Journal of Biometrics & Biostatistics
ISSN: 2155-6180

Nevis et al., J Biom Biostat 2015, 6:2
DOI: 10.4172/2155-6180.1000230

Research Article Open Access

Quality Control Tool for Screening Titles and Abstracts by second Reviewer: QCTSTAR

Nevis IF*, Sikich N, Ye C and Kabali C

Health Quality Ontario, Canada

Abstract

Background: Systematic reviews (SRs) remain the core of evidence based medicine. Routinely, two reviewers are required to screen titles and abstracts in SRs. Nonetheless, many organizations may use a single reviewer due to restricted time and resources. In situations where there is only a single reviewer, we propose a sampling method and assessed its performance in a simulation study.

Methods: We described the sampling process guided by a set of instructions. For validation, we generated 20,000 citations from a skewed normal distribution and assigned a score of raters' agreement. From these, we randomly selected a fixed number of citations, whose probability of selection was determined by a uniform distribution, and repeated the iteration 1000 times. In each iteration set, the sample size was fixed at 50, 100, and 200.

Results: We evaluated the sampling performance and proposed the appropriate sample size formula. Of the 20,000 citations, 86.7% fell into the category of "both reviewers have an agreement". The sampling performance was optimal. On average, the percent of agreement for samples of size 50, 100, and 200 were 86.7% (95% CI 76% to 96%), 86.7% (95% CI 79% to 93%), and 86.8% (95% CI 81.5% to 91.5%) respectively. When comparing the performance of sample size formula with simulations, we obtained identical results.

Conclusions: We propose a reliable and valid sampling methodology for screening titles and abstracts This method may be used in resource constrained environments conducting SRs.

Keywords: Duplicate review; Screening; Systematic reviews; Quality control; Second reviewer

Background

Systematic reviews (SRs) remain the core of evidence based medicine [1,2]. Many clinical and public policy decisions are made based on best available research synthesized by SRs [3]. SRs have formal approaches for appraising and collating evidence into a useable and reliable format [4]. They provide a comprehensive and valid synthesis of evidence for the research question posed. The reliability and validity are usually addressed by the numerous strategies that limit bias and random errors [2]. These strategies include a focused research question, a comprehensive search, and the use of explicit criteria for inclusion of studies in the review [2]. Other strategies to minimize errors in SRs, in particular selection bias, include using two reviewers at various stages of the process [5]. Screening of titles and abstracts of citations, full text eligibility, validity assessment, data abstraction and analysis are often done in duplicate to identify areas where two reviewers are interpreting questions and criteria differently [6].

Nonetheless, organizations may use only a single reviewer to synthesize evidence. This is primarily because of limited time and resources as well as to align with timely execution of policy recommendations. As aforementioned, this approach may potentially lead to bias if no quality check mechanism is in place. As a remedy, a technical report by the National Center for the Dissemination of Disability Research (NCDDR) has recommended 20–30% of all citations initially screened by a single reviewer, be screened by a second reviewer [7]. In addition, to the best of our knowledge, this cut-off point has not been proved to be optimal in terms of efficiency. Also, a 20–30% cut-off may be too high for systematic reviews with large number of citations to review, and a smaller sampling fraction with a satisfactory level of accuracy would be more desirable. To address this, we propose a sampling methodology which can be used by researchers working in a resource-constrained environment. We evaluate the reliability and validity of this approach using statistical simulations and suggest a valid sample size formula, commonly used in survey methods [2].

Methods

The sampling methodology

We propose a flexible sampling methodology to be used in the context where it is impractical to have two reviewers screening all citations at the title and abstract stage of a systematic review. The method is explicated as follows (Figure 1). Once all the relevant literature has been identified from a database, the primary reviewer selects studies that meet the inclusion criteria. The criteria may include (but are not limited to) a focused research question, valid study population and design, study size, and vulnerability to bias. Studies are then categorized into those meeting or not meeting the criteria.

As a next step, the entire original list of relevant literature is passed on to the second reviewer. The second reviewer computes the minimum required sample size using formula [1] below,

$$n = \frac{z_{\alpha/2}^{\ 2}\hat{p}(1-\hat{p})/\Delta^2}{\frac{N-1}{N} + \frac{z_{\alpha/2}^{\ 2}\hat{p}(1-\hat{p})}{N\Delta^2}} \quad (1)$$

***Corresponding author:** Nevis IF, Health Quality Ontario, Canada, Tel: 416-323-6868; Fax: 416-323-9261; E-mail: immaculatenevis@yahoo.ca

Received April 28, 2015; **Accepted** June 01, 2015; **Published** June 08, 2015

Citation: Nevis IF, Sikich N, Ye C, Kabali C (2015) Quality Control Tool for Screening Titles and Abstracts by second Reviewer: QCTSTAR. J Biom Biostat 6: 230. doi:10.4172/2155-6180.1000230

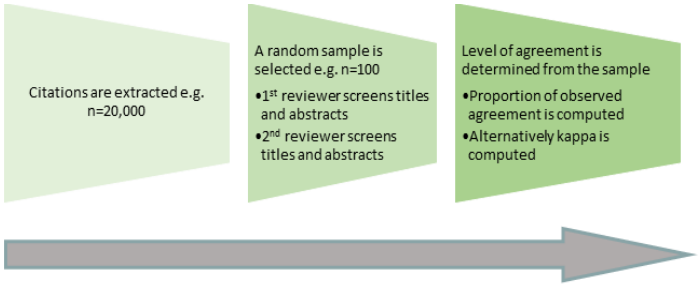

Figure 1: Proposed screening process.

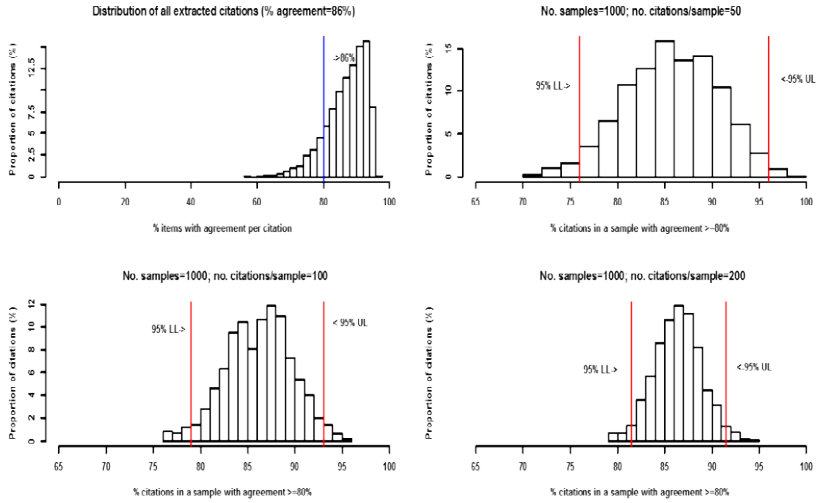

Figure 2: Distribution of simulated citations extracted from the database. The top left panel shows the distribution of all 20,000 citations. The blue line represents the threshold of consensus per citation. The other panels show the distribution of consensus across 1000 samples with varying sample sizes of 50, 100, and 200.

where the parameters Δ, α, $\hat{p}$, $z_{\alpha/2}$, n, and N represent the margin of error, type I error, proportion of citations with agreement, critical value in the standard normal distribution, sample size, and population size respectively. The derivation of this formula is shown in the appendix. More often $\hat{P}$ is unknown to reviewers, so to obtain the conservative for n, $\hat{p}$ can be set at 50%.

The second reviewer can then use any accessible software to select a random sample of citations from the entire original list. Ideally, the samples will be drawn from a uniform distribution using a simple random sampling design without replacement. In situations where it is possible to group studies based on characteristics that influence selection, a stratified random sampling approach can also be used. The second reviewer then reviews all the studies in the sample drawn, and categorizes them into those meeting or not meeting inclusion criteria. At this point, both reviewers can determine the percentage of studies in a sample in which they reached the same decision with regard to inclusion or exclusion, and compute the level of agreement. As we demonstrate in succeeding sections, the derived statistic from this sample should accurately reflect the actual level of agreement that would have been observed if all studies were reviewed by two reviewers.

Validation from a simulation study

We generated 20,000 simulations from a skewed normal distribution [8] with shape, scale, and location parameters of -20, 10, and 95 respectively. Given that moderate to high levels of agreement are expected for well-trained reviewers, the shape parameter was deliberately set to negative to skew the distribution to the left, reflecting that very low levels of agreement are unlikely. We also assumed that raters may decide to use a checklist with items indicating which sections in a citation they had an agreement. If the percentage of items with agreement in the checklist exceeded 80%, we classified the citation as meeting the criteria for agreement. The 80% threshold was chosen arbitrarily for illustration purpose but our sampling process applies to any inclusion criteria.

Simulations yielded 86.7% citations meeting the criteria of agreement (Figure 2a). From this distribution we randomly drew a sample of 50 citations and determined the percentage of citations meeting the threshold of agreement. We iterated this process 1000 times and plotted the histogram (Figure 2b). The value 86.7% (henceforth the population parameter) was well within the 95% confidence interval (CI), which can crudely be interpreted as: "had we to randomly select 50 citations from this pool and compute 95% CI for the level of agreement, the resulting interval will contain the population parameter not less than 95% of the time". The caveat of drawing a small sample is that the CIs will tend to be too wide reflecting large sampling errors. For example, in this scenario where only 50 samples were picked, the resulting 95% CI (76% to 96%) may seem too wide and hence less informative, warranting a larger sample. Therefore, we doubled the

sample to 100 and repeated the same iteration process (Figure 2c). The population parameter still remained within the 95% CI (79% to 93%) indicating a good coverage, but the width of the CI was reduced as most of the individual sample estimates were closer to the population parameter. A substantial gain in precision was observed when the sample size was further increased to 200 citations (Figure 2d), with the population parameter remaining within the 95% CI (82% to 92%).

Evaluation of statistical bias

We examined the presence and magnitude of bias from our sampling process by averaging the proportion of agreement across all samples and subtracted the resulting value from the population parameter. In essence, any departure of this difference from zero would indicate the extent of bias in our sampling strategy. We detected virtually no bias across all iteration sets (Table 1). However the precision as measured by the inverse of standard error was doubled as the sample quadrupled from 50 to 200.

Relationship between sampling fraction and standard error

To demonstrate the relationship between sampling fraction and standard error, we generated curves of standard errors for samples drawn from citations of size 5000, 10000, and 20000 (Figure 3) using the previously mentioned simulation procedure. For the purpose of this simulation, formula (1) has been simplified to formula (2). Please see the appendix for more details on the relationship between the two formulas. The curves show that whenever the sampling fraction (f) is held constant, the precision improves with an increase in the number of citations extracted. However this difference becomes less pronounced when the sampling fraction enlarges.

$$SE \approx \sqrt{\frac{\hat{p}(1-\hat{p})(1-f)}{n}}, \qquad (2)$$

Further, when we plotted the curve of standard error versus sampling fraction, fixing the total number of citations at 20,000 and superimpose it onto the one derived from simulations we almost had a perfect match for the two curves (Figure 4), validating the use of formula [1] for sample size calculations.

Extension to index measures of agreement

The simulation results described in previous sections hold true for index measures of agreement as Cohen's Kappa, Phi, and Kendall's Tau etc. To illustrate this we used Cohen's Kappa (hereafter kappa) as the parameter of interest for the distribution described in Figure 2a. To simulate Kappa values from random samples, additional steps had to be done.

First, we organized citations in a 2 by 2 table according to hypothesized reviewers' agreement.

This is summarized in Table 2, where n is the total number of citations (in our case 20000), n11, n12, n21, and n22 are the number of citations where both reviewers accept, 1st reviewer accept and 2nd reviewer reject, 1st reviewer reject and 2nd reviewer accept, and both reviewers reject respectively, n+1 and n+2 are column sums, and n1+ and n2+ are row sums. From this, we computed the Kappa statistic using the formula (B) [9] as below,

$$\frac{Observed\ proportion\ agreement - exp\ ected\ proportion\ agreement}{1 - exp\ ected\ proportion\ agreement} \qquad (B)$$

The parameters used in formula (B) were determined as follows. As we can recall from Figure 2a, the actual proportion of agreement was assumed to be 86.7%; hence the corresponding number of citations with agreement (n11+n22) should equal 0.867*20,000 (=17,340). The expected proportion of agreement under the hypothesis of no agreement (i.e. if agreement is determined by a coin toss) would be given as $\frac{n1+}{n} * \frac{n+1}{n} + \frac{n2+}{n} * \frac{n+2}{n}$.

If in addition we randomly select n1+ and n11 under the condition that the total number of agreement and citations is fixed at 17,340 and 20,000 respectively, then the Kappa value can uniquely be determined from simulation results.

Percent of citations with reviewers agreement	Number of samples extracted	Number of repeated iterations	Average agreement across citations (%)	Standard error	Bias (%)*
	50	1000	86.7	5	0
86.7	100	1000	86.7	3.5	0
	200	1000	86.8	2.5	0.1

*Bias=Percent citations with reviewers agreement minus average agreement across citations.

Table 1: Evaluation of statistical bias on sampling strategies.

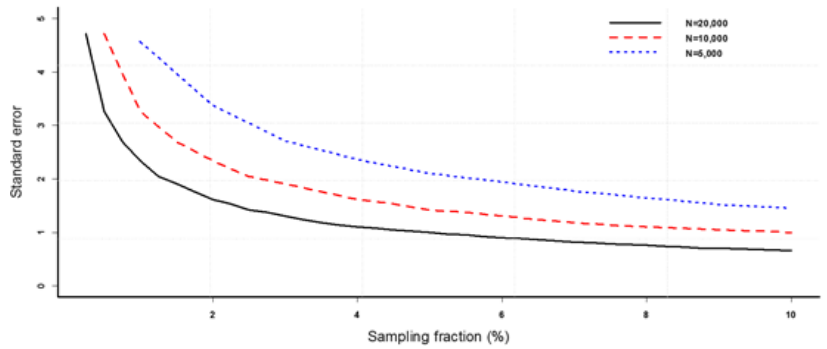

Figure 3: Comparison of standard error with sampling fraction when the total number of extracted citations is fixed at 5,000; 10,000 and 20,000.

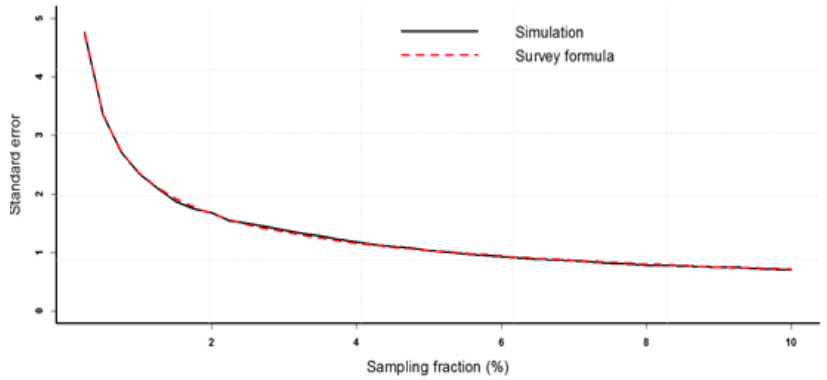

Figure 4: Comparison of simulation results with binomial formula on a plot of standard error versus sampling fraction.

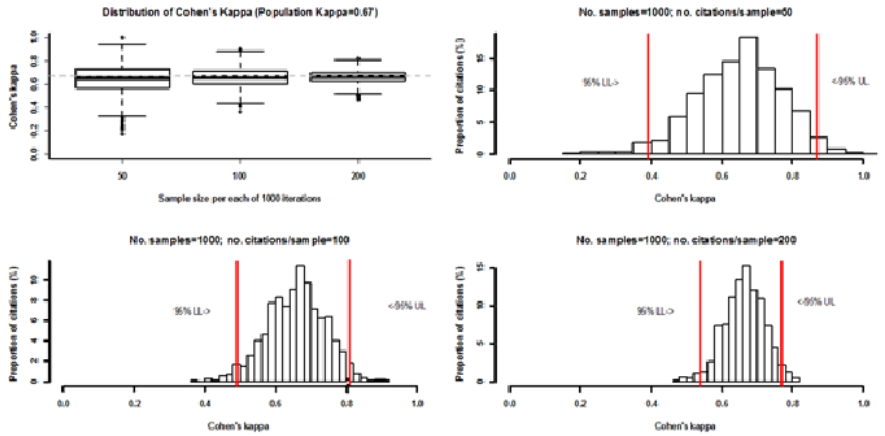

Figure 5: Distribution of Kappa, overall (a), and by sampling distribution (b, c and d).

		Reviewer 2		
		Yes	No	Total
Reviewer 1	Yes	n11	n12	n1+
	No	n21	n22	n2+
	Total	n+1	n+2	n

Table 2: Distribution of agreement between two reviewers.

Thus, we randomly selected 50 citations out of 20,000 and computed the Kappa value, repeating the iteration process 1,000 times. We run the same procedure for samples of size 100 and 200, and presented the results in Figure 5. As was the case for Figure 2, the population Kappa value (0.67) was covered by 95% CI across all samples. As the sample size increased, samples became more clustered to the population Kappa, thereby increasing precision. The statistical bias was negligible regardless of the choice of sample size although the precision doubled when the sample size increased from 50 to 200 (Table 3).

Discussion

We have proposed a sampling methodology for evaluating the quality of screening studies in SRs. Through statistical simulations, we examined some properties of the proposed sampling approach. Specifically, we showed that, for sampling fractions that are smaller than 20%, the sampling distribution produced by the proposed methodology covered the theoretical agreement of included citations with a relatively high degree of precision. The resultant 95% coverage became more precise as the sampling fraction increased. This suggests that the proposed sampling methodology is practical in situations where reviews are conducted within a short timeframe. We showed that the proposed sampling methodology produced an unbiased estimate of the theoretical agreement of included citations across different sampling fractions. In addition, we showed that different sampling fractions were consistent with the sample sizes calculated from the formula for the binomial sample with finite population correction. Finally, we demonstrated that the proposed sampling methodology worked well in cases where index measures of agreement are used.

We proposed the sample size formula to facilitate the sampling process. The formula requires an estimate of the proportion of agreement to be provided. Such estimates can be sought from past reviews. If there is uncertainty as to what estimate to be supplied, a conservative value of 50% can be used. For fixed SE, this would result in the largest sample size needed.

There are several limitations. First, we presented one scenario that mimic the empirical distribution of agreement. We believe that this scenario would provide a reasonable reflection of agreement whenever

Kappa value from all citations	No. samples extracted	No. repeated iterations	Average Kappa across citations	Standard error	Bias (%)*
	50	1000	0.65	0.12	0.02
	100	1000	0.66	0.08	0.01
0.67	200	1000	0.66	0.06	0.01

*Bias=kappa value from all citations minus average kappa across citations.

Table 3: Evaluation of statistical bias on the sampling distributions of Kappa values.

reviewers have an adequate level of training and a well-written screen in protocol. This claim is consistent with findings from previous study [7] that observed a minimum of 89% interrater agreement for including an article for full text review in 183 studies. So, our hypothesized value of agreement of 86.7% is not far off from what is observed in practice. Nevertheless, we also tried other scenarios of agreement and the conclusion did not differ. Second, we made an assumption that the reviewers are not influenced or biased by the presentation style of citations. For instance, one study found that the reviewer's decision with regard to title selection was influenced by the length of the title, presence of colon or acronym and reference to a specific country [10]. In such instances where certain factors are suspected to bias reviewer's judgment a stratified random sampling approach can be used as an alternative [4]. Finally, one might argue that reviewers may differ in experience which can influence the screening procedure and therefore the level of agreement. However, there is evidence that the level of experience of reviewers may not have an impact on the level of error, rather, adequate instruction and training can mitigate this bias [11]. Despite these limitations, our proposed sampling methodology has desired properties of reliability and validity for screening papers when resources are constrained. This methodology is easy to implement, pragmatic, and provides precise unbiased estimate of the theoretical agreement to full screening of citations at the titles and abstract stage of the systematic review.

Conclusion

In conclusion, we have shown that an alternate random sampling approach for the second reviewer citation selection at the titles and abstract stage in the systematic review process is both reliable and valid. It may be used in resource constrained environments instead of the conventional approach. Sampling fractions that are smaller than the minimum recommendation of 20% are often just as accurate, and can be used when resources are limited. However, care should be taken that this fraction reflects the magnitude of sampling error that researchers are willing to tolerate.

Acknowledgment

We would like to thank all staff at Health Quality Ontario for their support and encouragement throughout this project especially Shamara Baidoobonso, Les Levine and Ba' Pham for their contributions during the genesis of the ideas presented in this manuscript.

References

1. Moher D, Tetzlaff J, Tricco AC, Sampson M, Altman DG (2007) Epidemiology and reporting characteristics of systematic reviews. PLoSMed 4: e78.
2. Cochrane W (1977) Sampling Techniques (3rdedn.) John Wiley & Sons.
3. Tranfield D, Denyer D, Smart P (2003) Towards a Methodology for Developing Evidence-Informed Management Knowledge by Means of Systematic Review. British Journal of Management 14: 207-222.
4. Cook DJ, Mulrow CD, Haynes RB (1997) Systematic reviews: synthesis of best evidence for clinical decisions. AnnInternMed 126: 376-380.
5. Edwards P, Clarke M, DiGuiseppi C, Pratap S, Roberts I, et al. (2002) Identification of randomized controlled trials in systematic reviews: accuracy and reliability of screening records. Stat Med 21:1635-1640.
6. Zaza S, Wright-De Aguero LK, Briss PA, Truman BI, Hopkins DP, et al. (2000) Data collection instrument and procedure for systematic reviews in the Guide to Community Preventive Services. Task Force on Community Preventive Services. Am J Prev Med 18: 44-74.
7. Mateen FJ, Oh J, Tergas AI, Bhayani NH, Kamdar BB (2013) Titles versus titles and abstracts for initial screening of articles for systematic reviews. Clin Epidemiol 5: 89-95.
8. Azzalini A (1985) A class of distributions which includes the normal ones. Scandinavian Journal of Statistics 12: 171-178.
9. Maclure M, Willett WC (1987) Misinterpretation and misuse of the kappa statistic. AmJEpidemiol 126: 161-169.
10. Jacques TS, Sebire NJ (2010) The impact of article titles on citation hits: an analysis of general and specialist medical journals. JRSM Short Rep1: 2.
11. Horton J, Vandermeer B, Hartling L, Tjosvold L, Klassen TP, et al. (2010) Systematic review data extraction: cross-sectional study showed that experience did not increase accuracy. J Clin Epidemiol 63: 289-298.