Identification of randomized controlled trials in systematic reviews: accuracy and reliability of screening records

Phil Edwards¹,∗,†, Mike Clarke², Carolyn DiGuiseppi³, Sarah Pratap⁴, Ian Roberts¹ and Reinhard Wentz⁴

¹Public Health Intervention Research Unit, Department of Epidemiology and Population Health, London School of Hygiene & Tropical Medicine, 49–51 Bedford Square, London WC1B 3DP, U.K.
²The UK Cochrane Centre, NHS R&D Programme, Summertown Pavilion, Middle Way, Oxford OX2 7LG, U.K.
³Department of Preventive Medicine & Biometrics, University of Colorado Health Sciences Center, Campus Box C245, 4200 East Ninth Avenue, Denver, CO 80262, U.S.A.
⁴Cochrane Injuries Group, Public Health Intervention Research Unit, Department of Epidemiology and Population Health, London School of Hygiene & Tropical Medicine, 49–51 Bedford Square, London WC1B 3DP, U.K.

SUMMARY

A study was conducted to estimate the accuracy and reliability of reviewers when screening records for relevant trials for a systematic review. A sensitive search of ten electronic bibliographic databases yielded 22,571 records of potentially relevant trials. Records were allocated to four reviewers such that two reviewers examined each record and so that identification of trials by each reviewer could be compared with those identified by each of the other reviewers. Agreement between reviewers was assessed using Cohen’s kappa statistic. Ascertainment intersection methods were used to estimate the likely number of trials missed by reviewers. Full copies of reports were obtained and assessed independently by two researchers for eligibility for the review. Eligible reports formed the ‘gold standard’ against which an assessment was made about the accuracy of screening by reviewers. After screening, 301 of 22,571 records were identified by at least one reviewer as potentially relevant. Agreement was ‘almost perfect’ (κ > 0.8) within two pairs, ‘substantial’ (κ > 0.6) within three pairs and ‘moderate’ (κ > 0.4) within one pair. Of the 301 records selected, 273 complete reports were available. When pairs of reviewers agreed on the potential relevance of records, 81 per cent were eligible (range 69 to 91 per cent). If reviewers disagreed, 22 per cent were eligible (range 12 to 45 per cent). Single reviewers missed on average 8 per cent of eligible reports (range 0 to 24 per cent), whereas pairs of reviewers did not miss any (range 0 to 1 per cent). The use of two reviewers to screen records increased the number of randomized trials identified by an average of 9 per cent (range 0 to 32 per cent). Reviewers can reliably identify potentially relevant records when screening thousands of records for eligibility. Two reviewers should screen records for eligibility, whenever possible, in order to maximize ascertainment of relevant trials. Copyright © 2002 John Wiley & Sons, Ltd.

KEY WORDS: systematic reviews; screening; inter-observer reliability; ascertainment intersection

∗Correspondence to: Phil Edwards, Public Health Intervention Research Unit, Department of Epidemiology and Population Health, London School of Hygiene & Tropical Medicine, 49–51 Bedford Square, London WC1B 3DP, U.K.
†E-mail: phil.edwards@lshtm.ac.uk

Contract/grant sponsor: The BUPA Foundation

Copyright © 2002 John Wiley & Sons, Ltd.
1. INTRODUCTION

One of the most important steps in any systematic review of randomized controlled trials is to identify and include all, or nearly all, of the relevant trials. Inclusion of all relevant trials reduces random error in meta-analyses and, because ease of identification of trials is associated with treatment effects, complete ascertainment may also reduce bias [1]. To identify as many of the relevant trials as possible, reviewers will typically conduct sensitive searches of relevant electronic bibliographic databases. These searches might yield thousands of records of potentially relevant reports that must then be screened to determine whether they are eligible for the review. Records that are excluded during screening are not usually considered again. Although screening records for eligibility may impact importantly on both precision and bias in systematic reviews, there is little information on the accuracy and reliability of the screening process.

Previous studies [2, 3] have examined agreement between reviewers but have not estimated the number of relevant trials missed by reviewers, or the increase in the number of trials identified using a second reviewer. We conducted a study to estimate the accuracy and reliability of screening records for relevant trials for a systematic review, using records retrieved by a search of ten databases.

2. METHODS

A study was designed to assess the accuracy and reliability of reviewers when screening records. The records used were those identified by the search strategy for a systematic review of methods to influence response to postal questionnaires [4]. This systematic review sought to identify all reports of randomized controlled trials of strategies to influence response to postal questionnaires, with no restriction by type of strategy, questionnaire topic, target population or language. A search of ten electronic bibliographic databases yielded 26937 records of potentially relevant reports. Records including titles and abstracts were downloaded into a ProCite reference database. After removing duplicate records, there were 22571 records of potentially relevant reports. Records were sorted and divided into six approximately equal sets (A to F).

Four reviewers took part in the study, each being a co-reviewer on the systematic review [4]. Each reviewer had substantial experience of screening records for systematic reviews, apart from reviewer 2 who was relatively inexperienced. Each was allocated three of the sets to screen (that is, each reviewer was to screen approximately 11286 records). The six sets were allocated such that two reviewers examined each record, and so that identification of trials by each reviewer could be compared with each of the other reviewers (reviewer 1: A, B, C; reviewer 2: A, D, E; reviewer 3: B, D, F; reviewer 4: C, E, F). Prior to screening records, the reviewers met to discuss the criteria for trials to be eligible for inclusion in the systematic review. Reviewers then worked independently, screening the title and abstract of each record in their sets, selecting those they considered being potentially relevant.

Agreement between reviewers on the relevance of records was assessed using Cohen’s kappa statistic (κ), which adjusts the proportion of records for which there was agreement by the amount of agreement expected by chance alone [5]. We calculated 95 per cent confidence intervals for kappa as ±1.96 times the standard error of kappa [6].
Ascertainment intersection methods, analogous to capture–recapture methods used in ecology, were used to estimate the likely number of relevant records missed by pairs of reviewers [7, 8]. For example, the maximum likelihood estimate for the number of relevant records missed by reviewer 1 and reviewer 2 during screening is

\[ \frac{bc}{a + 1} \]

where \( b \) is the number of relevant records identified only by reviewer 1, \( c \) is the number identified only by reviewer 2, and \( a \) is the number identified by both reviewers. This method produces an unbiased estimate of relevant records missed by both reviewers [7]. Confidence intervals for the estimate were calculated using the goodness-of-fit model [9].

When screening was complete, full copies of the reports identified by at least one reviewer as potentially relevant were requested. Two researchers independently assessed each report for eligibility for inclusion in the systematic review. Disagreements about eligibility were referred to a third researcher. The eligible reports were used as the ‘gold standard’ against which an assessment was made about the accuracy of screening by reviewers.

3. RESULTS

After screening, 301 of 22,571 records had been selected by at least one of the four reviewers as being potentially relevant. Reviewer 1 selected 102 records, reviewer 2 selected 74 records, reviewer 3 selected 95 records and reviewer 4 selected 163 records. Of the six possible comparisons between reviewers, kappa coefficients of agreement ranged from 0.59 (95 per cent CI 0.56 to 0.62) to 0.93 (95 per cent CI 0.90 to 0.96) (Table I). Agreement was ‘almost perfect’ (\( \kappa > 0.8 \)) within two pairs, ‘substantial’ (\( \kappa > 0.6 \)) within three pairs, and ‘moderate’ (\( \kappa > 0.4 \)) within one pair [10]. Ascertainment intersection methods estimate that single reviewers missed on average 22 per cent (range 3 to 55 per cent) of potentially relevant records. Pairs of reviewers missed 4 per cent (range 0 to 6 per cent) of potentially relevant records.

Of the 301 records selected, we were able to obtain 273 complete reports for our assessment of accuracy (the remainder being unobtainable despite extensive attempts through the British Library and other sources) (Table II). Of these, 156 (57 per cent) met the inclusion criteria.
Table II. Numbers of records selected by four reviewers as potentially relevant (numbers subsequently found to be eligible). Ascertainment intersection estimates* of reports missed by reviewers are shown in bold.

<table>
<thead>
<tr>
<th>Reviewer 1</th>
<th>Reviewer 2</th>
<th>Reviewer 3</th>
<th>Reviewer 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Yes</td>
<td>22 (20)</td>
<td>34 (28)</td>
<td>35 (26)</td>
</tr>
<tr>
<td>No</td>
<td>3 (0)</td>
<td>1 (0)</td>
<td>40 (3)</td>
</tr>
<tr>
<td>Reviewer 2</td>
<td>Yes</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>19 (17)</td>
<td>4 (1)</td>
<td>25 (22)</td>
</tr>
<tr>
<td>No</td>
<td>8 (2)</td>
<td>2 (0)</td>
<td>24 (7)</td>
</tr>
<tr>
<td>Reviewer 3</td>
<td>Yes</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td></td>
<td></td>
<td>26 (18)</td>
</tr>
<tr>
<td>No</td>
<td></td>
<td></td>
<td>13 (5)</td>
</tr>
</tbody>
</table>
| Figures exclude 28 identified reports that were not available. *Ascertainment intersection estimates have been rounded to integers.

for the systematic review. Reviewer 1 identified 79 eligible reports, reviewer 2 identified 60 eligible reports, reviewer 3 identified 67 eligible reports and reviewer 4 identified 81 eligible reports. When pairs of reviewers agreed about the potential relevance of records, the majority (81 per cent) was found to be eligible (131 of 161 records). This proportion ranged across the six pairings of reviewers from 69 per cent eligible (18 of 26 records) to 91 per cent eligible (20 of 22 records). If two reviewers disagreed on the potential relevance of records, 22 per cent were still found to be eligible (25 of 112 records). This proportion ranged from 12 per cent eligible (1 of 8 records) to 45 per cent eligible (5 of 11 records). Ascertainment intersection methods estimate that single reviewers missed on average 8 per cent (range 0 to 24 per cent) of eligible reports and that pairs of reviewers did not miss any (range 0 to 1 per cent). The average increase in the total number of randomized trials identified by using a second reviewer was 9 per cent, ranging across pairs from 0 per cent (both reviewers individually identified 29 trials) to 32 per cent (a second reviewer identified an additional seven trials increasing the total identified to 29 trials).

4. DISCUSSION

Screening many thousands of records might be expected to be unreliable. This study has shown, however, that pairs of reviewers can reliably identify potentially relevant records and that together they can ensure that few eligible records are excluded. Our results on the inter-observer reliability of screening records were consistent with those reported in previous studies [2, 3]. There are, however, a number of methodological issues that may have a bearing on the validity of these, and our other results.

First of all, the eligibility of selected records was determined by two researchers. Although this was done using the full reports, some studies may nevertheless have been misclassified. Any misclassification may have led us to underestimate the number of studies missed by reviewers. Secondly, for the ascertainment intersection estimates to be accurate, the sources
used (reviewers screening records) must be independent. In this study, independence was ensured by having reviewers screen records in different locations, remaining blind to the results of other reviewers until all screening was complete.

Another requirement for the ascertainment intersection estimates to be accurate is for the probability with which relevant records are selected to be constant for any reviewer. Some variation in the ease with which relevant records can be identified, and in the skills of reviewers in identifying potentially relevant records, is likely. In this study, prior to screening, the criteria for records to be eligible were discussed, in an attempt to achieve consistency between reviewers. However, it is possible that their competence in applying the criteria improved with practice. In addition, decreased attentiveness during long periods of screening may also have influenced the probability of selection of relevant records. This was not supported by the results, however, where reviewer 2 took over seven days to screen records and reviewer 1 screened in less than two days, and yet agreement between them was ‘almost perfect’. There was considerable variation in the kappa coefficients none the less. It might be thought that greater experience in screening would lead to greater agreement, but the kappa coefficients of agreement for the reviewer with the least experience (reviewer 2) were higher, on average, than those for reviewer 4. These results provide some evidence of variation in the probability of ascertainment of relevant records by reviewers. Although any resulting bias is likely to be relatively small [11], estimates of the numbers of reports missed by pairs of reviewers must be treated with a degree of caution.

The limitations of the kappa coefficient of agreement have been discussed elsewhere [12]. Of particular relevance is the fact that the kappa statistic can vary considerably depending on the proportion of records that is eligible. In this study approximately 1 per cent of records was selected as potentially relevant. Similar levels of agreement between reviewers in other studies may not produce similar values of kappa if a greater proportion of records is eligible.

For this study we used the records identified by a search for randomized controlled trials of methods to influence response to postal questionnaires. The records came from searches of ten bibliographic databases, which included the major general medical databases [4]. Although the reports identified by the four reviewers were from journals, books, conference proceedings and other sources, covering a wide range of academic disciplines, the extent to which these results can be generalized (for example, to searches for randomized controlled trials of medical interventions only), is a matter for judgement.

Screening records for eligibility may impact importantly on both precision and bias in systematic reviews. Although a single reviewer is likely to identify the majority of relevant trials, we found that a second reviewer can increase the number identified by as much as one-third. Two reviewers should therefore screen records for eligibility whenever possible in order to maximize ascertainment of relevant trials. Further research could examine how aspects of electronic records influence their ease of identification for systematic reviews.

REFERENCES