An introduction to systematic reviews in respiratory medicine

P.W. Jones

Division of Physiological Medicine, St George’s Hospital Medical School, London, UK

Abstract

Reviews are an essential component of modern medicine. The volume of literature is large, even about a single treatment for pulmonary disease. The task of retrieving all the relevant papers, then assessing the evidence to reach a valid conclusion is very time consuming. At every stage there is the risk of sampling error (failure to get all the evidence) and bias (a systematic distortion of the results due to a weakness in the methodology). There are essentially two types of review: narrative reviews that follow no rules, exposing them to sampling error and bias; and systematic reviews that attempt to minimise these effects by following an explicit structure for retrieving all of the evidence and attempting an objective synthesis of the results from the different trials. A good review can serve a number of purposes including: assembling all the relevant evidence in one place, providing a valid estimate of the overall effect of treatment, producing guidance for clinical practice and generating hypotheses for further trials about patients or settings in which the treatment effect may be less or more effective.

INTRODUCTION

This journal has created a new section that will contain reviews and overviews of topics in respiratory medicine, usually but not exclusively concerning treatment. This paper is designed to serve as an introduction and guide.

Reviews of treatment are needed. Most of us scarcely have time to read the last sentence of the conclusion in the abstract of a paper, never mind critically appraise its content. There is a large and ever-growing literature concerning therapy for pulmonary disease. There is usually a number of relevant studies for a single treatment and it is not possible for an individual clinician to keep up with all developments. Not only is the volume of published work large, it is also complex. Clinical trials have different designs, inclusion and exclusion criteria and often use different measures of treatment efficacy. Readers of papers that describe clinical trials require not only the time to devote to the enterprise, but also the appropriate critical appraisal skills needed to assess with confidence the quality of the data presented, the strength of the evidence and the validity of the conclusions drawn by the paper’s authors.

Even the task of finding the relevant papers is not as simple as it first appears. We are all familiar with carrying out electronic searches for papers either by ourselves or with the help of a medical librarian, and we all know how over-inclusive such searches can be. Put they often turn up some papers that we need and many that we do not. To make life easier, we are often tempted to limit our searches using all sorts of devices—searching only the last 2 years, only English language papers, etc. One useful device for restricting the search is use of the ‘AND’ operator in the search term. This can increase the specificity of the search but has the potential to sacrifice sensitivity for specificity, so must be used with care.

Having found all the potentially relevant papers it is then necessary to screen out those that are not appropriate. Not all papers that we want will be readily available in the local library or electronically and may be quite expensive to obtain through inter-library loan services, so one only reads those that are most accessible.

Having obtained the papers and found time to read them, it will be necessary to synthesise all that knowledge into an assessment of the value of the treatment. That may be easier said than done. Clinical trials
differ in many ways (Table 1). Somehow the reader has to take into consideration, different pieces of information that suggest “On the one hand this . . . . and on the other that . . . .” It may often be necessary to tabulate findings from the various papers to enable a reasonable overview of the results from the different studies. Quite clearly, if a clinician is interested in getting a balance view of the overall efficacy of even a single treatment, this will involve a lot of work. As a result, in a busy world, most clinicians’ considerations of a treatment’s efficacy will be based on a review of some form.

**SAMPLING ERRORS AND BIAS**

All scientific assessments are prone to sampling error and biases that may result in an incorrect conclusion. Reading (reviewing) the literature is subject to these factors in the same way as the measurement of the basic observations upon which the publications are based. Sampling errors may be random, often due to an inadequate number of observations being made, or in the case of reviews, papers that are read. These errors may also be non-random, in which case the term ‘bias’ is used. Issues concerning bias will appear recurrently in this paper. Bias is defined along these lines: ‘a systematic distortion of a statistical result due to a variable that has not been considered or accounted for in its derivation’. In the context of obtaining results from clinical trials, there are many potential sources of bias, apart from sampling error due to attempts to reduce the work of reviewing a topic to a manageable load.

Systematic biases include publication bias. A well-known example being trials with negative results that are being published in abstract form only or in non-English language journals (1,2). A large study of which I was a principal investigator was almost rejected by a reviewer for a major respiratory journal because it showed that the treatment was only modestly effective—so was judged to be ‘uninteresting’. Journals need high citation ratings—negative results are cited less often than positive ones. Other biases may creep in with the construction of the search routine and most especially with retrieval. For example, pharmaceutical companies can often supply to clinicians reprints of studies that report favourably upon their product, negative trails are not so easy to come by. Within the Cochrane Airways Group, staff at the Editorial Base of the Airways Review Group spend a long time at the British Library chasing down clinical trials published in many different journals that are not available, even in a large medical school library.

**AVOIDING BIAS IN REVIEWS**

It is clear from the foregoing that, to obtain the least-biased estimate of the efficacy of a treatment, a review must be created using a very systematic approach. This requires a set of rules that are similar, in many respects, to those used for randomised controlled clinical trials (3).

Reviews that follow this structured approach are termed Systematic Reviews. This term is not synonymous with the term ‘meta-analysis’ which is a group of specific statistical techniques used to aggregate results from different trials. Whilst it may not be possible always to aggregate results using statistics, this does not negate the use of a systematic approach for the rest of the review process. The alternative to a systematic review is the so-called ‘Narrative Review’. Such reviews are more common than Systematic Reviews, since they do not have to use a rigorous and time-consuming methodology. However, they are open to many potential sources of bias and may reach erroneous conclusions, not only because they may not have identified all of the relevant trials, but also because of a flawed analysis of the trials.
that were retrieved. The following sections describe some of the types of bias that may occur in any review, whether systematic or narrative, but are more likely to occur with the latter.

WHICH TYPE OF TRIAL TO INCLUDE IN A REVIEW?

The randomised controlled double-blind trial (RCT) design forms the cornerstone of the evaluation of treatment effects in lung disease. The reasons for this are very simple. New treatments must be compared with another treatment to know whether it has any beneficial effect. The baseline state can be used for this purpose, but this is fraught with problems. Not least is the study effect; in which patients included in a clinical trial improve spontaneously. This is not a 'placebo' effect, it is due to patient factors such as improved adherence or use of usual therapy, and physician factors such as better clinical supervision and faster response to changes in state. Thus, it is necessary to have a contemporaneous control group who have 'usual care' or another comparator treatment. Randomisation is used in an attempt to ensure that the two treatment groups are matched at baseline. There are always differences between patients and usually the most secure way of ensuring that each limb of the study contains patients with same characteristics is to allocate them randomly. The success of this will depend on chance and the size of the study. The bigger the patient population, the lower the chance of random sampling errors.

The third factor in the RCT is the double-blind component. This is needed to ensure that neither patient nor trialist knows which treatment is being given. The ethics behind this are that at the outset of the study the trialist should be in 'equipoise'—i.e. they do not know which limb of the study will have the better outcome. There are too many important issues around this to be discussed here, apart from the fact that if the trial is ethical, then it is probably unethical not to conceal the treatment allocation. The reason being that the study's purpose is to produce the most reliable estimate of the efficacy of the treatment under test. There is good evidence that lack of treatment allocation can lead to a large over-estimate of the treatment's efficacy. The causes of this are multi-factorial, but include treatment bias in which the clinician and patient make other treatment decisions in the knowledge of which trial therapy they are receiving, and bias in the measurement of treatment outcome.

The RCT should not be thought of as being the 'best' trial, but rather the trial design in which biases are minimised to the greatest degree. Clearly not all of the requirements are achievable in all trials. For example, sham lung-reduction surgery (thoracotomy with no surgery to the lungs) is not possible; nor is it possible to prevent a patient knowing that they have received physical rehabilitation for COPD, but this should not prevent the use of design that will minimise identifiable sources of bias. It is for reasons of minimising bias, the Cochrane Collaboration has focussed its efforts upon reviewing RCTs or at the least trials that are adequately controlled.

ASKING THE RIGHT QUESTION IN THE REVIEW

The first headings in Table 2 is ‘Define the question’. This is the starting point for the review (4). In fact, this breaks down into three components:

- Which treatment is being assessed?
- In whom?
- In what setting?

It may seem surprising that it is necessary to specify the treatment, since that ought to be apparent from the title. However, take for example a narrative review entitled “Beta-adrenergic bronchodilators” (5). This could include short- and long-acting drugs, acute or stable chronic asthma, and several different modes of drug delivery. Contrast this with “Nebulised beta2-agonists for the treatment of acute asthma in the emergency department” (6). Note that this title specifies the reason for the treatment being given and the setting. The latter is important. There is still no agreed method of assessing the severity of acute asthma, but on examining trials in acute severe asthma, clear differences in admission rate are apparent between studies of patients reported to have severe acute asthma. Thus, it is important to specify whether the review covers treatment for acute asthma in primary care, the emergency room or in patients admitted to hospital.

The question “In whom?” addresses the inclusion and exclusion criteria for the review. This defines the limits of generalisability of the conclusions drawn from the review. For example, the NHLBI/WHO Global Initiative on Obstructive Lung Disease (GOLD) defines COPD as

<table>
<thead>
<tr>
<th>Table 2. The basic components of a systematic review</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Define the question to be addressed</td>
</tr>
<tr>
<td>- Obtain all the randomised controlled trial data</td>
</tr>
<tr>
<td>- Use explicit inclusion and exclusion criteria</td>
</tr>
<tr>
<td>- Extract data carefully (obtaining data from 17 authors)</td>
</tr>
<tr>
<td>- Identify subgroups (a priori)</td>
</tr>
<tr>
<td>- Assemble all the data</td>
</tr>
<tr>
<td>- Aggregate the data using meta-analysis (if possible)</td>
</tr>
<tr>
<td>- Test for heterogeneity (ie differences between trials)</td>
</tr>
<tr>
<td>- Draw valid conclusions</td>
</tr>
<tr>
<td>- Distinguish between</td>
</tr>
<tr>
<td>- No evidence of effect</td>
</tr>
<tr>
<td>- Evidence of no effect</td>
</tr>
</tbody>
</table>
a disease characterised by airflow limitation that is not
fully reversible (7). A degree of reversibility is usually
present, however, and some patients with COPD have a
clearly reversible element to their condition. This inter-
national definition of COPD is relatively new, but some
trialists and authors of reviews use a tighter criterion
for the definition, one that specifies that the patient's
FEV₁ should not reverse more than a certain amount.
Results from studies/reviews with such criteria may not
be generalisable to COPD patients identified using a
broader definition.

SELECTING THE PRIMARY MEASURE
OF TREATMENT EFFECTIVENESS

In a clinical trial carried out to provide evidence for the
registration of a new drug, the regulatory authorities re-
quire the trialists to specify their primary endpoint, i.e.
the outcome on which they are going to base their claim
for the efficacy of the drug. The reason for this is that
there are many well-established measures of treatment
outcome and most clinical trials incorporate a number of
these. The disadvantage is that, by random chance, one
or more of these will show an effect in favour of one or
other treatment, so it is necessary to specify the pri-
mary outcome.

The same discipline should apply to reviews, since the
reviewer will have a large number of possible outcomes
to choose from—well over 20 in some cases. There is
currently no agreement as to which is the most impor-
tant outcome in each of the major respiratory diseases.
Furthermore, there are different treatment objectives
for each disease. For example, in COPD these include re-
lief of symptoms, reduction in mortality and slowing dis-
ease progression. Thus, the outcome should match the
purpose of the study—which means that to three com-
ponent question listed above (Which treatment, in
Whom and Where ?) should be added “To what end or
Why ?”

THE ‘AVERAGE’ RESULT

Clinical trials are carried out in groups of patients re-
cruited because they are believed to be representative
of the population for whom the treatment is intended.
However, the trial population is not homogenous, it will
include patients with different levels of disease severity.
Furthermore within the population, there may be differ-
ent (as yet unidentified) genetic factors that will influ-
ence the response to therapy. Results of the trial are
given as a mean together with a measure of the distribu-
tion around that mean (a standard error, standard devia-
tion or confidence intervals). But who remembers the
distribution statistic? Only the mean result is recalled
and used as an example of the treatment’s effect.

For some reason this issue is seen to be a particular
problem with meta-analyses. The pejorative phrase
“combining apples and oranges” is often used in the con-
text of such analyses, but rarely when the original clinical
trial is considered, even though the inclusion criteria for
the latter are usually wide. It is often implied that meta-
analysis hides important treatment effects by averaging,
whilst it is forgotten that a clinical trial can have exactly
the same effect. Figure 1 illustrates different ways of illus-
trating the distribution statistics around the mean im-
provement in FEV₁ in patients treated with a long-acting
beta2-agonist for COPD (8). The difference between the
maximum and minimum actually measured is much
greater than one would imagine from looking at the
standard deviation or the standard error.

Two main factors lie behind these concerns about the
possibility that meta-analysis may hide a group of ‘good’
responders. One is the recognition that clinicians treat
individuals not groups of patients or even the ‘average’ pa-
tient. The second is the belief that a cherished treatment
works well in some patients. Of course it does, Fig. 1
shows that some patients have an apparently large effect
(which, of course includes a true biological effect, sponta-
nous variation and measurement error). Such big ef-
fects stick in the mind of clinicians, not unreasonably.
Figure 1 also shows some quite large negative effects,
but unless these are positively harmful, the clinician will
ignore these in practice and just conclude that the treat-
ment did not work in that particular individual.

It would be inappropriate to leave this section without
a further word about meta-analysis. This approach has
two main purposes. First, it can provide an average esti-
mate of the efficacy of a treatment across a number of
trials (the ‘on average’ effect). Second, it can increase the power of the analysis by combining results from multiple small trials. This illustrated by the meta-analysis in Fig. 2, which shows that two trials of written self-management plans coupled with regular clinician review showed a statistically significant effect, whilst two did not (9). Taking the results from all four trials overall, there is a significant effect. It would also be appropriate to warn that meta-analysis can also under-estimate the level of statistical significance from an individual trial’s results. In its most widely used form, a meta-analysis is related to the simple unpaired t-test, but weighted to take into account the size of the study and the amount of variation around the mean result. The unpaired test is a weaker test than a paired test because it contains both the variation within a patient (i.e. the treatment effect) and baseline variations between patients. These baseline differences are effectively ‘noise’ if one is only interested in the change with treatment. Paired analyses can remove the effects of between-patient differences. A similar effect can be achieved by using differences in change from baseline in the two groups as the outcome measure in the meta-analysis, rather than the difference between the two treatments groups at the end of the study. Modern clinical trials often use a more sophisticated form of paired analysis termed analysis of covariance. This can remove even more of the baseline differences between the patients and strengthen the statistical significance of the treatment effect. Unfortunately, it is not always possible to replicate this sophisticated analysis when the reported data from the trial are incorporated into a meta-analysis. This can be confusing and lead to loss of confidence in the meta-analysis since the trial, when analysed in a meta-analysis, may appear to be not statistically significant whereas it was when reported in the primary paper. There are ways around this problem, but only if the primary trialists report the ‘adjusted’ means and standard deviations.

IDENTIFYING POTENTIAL ‘RESPONDERS’

Average treatment effects, whether from a clinical trial or meta-analysis can provide only the starting point for a treatment decision. It is an interesting phenomenon, that immediately on hearing the average result from a study, most clinicians want to know in whom the treatment works best or least. Unfortunately this type of post hoc analysis is open to many challenges, not least the risk of chance findings resulting from multiple analyses of the same data set. The more post hoc analyses that are undertaken, the greater the likelihood of a ‘chance’ effect being mistakenly taken for a ‘real’ effect.

It may be possible to pick up clues to the existence of a subpopulation of patients who respond more than others. This is illustrated in Fig. 3. Panel (A) shows the most usual result from a large trial—a mean with a normal distribution of data around it. The ‘good responders’ are clearly just the tail of the distribution and not separate from the rest of the population. Panel (B) shows a bimodal pattern of distribution in the results which would provide very strong evidence for the existence of a subgroup of responders. This is extremely rare, if it occurs at all. In practice, the strongest hint for such patients may lie in a skewed distribution in which there is an extended ‘tail’ of patients who have an apparently large response (Panel C), however even this has to be interpreted with caution.

As pointed out above, enthusiasts who search for ‘responders’ ignore the ‘deterioraters’. A big response is most likely to be just one that lies in the tail of the normal distribution that lies around any mean effect. Post hoc analyses of this type are justifiable, however, if done with care. Many developments in medicine have resulted from serendipitous chance observations. In the context of a clinical trial results, patterns in the results that suggest ‘good’ or ‘poor’ responders are hypothesis-generating, and will depend on the identification of features that characterise such patients before they are given the treatment. Once this can be done, the whole picture changes because subgroup analysis then become possible.

SUBGROUP ANALYSES

The most important aspect of a subgroup analysis is that it is pre-specified and based upon a biologically plausible difference between patients. For example, children may respond differently to adults; or COPD patients with a poor response to short-acting bronchodilators may have
a similarly weak response to longer acting agents. A good example of subgroup analysis is shown in the meta-analysis in Fig. 4, in which it appears that patients with less severe acute exacerbations of asthma appeared not to respond to intravenous magnesium, whereas more severe patients did (10). This is not the place to discuss possible mechanisms, but this analysis illustrates an example of hypothesis generation and subgroup analysis that would never have been possible if the authors of the review had not attempted to aggregate all the data together into a meta-analysis. When the analysis was first carried out, without subgroups, a significant amount of difference (or heterogeneity) was found in the size of the treatment effect between the different studies. When this was seen, the authors hypothesised that it could be due to differences in severity so they carried out this subgroup analysis. The result should be thought of as being hypothesis generating, rather than clear evidence that patients with less severe attacks of asthma do not need magnesium. However, the admission rate in the control patients in the less severe group was lower than in the more severe, lending support to the hypothesis and encouraging the design of trials to test this.

**SUMMARY**

Systematic reviews bring together and summarise in one place, all the available evidence for a treatment’s efficacy. They should give the reader enough evidence with which to judge the reviewer’s interpretation of the data and the implications for clinical practice. Meta-analysis can identify treatment effects that may be missed by inadequate trial size and occasionally can generate hypotheses concerning patients who may or may not benefit from treatment. The identification of ‘responders’ is not often possible and we are usually left with the ‘on average’ result. This is perhaps the most important point of evidence-based medicine—it provides the starting point for intelligent and thoughtful practice, it is not a straight-jacket of rigid rules.

**REFERENCES**

1. Dickersin K, Min YI, Meinert LL. Factors influencing publication of research: follow up of applications to two institutional review boards. JAMA 1997; 263: 374–378.


