------
    Total |        5          0 |         5

             Expected
Agreement   agreement     Kappa   Std. err.         Z      Prob>Z
-----------------------------------------------------------------
  20.00%      20.00%     0.0000     0.0000          .           .
```

Gwet's AC1 with 95% CI: -0.54 (-1.42 to 0.34)

SEVERE HYPOGLYCAEMIA: perfect agreement

CARDIOVASCULAR MORTALITY

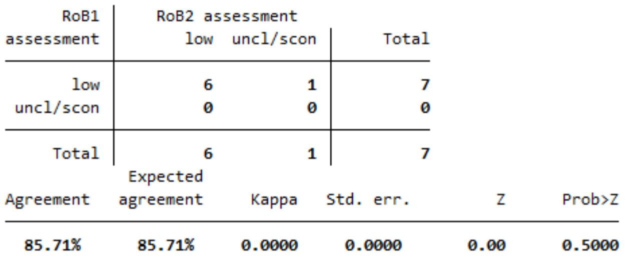


```
     RoB1 |   RoB2 assessment
assessment|      low  uncl/scon |     Total
----------+---------------------+----------
      low |       23          1 |        24
uncl/scon |        0          0 |         0
----------+---------------------+----------
    Total |       23          1 |        24

             Expected
Agreement   agreement     Kappa   Std. err.         Z      Prob>Z
-----------------------------------------------------------------
  95.83%      95.83%     0.0000     0.0000       0.00      0.5000
```

Gwet's AC1 with 95% CI: 0.96 (0.87 to 1.04)

NON-FATAL MYOCARDIAL INFARCTION/STROKE

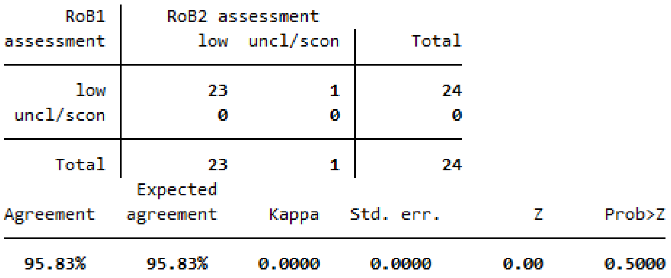


```
     RoB1 |  RoB2 assessment
assessment|      low  uncl/scon |     Total
----------+---------------------+----------
       low|        6          1 |         7
 uncl/scon|        0          0 |         0
----------+---------------------+----------
     Total|        6          1 |         7
            Expected
Agreement   agreement     Kappa   Std. err.         Z      Prob>Z
-----------------------------------------------------------------
  85.71%      85.71%     0.0000     0.0000       0.00      0.5000
```

Gwet’s AC1 with 95% CI: 0.84 (0.49 to 1.18)

END-STAGE RENAL DISEASE/BLINDNESS: perfect agreement (1 study)

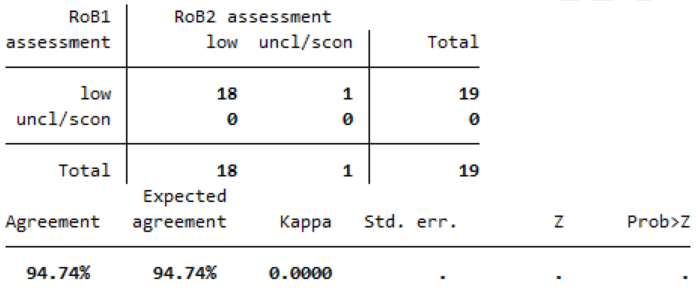
SERIOUS ADVERSE EVENTS



```
     RoB1 |  RoB2 assessment
assessment|      low  uncl/scon |     Total
----------+---------------------+----------
       low|       23          1 |        24
 uncl/scon|        0          0 |         0
----------+---------------------+----------
     Total|       23          1 |        24
            Expected
Agreement   agreement     Kappa   Std. err.         Z      Prob>Z
-----------------------------------------------------------------
  95.83%      95.83%     0.0000     0.0000       0.00      0.5000
```

Gwet’s AC1 with 95% CI: 0.96 (0.87 to 1.04)

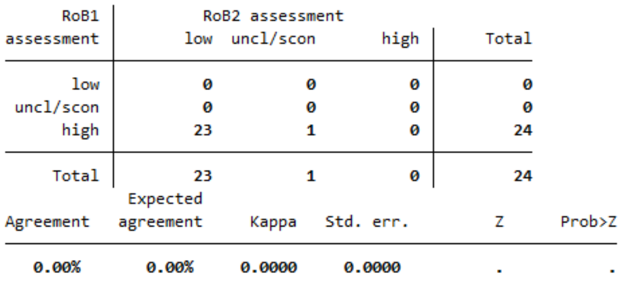
DIABETIC KETOACIDOSIS



```
     RoB1 |  RoB2 assessment
assessment|      low  uncl/scon |     Total
----------+---------------------+----------
       low|       18          1 |        19
 uncl/scon|        0          0 |         0
----------+---------------------+----------
     Total|       18          1 |        19
            Expected
Agreement   agreement     Kappa   Std. err.         Z      Prob>Z
-----------------------------------------------------------------
  94.74%      94.74%     0.0000          .          .           .
```

Gwet’s AC1 with 95% CI: 0.94 (0.83 to 1.06)

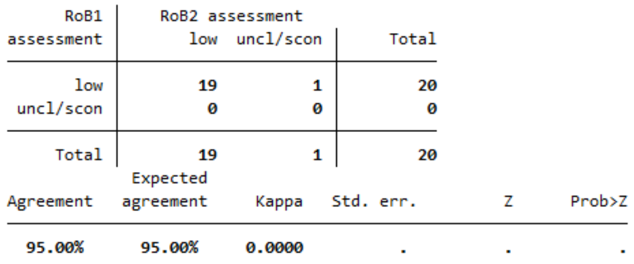
NON-SERIOUS ADVERSE EVENTS



```
     RoB1 |         RoB2 assessment
assessment|      low  uncl/scon       high |     Total
----------+--------------------------------+----------
       low|        0          0          0 |         0
 uncl/scon|        0          0          0 |         0
      high|       23          1          0 |        24
----------+--------------------------------+----------
     Total|       23          1          0 |        24
            Expected
Agreement   agreement     Kappa   Std. err.         Z      Prob>Z
-----------------------------------------------------------------
   0.00%       0.00%     0.0000     0.0000          .           .
```

Gwet’s AC1 with 95% CI: -0.35 (-0.38 to -0.32)

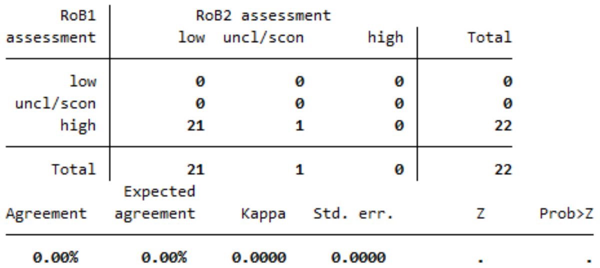
SEVERE NOCTURNAL HYPOGLYCAEMIA



```
      RoB1 |   RoB2 assessment
assessment |      low  uncl/scon |     Total
-----------+----------------------+----------
       low |       19          1 |        20
 uncl/scon |        0          0 |         0
-----------+----------------------+----------
     Total |       19          1 |        20
             Expected
Agreement   agreement     Kappa   Std. err.         Z      Prob>Z
-----------------------------------------------------------------
  95.00%      95.00%     0.0000          .          .          .
```

Gwet's AC1 with 95% CI: 0.95 (0.84 to 1.05)

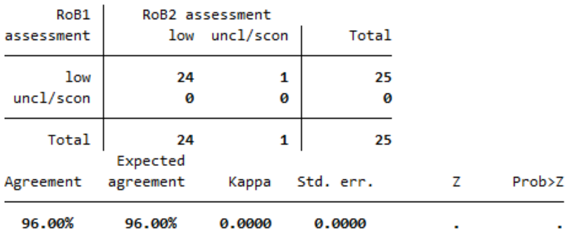
MILD/MODERATE HYPOGLYCAEMIA



```
      RoB1 |        RoB2 assessment
assessment |      low  uncl/scon       high |     Total
-----------+---------------------------------+----------
       low |        0          0          0 |         0
 uncl/scon |        0          0          0 |         0
      high |       21          1          0 |        22
-----------+---------------------------------+----------
     Total |       21          1          0 |        22
             Expected
Agreement   agreement     Kappa   Std. err.         Z      Prob>Z
-----------------------------------------------------------------
   0.00%       0.00%     0.0000     0.0000          .          .
```

Gwet's AC1 with 95% CI: -0.35 (-0.39 to -0.32)

HBA1C



```
      RoB1 |   RoB2 assessment
assessment |      low  uncl/scon |     Total
-----------+----------------------+----------
       low |       24          1 |        25
 uncl/scon |        0          0 |         0
-----------+----------------------+----------
     Total |       24          1 |        25
             Expected
Agreement   agreement     Kappa   Std. err.         Z      Prob>Z
-----------------------------------------------------------------
  96.00%      96.00%     0.0000     0.0000          .          .
```

Gwet's AC1 with 95% CI: 0.96 (0.88 to 1.04)

HBA1C <7% WITHOUT SEVERE HYPOGLYCAEMIA: perfect agreement (4 studies)

Incomplete outcome data vs missing outcome data

ALL-CAUSE MORTALITY: perfect agreement

HEALTH-RELATED QUALITY OF LIFE: perfect agreement

SEVERE HYPOGLYCAEMIA: perfect agreement

CARDIOVASCULAR MORTALITY: perfect agreement

NON-FATAL MYOCARDIAL INFARCTION/STROKE: perfect agreement

END-STAGE RENAL DISEASE/BLINDNESS: perfect agreement

SERIOUS ADVERSE EVENTS: perfect agreement

DIABETIC KETOACIDOSIS: perfect agreement

NON-SERIOUS ADVERSE EVENTS: perfect agreement

SEVERE NOCTURNAL HYPOGLYCAEMIA: perfect agreement

MILD/MODERATE HYPOGLYCAEMIA: perfect agreement

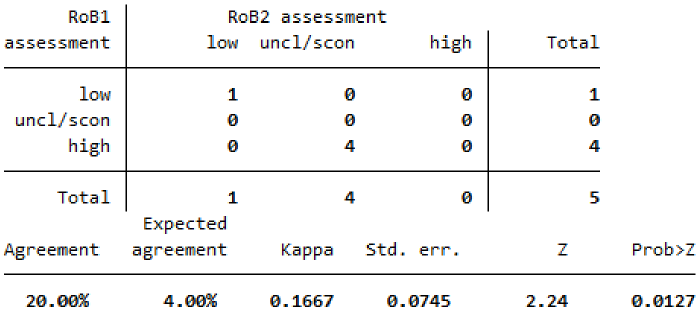
HBA1C: perfect agreement

HBA1C < 7% WITHOUT SEVERE HYPOGLYCAEMIA: perfect agreement

Blinding outcome assessment vs measurement of the outcome

ALL-CAUSE MORTALITY: perfect agreement

HEALTH-RELATED QUALITY OF LIFE



```
        RoB1 |          RoB2 assessment
  assessment |       low  uncl/scon       high |     Total
-------------+-------------------------------+----------
         low |         1          0          0 |         1
   uncl/scon |         0          0          0 |         0
        high |         0          4          0 |         4
-------------+-------------------------------+----------
       Total |         1          4          0 |         5
             Expected
Agreement   agreement     Kappa   Std. err.         Z      Prob>Z
-----------------------------------------------------------------
  20.00%       4.00%     0.1667     0.0745       2.24      0.0127
```

95% CI for kappa: -0.08 to 0.41
Gwet's AC1 with 95% CI: -0.18 (-0.57 to 0.22)

SEVERE HYPOGLYCAEMIA: perfect agreement

CARDIOVASCULAR MORTALITY: perfect agreement

NON-FATAL MYOCARDIAL INFARCTION/STROKE: perfect agreement

END-STAGE RENAL DISEASE/BLINDNESS: perfect agreement

SERIOUS ADVERSE EVENts: perfect agreement

DIABETIC KETOACIDOSIS: perfect agreement

NON-SERIOUS ADVERSE EVENTS: disagreement (RoB1 all high vs RoB 2 all unclear)

SEVERE NOCTURNAL HYPOGLYCAEMIA: perfect agreement

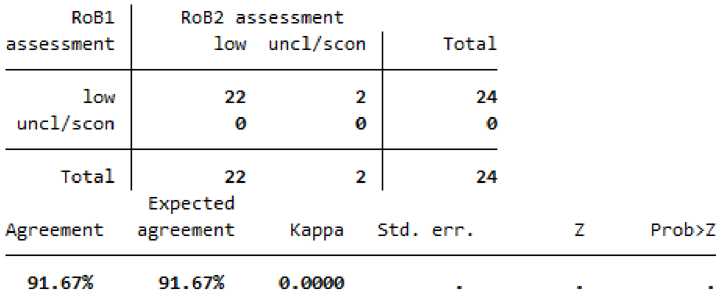
MILD/MODERATE HYPOGLYCAEMIA: disagreement (RoB1 all high vs RoB 2 all unclear)

HBA1C: perfect agreement

HBA1C < 7% WITHOUT SEVERE HYPOGLYCAEMIA: perfect agreement

Selective reporting vs selection of the reported result (exercise only)

Example for NON-SERIOUS ADVERSE EVENTS:

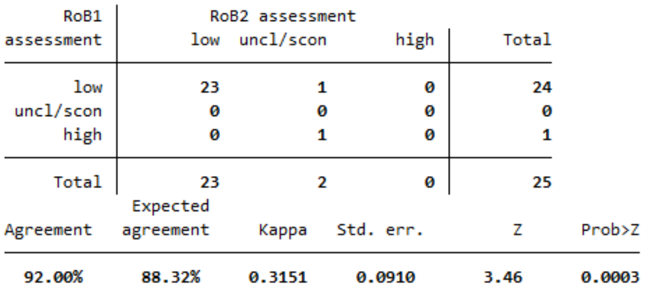


```
      RoB1  |   RoB2 assessment
assessment  |     low  uncl/scon |     Total
------------+--------------------+----------
       low  |      22          2 |        24
 uncl/scon  |       0          0 |         0
------------+--------------------+----------
     Total  |      22          2 |        24

             Expected
Agreement    agreement     Kappa   Std. err.         Z      Prob>Z
-----------------------------------------------------------------
  91.67%      91.67%      0.0000          .          .          .
```

Gwet's AC1 with 95% CI: 0.91 (0.78 to 1.04)

Example for HBA1C:



```
      RoB1  |          RoB2 assessment
assessment  |     low  uncl/scon       high |     Total
------------+-------------------------------+----------
       low  |      23          1          0 |        24
 uncl/scon  |       0          0          0 |         0
      high  |       0          1          0 |         1
------------+-------------------------------+----------
     Total  |      23          2          0 |        25

             Expected
Agreement    agreement     Kappa   Std. err.         Z      Prob>Z
-----------------------------------------------------------------
  92.00%      88.32%      0.3151     0.0910       3.46      0.0003
```

95% CI for kappa: 0.01 to 0.62
Gwet's AC1 with 95% CI: 0.92 (0.80 to 1.03)

Overall bias (RoB2)

( ) = number of studies

All-cause mortality (n = 24): 92% low; 8% some concerns
Health-related quality of life (n = 5): 20% low; 80% some concerns
Severe hypoglycaemia (n = 24): 87.5% low; 12.5% some concerns
Cardiovascular mortality (n = 24): 92% low; 8% some concerns
Non-fatal myocardial infarction/stoke (n = 7): 86% low; 14% some concerns
End-stage renal disease/blindness (n = 1): 100% low
Serious adverse events (n = 24): 92% low; 8% some concerns
Diabetic ketoacidosis (n = 19): 95% low; 5% some concerns
Non-serious adverse events (n = 24): 100% some concerns
Sever nocturnal hypoglycaemia (n = 20): 95% low; 5% some concerns
Mild/moderate hypoglycaemia (n = 24): 100% some concerns
HbA1c (n = 25): 88% low; 12% some concerns
HbA1c <7% without severe hypoglycaemia (n = 4): 100% low