Clinical Epidemiology

Dovepress
open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

Quality assessment of observational studies in a drug-safety systematic review, comparison of two tools: the Newcastle–Ottawa Scale and the RTI item bank

This article was published in the following Dove Press journal:
Clinical Epidemiology
10 October 2014
Number of times this article has been viewed

Andrea V Margulis[1]
Manel Pladevall[1]
Nuria Riera-Guardia[1]
Cristina Varas-Lorenzo[1]
Lorna Hazell[2,3]
Nancy D Berkman[4]
Meera Viswanathan[4]
Susana Perez-Gutthann[1]

[1]RTI Health Solutions, Barcelona, Spain; [2]Drug Safety Research Unit, Southampton, UK; [3]Associate Department of the School of Pharmacy and Biomedical Sciences, University of Portsmouth, Portsmouth, UK; [4]RTI International, Research Triangle Park, NC, USA

Background: The study objective was to compare the Newcastle–Ottawa Scale (NOS) and the RTI item bank (RTI-IB) and estimate interrater agreement using the RTI-IB within a systematic review on the cardiovascular safety of glucose-lowering drugs.

Methods: We tailored both tools and added four questions to the RTI-IB. Two reviewers assessed the quality of the 44 included studies with both tools, (independently for the RTI-IB) and agreed on which responses conveyed low, unclear, or high risk of bias. For each question in the RTI-IB (n=31), the observed interrater agreement was calculated as the percentage of studies given the same bias assessment by both reviewers; chance-adjusted interrater agreement was estimated with the first-order agreement coefficient (AC1) statistic.

Results: The NOS required less tailoring and was easier to use than the RTI-IB, but the RTI-IB produced a more thorough assessment. The RTI-IB includes most of the domains measured in the NOS. Median observed interrater agreement for the RTI-IB was 75% (25th percentile [p25] =61%; p75 =89%); median AC1 statistic was 0.64 (p25 =0.51; p75 =0.86).

Conclusion: The RTI-IB facilitates a more complete quality assessment than the NOS but is more burdensome. The observed agreement and AC1 statistic in this study were higher than those reported by the RTI-IB's developers.

Keywords: systematic review, meta-analysis, quality assessment, AC1

Introduction

The quality assessment of studies included in systematic reviews is fundamental for the interpretation of those reviews and more so when analyses based on study quality are conducted.[1–3] Quality assessment of observational studies is challenging due to their methodological intricacies, frequent use of data originally collected for purposes other than research, and subjective nature of the quality evaluation.[4] The difficulty of the quality assessment and the lack of agreement on how best to perform it and which domains should be included is reflected by the large number of available tools, of which 61,[4] 86,[5] and 194[6] have been recently reviewed. No tool covered all the domains considered of importance in drug safety studies.[4]

The Cochrane Collaboration endorsed the use of the Newcastle–Ottawa Scale (NOS)[7] to assess the quality of observational studies in its 2011 handbook,[1] which is the reason why the NOS was selected for use in the systematic review component of the SOS Project, a large study on the safety of nonsteroidal anti-inflammatory agents sponsored by the European Commission.[8–11] While working on the SOS Project,[8]

Correspondence: Andrea V Margulis
RTI Health Solutions, Travessera de Gracia 56, Atico 1, Barcelona 08006, Spain
Tel +34 933 622 806 or +34 693 822 166
Fax +34 93 414 2610
Email amargulis@rti.org

researchers in our group felt that, although widely used, the NOS enabled only a limited exploration of the quality of included studies. Therefore, when the group later conducted a systematic review for the European Commission–sponsored Safety Evaluation of Adverse Reactions in Diabetes (SAFEGUARD),[12] an additional tool, the RTI item bank,[13,14] was used for a second assessment of the quality of the included studies to better understand the risk of bias associated with each study.

This report describes our experience using the two tools to assess the quality of observational studies. Because the individual items in the RTI item bank have not been validated outside its development process, we also evaluated and reported on the interrater agreement of this tool.

Materials and methods

Parent project: SAFEGUARD

This work is a part of SAFEGUARD,[12] a large multinational research project conducted by 14 research partners and funded by the European Commission to study the safety of blood glucose-lowering medications in type 2 diabetes mellitus patients. SAFEGUARD includes mechanistic, pharmacovigilance, database-based observational studies and systematic reviews on the pancreatic and cardiovascular safety of these drugs. We present here the methodological evaluation of the quality assessment of the observational studies included in the systematic review on cardiovascular safety.

Briefly, we searched PubMed, Embase, and the Cochrane Library for case-control and cohort studies reporting on blood glucose-lowering drugs and cardiovascular or cerebrovascular outcomes (ie, acute myocardial infarction, acute coronary syndrome, stroke, heart failure, and cardiovascular mortality). Ultimately, 44 studies were selected and assessed for quality: 35 cohort and nine case-control studies. Further details on the study selection process will be published along with the main results of the systematic review. Preliminary information has been published as conference abstracts.[12,15,16] The quality assessment was performed simultaneously with the data extraction and is described in detail in the sections below.

Quality assessment tools

Newcastle–Ottawa Scale

The NOS was developed jointly by the University of Newcastle (Australia) and the University of Ottawa (Canada) to assess the quality of nonrandomized studies to be included in systematic reviews.[7] It has been widely used since at least 2004,[17] and results from several validation studies have been published.[2,18–20]

The NOS has a version for case-control studies and one for cohort studies. Each version includes a set of questions, or scale, and a short manual, along with an explanatory slide presentation; these resources are available for free download from the website of the Ottawa Hospital Research Institute.[7] Both versions of the scale consist of eight multiple-choice questions that address subject selection and comparability (of cases and controls in case-control studies, of cohorts in cohort studies) and the assessment of the outcome (in case-control studies) or exposure (in cohort studies). A few questions require adaptation to the systematic review to which the NOS is being applied (eg, "Select the most important factor" for comparability of cases and controls requires the investigator to select the most important factor for that particular systematic review). The number of possible answers per question ranges from two to five. High-quality responses earn a star, totaling up to nine stars (the comparability question earns up to two stars). The results of application of the NOS have been conveyed with varying level of detail: from the answer to each question for each study[21,22] (maximum detail) to a summary score equal to the number of stars earned by each study (minimum detail).[23,24] Our group presented a partial score summarizing the number of stars earned by each study in each domain.[9,10]

Although the Cochrane Collaboration[1] endorses the NOS, it acknowledges that researchers may want to assess study quality based not only on the quality of the analysis, covered by the NOS, but also on the quality of the reporting of the study, which is not included in this tool.

RTI item bank

The RTI item bank was developed by RTI-University of North Carolina Evidence-based Practice Center under sponsorship of the US Agency for Health Care Research and Quality; the project's objective was to create a set of questions or items to evaluate the conduct of observational studies included in systematic reviews, with a focus on bias and precision.[13,14] The item bank consists of 29 multiple-choice questions or items that can be applied to multiple study designs and covers eleven domains: sample definition and selection, interventions/exposure, outcomes, creation of treatment groups, blinding, soundness of information, follow-up, analysis comparability, analysis outcome, interpretation, and presentation and reporting. Most sets of possible responses are combinations of "Yes," "No," "Partially," "Cannot determine," and "Not applicable." The RTI item bank usually requires extensive tailoring, which may involve selecting the items appropriate to each systematic review. Items and instructions

are available for free download.[13] The developers do not offer suggestions as to how to convey the results, but the structure of the evaluation is amenable to graphic layouts as recommended in the Cochrane Handbook for Systematic Reviews of Interventions,[1] where green or a "+" sign would reflect an answer with low risk of bias, red or "–" would reflect high risk of bias, and yellow or "?" would reflect an unclear risk of bias.[25] After our quality assessment had been performed, a revised and shorter version of the tool, consisting of 13 items considered essential by a working group of six reviewers, was published.[26] To our knowledge, the RTI item bank has been used in two published systematic reviews since its publication in 2011.[27,28] The first review presents the answer to each item in each study (maximum detail) and a summary score for each study – number of items with a low-risk-of-bias response divided by the number of items applicable in each study (minimum detail).[27] The second review presents the bias appraisal associated with each item in each study (ie, low, unclear, or high risk of bias) and the overall bias appraisal for each study.[28]

Study quality assessment

For our systematic review on the cardiovascular safety of glucose-lowering drugs, we modified the NOS so that case-control studies earned a star when the case definition was based on record linkage to liken the evaluation of case-control studies to that of cohort studies. Our adapted version is provided in the online supplementary material.

The study team discussed the adaptations needed to apply the RTI item bank (eg, converting items with two components into two separate items, dropping four items that were not applicable in this setting, and adding a "Not applicable" response to some items) and created a document with details and decision-making rules that complemented the original RTI item bank instructions. These detailed instructions had to be revised as the full-text review of publications progressed and the need for more detailed guidance was noted; for two items, instructions changed substantially between the first and the second reviewers' appraisals. Our version of the items and instructions are included in the online supplementary material.

We identified four important pharmacoepidemiologic issues that were not covered explicitly by either tool and developed very specific questions to address them following the RTI item bank structure. These items focused on immortal time bias,[29] formulary restrictions (ie, restricted access to medications based on the health care system or plan drug formulary; patient access to restricted drugs is sometimes authorized after failure of the first-line treatment), confounding by indication,[30] and unmeasured confounding. The domains and questions or items in both tools are listed side by side in the online supplementary material for the purpose of comparison.

One investigator (AVM or NR-G) performed the quality assessment with the NOS and the RTI item bank simultaneously with the data extraction. A second investigator (AVM, MP, or NR-G) reviewed the quality assessment with the NOS and performed a second assessment using the RTI item bank. Thus, the RTI item bank was applied independently by the two reviewers. Disagreements were resolved by consensus. We then recategorized star-earning responses in the NOS as "low risk of bias" and all other responses as "high risk of bias." In the RTI item bank, responses that reflected high quality were interpreted as conveying a low risk of bias; "cannot determine" and "partially" were reclassified as "unclear risk of bias," and low-quality responses were interpreted as conveying a high risk of bias (details are provided in the online supplementary material).

Correlation between study quality assessments with both tools

To compare quality assessments with both tools, for each study, we calculated the percentage of responses in the NOS indicating high risk of bias (A), the percentage of responses in the RTI item bank indicating high risk of bias (B), and the percentage responses in the RTI item bank indicating high or unclear risk of bias (C). We then calculated the Spearman's rank correlation coefficient for the comparisons of A versus B and A versus C.

Observed interrater agreement and reliability of the RTI item bank

We assessed the observed agreement and the interrater reliability in terms of risk of bias. For example, for item 2 – "Are critical inclusion/exclusion criteria clearly stated (does not require the reader to infer)?" – for which the three possible responses denote low, unclear, or high risk of bias, agreement was based on a three-by-three table. For each item in the tool, the observed agreement was calculated as the percentage of studies to which both reviewers gave the same risk-of-bias assessment. The observed agreement includes the agreement due to chance (ie, one or more reviewers gave a random response to a question and, as a consequence, reviewers' responses agree). From among the chance-corrected statistics, we chose the first-order agreement coefficient (AC1) statistic to make our results comparable with the results reported by the RTI item bank developers. Further, the AC1

is not affected by a paradox that acts on the more widely used kappa statistic: the counterintuitive low value of the statistic when the observed agreement is high but the prevalence of the condition as assessed by the reviewers is either low or high.[13,31–33] Higher values of the AC1 reflect better agreement.

In this report, we present a brief summary of the results of the application of the NOS and the RTI item bank to the 44 studies included in our systematic review; the correlation between the quality assessments with both tools; the observed agreement and AC1 statistic for each question in the RTI item bank, including the seven questions that were considered to require a very subjective assessment by the raters (eg, "Are results believable taking study limitations into consideration?"); and the two questions whose instructions were modified while the quality assessment was underway (these items are noted in Table 1). We also discuss our experience with both tools for this drug-safety systematic review. Study-specific quality-assessment results will be presented along with the systematic review results in future publications.

Analyses were performed and figures were drawn with R (The R Project for Statistical Computing; http://www.r-project.org/).[34] For quality-control purposes, results were replicated with SAS macro AC1 (SAS Institute Inc., Cary, NC, USA).[35]

Results

NOS scores for the observational studies included in our systematic review ranged from 5 to 9 (the range of possible scores goes from 0 through 9), with a median and mode of 8 (25th percentile [p25] =7; p75 =8). A summary of the risk of bias as assessed using the NOS for case-control and cohort studies is shown in Figures 1 and 2. All studies earned a star for comparability with regards to age and sex, which we considered the most important factors for adjustment because

Table 1 Interrater agreement and AC1 statistic by item, all studies (n=44)

Item	Item description	Observed agreement, % (95% confidence interval)	AC1 statistic (95% confidence interval)
1a	Prospective/retrospective design: potential for recall bias	100% (91.97%–100%)	NA
1b	Prospective/retrospective design: tailored data collection	100% (91.97%–100%)	NA
2	Critical inclusion/exclusion criteria: clearly stated?	79.55% (65.5%–88.85%)	0.76 (0.61–0.91)
3	Critical inclusion/exclusion criteria: valid and reliable measures?	40.91% (27.69%–55.59%)	0.3 (0.11–0.49)
4	Critical inclusion/exclusion criteria: applied uniformly?	88.64% (76.02%–95.05%)	0.88 (0.78–0.98)
5	Strategy for recruitment: same across study groups	90.91% (78.84%–96.41%)	0.9 (0.81–1)
6[a]	Precision	77.27% (63.01%–87.16%)	0.66 (0.45–0.88)
7	Level of detail in describing the exposure	65.91% (51.14%–78.12%)	0.52 (0.32–0.73)
8	Important outcomes prespecified?	100% (91.97%–100%)	1 (1–1)
9[a]	Selection of the comparison group adequate?	93.18% (81.77%–97.65%)	0.93 (0.85–1)
10	Allocation between the groups: balance	70.45% (55.78%–81.84%)	0.59 (0.35–0.83)
11[b]	Isolation from unintended exposures	54.55% (40.07%–68.29%)	0.42 (0.2–0.63)
12	Outcome validation independent of exposure status	88.64% (76.02%–95.05%)	0.88 (0.78–0.98)
13	Exposures: valid and reliable measures, consistently implemented?	77.27% (63.01%–87.16%)	0.74 (0.58–0.9)
14a	Outcomes: valid and reliable measures?	63.64% (48.87%–76.22%)	0.51 (0.31–0.71)
14b	Outcomes: measures consistently implemented?	88.64% (76.02%–95.05%)	0.88 (0.78–0.98)
15[b]	Length of follow-up: same for all groups?	43.18% (29.68%–57.78%)	0.17 (0–0.39)
16	Length of follow-up: long enough?	75% (60.56%–85.43%)	0.73 (0.58–0.88)
17	Attrition: different across exposure groups	61.36% (46.62%–74.28%)	0.55 (0.36–0.73)
18	Control for baseline differences	97.73% (88.19%–99.6%)	0.98 (0.93–1)
19	Confounding: valid and reliable measures, consistently implemented?	86.36% (73.29%–93.6%)	0.86 (0.75–0.97)
20	Confounding, effect modification: important variables were considered?	47.73% (33.75%–62.06%)	0.28 (0.07–0.49)
21	Loss to follow-up: assessment of impact?	31.82% (20%–46.56%)	0.12 (0–0.3)
22	Intermediate variables not controlled for?	61.36% (46.62%–74.28%)	0.54 (0.36–0.73)
23[a]	Statistical methods appropriate?	56.82% (42.22%–70.32%)	0.48 (0.29–0.66)
24[a]	Results: believable?	50% (35.83%–64.17%)	0.31 (0.08–0.54)
25	Source of funding identified?	88.64% (76.02%–95.05%)	0.83 (0.69–0.98)
26[a]	Potential for immortal time bias	63.64% (48.87%–76.22%)	0.51 (0.3–0.71)
27	Formulary restrictions present?	84.09% (70.63%–92.07%)	0.81 (0.67–0.94)
28[a]	Confounding by indication present?	61.36% (46.62%–74.28%)	0.51 (0.31–0.7)
29[a]	Unmeasured confounding present?	77.27% (63.01%–87.16%)	0.64 (0.41–0.87)

Notes: [a]Items that call for a very subjective appraisal; [b]items whose instructions were substantially revised as the review of full-text publications progressed.
Abbreviations: AC1, first-order agreement coefficient; NA, not applicable.

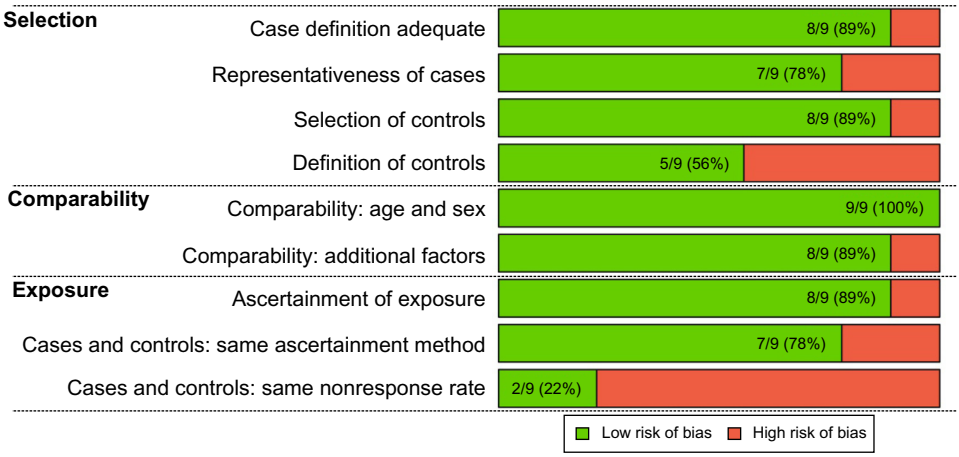

Figure 1 Risk of bias by domain (in bold) and question in nine case-control studies using the Newcastle–Ottawa Scale.
Note: Numbers on the green bar represent the number of studies with low risk of bias over the number of studies assessed.

our study eligibility criteria required, as a minimum, adjustment for age and sex. All except one included study earned a second star for additional adjustment. Among case-control studies, evaluation of the nonresponse rate was the question with the lowest count of stars, with only 22% of the studies having a low risk of bias. Among cohort studies, the lowest count of stars was for the question evaluating the presence of the outcome at the start of follow-up, with 46% of studies showing low risk of bias.

Results of the application of the RTI item bank cannot be summarized as easily as results from the NOS because there is no recommendation to aggregate RTI item bank results into a summary score. In the extremes, questions on study design and on whether the study outcome had been prespecified showed low risk of bias for 100% of the 44 studies, whereas the question on whether outcome validation had been independent of exposure status showed high risk of bias in 100% of the two studies to which the item was applicable (Figure 3).

In testing the correlation between study quality assessments with both tools, the Spearman's rank correlation coefficient was 0.35 for high risk of bias with the NOS and high or unclear risk of bias with the RTI item bank, and 0.38 for high risk of bias with the NOS and high risk of bias with the RTI item bank (Figure 4).

The observed agreement and the AC1 statistics for each item in the RTI item bank are shown in Table 1. Excluding items 1a and 1b (which would trivially increase agreement because they had a single response category, low risk of bias), the observed agreement between raters had a median of 75% (p25 =61%; p75 =89%) and the AC1 statistic had a median of 0.64 (p25 =0.51; p75 =0.86). In the seven most subjective

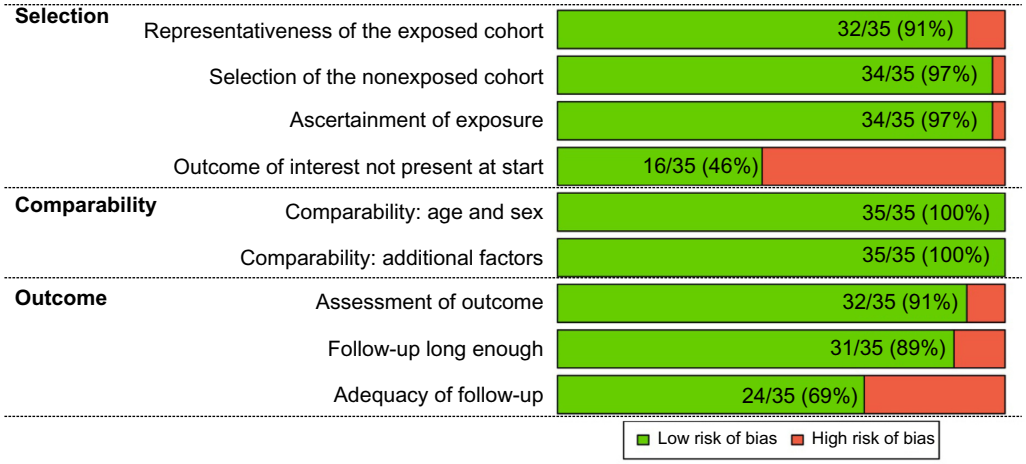

Figure 2 Risk of bias by domain (in bold) and question in 35 cohort studies using the Newcastle–Ottawa Scale.
Note: Numbers on the green bar represent the number of studies with low risk of bias over the number of studies assessed.

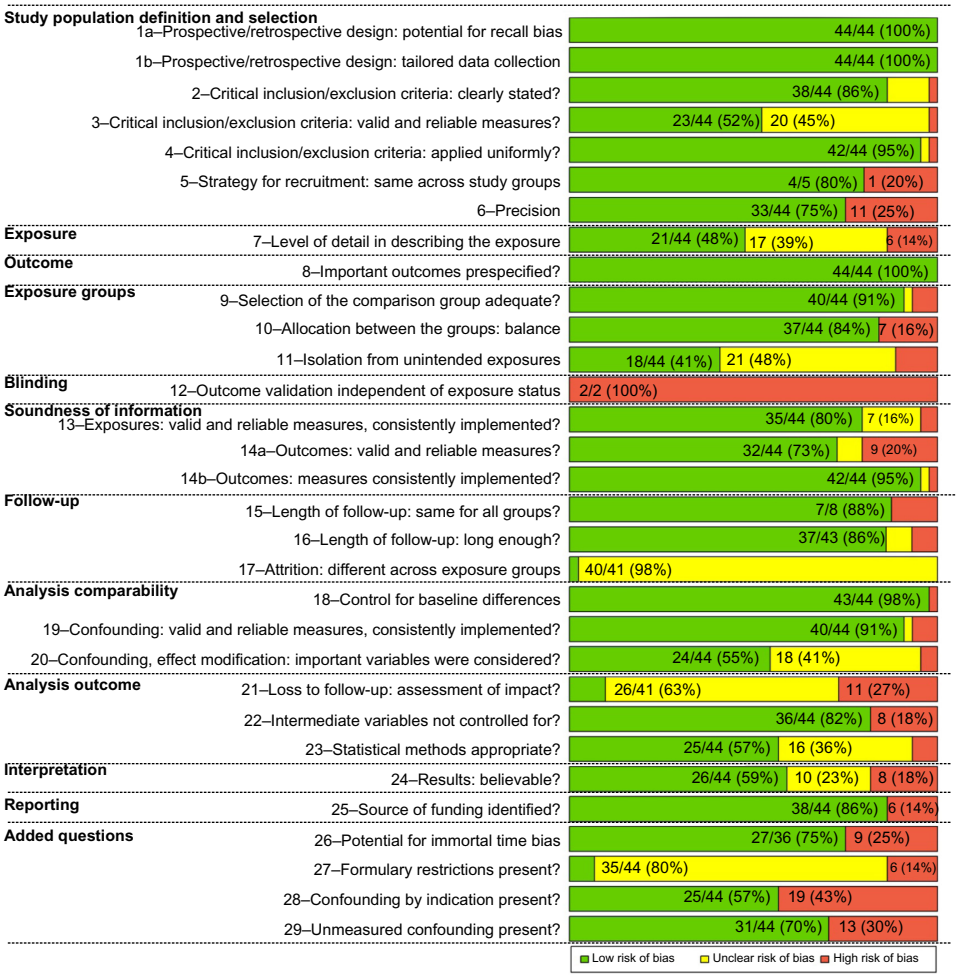

Figure 3 Risk of bias by domain (in bold) and item in all studies (n=44) using the RTI item bank.
Notes: Numbers on the bars represent the number of studies with low risk of bias (green bars), unclear risk of bias (yellow bars), or high risk of bias (red bars) over the number of studies for which the item was applicable, and percentages. Due to limited room, some of these numbers were not included. Percentages were calculated over the studies for which the items were applicable. In the first domain, we replaced "Sample" with "Study population," and removed "Interventions" from the domain "Interventions/exposure."

items, the observed agreement between raters had a median of 64% (p25 =59%; p75 =77%) and the AC1 statistic had a median of 0.51 (p25 =0.49; p75 =0.65). Regarding the two questions whose instructions we modified during the course of the quality assessment, the values for observed agreement between raters were 43% and 55% and the values for the AC1 statistic were 0.17 and 0.42.

Discussion

We found the NOS easier to apply than the RTI item bank, but more limited in scope. Most aspects covered by the NOS are also covered by the RTI item bank. The RTI item bank helped us undertake a more thorough assessment of study quality than the NOS. The observed agreement between raters and the inter-rater reliability AC1 statistic were fair and generally higher in questions that were less subjective and whose instructions remained unchanged during the review process.

The NOS, widely used and endorsed by the Cochrane Collaboration, has been criticized for its somewhat arbitrary selection of the best-quality answer, which will be assigned a star.[17] Also, it has been argued that quality summary scores may mask variations in quality by domain and use an unclear and often implicit weighting scheme.[3,5,13,36] The interrater reliability or agreement of the NOS has been previously evaluated. In one study, five pairs of neuroscience undergraduate and graduate students were asked to assess the quality of studies on electroconvulsive therapy and cognitive impairment. The interrater agreement was evaluated with the kappa statistic (κ). The kappa statistic had low values for most questions in both the case-control and the cohort versions: reliability was categorized as poor ($\kappa<0.21$) for eleven of the 16 questions (five questions in the version for case-control studies and six questions in the version for cohort studies) and fair ($0.21\leq\kappa\leq0.40$) for an additional

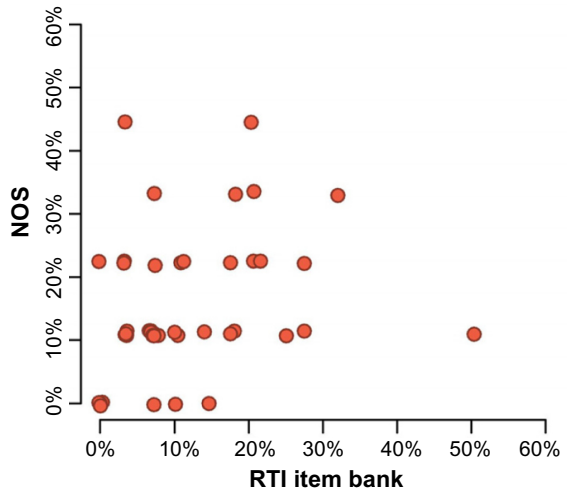

Figure 4 Correlation between high risk of bias with the Newcastle–Ottawa Scale and with the RTI item bank.
Notes: Each dot represents an observational study included in the systematic review. The x value for each dot represents the percentage of items in the RTI item bank with high risk of bias. The y value represents the percentage of questions in the NOS with high risk of bias. Horizontal and vertical jittering (the addition of a random number from a uniform distribution between –0.4% and 0.4%) was applied to avoid the masking of overlapping points in the plot.
Abbreviation: NOS, Newcastle–Ottawa Scale.

three questions.[18] The most salient characteristic of this validation study is that the reviewers, who were students, likely had limited experience in the subject matter area. Another study involved 16 reviewers, with widely varying levels of expertise in the subject matter area and in systematic reviews, who applied the NOS to 131 cohort studies (two reviewers per paper); interrater agreement for the overall score, assessed with a weighted kappa statistic, was κ=0.29. In that study, the researchers developed instructions to complement the NOS manual after a small pilot application on three studies.[2,19] In another study, the quality of the 46 cohort studies published in the *Journal of Pediatric Surgery* in relation to the 1998–2007 annual meetings of the Canadian Association of Pediatric Surgeons was assessed by pediatric surgeons and a clinical epidemiology researcher using the NOS. Interrater agreement assessed with the intraclass correlation coefficient on the overall score was 94%.[20]

To our knowledge, interrater agreement with the RTI item bank was explored in a single systematic review that used the tool to evaluate the quality of the included observational studies.[27] The authors reported an agreement of 93.5% and a kappa statistic of 0.88 for all items combined. We estimated the observed and the interrater agreement for each item, because we felt it would be more informative, and to be able to contrast our results with the ones obtained during the tool development. In comparison with the RTI item bank developers, where 12 scientists with expertise in different fields evaluated ten papers on various topics and with different designs with limited instructions,[13,14] our agreement results measured with the AC1 statistic were higher in 18 of the 25 items that were not substantially modified from the original version. Reasons for this may be that all the studies we evaluated pertained to the same area of knowledge, on which the reviewers had expertise, and had designs with which the reviewers were very familiar, as opposed to the situation in the developers' setting. Furthermore, we wrote item-specific instructions, which we found very helpful and likely increased the agreement for most items, except for the two items whose instructions were modified during the course of the quality assessment (in some cases between the first and second rating for individual studies). When applying the RTI item bank, the reviewers often resorted to the "partially" or "cannot determine" options (later interpreted it as "unclear risk of bias"), which supports that not having those options is a potential limitation in the NOS.[2]

Another example of interrater agreement on risk of bias comes from the evaluation of the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) system for rating the quality of a body of evidence and the strength of recommendations.[37] In a recent validation effort,[38] researchers used this system to rate the overall quality of the body of evidence presented in four published meta-analyses as high, moderate, low, or very low quality. The intraclass correlation coefficient for pairs of assessments was 0.72 for raters who were highly trained in the GRADE system and 0.66 for less experienced raters.

Because in pharmacoepidemiology, both prospective and retrospective designs are considered appropriate, we did not associate any of the categories (ie, prospective, retrospective, or mixed) with an increased risk of bias; therefore, these items were not useful for our quality assessment. Items that were not useful in our assessment due to lack of information in the included studies were those that evaluated attrition/missing data and formulary restrictions (the latter was an item we added to the tool). Further, items applicable only to field studies or studies in which internal outcome validation was conducted were not useful due to the very small number of such studies. For the shortened and updated version of the RTI item bank that was published after our quality assessment process was completed, six epidemiologists and trialists selected 13 essential items from the original 29 and revised some of them.[26] This reduction in the number of items was based on the types of bias to be evaluated in the quality assessment of observational studies with varied designs and was performed outside the setting of a systematic review.

In contrast, our adaptation was specific to our systematic review and was aimed at a thorough quality assessment. As a consequence, the 13-item RTI item bank and our adapted RTI item bank are dissimilar.

When comparing the two tools, we found the NOS easier to apply. It requires little tailoring, whereas the RTI item bank requires extensive and, in our experience, iterative adaptations. The more time-consuming evaluation called for by the RTI item bank goes in parallel with a more detailed evaluation that involves more domains and possibly more detail in the domains evaluated by both tools. Most aspects included in the NOS were also included in the RTI item bank, although not in an identical manner (Table S1). The exception is the NOS question on the presence of an outcome at the start of the study, which is not explicitly included in the RTI item bank and was relevant in our assessment. In the application of the NOS, disagreement between the reviewers, although not explicitly quantified in this study, seemed to be rare. The overlap between the two tools and the wider scope of the RTI item bank explain why the correlation between quality assessments with both tools was positive, but moderate at best. We felt that our objective for assessing the quality of the observational studies, which was to understand each study's risk of bias, was better met by the RTI item bank.

Aspects that are vital in pharmacoepidemiologic research, such as confounding by indication, are not explicitly covered by either tool, although the RTI item bank can accommodate them with appropriate adaptations. In the present systematic review, the assessment of these aspects was very subjective, and interrater agreement was lower than it was for most other questions. Another aspect important for systematic reviews but not covered by either tool is the selective failure of researchers to report results that are not statistically significant: sometimes analyses are described and performed, but numerical results that are not statistically significant are often not reported and therefore cannot be incorporated into meta-analyses. We identified this aspect too late to incorporate it in the quality assessment; although it does not represent a threat to internal validity, we recommend considering the inclusion of this aspect in future uses of the tool.[37] Although it has been noted that industry-sponsored studies are sometimes biased toward the product manufactured by the sponsor,[39–43] there are voices favoring[44] but also opposing[45] the addition of such a question into the Cochrane risk of bias tool. One of the major concerns reported is the selection of an inappropriate comparator,[39,45] which is evaluated in the RTI item bank (in the version first published, in our adaptation, and in the updated 13-item version). Further, a question on sponsorship was included in the early stages of the development of the tool, but it was later removed as the panel of experts involved in the selection of items felt it was not essential or useful for evaluating the risk of bias.[13]

An important yet subtle issue that arose during the quality assessment is whether studies should be evaluated against the best possible study given the data source limitations or against the ideal study to answer the question at hand. The Cochrane tool for the quality assessment of nonrandomized studies, currently under development, will support the value of evaluating an observational study's risk of bias in comparison to that of a target trial.[46] A related question has been posed previously: Should each observational study be assessed based on the question under evaluation in the systematic review or in relation to the observational study's objectives?[2] Another issue we found, also previously reported,[2] is whether to assess the validity of the study as described in the publication included in the systematic review or to take into consideration other publications on the same study population to complete information gaps which are common.

Conclusion

In conclusion, the RTI item bank was more useful to evaluate the quality of the observational studies and to detect variation in study quality but was more burdensome than the NOS. Most aspects included in the NOS are covered by the RTI item bank, and the RTI item bank produced a more thorough quality evaluation. However, the NOS continues to be the most frequently used tool in practice, as conveyed by the large number of publications in which it has been employed. Contrary to the findings of the RTI item bank's developers, interrater agreement was not low, except perhaps for the questions whose answer was highly dependent on subjective appreciations. Some aspects of study quality that are important in pharmacoepidemiology are not covered by either tool and should be incorporated in future applications in this area.

Acknowledgments

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007–2013) under grant agreement number 282521 – the SAFEGUARD project.

We would like to thank Mark Howell, information services specialist, for his support and contributions in the development of the literature search strategy, and Adele C Monroe, medical editor, who provided valuable editing of the manuscript.

Disclosure

RTI Health Solutions employees work on projects funded by pharmaceutical companies including manufacturers of treatments for patients with diabetes. Manel Pladevall, Susana Perez-Gutthann, and Cristina Varas-Lorenzo as employees of RTI Health Solutions also participate in advisory boards funded by pharmaceutical companies.

While authors of the RTI item bank participated in later phases of this effort, they were not involved in the decision to use the scale nor in the evaluation itself.

A previous version of this work was presented as a poster in the 2013 annual conference of the International Society of Pharmacoepidemiology.[47]

References

1. Higgins JPT, Green S, editors. Section 13.5.2.3. Tools for assessing methodological quality or risk of bias in non-randomized studies. In: *Cochrane Handbook for Systematic Reviews of Interventions*, *Version 5.1.0*. London: The Cochrane Collaboration; 2011.
2. Hartling L, Hamm M, Milne A, et al. *Validity and Inter-Rater Reliability Testing of Quality Assessment Instruments (Report No*: *12-EHC039-EF)*. Rockville, MD: Agency for Healthcare Research and Quality (US); 2012. Available from: http://www.ncbi.nlm.nih.gov/books/NBK92293/. Accessed June 24, 2013.
3. Greenland S, O'Rourke K. On the bias produced by quality scores in meta-analysis, and a hierarchical view of proposed solutions. *Biostatistics*. 2001;2(4):463–471.
4. Neyarapally GA, Hammad TA, Pinheiro SP, Iyasu S. Review of quality assessment tools for the evaluation of pharmacoepidemiological safety studies. *BMJ Open*. 2012;2(5).
5. Sanderson S, Tatt ID, Higgins JP. Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: a systematic review and annotated bibliography. *Int J Epidemiol*. 2007;36(3):666–676.
6. Deeks JJ, Dinnes J, D'Amico R, et al; International Stroke Trial Collaborative Group; European Carotid Sugery Trial Collaborative Group. Evaluating non-randomised intervention studies. *Health Technol Assess*. 2003;7(27):iii–x, 1–173.
7. Wells GA, Shea B, O'Connell D, et al. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses [webpage on the Internet]. Ottawa, ON: Ottawa Hospital Research Institute; 2011. Available from: http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp. Accessed February 5, 2013.
8. SOS. Safety of NSAIDs [homepage on the Internet]. 2013. Available from: http://www.sos-nsaids-project.org/. Accessed June 24, 2013.
9. Castellsague J, Riera-Guardia N, Calingaert B, et al; Safety of Non-Steroidal Anti-Inflammatory Drugs (SOS) Project. Individual NSAIDs and upper gastrointestinal complications: a systematic review and meta-analysis of observational studies (the SOS project). *Drug Saf*. 2012;35(12):1127–1146.
10. Varas-Lorenzo C, Riera-Guardia N, Calingaert B, et al. Stroke risk and NSAIDs: a systematic review of observational studies. *Pharmacoepidemiol Drug Saf*. 2011;20(12):1225–1236.
11. Varas-Lorenzo C, Riera-Guardia N, Calingaert B, et al. Myocardial infarction and individual nonsteroidal anti-inflammatory drugs meta-analysis of observational studies. *Pharmacoepidemiol Drug Saf*. 2013;22(6):559–570.
12. SAFEGUARD. Safety evaluation of adverse reactions in diabetes [homepage on the Internet]. 2011. Available from: http://www.safeguard-diabetes.org/. Accessed June 24, 2013.
13. Viswanathan M, Berkman ND. Development of the RTI item bank on risk of bias and precision of observational studies [webpage on the Internet]. Rockville, MD: Agency for Healthcare Research and Quality; 2011. Available from: http://effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productid=784. Accessed February 5, 2013.
14. Viswanathan M, Berkman ND. Development of the RTI item bank on risk of bias and precision of observational studies. *J Clin Epidemiol*. 2012;65(2):163–178.
15. Varas-Lorenzo C, Riera-Guardia N, Calingaert B, et al. Cardiovascular risk with glitazones and metformin: results from a systematic review of observational studies. *Pharmacoepidemiol Drug Saf*. 2013;22(Suppl 1):S68.
16. Pladevall M, Varas-Lorenzo C, Margulis AV, et al. Rosiglitazone vs pioglitazone and acute myocardial infarction: systematic review and meta-analysis of observational studies. *Pharmacoepidemiol Drug Saf*. 2013;22(Suppl 1):S181–S182.
17. Stang A. Critical evaluation of the Newcastle-Ottawa Scale for the assessment of the quality of nonrandomized studies in meta-analyses. *Eur J Epidemiol*. 2010;25(9):603–605.
18. Oremus M, Oremus C, Hall GB, McKinnon MC; ECT and Cognition Systematic Review Team. Inter-rater and test-retest reliability of quality assessments by novice student raters using the Jadad and Newcastle-Ottawa Scales. *BMJ Open*. 2012;2(4).
19. Hartling L, Milne A, Hamm MP, et al. Testing the Newcastle Ottawa Scale showed low reliability between individual reviewers. *J Clin Epidemiol*. 2013;66(9):982–993.
20. Al-Harbi K, Farrokhyar F, Mulla S, Fitzgerald P. Classification and appraisal of the level of clinical evidence of publications from the Canadian Association of Pediatric Surgeons for the past 10 years. *J Pediatr Surg*. 2009;44(5):1013–1017.
21. Tricco AC, Lillie E, Soobiah C, Perrier L, Straus SE. Impact of H1N1 on socially disadvantaged populations: systematic review. *PLoS One*. 2012;7(6):e39437.
22. Sunkara SK, Khairy M, El-Toukhy T, Khalaf Y, Coomarasamy A. The effect of intramural fibroids without uterine cavity involvement on the outcome of IVF treatment: a systematic review and meta-analysis. *Hum Reprod*. 2010;25(2):418–429.
23. Huang Y, Li YL, Huang H, Wang L, Yuan WM, Li J. Effects of hyperuricemia on renal function of renal transplant recipients: a systematic review and meta-analysis of cohort studies. *PLoS One*. 2012;7(6):e39457.
24. Uçeyler N, Häuser W, Sommer C. Systematic review with meta-analysis: cytokines in fibromyalgia syndrome. *BMC Musculoskelet Disord*. 2011;12:245.
25. Higgins JP, Altman DG, Gøtzsche PC, et al; Cochrane Bias Methods Group; Cochrane Statistical Methods Group. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ*. 2011;343:d5928.
26. Viswanathan M, Berkman ND, Dryden DM, Hartling L. Assessing risk of bias and confounding in observational studies of interventions or exposures: further development of the RTI item bank [webpage on the Internet]. Rockville, MD: Agency for Healthcare Research and Quality (US); 2013. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24006553. Accessed September 6, 2013.
27. Al-Saleh MA, Armijo-Olivo S, Thie N, et al. Morphologic and functional changes in the temporomandibular joint and stomatognathic system after transmandibular surgery in oral and oropharyngeal cancers: systematic review. *J Otolaryngol Head Neck Surg*. 2012;41(5):345–360.
28. Ruhe A, Fejer R, Walker B. Does postural sway change in association with manual therapeutic interventions? A review of the literature. *Chiropr Man Therap*. 2013;21(1):9.
29. Suissa S. Immortal time bias in observational studies of drug effects. *Pharmacoepidemiol Drug Saf*. 2007;16(3):241–249.
30. Walker AM. Confounding by indication. *Epidemiology*. 1996;7(4):335–336.
31. Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. *Br J Math Stat Psychol*. 2008;61(Pt 1):29–48.

32. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. *Fam Med*. 2005;37(5):360–363.
33. Wongpakaran N, Wongpakaran T, Wedding D, Gwet KL. A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. *BMC Med Res Methodol*. 2013;13:61.
34. Brasil PE. Gwet's AC1 interrater reliability [webpage on the Internet]. R-sig-Epi; 2012. Available from: https://stat.ethz.ch/pipermail/r-sig-epi/2012-May/000273.html. Accessed June 12, 2014.
35. Blood E, Spratt KF. *Disagreement on Agreement: Two Alternative Agreement Coefficients*. SAS Global Forum 2007: Statistics and Data Analysis. Paper 186-2007. Available from: http://www2.sas.com/proceedings/forum2007/186-2007.pdf. Accessed October 15, 2012.
36. da Costa BR, Hilfiker R, Egger M. PEDro's bias: summary quality scores should not be used in meta-analysis. *J Clin Epidemiol*. 2013;66(1):75–77.
37. Guyatt GH, Oxman AD, Vist G, et al. GRADE guidelines: 4. Rating the quality of evidence – study limitations (risk of bias). *J Clin Epidemiol*. 2011;64(4):407–415.
38. Mustafa RA, Santesso N, Brozek J, et al. The GRADE approach is reproducible in assessing the quality of evidence of quantitative evidence syntheses. *J Clin Epidemiol*. 2013;66(7):736–742; quiz 742. e1.
39. Lexchin J, Bero LA, Djulbegovic B, Clark O. Pharmaceutical industry sponsorship and research outcome and quality: systematic review. *BMJ*. 2003;326(7400):1167–1170.
40. Sismondo S. Pharmaceutical company funding and its consequences: a qualitative systematic review. *Contemp Clin Trials*. 2008;29(2):109–113.
41. Golder S, Loke YK. Is there evidence for biased reporting of published adverse effects data in pharmaceutical industry-funded studies? *Br J Clin Pharmacol*. 2008;66(6):767–773.
42. Schott G, Pachl H, Limbach U, Gundert-Remy U, Ludwig WD, Lieb K. The financing of drug trials by pharmaceutical companies and its consequences. Part 1: a qualitative, systematic review of the literature on possible influences on the findings, protocols, and quality of drug trials. *Dtsch Arztebl Int*. 2010;107(16):279–285.
43. Lundh A, Sismondo S, Lexchin J, Busuioc OA, Bero L. Industry sponsorship and research outcome. *Cochrane Database Syst Rev*. 2012;12:MR000033.
44. Bero LA. Why the Cochrane risk of bias tool should include funding source as a standard item. *Cochrane Database Syst Rev*. 2013;12: ED000075.
45. Sterne JA. Why the Cochrane risk of bias tool should not include funding source as a standard item. *Cochrane Database Syst Rev*. 2013;12:ED000076.
46. Hernán MA. With great data comes great responsibility: publishing comparative effectiveness research in epidemiology. *Epidemiology*. 2011;22(3):290–291.
47. Margulis AV, Pladevall M, Riera-Guardia N, Varas-Lorenzo C, Hazell L, Perez-Gutthann S. *Quality assessment of observational studies in a drug- safety systematic review, comparison of two tools: the Newcastle-Ottawa Scale and the RTI Item Bank. Presented at the 29th Annual ICPE meeting, Montreal, Canada. 2013*. Pharmacoepidemiol Drug Saf. 2012;22(Suppl 1):S226.