Agreement in primary study selection between systematic reviews

Martin Bland

Dept. of Health Sciences
University of York
YO10 5DD

With Mark Rodgers [b], Debra Fayter [b], Amanda Sowden [b], Robert Lewin [a]

[b] Centre for Reviews and Dissemination
University of York
YO10 5DD

[a] Dept. of Health Sciences
University of York
YO10 5DD

This talk was first presented to the Department of Health Sciences, University of York, 8th March 2006.

Introduction

Systematic reviews of the same question sometimes come to different conclusions. There are two plausible explanations of this: the reviews may use different analytical techniques, or they may take data from different primary studies. This paper is concerned with the latter possibility. We want to assess the agreement between systematic reviews as to which primary studies they include.

There are two problems in doing this. First, we only know that a primary study is eligible for inclusion in a review because one of the existing reviews selected it. We do not know about studies which were not identified by any review. Second, reviews will have been carried out at different times and so could not be expected to select the same studies. Some papers are available to be selected by more reviews than others. The earliest papers could be selected by all reviews, as all the reviews came after them. Recently published papers could not be selected by the early reviews, only by later ones.

In this paper we describe a method for estimating the agreement between reviews as to which primary studies should be included, which takes account of both problems. We illustrate the method using a series of reviews of complex psychosocial interventions in heart disease (Rodgers *et al.*, 2006).

All analyses were done using Stata 8 (Stata Corp., College Station, Texas).

Proposed method

We can deal with the first problem using a method developed by Markus *et al.* (1996) and described in more detail by Bland (2004). These authors were looking at agreement in the detection of signals. They had no way of knowing how many signals were undetected by all observers. They could not, therefore, use kappa statistics to describe the agreement. Instead, they estimated the probability that if one observer detected a signal another observer would also detect the signal. In the present application, we can estimate the probability that if one review selects a paper, another review will also select the paper.

To estimate this probability, all we need are the numbers of reviews selecting a primary study for each study selected in any of the reviews. Denote the numbers of reviews by n, and the number of reviews selecting study i by r_i. For each review selecting study i, there are n–1 other reviews, r_i–1 of which select the study. Hence the proportion of other reviews selecting the study is $(r_i-1)/(n-1)$. This proportion will be the same for all the r_i selections of study i. The total number of selections of primary studies is Σr_i and the average proportion of further reviews which select a study, averaged over all selections, is

$$p_{select} = (\Sigma r_i(r_i-1)/(n-1))/(\Sigma r_i) = (\Sigma r_i^2 - \Sigma r_i)/(n-1)\Sigma r_i$$

The second problem, that not all reviews could select all primary studies, can be dealt with as follows. For each year when an included study is published, we estimate the probability that another review would select the study for all the studies published in that year only. The reviews that could select them will be those reviews published one or more years later. Reviews published before then would not be able to select these primary studies. This gives us an estimate of the probability for each year. We then find an average of these. Because some years will have more publications of primary studies than others and so will contribute more information, for the calculation of the average probability that another review would select the study the probability for the year will be weighted by the number of studies published in that year. This will also give us a standard error and a 95% confidence interval for the probability. This works because each study is considered only once and so the estimates of probability for each year are independent.

Hence to carry out the calculation what we need is for each primary study the number of reviews which have referenced it and the year of publication of the study, plus the number of reviews able to select studies published each year.

Example

Table 1 shows the primary studies included in a series of systematic reviews of complex psychosocial interventions in heart failure (Rodgers *et al.*, 2006). For the purposes of this illustration, reviews are identified by a letter code and primary studies by a numeric code. Table 2 shows for each year of publication the primary studies published and the number of reviews which cited each of them.

Table 2 contains all the data necessary to estimated p_{select} for each year, the number of reviews which could cite the study and the number of citations for each study. From these we calculate the sum of the numbers of citations and the sum of the numbers of citations squared and hence p_{select}. For years when each of the studies was cited by only one review, such as 1972, p_{select} = zero. For years when only one review could have cited the studies, such as 2001, no estimate of p_{select} is possible as there is no other review to cite the study.

We average the p_{select} estimates for each year when estimates exist, weighted by the number of primary studies published in the year. This gives the estimated probability that if a study is selected by a review, a further review would also select it. For Table 2 this is 0.110, SE =

0.019, 95% confidence interval 0.071 to 0.148. We estimate that if a study is selected by a review, the probability that another review would also select it is estimated to be between 7% and 15%. Despite their reporting of similar research questions, there is very little consistency in the selection of studies in these reviews.

Discussion

We have demonstrated a method to quantify the agreement between systematic reviews in their selection of primary studies. This overcomes two problems: that we only know that a primary study is eligible for inclusion because one of the studies selected it and that reviews can select only primary studies which precede them in time.

We have not found any other discussion of this problem. The nearest is the use of capture-recapture methods to estimate the total number of eligible studies when two different search methods are employed (Spoor *et al.*, 1996, Bennett *et al.*, 2004). Considerable modification would be required to adapt these methods to deal with the time factor here, where primary studies could be captured only by reviews later in time. It may be possible to adapt such techniques to estimate the number of excluded studies and hence produce a kappa statistic of some sort.

As the number of systematic reviews increases and the techniques is extended to more complex questions, we may expect contradictory results to become more frequent and we hope that the method described here will help to resolve these contradictions.

References

Rogers M, Fayter D, Sowden A, Bland JM, Lewin M. (2006) Systematic reviews of psychological and behavioural interventions in cancer and heart disease: how reliable and useful are the findings? Submitted for publication.

Markus H, Bland JM, Rose G, Sitzer M, Siebler M. (1996) How good is intercenter agreement in the identification of embolic signals in carotid artery disease? *Stroke*, **27**: 1249-1252.

Bland JM. (2004) How do I measure observer agreement when only positive observations are recorded? http://www-users.york.ac.uk/~mb55/meas/nonos.

Spoor P, Airey M, Bennett C, Greensill J, Williams R. (1996) Use of the capture-recapture technique to evaluate the completeness of systematic literature searches. *Brit Med J*, **313**: 342-343.

Bennett DA, Latham NK, Stretton C, Anderson CS. (2004) Capture-recapture is a potentially useful method for assessing publication bias. *J Clin Epidem* **57**: 349–357.

Table 1. Primary studies included in a series of systematic reviews of complex interventions in heart failure (ref Lewin)

Review	Year	Serial numbers of included papers
A	1987	1, 40, 54, 58, 62, 71, 72, 82, 83, 112, 120, 123, 130, 133, 143, 145, 154, 164
B	1989	20, 25, 36, 37, 38, 41, 53, 64, 67, 68, 71, 77, 78, 91, 94, 95, 108, 118, 123, 125, 135, 166, 167
C	1992	6, 8, 12, 20, 23, 28, 30, 32, 33, 40, 43, 57, 63, 66, 71, 79, 82, 86, 90, 93, 94, 98, 103, 109, 110, 123, 129, 131, 134, 139, 141, 163, 166, 170
D	1996	1, 14, 16, 19, 24, 40, 44, 51, 58, 60, 69, 101, 121, 122, 123, 137, 143, 149, 150, 155, 156, 157, 158
E	1997	2, 4, 7, 9, 10, 11, 22, 27, 32, 49, 55, 56, 93, 100, 115, 116, 117, 126, 136, 138, 142
F	1999	12, 29, 31, 34, 35, 39, 42, 43, 47, 49, 55, 65, 70, 71, 76, 77, 78, 87, 93, 95, 99, 105, 106, 121, 122, 134, 141, 143, 144, 147, 148, 156, 160, 161, 170
G	2000	18, 19, 21, 49, 50, 51, 60, 77, 80, 81, 91, 104, 111, 112, 113, 124, 127, 128, 129, 139, 140, 146, 147, 162, 165, 166, 168
H	2003	3, 5, 13, 15, 16, 17, 18, 26, 31, 35, 36, 42, 44, 45, 46, 47, 48, 49, 51, 52, 59, 69, 71, 73, 74, 75, 76, 84, 85, 88, 89, 96, 97, 99, 102, 105, 107, 113, 114, 119, 120, 123, 132, 143, 149, 150, 151, 152, 153, 156, 157, 158, 159, 161, 162, 169

Table 2. Primary studies published and the number of reviews which cited them, for each year of publication.

Year	**Number of reviews which could select these papers, *n***	**Paper serial no. (number of reviews, r_i)**	$\sum r_i$	$\sum r_i^2$	p_{select}
1968	8	1 (2)	2	4	0.1428571
1971	8	129 (1)	1	1	0.0000000
1972	8	80 (1), 167 (1)	2	2	0.0000000
1973	8	79 (1)	1	1	0.0000000
1974	8	20 (2), 23 (1), 71 (5)	8	30	0.3928571
1975	8	58 (2), 122 (2), 166 (3)	7	17	0.2040816
1977	8	91 (1), 121 (2), 164 (1)	4	6	0.0714286
1978	8	22 (1), 109(1), 145 (1)	3	3	0.0000000
1979	8	39 (1), 72 (1), 77 (3), 86 (1), 123 (5), 130 (1)	12	38	0.3095238
1980	8	7 (1), 40 (3), 133 (1), 135 (1), 154 (1)	7	13	0.122449
1981	8	33 (1), 67 (1), 78 (2), 94 (2), 139 (2),	8	14	0.1071429
1982	8	21 (1), 32 (2), 51 (2), 57 (1), 59 (1), 82 (2), 115 (1), 148 (1), 170 (2)	13	21	0.0879121
1983	8	8 (1), 12 (2), 35 (2), 38 (1), 63 (1), 68 (1), 83 (1), 95 (2), 110 (1), 112 (2), 124 (1), 128 (1), 141 (2), 143 (4), 162 (2), 163 (1)	25	49	0.1371429
1984	8	14 (1), 31 (1), 50 (2), 62 (1), 70 (2), 103 (1), 120 (2), 125 (1), 131 (1), 168 (1)	13	19	0.0659341
1985	8	28 (1), 30 (2), 37 (1), 43 (2), 45 (1), 54 (1), 92 (1), 105 (2), 134 (2)	13	21	0.0879121
1986	8	6 (1), 25(1), 36 (2), 49 (3), 53 (1), 66 (1), 93 (3), 108 (1), 118 (1), 137 (1), 144 (1)	16	30	0.1250000
1987	7	4 (1), 16 (2), 46 (1), 64 (1), 65 (1), 90 (1), 142 (1), 157 (2)	8	10	0.0416667
1988	7	41 (1), 56 (1), 98 (1), 101 (1), 111 (1), 119 (1), 146 (1)	7	7	0.0000000
1989	6	11 (1), 44 (2), 60 (1), 106 (1), 116 (1), 136 (1), 138 (1), 149 (2), 156 (3)	13	23	0.1538462
1990	6	113 (2), 117 (1), 147 (2), 150 (2), 152 (1), 158 (2)	10	18	0.1600000
1991	6	27 (1), 47 (2), 97 (1), 132 (1), 151 (1)	6	8	0.0666667
1992	5	24 (1), 34 (1), 48 (1), 81 (1), 84 (1), 99 (2), 104 (1), 126 (1), 140 (1), 165 (1)	11	13	0.0454545
1993	5	15 (1), 55 (2), 61 (1)	4	6	0.1250000

1994	5	19 (2), 29 (1), 127 (1), 160 (1), 161 (2)	7	11	0.1428571
1995	5	9 (1), 10 (1), 17 (1), 87 (1), 107 (1), 155 (1)	8	10	0.0625000
1996	4	2 (1), 18 (2), 76 (2), 85 (1), 100 (1)	7	11	0.1904762
1997	3	42 (2), 52 (1)	3	5	0.3333333
1998	3	13 (1), 74 (1), 114 (1), 153 (1)	4	4	0.0000000
1999	2	69 (1), 73 (1), 75 (1), 88 (1), 159 (1)	5	5	0.0000000
2000	1	3 (1), 5 (1)	2	2	*
2001	1	26 (1), 96 (1), 102 (1)	3	3	*

* no estimate possible because only one review could cite these papers