The Royal Society of Medicine

Journal of the Royal Society of Medicine; 0(0) 1–6
DOI: 10.1177/01410768221140733

Key concepts for informed health choices. 2.1: comparisons of treatments should be fair

AD Oxman[1], I Chalmers[2] and A Dahlgren[3]
[1]Centre for Epidemic Interventions Research, Norwegian Institute of Public Health, 0213 Oslo, Norway
[2]Centre for Evidence-Based Medicine, University of Oxford, OX2 6GG, UK
[3]Faculty of Health Sciences, Oslo Metropolitan University, 0130 Oslo, Norway
Corresponding author: AD Oxman. Email: oxman@online.no

Consider whether evidence from comparisons is trustworthy

To identify treatment effects, studies should make fair comparisons, designed to minimise the risk of systematic errors (biases) and random errors (the play of chance).

As explained in previous essays in this series, it is impossible to know what the effects of a treatment are without comparing it to what would have happened without the treatment. Predicting what the likely effects of a treatment will be depends on treatment comparisons – typically comparisons between groups of people who are treated differently. The trustworthiness of estimates of treatment effects from treatment comparisons depends on the extent to which the design, conduct, analysis, reporting and interpretation of the comparisons have minimised the risk of systematic errors (biases) that distort effect estimates away from the actual effects. Because it is generally not possible to know the degree to which an effect estimate is biased, judgements must be made about the risk of bias using criteria that assess sources of bias. In this essay, we explain seven sources of bias that should be considered when making judgements about the trustworthiness of treatment comparisons:

- dissimilar comparison groups,
- dissimilar care,
- people knowing which treatments they received,
- dissimilar assessment of outcomes,
- unreliable assessment of outcomes,
- outcomes not assessed in all (or nearly all) of the people being compared and
- people's outcomes not analysed in the group to which they were allocated.

The basis for these concepts is described elsewhere.[1]

In the three essays (nos. 2.2, 2.3 and 2.4) after this one, we explain three other factors, besides the risk of bias, that should be considered when making judgements about the trustworthiness of effect estimates from treatment comparisons:

- the trustworthiness of reviews of treatment effects,
- the reporting and interpretation of effect estimates, and
- the reporting and interpretation of the risk of being misled by random errors (the play of chance).

Consider whether the people being compared were similar

If people in treatment comparison groups differ in ways other than the treatments being compared, the apparent effects of the treatments might reflect those differences rather than actual treatment effects. Differences in the characteristics of the people in the comparison groups at the beginning of the comparison might result in estimates of treatment effects that appear either larger or smaller than they actually are. A method such as allocating people to different treatments by assigning them random numbers (the equivalent of flipping a coin) is the best way to ensure that the groups being compared are similar in terms of both measured and unmeasured characteristics.

If people are not randomly allocated to treatment comparison groups, differences between the groups other than the treatments may result in estimates of treatment effects appearing larger or smaller than they actually are because of confounders or other differences. For example, patients who are most ill

(e.g. have severe pain) may be more likely to be given a new treatment than patients who are less ill. There may appear to be a sharp response to treatment in the most-ill patients because of regression to the mean. If they are compared to patients who are less ill and receive an older treatment, the new treatment may appear to be more effective than it actually is compared to the older treatment. Differences in recall ('recall bias') can also lead to over- or underestimates of effects in case–control and retrospective cohort studies that are based on recollection of exposure to a treatment.

The effect of hormone replacement therapy (HRT) on cardiovascular disease (CVD) is an example of an overestimate of a treatment effect in non-randomised studies. For many years, experts and doctors believed – based on non-randomised studies – that HRT reduced the risk of CVD, but the results of large, randomised trials provided no support for this belief and sometimes suggested an increased risk of CVD in women assigned to HRT. This may be because women of lower socioeconomic status are more likely to have CVD and less likely to take HRT. So, a reason for the apparent beneficial effect of HRT on CVD in non-randomised studies is the difference in socioeconomic status between the comparison groups, rather than the difference in whether they took HRT or not.[2]

Quinidine is an example of a treatment for which a beneficial effect appeared smaller in non-randomised studies when compared to those in randomised studies. Quinidine was frequently used to treat heart rhythm abnormalities (atrial fibrillation). Although quinidine was effective for maintaining a normal heart rhythm, it has been replaced by safer and more effective medicines. A systematic review of randomised and non-randomised studies found that the beneficial effect of maintaining a normal heart rhythm was 54% less after three months and 76% less after 12 months in non-randomised studies when compared with randomised studies.[3] One possible explanation for the apparently smaller effects in the non-randomised studies is that patients with the most symptoms and the highest risk may have been more likely to receive quinidine in the non-randomised studies.

Aspirin is an example of a treatment with which a harmful effect appeared larger in non-randomised studies when compared to the results of randomised studies. Randomised studies have shown that low-dose aspirin reduces the risk of stroke in people at high risk (with symptoms and signs of vascular disease) but not in people at low risk. A systematic review of randomised and non-randomised studies found an increased risk of stroke in people at low risk who took aspirin, whereas randomised studies did not find an increased risk.[4] Aspirin use in the non-randomised studies was largely self-selected and it is possible that people who chose to take aspirin had a higher risk of stroke than those who did not, even after statistical adjustment for risk factors that were known and had been measured.

Consider whether the people being compared were cared for similarly

If people in one treatment comparison group receive additional treatments or more care and attention ('co-intervention') than people in the other comparison group, differences in outcomes may reflect those differences rather than the effects of the treatments being compared. For example, in a randomised trial of cognitive behavioural therapy (CBT) for hypochondriasis (persistent fear or belief that one has a serious, undiagnosed illness) compared with no cognitive therapy, a detailed letter of advice was sent to the primary care physicians whose patients were allocated to receive CBT.[5,6] Thus, it was not possible to attribute any differences in outcomes to CBT alone since the letter could have altered how the primary care physicians managed patients allocated to CBT. In addition, patients in the CBT group received more attention than those who did not receive CBT. So, it is uncertain how much of the observed difference in outcomes was due to non-specific attention, support, concern and positive expectation, and not specifically to CBT.

Treatment providers who are aware of the treatment to which people are allocated may treat people differently based on their beliefs about the effectiveness of the treatments that are being compared. Their inclinations for or against the treatment can influence the people receiving care and this could have an impact on the outcome of interest. One way of preventing co-intervention is to keep treatment providers and patients unaware of ('blind' to) which people have been allocated to which treatment. However, this is not always possible. For example, a randomised comparison of acupuncture to relieve symptoms of irritable bowel syndrome compared three groups prior to administering genuine acupuncture to two of the groups.[7] Two groups received sham acupuncture. This blinded the recipients of care, but not the providers. To assess the impact of the providers' attitudes about the treatment, in one group, the providers were instructed to interact minimally with the patients, explaining that it was 'a scientific study' for which they had been 'instructed not to converse with patients'. In the other group,

they communicated with the patients in a warm, friendly manner, actively listened, showed empathy, and communicated confidence and positive expectation. Patients in the third group were added to a waiting list. The proportion of patients reporting adequate relief was 28% in the waiting list group, 44% in the sham acupuncture + minimal interaction group and 62% in the sham acupuncture + positive communication group.

Consider whether the people being compared knew which treatments they received

People in a treatment group may behave differently or experience improvements or deterioration because they know the treatment to which they have been assigned. If this phenomenon is associated with an improvement in their symptoms, it is known as a placebo effect; if it is associated with a harmful effect, it is known as a nocebo effect. If individuals know that they are receiving a treatment that they believe is either better or worse than an alternative (that is, they are not 'blinded'), some or all the apparent treatment differences may be due either to placebo or nocebo effects. For example, a systematic review found 10 randomised trials of acupuncture that included both a 'no acupuncture' group and a 'sham acupuncture' (placebo) group.[8] The non-blinded comparison (of acupuncture compared to no acupuncture) resulted in an overestimate of the effect of acupuncture compared to the blinded comparison (of acupuncture compared to sham acupuncture).

Patients who are aware of the treatment to which they are allocated may also seek additional care or behave differently based on which treatment they receive and their prior beliefs about the effectiveness of the treatment. If they believe a treatment is effective and they are allocated to 'no treatment', they may decide to use the treatment anyway (resulting in 'contamination'), to use some other treatment or to withdraw from the study (resulting in 'attrition bias'). For example, in a randomised trial, a new type of counselling to help people lose weight was compared to 'usual care'. People allocated to the counselling were satisfied with their allocation, whereas those allocated to usual care were disappointed.[9] Their disappointment may have led some participants to 'take control' and change their diet or to seek support elsewhere. This could have resulted in underestimating the effect of the counselling compared to usual care.

Consider whether outcomes were assessed similarly in the people being compared

If a possible treatment outcome is assessed differently in two treatment comparison groups, differences in that outcome may be due to *how* the outcome was assessed rather than *because* of the treatments received by people in each group. For example, if outcome assessors believe that a particular treatment works and they know which patients have received that treatment, they may be more likely to record better outcomes in those who have received the treatment. One way of preventing this is to keep outcome assessors unaware of ('blind' to) which people have been allocated to which treatment.

For example, a randomised trial compared laser surgery to medical treatment for patients with angina (chest pain caused by reduced blood flow to the heart).[10] The severity of angina after one year was assessed by the investigators who were aware of treatment assignment (i.e. unblinded) and by trained interviewers who were not aware (blinded). Comparison of the non-blinded investigators' assessments to the blinded interviewers' assessments showed that the investigators assessed the angina as being less severe much more often in the laser surgery group than in the medical treatment group. Of the apparent angina improvement, 28% could be attributed to bias.

Systematic differences in outcome assessment ('measurement bias') can make treatment effects appear either larger or smaller than they actually are. Blinding is less important for 'objective' outcomes, like death, than for 'subjective' outcomes, like pain.

Consider whether outcomes were assessed reliably

Some outcomes are easy to assess, such as births and deaths. Others are more difficult, such as depression or quality of life. For treatment comparisons to be meaningful, outcomes that are meaningful to people should be assessed using methods that have been shown to be reliable.

Unreliable outcome measures result in outcome misclassification or measurement error. When misclassification is similar in the groups of people being compared ('non-differential'), this tends to lead to underestimation of effects. For example, a vaccine cannot be expected to protect against infections other than those for which it was developed. So, for example, influenza vaccines are less effective for preventing 'influenza-like' illness (much of which is not caused by influenza viruses) than for preventing

Table 1. Total number of deaths after five years in the HIP randomised trial of breast cancer screening.[a]

Comparison group	Group size	Deaths per 1000 women
Offered screening	31,000	28
Chose to be screened	20,200	22
Chose not to be screened	10,800	40
Not offered screening	31,000	30

[a]Data from Table 1 in Freedman et al.[14]

influenza that is confirmed by a laboratory test.[11] As the proportion of influenza-like illnesses that are caused by influenza viruses decreases, the difference will increase between the effects of vaccines on influenza-like illness and laboratory-confirmed influenza.

Consider whether outcomes were assessed in all (or nearly all) the people being compared

People in treatment comparisons who are not followed up to the end of the study may have worse outcomes than those who completed follow-up. For example, they may have dropped out because the treatment was not working or because of side effects. If those people are excluded from the comparison, the findings of the study may be misleading.

For example, in a randomised trial of hip protectors for preventing hip fracture, about 20% of participants were lost to follow-up.[12] The authors dealt with this problem for the main outcome (hip fracture) by accessing the general practice records of patients who were lost to follow-up. However, for other outcomes, such as quality of life, the necessary information had not been recorded, so this was not possible. Therefore, effect estimates for those outcomes could be misleading. Slightly more participants were lost to follow-up in the group assigned to use hip protectors than in the group assigned not to use hip protectors (28% versus 22%). This difference increased the likelihood that participants in the comparison groups were no longer similar, even though they were similar at the start of the trial, as would be expected with random allocation. By looking at the baseline characteristics of study participants, one can see, for example, that more volunteers – people with poor or fair health – and people with a previous fracture had been lost from the control group than had been lost from the intervention group. It is possible to adjust for those variables in the statistical analyses of the results. However, because differences in attrition are difficult to predict, such analyses are rarely planned. Moreover, adjustment can only be made for variables (potential confounders) that have been measured at baseline. Thus, the apparent effect of hip protectors on quality of life is far less certain than the effect on hip fractures.

Consider whether people's outcomes were analysed in the group to which they were allocated

Random allocation to treatment comparison groups helps to ensure that people in the comparison groups have similar characteristics before they receive treatment. However, people sometimes do not receive or take the treatment allocated to them. The characteristics of such people often differ from those who do take the treatments allocated to them. Excluding from the analysis people who did not receive the treatments allocated to them may mean that like is no longer being compared with like. This may lead to an underestimate or an overestimate of treatment differences relative to what would have been the case if everyone had received treatment that had been intended for them.

For example, in a comparison of surgery and drug treatments, people who die while waiting for surgery should be counted in the surgery group, even though they did not receive surgery. This may seem counter-intuitive, but if they are excluded and people who die during the same time in the drug group are not excluded, it will not be a fair comparison.

The New York Health Insurance Plan (HIP) randomised trial of screening for breast cancer provides a striking illustration of how people who comply with a treatment (in this case, screening mammography) may be different from those who do not. The study found similar numbers of deaths after five years

among women offered screening and those who were not offered screening (Table 1).[13] Some women offered screening chose not to be screened. If those women are excluded from the comparison, it appears that there were fewer deaths in the screened group compared to the women who were not offered screening (22 versus 30 per 1000 women). However, that comparison is misleading because there were important differences between the women offered screening who chose to be screened and those who chose not to be screened. Those differences resulted in almost twice as many deaths among women who chose not to be screened compared to women who chose to be screened (40 versus 22 per 1000 women).

Implications

- Be cautious about relying on the results of non-randomised treatment comparisons (for example, if the people being compared chose which treatment they received). Be particularly cautious when you cannot be confident that the characteristics of the comparison groups are similar. If people were not randomly allocated to treatment comparison groups, ask if there were important differences between the groups that might have resulted in the estimates of treatment effects appearing either larger or smaller than they actually are.
- Be cautious about relying on the results of treatment comparisons if people in the groups that are being compared were not cared for similarly (apart from the treatments being compared).
- Be cautious about relying on the results of treatment comparisons if the participants knew which treatment they had received. This may have affected their expectations or behaviour.
- Be cautious about relying on the results of treatment comparisons if outcomes were not assessed in the same way in the different treatment comparison groups.
- Be cautious about relying on the results of treatment comparisons if outcomes have not been assessed using methods that have been shown to be reliable.
- Be cautious about relying on the results of treatment comparisons if many people were lost to follow-up, or if there was a big difference between the comparison groups in the proportions of people lost to follow-up.
- Be cautious about relying on the results of treatment comparisons if patients' outcomes have not been counted in the group to which the patients were allocated.

Declarations

Competing Interests: None declared.

Funding: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Research Council of Norway (Project numbers 220603/H10 and 284683). The funder had no role in the decision to publish, or preparation of the manuscript.

Ethics approval: Not applicable.

Guarantor: ADO.

Contributorship: ADO, IC and AD conceptualised, reviewed and edited drafts of this essay. ADO prepared the first draft.

Provenance: Not commissioned; invited article from the James Lind Library.

Note: Additional material for this article is available from the James Lind Library website [www.jameslindlibrary.org], where it was previously published.

References

1. Oxman AD, Chalmers I, Dahlgren A and the Informed Health Choices Group. Key concepts for informed health choices: a framework for enabling people to think critically about health claims (Version 2022). *IHC Working Paper*, 2022. http://doi.org/10.5281/zenodo.6611932
2. Humphrey LL, Chan BK and Sox HC. Postmenopausal hormone replacement therapy and the primary prevention of cardiovascular disease. *Ann Intern Med* 2002; 137: 273–284.
3. Reimold SC, Chalmers TC, Berlin JA and Antman EM. Assessment of the efficacy and safety of antiarrhythmic therapy for chronic atrial fibrillation: observations on the role of trial design and implications of drug-related mortality. *Am Heart J* 1992; 124: 924–932.
4. Hart RG, Halperin JL, McBride R, Benavente O, Man-Son-Hing M and Kronmal RA. Aspirin for the primary prevention of stroke and other major vascular events: meta-analysis and hypotheses. *Arch Neurol* 2000; 57: 326–332.
5. Barsky AJ and Ahern DK. Cognitive behavior therapy for hypochondriasis: a randomized controlled trial. *JAMA* 2004; 291: 1464–1470.
6. Thomson AB and Page LA. Psychotherapies for hypochondriasis. *Cochrane Database Syst Rev* 2007; 2007: CD006520.
7. Kaptchuk TJ, Kelley JM, Conboy LA, Davis RB, Kerr CE, Jacobson EE, et al. Components of placebo effect: randomised controlled trial in patients with irritable bowel syndrome. *BMJ* 2008; 336: 999–1003.
8. Hróbjartsson A, Emanuelsson F, Skou Thomsen AS, Hilden J and Brorson S. Bias due to lack of patient blinding in clinical trials. A systematic review of trials randomizing patients to blind and nonblind sub-studies. *Int J Epidemiol* 2014; 43: 1272–1283.

9. McCambridge J, Sorhaindo A, Quirk A and Nanchahal K. Patient preferences and performance bias in a weight loss trial with a usual care arm. *Patient Educ Couns* 2014; 95: 243–247.
10. Oesterle SN, Sanborn TA, Ali N, Resar J, Ramee SR, Heuser R, et al. Percutaneous transmyocardial laser revascularisation for severe angina: the PACIFIC randomised trial. Potential class improvement from intramyocardial channels. *Lancet* 2000; 356: 1705–1710.
11. Demicheli V, Jefferson T, Ferroni E, Rivetti A and Di Pietrantonj C. Vaccines for preventing influenza in healthy adults. *Cochrane Database Syst Rev* 2018; 2: CD001269.
12. Dumville JC, Torgerson DJ and Hewitt CE. Reporting attrition in randomised controlled trials. *BMJ* 2006; 332: 969–971.
13. Shapiro S. Evidence on screening for breast cancer from a randomized trial. *Cancer* 1977; 39(6 Suppl): 2772–2782.
14. Freedman DA, Petitti DB and Robins JM. On the efficacy of screening for breast cancer. *Int J Epidemiol* 2004; 33: 43–55.