Reporting of outcomes in arthritis trials measured on ordinal and interval scales is inadequate in relation to meta-analysis

P C Gøtzsche

Abstract

Objectives—To study whether the reporting of clinical outcomes in arthritis trials measured on ordinal and interval scales is adequate in relation to meta-analysis.

Methods—Systematic review of randomised trials of non-steroidal anti-inflammatory drugs in patients with rheumatoid arthritis. Optimal reporting was defined as data in the original ordered categories for global evaluation and pain, and as mean and SD for number of tender joints and grip strength, and if a visual analogue scale had been used to measure pain.

Results—A total of 144 trials were included. The median sample size was 60 patients. The quality of the reporting increased over time for three of the four variables. Global evaluation was optimally reported in 52 of the 127 trials (41%) in which it was recorded. Pain was optimally reported in 27 of 98 trials (28%), number of tender joints in 41 of 123 trials (33%), and grip strength in 34 of 124 trials (27%). Even if rather broad criteria are adopted, only about half of the data were reported in a potentially useful way for a meta-analysis.

Conclusions—Arthritis trials have been reported inadequately in relation to meta-analysis. As most trials are underpowered, meta-analysis is indispensable and the deficit therefore needs urgent improvement. Investigators should specify a priori what constitutes an important treatment effect and report numbers of patients improved.

(Ann Rheum Dis 2001;60:349–352)

It is an important challenge to the medical community how best to extract and combine the information that is contained in reports of randomised clinical trials. The total number of trials has proved to be much larger than was expected just a few years ago. In 1994, only 19 000 citations on Medline were identifiable as randomised or controlled clinical trials.¹ Largely because of the systematic electronic and hand searches performed by members of the Cochrane Collaboration, this number has increased to more than 180 000, and the most comprehensive source of citations to trials, The Cochrane Controlled Trials Register, currently contains more than 250 000 references.²

As most trials are too small to answer the question they address,^{3 4} it is usually necessary

to aggregate the results in meta-analyses. In most trials, the primary outcomes have been measured on an ordinal scale with graded classes—for example, severity of pain—or on an interval scale—for example, blood pressure.⁴ This presents a challenge for meta-analysts, because the reporting of such data is often poor.⁵ It may therefore not be possible to use standard methods for meta-analysis which require that numbers of patients in ordered or binary categories or means and standard deviations are reported.

To make the most of the available information for clinical decision making and research planning, it may be necessary to use more crude, less ideal meta-analytical methods. As a starting point for this work, it is useful to know what the main problems are. I studied the reporting of data measured on ordinal and interval scales in a systematic review of trials of non-steroidal anti-inflammatory drugs (NSAIDs) in patients with rheumatoid arthritis.

Methods

SEARCH STRATEGY

I collected reports on group comparative randomised double blind trials that compared two of the 22 NSAIDs currently marketed in Denmark. The trials were of tablets or capsules, given in repeated doses to patients with rheumatoid arthritis, and published before 1998. I excluded trials on several diseases if the outcomes were not reported for rheumatoid arthritis separately, and trials published as abstracts. A Medline search was carried out in August 1998. “Arthritis, rheumatoid” (exploded) was combined with (review in publication type or comparative-study in check tag or “dose-response relationship, drug” (exploded)) and with one or more of the drugs as text words, apart from aspirin which was exploded. I scanned the reference lists of the collected articles. In the case of repetitive publications, the most informative was chosen. The searches identified 156 potentially eligible trials. I excluded five trials in Japanese, one in Russian and one in Serbo-Croat because of the language. Another five trials were excluded as the reports could not be obtained via the university library. The final sample consisted of 144 trials in which one or more of the four selected outcomes mentioned below had been measured (references are available on request).

DATA EXTRACTION

Data were extracted on year of publication, number of patients, and on the following clinical outcomes: global evaluation of the effect,

The Nordic Cochrane Centre, Rigshospitalet, Department 7112, Blegdamsvej 9, DK-2100 Copenhagen Ø, Denmark
P C Gøtzsche

Correspondence to:
Dr Gøtzsche
p.c.gotzsche@cochrane.dk

Accepted 12 September
2000

Table 1 Proportion of trials with optimal and usable reporting for meta-analysis (see definitions in text). Test for positive trend

	1966–77	1978–82	1983–87	1988–97	Trend (p value)
Global evaluation					
No of trials	35	31	33	28	
Optimal reporting (%)	43	52	33	36	0.32
Usable reporting (%)	71	81	67	57	0.15
Pain					
No of trials	16	27	25	30	
Optimal reporting (%)	13	15	28	47	0.003
Usable reporting (%)	50	30	48	63	0.09
No of tender joints					
No of trials	31	30	31	31	
Optimal reporting (%)	19	20	39	55	0.001
Usable reporting (%)	29	27	45	58	0.007
Grip strength					
No of trials	40	28	30	25	
Optimal reporting (%)	18	25	33	40	0.03
Usable reporting (%)	25	29	40	44	0.07

pain, number of tender joints, and grip strength. These four outcomes are commonly used in NSAID trials³ and they are measured on an ordinal scale or on an interval scale.

I noted whether the reported data could be used in a standard meta-analysis and whether the reporting improved over time. To allow time trend analyses, I separated the trials into four groups of approximately equal size.

To be optimally reported, data on global evaluation should be available in the original ordered categories. Thus, if a five point scale had been used but only numbers of patients who had improved were reported, this outcome was not optimally reported. Grip strength should be reported as mean after treatment or mean change, and standard deviations should be available for absolute values or for changes, as appropriate, or should be estimable—for example, from exact p or t values. The same applied to number of tender joints. Pain data should be available in the original ordered categories; if a visual analogue scale had been used, mean and SD were also accepted.

In a secondary analysis, I noted whether the data were potentially usable for a meta-analysis in a rather broad sense. I only required that the data were reported in two or more ordered categories, or as mean or mean change with corresponding SD.

The Jonckheere-Terpstra test was used for time trend analysis of sample size (MEDSTAT program); proportions were analysed with a χ^2 test for trend (BMDP program).

Results

The 144 trials included comprised a total of 16 659 patients. The median sample size was 60 patients; it increased over time (p=0.01) and was 110 in the most recent time period, 1988–1997.

Table 1 shows the usefulness of the reported outcomes for meta-analysis. Global evaluation was recorded in 127 trials. It was optimally reported in 52 trials (41%) and usefully reported in 88 trials (69%). There was a non-significant trend towards poorer reporting in the most recent years (p=0.32 for optimal reporting and p=0.15 for useful reporting).

Pain was recorded in 98 trials. It was optimally reported in 27 trials (28%) and usefully reported in 47 trials (48%). There was a significant trend towards better reporting over time (p=0.003 and p=0.09 respectively).

Number of tender joints, or closely similar measures such as Ritchie’s joint count index, was recorded in 123 trials. It was optimally reported in 41 trials (33%) and usefully reported in 49 trials (40%). There was a significant trend towards better reporting over time (p=0.001 and p=0.007 respectively).

Grip strength was recorded in 124 trials. It was optimally reported in 34 trials (27%) and usefully reported in 41 trials (33%). There was a significant trend towards better reporting over time (p=0.03 and p=0.07 respectively).

Table 2 shows the mode of reporting in detail. The most common problem was the lack of standard deviations of the measurements. Confidence intervals, standard errors, exact p or t values, or other relevant test statistics that would have allowed the calculation of standard deviations, were also missing. Twenty out of 127 reports (16%) provided no data at all on global evaluation; the same applied for 21 out of 123 (18%) reports on pain.

Table 2 Method of data presentation in 144 reports of trials comparing two non-steroidal anti-inflammatory drugs

Method of data presentation	Global evaluation	Pain	Number of tender joints	Grip strength
Usable reporting				
Numbers of patients in original ordered categories	52	7		
Numbers in reduced ordered categories, at least 3	12	5	4	3
Numbers in only 2 (reduced) ordered categories	14	5	4	4
Mean and SD, before and after treatment	4	26	35	29
Mean and SD before, mean after treatment	1	1	2	
Mean and SD after treatment	4	1	1	
Mean change and SD for change	1	2	3	5
Unusable reporting				
Mean before and after treatment, no SD	8	26	46	44
Mean after treatment, no SD	6		1	
Mean change, no SD	5	6	9	9
Percentage change, no SD		1	4	2
Median and IQ* range, before and after treatment			1	2
Median change and IQ range				1
Geometric mean percentage change, no SD		3	4	1
Median change, no percentiles or SD		2	2	2
Median before and after treatment, no centiles			1	
No data shown	20	13	6	21
Total number of trials reporting the variable	127	98	123	123

*IQ = Interquartile.

Discussion

The reporting quality improved over time, but even in the most recent time period, and even if rather broad criteria were adopted, only about half of the data were reported in a way that made them potentially useful for a meta-analysis. This deficit in reporting was similar in small and large trials. For pain, for example, 48% of the trial reports were useful, and if the unit of analysis was patients, this proportion was very similar, 49%.

It was surprising that not a single confidence interval was presented in any of the 34 most recent trial reports, as all the reviewed trials aimed to compare the effect and tolerability of two similar drugs. Such trials are pragmatic, as the questions they address are directly relevant for clinical decision making. The 95% confidence interval allows the reader to judge whether a clinically relevant difference can be ruled out with reasonable certainty. Without confidence intervals, the non-statistically minded clinician may be misled into believing curious claims such as the new drug is "at least as effective"⁵ as the standard drug. As such claims are often accompanied by doubtful or biased statements about less toxicity,⁵ these trials seem primarily to have served marketing purposes.

Most of the trials were considerably underpowered, even if the sample size estimation is based on the patient's global evaluation of the effect, which is the most sensitive outcome.⁶ If, in accordance with the examples above, one wishes not to overlook the possibility that a new drug may be only half as effective as the standard drug, the necessary sample size is 290 patients for a power of 80%.^{6,7} As NSAIDs are effective drugs,⁶ a decrease in effect of 25% could also be clinically relevant to detect, in which case 1160 patients would be needed.

When sample sizes are too small, it is particularly important that the results are reported in a way that facilitates meta-analysis. Even though small trials tend to exaggerate the treatment effect,⁸ they can sometimes be useful for meta-analysis if the aggregated result is interpreted cautiously and the likely magnitude and direction of bias is taken into account. For example, it has been shown in meta-analyses of small trials with ordinal and interval scale outcomes that low dose prednisolone is highly effective in rheumatoid arthritis⁹ and, conversely, that chemical and physical measures to decrease exposure to house dust mites are ineffective in patients with asthma.¹⁰

For a symptomatic treatment such as an NSAID, the global evaluation of the effect is an important outcome because it allows the patient to express the various benefits and adverse effects of the drug in a simple measure. As this outcome is almost always recorded on a four or five point ranking scale, it was surprising that the data were reported in the original ordered categories in only 52 (41%) of the trials (table 2). In 26 trials (20%), the data were reduced to a scale with only two or three categories before reporting, which creates a risk of bias, as the decision to choose new categories and cut offs between them may not be

independent of which drug will be favoured. It is also problematic for the validity of a meta-analysis that no data were shown for global evaluation in as many as 16% of the trials, as one would not expect such omissions to occur when the result favours the new drug over the standard.⁵

These problems could be minimised if research ethics committees ensured that raw trial data became publicly available after a period of time that allows the investigators to publish their results. By making unpublished data available, publication bias could also be avoided. As long as this has not happened, however, it remains an important challenge for meta-analysts to determine how to take stock of all the imperfectly reported results. For example, trials could be weighted by their sample size rather than by their inverse variance,¹¹ and missing standard deviations could be estimated from similar trials. It is not the purpose of this article, however, to explore these possibilities.

Previous surveys of statistical reporting of randomised trials seem not to have addressed the objective of this survey—that is, whether the reporting allows the data to be used for meta-analysis. It is a limitation of the survey that only one drug category was included. On the other hand, it is likely that the results can be generalised. Most clinical trials are drug trials supported by industry and many of the largest companies have developed and marketed NSAIDs. A similar reporting pattern would therefore be expected for many other trials sponsored by drug companies. Furthermore, the types of outcomes I have reviewed are very common; in fact, most trials in health care have a primary outcome that has been measured on an ordinal or an interval scale.⁴

It can therefore be assumed that a large body of trials of healthcare interventions have been reported inadequately in relation to meta-analysis. This deficit needs urgent improvement. It would also be helpful if investigators always specified a priori what constituted an important treatment effect and reported numbers of patients who improved. This would lead to a minor decrease in statistical power which seems acceptable, in particular as there is an important gain in clinical relevance. It is difficult to know whether an improvement of 1.5 cm on a 10 cm pain scale or some points on a composite score matters, whereas it is obviously relevant if a severe pain or depression has gone.

The study was funded by the Danish Medical Research Council.

- 1 Lefebvre C, Clarke M. Identifying randomised trials for systematic reviews. In: Egger M, Davey Smith G, Altman DG, eds. *Systematic reviews*. 2nd ed. London: BMJ Publications (in press).
- 2 The Cochrane Library. Oxford: Update Software. Updated quarterly Issue 2, 2000.
- 3 Freiman JA, Chalmers TC, Smith H, Kuebler RR. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial: survey of 71 "negative" trials. *N Engl J Med* 1978;299: 690-4.
- 4 Mulward S, Gøtzsche PC. Sample size of randomized double-blind trials 1976-1991. *Dan Med Bull* 1996;43: 96-8.
- 5 Gøtzsche PC. Methodology and overt and hidden bias in reports of 196 double-blind trials of nonsteroidal, anti-inflammatory drugs in rheumatoid arthritis. *Control Clin Trials* 1989;10:31-56.

6 Gotzsche PC. Sensitivity of effect variables in rheumatoid arthritis: a meta-analysis of 130 placebo controlled NSAID trials. *J Clin Epidemiol* 1990;43:1313-18.

7 Pocock SJ. *Clinical trials: a practical approach*. Chichester: John Wiley & Sons, 1983.

8 Egger M, Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *BMJ* 1997;315:629-34.

9 Gotzsche PC, Johansen HK. Meta-analysis of short term low dose prednisolone versus placebo and non-steroidal, anti-inflammatory drugs in rheumatoid arthritis. *BMJ* 1998;316:811-18. Also published in The Cochrane Library, Issue 3, 2000. Oxford: Update Software.

10 Gotzsche PC, Hammarquist C, Burr M. House dust mite control measures in the management of asthma: meta-analysis. *BMJ* 1998;317:1105-10. Also published in The Cochrane Library, Issue 3, 2000. Oxford: Update Software.

11 Laird NM, Mosteller F. Some statistical methods for combining experimental results. *Int J Technol Assess Health Care* 1990;6:5-30.