Methods in health services research: Interpreting the evidence: choosing between randomised and non-randomised studies

Martin McKee, Annie Britton, Nick Black, Klim McPherson, Colin Sanderson and Chris Bain

BMJ 1999;319;312-315

Updated information and services can be found at:
http://bmj.com/cgi/content/full/319/7205/312

These include:

Data supplement
“further references”
http://bmj.com/cgi/content/full/319/7205/312/DC1

References
This article cites 21 articles, 8 of which can be accessed free at:
http://bmj.com/cgi/content/full/319/7205/312#BIBL
16 online articles that cite this article can be accessed at:
http://bmj.com/cgi/content/full/319/7205/312#otherarticles

Rapid responses
3 rapid responses have been posted to this article, which you can access for free at:
http://bmj.com/cgi/content/full/319/7205/312#responses
You can respond to this article at:
http://bmj.com/cgi/eletter-submit/319/7205/312

Email alerting service
Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article

Topic collections
Articles on similar topics can be found in the following collections

Systematic reviews (incl meta-analyses): descriptions (129 articles)
Other Statistics and Research Methods: descriptions (568 articles)

Notes

To order reprints of this article go to:
http://www.bmjjournals.com/cgi/reprintform

To subscribe to BMJ go to:
http://bmj.bmjjournals.com/subscriptions/subscribe.shtml
Methods in health services research

Interpreting the evidence: choosing between randomised and non-randomised studies

Martin McKee, Annie Britton, Nick Black, Klim McPherson, Colin Sanderson, Chris Bain

This is the evaluations of healthcare interventions can either randomise subjects to comparison groups, or not. In both designs there are potential threats to validity, which can be external (the extent to which they are generalisable to all potential recipients) or internal (whether differences in observed effects can be attributed to differences in the intervention). Randomisation should ensure that comparison groups of sufficient size differ only in their exposure to the intervention concerned. However, some investigators have argued that randomised controlled trials (RCTs) tend to exclude, consciously or otherwise, some types of patient to whom results will subsequently be applied. Furthermore, in unblinded trials the outcome of treatment may be influenced by practitioners' and patients' preferences for one or other intervention. Though non-randomised studies are less selective in terms of recruitment, they are subject to selection bias in allocation if treatment is related to initial prognosis.

These issues have led to extensive debate, although empirical evidence is limited. This paper is a brief summary of a more detailed review of the impact of these potential threats.

Summary points

Treatment effects obtained from randomised and non-randomised studies may differ, but one method does not give a consistently greater effect than the other.

Treatment effects measured in each type of study best approximate when the exclusion criteria are the same and where potential prognostic factors are well understood and controlled for in the non-randomised studies.

Subjects excluded from randomised controlled trials tend to have a worse prognosis than those included, and this limits generalisability.

Subjects participating in randomised controlled trials evaluating treatment of existing conditions tend to be less affluent, educated, and healthy than those who do not; the opposite is true for trials of preventive interventions.
Intervention was evaluated by both methods (a full list
Eighteen papers were identified where a single
studies
Comparing results of RCTs and non-randomised
outcome?
ment they would not choose, how does this affect their
intervention and, if patients are randomised to a treat-
x
not allocated randomly?
unknown baseline differences between groups that are
x
different measurements of treatment effect from
x
RCTs?

Findings

Nature of the evidence

The review focused on threats to internal and external
validity of evaluations of effectiveness and on the
strategies proposed to overcome them (table). Various
factors act through their effect on the distribution of
the potential to benefit among different groups. This
can be illustrated schematically (fig 1). The reference
population is defined by an envelope, represented here
as a triangle but potentially taking many shapes. At
some point, a threshold is reached, below which the
overall risks outweigh the benefits. As patients are
excluded or do not participate, the study population
becomes a progressively smaller subset of the
reference population, in principle increasing the scope
for selection bias and raising the question of whether it
is valid to apply the results obtained to the reference
population.

We used systematic reviews to explore the potential
and actual importance of factors that lead to selective
recruitment, examining four questions:

● Do non-randomised studies give systematically
different measurements of treatment effect from
RCTs?

● Are there systematic differences between the
subjects included in or excluded from studies, and do
these influence the measured treatment effect?

● To what extent is it possible to overcome known or
unknown baseline differences between groups that are
not allocated randomly?

● How important are patients’ preferences for an
intervention and, if patients are randomised to a treat-
ment they would not choose, how does this affect their
outcome?

Findings

Comparing results of RCTs and non-randomised
studies

Eighteen papers were identified where a single
intervention was evaluated by both methods (a full list
is available on the BMJ’s website). A review was
published just after our original report; on the basis of
eight comparisons it found that, on average, non-
randomised studies overestimate effect size. In
contrast, of the seven studies in our review where the
two methods detected effects in the same direction, in
three the effect size was greater in the randomised trial
and in four it was greater in the non-randomised study.
The key finding in our study is that neither method
consistently gave larger estimates of treatment effect.

In addition to chance, there are several potential
explanations for different measurements of treatment
effects. The overall impact will reflect the relative
importance of each issue in a particular case. A
randomised controlled trial may produce a greater
effect if the patients enrolled in it receive higher quality
care or are selected so that they have greater capacity
to benefit than patients in non-randomised studies. But
it may produce a lower estimate of treatment effect for
several reasons:

● In non-randomised studies, patients tend to be allo-
cated to treatments that are correctly considered most
appropriate for their individual circumstances;

● Exclusions from a RCT create a sample with less
capacity to benefit than in a non-randomised study;

● An unblinded RCT fails to capture patients with
strong preferences for a particular treatment who show
an enhanced response to treatment;

● Non-randomised studies of preventive interven-
tions include disproportionate numbers of individuals
who, by virtue of their health related behaviour, have
greater capacity to benefit;

● Publication bias leads to negative results being less
likely to be published from non-randomised studies
than from RCTs.

The limited evidence indicates that the results of
non-randomised studies best approximate to results of
RCTs when both use the same exclusion criteria and
when potential prognostic factors are well understood,
measured, and appropriately controlled in non-
randomised studies.7

In summary, the results of RCTs and non-
randomised studies of similar patients may not, after
adjustment, be substantially different in relative effect
size. Any variations are often no greater than those
between different RCTs or among non-randomised
studies. Differences in effect sizes could be due to
chance or differences in populations studied, timing, or
nature of the intervention.

Exclusions

Randomised controlled trials vary widely in their
inclusiveness. Medical reasons cited for exclusion from
trials include a high risk of adverse effects and belief

---

**Table: Threats to validity of evaluative research and possible solutions**

<table>
<thead>
<tr>
<th>Type of validity</th>
<th>Threatening factor</th>
<th>Proposed solution</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Internal</strong></td>
<td>Allocation bias (risk of confounding)</td>
<td>Randomisation</td>
</tr>
<tr>
<td></td>
<td>Patient preference</td>
<td>Preference arms or adjustment for preference (design)</td>
</tr>
<tr>
<td><strong>External</strong></td>
<td>Exclusions (eligibility criteria)</td>
<td>Expand inclusion criteria</td>
</tr>
<tr>
<td></td>
<td>Non-participation (centres/practitioners)</td>
<td>Multicentre, pragmatic design</td>
</tr>
<tr>
<td></td>
<td>Not invited (practitioner preference or administrative oversight)</td>
<td>Encourage practitioners to invite all eligible patients</td>
</tr>
<tr>
<td></td>
<td>Non-participation (patients)</td>
<td>Less rigorous consent procedures</td>
</tr>
</tbody>
</table>

---

**Fig 1** Differences in inclusion and participation. Shaded areas represent the study population.
that benefit, or lack of it, has already been established for some groups.

Scientific reasons include greater precision in estimating treatment effects by having a homogeneous sample, and reduced risk of bias by excluding individuals most likely to be lost to follow up. In addition, many RCTs have blanket exclusions, the reasons for which are often unstated, of categories of patients such as the elderly, women, and ethnic minorities.

Few studies have examined differences in prognostic factors between included and excluded patients, but some have used clinical databases to examine this. The patients included in such databases tended to have a poorer prognosis than those in trials; in one study, a subset selected to meet eligibility criteria of RCTs produced treatment effects of similar size to those obtained from RCTs.

**Participation**

Evaluative research is undertaken predominantly in university or teaching centres, but non-randomised studies are more likely than RCTs to include non-teaching centres, and criteria for participation in RCTs may include the achievement of a specified level of clinical outcome. The available evidence suggests that this may exaggerate the measured treatment effect.

Most evaluative studies fail to document adequately the characteristics of eligible patients who do not participate. The effect of non-participation differs between RCTs that evaluate interventions designed to treat an existing condition and those directed at preventing disease (fig 2).

In contrast, participants in RCTs evaluating preventive interventions tend to be more affluent, better educated, and more severely ill than eligible patients who do not participate. In contrast, participants in RCTs evaluating preventive interventions tend to be more affluent, better educated, and more likely to have adopted a healthy lifestyle than those who decline. On the basis of the evidence from the comparisons discussed earlier, it is plausible that low participation in RCTs of treatment may exaggerate treatment effects by including more skilful practitioners and subjects with a greater capacity to benefit, while RCTs of prevention may underestimate effects as participants have selectively less capacity to benefit.

**Impact of patients’ preferences**

There is little empirical research on the impact on outcome of patients’ preferences. The four studies that attempted to measure preference effects either were small or have yet to report full results. In theory, preference could have an important impact on results of RCTs, especially where the true effect is small. Such effects could account for some observed differences between results of RCTs and non-randomised studies. There are methods that may detect preference effects reliably; though these may contribute to understanding this phenomenon, none provides a complete answer. This is mainly because randomisation between preferring a treatment and not is impossible, and confounding may bias any observed comparison.

**Adjustment for baseline differences in non-randomised studies**

Despite the evidence that the results of RCTs and non-randomised studies are often similar, differences in baseline prognostic factors clearly can be important. Absence of randomisation can produce groups that differ in important ways, and it is necessary to consider whether it is possible to adjust for such differences. Adjustment for imbalance in baseline prognostic factors between arms of non-randomised studies commonly changes the size of the measured treatment effect, but such changes are often small and inconsistent.

Overall, the limited evidence suggests that differences in the populations studied by RCTs and non-randomised studies are likely to be of at least as much importance in explaining any differences and that the two methods should be compared only after patients not meeting eligibility criteria for the RCT are excluded.

**Recommendations**

A large, inclusive, fully blinded RCT incorporating appropriate subgroup analysis is likely to provide the best possible evidence of effectiveness, but there will always be circumstances in which randomisation, especially on an inclusive basis, is unethical or impractical. In circumstances where there are genuine reasons for not randomising, non-randomised studies can provide useful evidence. In such studies, adjustment for baseline imbalances should always be attempted, as rigorously and extensively as possible, and the procedures should be reported explicitly to help readers’ evaluations. However, adjustment cannot be relied on to approximate the prognostic balance of randomisation, given unknown or unmeasurable confounding.

Investigators conducting evaluative research (using any design) must improve the quality of reporting. Authors should define the population to whom they expect their results to be applied; what has been done to ensure that the study population is representative of this wider population, and any evidence of how it differs; whether centres that participated differ from those that declined; and the numbers and characteristics of patients eligible to be included who either were not invited to do so or were invited and declined.

The findings of such studies have implications for the way in which evidence is interpreted. When faced with data from any source, whether randomisation has been used or not, it is important first to pursue alternative (non-causal) explanations thoroughly and examine the possible influence of chance, bias, and confounding, perhaps using sensitivity analyses where feasible.
Where only non-randomised data are available, the potential for allocation bias should be considered and any attempts at risk adjustment should be assessed.

Where only randomised trials are found, preference effects should also be considered. To obtain an uncontaminated estimate of the physiological effect of a treatment, RCTs should be blind to everyone involved, but for many interventions this will be impossible. Also, the advantages of narrowing inclusion criteria to ensure high participation in RCTs should be balanced by the potential need for subgroup analysis. It should not be assumed that a summary result applies to all potential patients.

When both randomised and non-randomised studies have been conducted it is important to ascertain whether estimates of treatment effect are consistent for patients at similar risk across studies. If so, it may be reasonable to accept the results of the less exclusive non-randomised studies. Differences in results cannot be assumed to be solely due to the presence or lack of randomisation—differences in study populations, characteristics of the intervention, and the effects of patients’ preferences may also affect the results.

Whichever design is used, generalisability needs attention. One approach involves examining the relation between reduction in relative risk (as a measure of effect size) against the percentage of events in the control arm (as an indirect measure of inclusiveness); this is sometimes referred to as metaregression. Where sufficient data are available from RCTs, it may be possible to identify separate measures of benefit and harm. If, as has been shown for giving anticoagulants to prevent stroke, the percentage reduction in relative risk remains constant at all levels of severity and the increase in absolute risk of an adverse effect remains constant, the reduction in absolute risk for a given patient can then be estimated.

In conclusion, RCTs and non-randomised studies can provide complementary evidence—but it is important that clinicians using this evidence are aware of the strengths and weaknesses of each method.


Funding: This work was supported by a grant from the NHS Health Technology Assessment Programme. Competing interests: None declared.

9 Stulenborg GJ. Comparison of carotid endarterectomy outcomes from randomized controlled trials and medicare administrative databases. Arch Neurol 1997;54:826-32.
21 Holme I. Cholesterol reduction and its impact on coronary artery disease and total mortality. Am J Cardiol 1995;76:10-76C.