Unpublished rating scales: a major source of bias in randomised controlled trials of treatments for schizophrenia

MAX MARSHALL, AUSTIN LOCKWOOD, CAROLINE BRADLEY, CLIVE ADAMS, CLAIRE JOY and MARK FENTON

Background A recent review suggested an association between using unpublished scales in clinical trials and finding significant results.

Aims To determine whether such an association existed in schizophrenia trials.

Method Three hundred trials were randomly selected from the Cochrane Schizophrenia Group's Register. All comparisons between treatment groups and control groups using rating scales were identified. The publication status of each scale was determined and claims of a significant treatment effect were recorded.

Results Trials were more likely to report that a treatment was superior to control when an unpublished scale was used to make the comparison (relative risk 1.37 (95% CI 1.12–1.68)). This effect increased when a 'gold-standard' definition of treatment superiority was applied (RR 1.94 (95% CI 1.35–2.79)). In non-pharmacological trials, one-third of 'gold-standard' claims of treatment superiority would not have been made if published scales had been used.

Conclusions Unpublished scales are a source of bias in schizophrenia trials.

Declaration of interest None. No funding.

Rating scales are widely used in clinical trials to measure 'intangible' but important outcomes such as symptoms, social functioning or quality of life (Bowling, 1991). Reviews have shown that a substantial minority of such scales are unpublished (Sanders *et al*, 1998; Marshall & Lockwood, 1999). A recent systematic review of assertive community treatment for people with severe mental disorders observed that comparisons based on data from unpublished scales were more likely to favour the treatment group than comparisons based on data from published scales (Marshall & Lockwood, 1999). The present study aimed to determine whether the same effect was present in a representative sample of all randomised controlled trials of treatments for schizophrenia.

METHOD

Selection of trials

An independent research assistant, who was not otherwise participating in the study, used computer-generated random numbers to select 300 trials from the Cochrane Schizophrenia Group's Register of randomised controlled trials (RCTs). The Register is the most comprehensive source of controlled trials involving people with schizophrenia. It contains reports of about 4000 RCTs, identified by systematic searching of a wide range of bibliographic databases, conference abstracts and unlisted journals (Adams, 1998). The Register records whether trials were of pharmacological or non-pharmacological treatments, and this information was used to stratify the random selection process, so that 150 trials were of pharmacological treatments and 150 of non-pharmacological treatments. Hard copies of the selected trials were extracted from the Cochrane Schizophrenia Group's files. There were not enough data to make a reliable formal power calculation, so the sample size was based on an estimate of the largest number of trials that could be handled in the time available to the researchers. This sample was expected to provide three to four times more observations than had been available to the earlier systematic review (Marshall & Lockwood, 1999).

Assessment of trials and rating scales

Two teams of three raters, who were trained before the study began, evaluated the trials. Each team was randomly allocated 150 trials (75 pharmacological, 75 non-pharmacological). Each team screened its allocated trials to determine which met the following criteria: (a) that the trial was available in English; (b) that it was an investigation of treatment effectiveness; and (c) that it was a true randomised controlled trial (i.e. not a quasi-experimental or case–control design).

Two raters from the same team examined each eligible trial. The first rater read the Method and Reference sections of the trial report, to identify any rating scales used and to determine whether they had been published. Two definitions of 'unpublished' were examined: (a) the scale had never been published in a peer-reviewed journal, indexed on one of the major electronic databases; and (b) the scale was unpublished (definition as above) at the time the trial report appeared. The judgement of publication status was made by checking each scale's citation in the reference section of the trial. Where a citation was missing or inadequate, the rater searched for a reference to the scale in: (a) EMBASE (01/1980–07/1998), MEDLINE (01/1966–07/1998) and PsycLIT (01/1887–07/1998); (b) the Cochrane Schizophrenia Group's database of instruments used in schizophrenia trials; and (c) the bibliography of psychiatric rating scales published by the Royal College of Psychiatrists (Royal College of Psychiatrists, 1994). Finally, the first rater recorded whether some measures had been taken to limit observation bias when the outcome was assessed, by using a single- or double-blind evaluation.

The second rater then examined the Abstract, Results and Conclusions of the trial report, to evaluate the outcome of comparisons between the treatment and the control groups that were based on data from the identified rating scales. The second rater evaluated the outcome of all

https://doi.org/10.1192/bjp.176.3.249 Published online by Cambridge University Press

comparisons according to broad ('face-value') and narrow ('gold-standard') definitions of a significant superiority of treatment over control. The 'face-value' definition was that: (a) the trialists claimed that the treatment group had a significantly better outcome; and (b) this claim was supported by a significant difference (at a 5% level of significance) between groups at some point in the trial as measured by the rating scale. This definition permitted the rater to take trialists' claims at face value, but had the disadvantage of including claims based on uncorrected multiple testing (such as occurs, for example, when separate statistical tests are applied to each scale item, and treatment superiority is claimed if any item shows a significant difference in favour of treatment).

The 'gold-standard' definition was that the treatment group had a significantly better outcome (5% level) on the overall (summary) score from the rating scale at the end of the trial. This definition required a more complex judgement on the part of the rater, but had the advantage of excluding all claims based on analysis of individual scale items or of interim results.

Throughout the rating process, the first rater (evaluating scales) was blind to the outcome of comparisons made by those scales, and the second rater (evaluating comparisons) was blind to whether the rating scales, used to make the comparisons, had been published. Data were entered on a customised Visual Basic program running on a Windows NT network.

Reliability of judgements

The reliability of the teams' judgements was assessed by duplicating 30 of the 300 selected trials. In total, therefore, each team actually assessed 165 trials, 15 of which were duplicates of studies already allocated to the other team. The duplicates were selected, copied and embedded by an administrator who was not a team member. Raters were blind to the identity of the duplicates. Cohen's kappa (Landis & Koch, 1977) was calculated for the following judgements: (a) the trial met the inclusion criteria; (b) the trial reported data from rating scales; (c) the rating of outcome in the trial was blind; (d) treatment was superior to control according to face-value definition; and (e) treatment was superior to control according to gold-standard definition.

Analysis

The main hypothesis was that comparisons based on data from unpublished scales would be more likely to show that treatment was superior to control for both face-value and gold-standard definitions of treatment superiority. This was tested by calculating the relative risk (RR) and the relevant 95% confidence intervals (CIs) of finding treatment superior to control, for comparisons based on unpublished *v.* published scales. Secondary hypotheses were that any association between using unpublished scales and finding treatment superior to control would: (i) remain even in trials where outcome was assessed blind to treatment allocation; and (ii) be present at the same level in pharmacological and non-pharmacological trials.

RESULTS

Characteristics of sample

Of the 300 trials, 107 were excluded for the following reasons: available in abstract only (46); non-therapeutic (50); unavailable in English (5); and not true RCT (6). Of 193 eligible trials, 44 did not

Table 1 Published *v.* unpublished scales: relative risk (RR) of finding a significant superiority of treatment over control group[1]

	Face-value claim of significance					Gold-standard claim of significance (based on overall score at trial end-point)				
	Unpublished scales		Published scales			Unpublished scales		Published scales		
	S	NS	S	NS	RR (95% CI)	S	NS	S	NS	RR (95% CI)
1. All trials										
Never published *v.* published	70	57	135	194	1.34 (1.00–1.81)	38	89	52	277	1.89 (1.40–2.56)
Unpublished at trial publication date *v.* published	74	59	131	192	1.37 (1.03–1.83)	40	93	50	273	1.94 (1.45–2.60)
2. Trials where patient rated blind to treatment allocation										
Never published *v.* published	32	34	87	127	1.19 (0.78–1.82)	23	43	42	172	1.78 (1.16–2.71)
Unpublished at trial publication date *v.* published	36	36	83	125	1.25 (0.84–1.86)	25	47	40	168	1.81 (1.21–2.69)
3. Pharmacological trials										
Never published *v.* published	27	27	71	107	1.25 (0.79–2.00)	16	38	35	143	1.51 (0.92–2.47)
Unpublished at trial publication date *v.* published	30	29	68	105	1.29 (0.83–2.01)	18	41	33	140	1.60 (1.01–2.53)
4. Non-pharmacological trials										
Never published *v.* published	43	30	64	87	1.39 (0.95–2.04)	22	51	17	134	2.68 (1.86–3.84)
Unpublished at trial publication date *v.* published	44	30	63	87	1.42 (0.97–2.08)	22	52	17	133	2.62 (1.83–3.76)

1. Comparison 1 was the main hypothesis of the study, whereas Comparisons 2–4 are secondary exploratory hypotheses.
S, significant; NS, not significant.

https://doi.org/10.1192/bjp.176.3.249 Published online by Cambridge University Press

report any rating scale data. The remaining 149 studies used 196 instruments to make 456 comparisons between treatment and control groups (multiple comparisons at different time points counted as one comparison). Of these comparisons, 224 came from non-drug trials and 232 came from drug trials.

Reliability of judgements

Inter-team reliability (Cohen's kappa) was as follows: meeting inclusion criteria (0.73, 95% CI 0.45–1); the trial provided rating scale data (1); the outcome was rated blind (0.73, 95% CI 0.45–1); meeting face-value definition of treatment superiority (0.86, 95%, CI 0.67–1); meeting gold-standard definition of superiority (0.43, 95% CI 0.04–0.84). In view of the low reliability of the final rating, it was repeated independently by the two most experienced raters and a second reliability analysis performed (kappa=1). In the results below, the judgement that a comparison meets the gold-standard definition of treatment superiority is based on the opinion of the two most experienced raters.

Usage and effects of unpublished scales

According to the face-value definition of treatment superiority, treatment was superior to control in 205 (44.9%) of 456 comparisons. Of these 205 comparisons, 74 (36.1%, 95% CI 29.5–42.6%) were based on data from unpublished scales. According to the gold-standard definition of treatment superiority, treatment was superior to control in only 90 of 456 comparisons. Of these 90 comparisons, 40 (44.4%, 95% CI 34.1–54.7) were based on data from unpublished scales.

Table 1 shows the relative risks of detecting treatment superiority. Comparisons based on data from unpublished scales were significantly more likely to meet both face-value and gold-standard definitions of treatment superiority (face-value RR 1.37, 95% CI 1.12–1.68; gold-standard RR 1.94, 95% CI 1.35–2.79). For both face-value and gold-standard definitions, this association was present at a significant level in non-pharmacological trials. For the gold-standard definition only, the association was also present at a significant level in trials where outcome was assessed blind to treatment allocation. Contrary to expectation, the association was greatest in non-pharmacological trials, where 56% (95% CI 42.3–69.7) of gold-standard claims of treatment superiority are based on data from unpublished scales (gold-standard RR for non-pharmacological trials 2.62, 95% CI 1.83–3.76).

DISCUSSION

Are unpublished scales a major source of bias?

This study has shown that a treatment for schizophrenia is more likely to be found effective when evaluated using an unpublished rating scale. This association suggests that unpublished scales may be a source of bias, because there is no good reason to suppose that they are better at detecting true treatment effects than published scales. Judgements of what non-pharmacological treatments are effective in schizophrenia may have been seriously distorted by the fact that about one-third (32.2% – calculated from relative risk) of all gold-standard claims of treatment superiority would not have been made if published scales had been used to measure outcome.

Why might unpublished scales be a source of bias?

This study has shown that the association between unpublished scales and significant treatment effects cannot be explained by a tendency for unpublished scales to be used in unblinded trials (Table 1); however, alternative explanations associated with trial design are possible. For example, it may be that unpublished scales tend to be used in small, poor-quality trials, and, although it seems implausible, it is possible that the size and/or quality of the trial could lead to the association and not the publication status of the scale itself.

There are two other possible explanations for the association, both of which assume that it is directly due to the publication status of the scale, rather than to some unknown confounding variable. The first explanation is that comparisons based on data from unpublished scales are less likely to be reported when they are not significant, as compared with comparisons based on data from published scales. This publication bias is compatible with the observations of other researchers who have noted discrepancies between the scales that trialists say they have used and the data which they actually report (Gotzsche, 1989). The second explanation is that there may have been *post hoc* 'adjustment' of the contents of unpublished scales by dropping unfavourable items in order to fabricate differences in favour of the treatment group. This procedure is unlikely to raise protests because unpublished scales usually belong to the trialists themselves or to their colleagues. If a scale is published, however, such 'adjustments' become more risky.

Why is the effect of unpublished scales greatest in trials of non-pharmacological treatments?

It was surprising that the association between unpublished scales and significant treatment effects was much greater in non-pharmacological trials, even though the rate of utilisation of unpublished scales was similar to that in pharmacological trials (in fact, for pharmacological trials the association is significant only for one of the two definitions of 'unpublished'). It may be that even more unknown confounding variables operate in non-pharmacological studies, or that bias linked to the use of unpublished scales is particularly potent in these studies. Significant treatment effects (detected using published scales) are about 50% less common in non-pharmacological trials. Thus, trialists in non-pharmacological trials may be more tempted to tamper with the contents of an unpublished scale in order to find a significant treatment effect that will increase their chances of publishing the study. Non-pharmacological trials also tend to use large multi-item unpublished scales, whereas pharmacological trials tend to use smaller scales (often single-item). Thus, the type of unpublished scale seen in non-pharmacological trials may be more suitable for *post hoc* adjustment of items.

Reducing potential bias due to unpublished scales

It is unlikely that the enhanced ability of unpublished scales to detect treatment effects occurs only in schizophrenia trials. We would suggest, therefore, that all trialists should be discouraged from using unpublished scales. This might be achieved by journals and systematic reviews refusing to accept data from unpublished scales and by including guidance on using unpublished scales in the CONSORT guidelines for the reporting of clinical trials (Begg *et al*, 1996).

https://doi.org/10.1192/bjp.176.3.249 Published online by Cambridge University Press

ACKNOWLEDGEMENTS

Caroline Bradley was supported by a Wellcome Trust Health Services Research Training Fellowship. Austin Lockwood was supported by a National Health Service North-West R&D grant.

REFERENCES

Adams, C. (1998) Cochrane Schizophrenia Group's Register of Randomised Controlled Trials. *The Cochrane Library*. Oxford: Update Software.

Begg, C., Cho, M. & Eastwood, S. (1996) Improving the quality of reporting of randomized controlled trials: the CONSORT statement. *Journal of the American Medical Association*, **276**, 637–639.

Bowling, A. (1991) *Measuring Health*. Milton Keynes: Open University Press.

Gotzsche, P. C. (1989) Methodology and overt and hidden bias in reports of 196 double-blind trials of nonsteroidal, antiinflammatory drugs in rheumatoid arthritis. *Controlled Clinical Trials*, **10**, 31–56.

Landis, J. R. & Koch, G. G. (1977) The measurement of observer agreement for categorical data. *Biometrics*, **33**, 159–174.

Marshall, M. & Lockwood, A. (1999) Assertive community treatment for people with severe mental disorders (Cochrane Review). In *The Cochrane Library*. Oxford: Update Software.

Royal College of Psychiatrists (1994) *Psychiatric Instruments and Rating Scales. A Select Bibliography (2nd edn)*. Occasional Paper OP23. London: Royal College of Psychiatrists.

Sanders, C., Egger, M., Donovan, J., *et al* (1998) Reporting on quality of life in randomised controlled trials: bibliographic study. *British Medical Journal*, **317**, 1191–1194.

CLINICAL IMPLICATIONS

- **In clinical trials, a treatment for schizophrenia is more likely to be found effective when evaluated using an unpublished scale.**
- **In trials of non-pharmacological treatments for schizophrenia, one-third of claims that treatment was superior to control would not have been made if published scales had been used to collect the data.**
- **Researchers should be discouraged from using unpublished scales in clinical trials.**

LIMITATIONS

- **It is not clear how far the findings of the study apply to other branches of medicine outside schizophrenia research.**
- **It is conceivable that some factor other than bias could explain the findings.**
- **The final part of the rating procedure (meeting the 'gold-standard' definition of treatment superiority) was of low reliability and had to be repeated by the two most experienced raters.**

MAX MARSHALL, MRCPsych, Senior Lecturer, AUSTIN LOCKWOOD, RMN, Research Nurse, University of Manchester, Department of Psychiatry; CAROLINE BRADLEY, MRCPsych, Wellcome Health Services Research Training Fellow, University of Oxford, Department of Psychiatry; CLIVE ADAMS, MRCPsych, CLAIRE JOY, BSc, Research Assistant, MARK FENTON, MA, Research Nurse, Cochrane Schizophrenia Group, Summertown Pavilion, Middleton Way, Oxford

Correspondence: Dr Max Marshall, University of Manchester, Department of Community Psychiatry, Guild Academic Centre, Royal Preston Hospital, Sharoe Green Lane, Preston PR2 9HT

(First received 20 April 1999, final revision 25 August 1999, accepted 1 September 1999)