

EDITORIAL

Inadequate comparators produce misleading results – the importance of good comparison practice

1. Inadequate comparison

While comparison is the basis of good clinical research, incorrect comparisons are likely to produce misleading results, and may cause unnecessary harm to patients. A well known example is that comparing a new drug with an inactive placebo may suggest added value, while comparison with already available drugs would have shown that the new one has no added value or is even less effective. Therefore, drug registration authorities should require head to head comparisons with the current standard of practice, based on comparative effectiveness research [1,2].

In comparing interventions it is also essential to use an appropriate time window, in order to ensure appropriate health care practice. For instance, in many prisons in the USA forced withdrawal from methadone among those who were receiving methadone treatment in the community is used as quickly efficacious methadone detox method. While this may de facto be the case during imprisonment, longer term post-incarceration comparison has shown negative health, economic, and ethical outcomes of this policy as compared with continued methadone use [3].

Furthermore, a seemingly correct comparison can, even in randomized trials, in fact be biased if the selection of the study population does not match the research question. An example is pretending testing effectiveness of a drug by starting from a patient group that has been using it already many years, and then compare the effects of randomly assigning continued use with placebo use, obviously with a quite likely (positive) result [4].

In deciding about what exactly should be compared, conceptual misunderstanding must be avoided. An instructive example is how the concept ‘placebo intervention’ should be elaborated in drug cessation trials, in which drug cessation is the active intervention to be tested. In contrast to what is often practiced - comparing continued use with placebo use, in fact studying the effect of pharmacological withdrawal- the comparator placebo intervention would not be introducing blinded fake medication, but ‘placebo cessation’ [5]. This would imply a visible act of cessation with the pharmacological intervention being invisibly continued. This approach, in which the effect of the act of cessation is also evaluated, needs a highly sophisticated approach in which the medication is administered unnoticed, for

example, via other, not deprescribed medication, or food. While, as far as we know, this design has not been applied yet, it is good to realize what the appropriate comparator would really be.

Finally, inappropriate comparison can relate to external validity. This in view of the fact that effectiveness of diagnostic and therapeutic interventions are sensitive to clinical spectrum [6,7] and—very important in pragmatic trials—to setting. For example, ‘usual care’ may strongly differ per country. If usual care serves as comparator and is of low standard, for example, in deprived settings, the added value of a new intervention may seem impressive, while it may be negligible if the usual care standard is very high. Evaluating such differences is crucial for meta-analyses and the inferences based on them, including setting-specific recommendations.

2. Good comparison practice

While biased comparisons may sometimes be an element of deliberate manipulation, most likely they are the consequence of well meant but inappropriate methodological choices. But one way or another, in terms of a general framework of ‘good comparison practice’, the only way to avoid biased comparisons is respecting the basic principle that a comparator, whether or not in a randomized context, should well match the clinical research question under study. This is a decisive step in designing a methodologically solid and externally valid study. Second, speaking about external validity, the pragmatic nature of the research must be taken into account to make sure that the study outcome is appropriate and acceptable for application in practice. From the comparison made it should be clear what should and what should not be concluded for applying the results in relation to characteristics such as clinical spectrum, age, gender, and setting, including health care system and national standards.

In this issue, the importance of appropriate comparisons is emphasized by Tsui *c.s.* (9841). They evaluated whether non-inferiority trials are designed to adequately preserve *the* historical treatment effect of their active comparators. For this purpose, they reviewed noninferiority trials published in high-impact medical journals. Only 15% of the trials were appeared to be designed so that interventions could only be

found noninferior if they preserved at least 50% of the active comparator's historical treatment effect, while 9% of all non-inferiority trials would have allowed the intervention to be declared noninferior even if it was worse than either placebo or another historical control. The investigators conclude that noninferiority trials published in major medical journals could allow erroneous declarations of noninferiority, and that the design of such trials must be improved to ensure that new interventions are sufficiently effective.

Freedland et al. (9832), in preparing recommendations of an NIH expert panel, addressed the specific challenges in selecting appropriate comparators for randomized controlled trials of non-pharmacological, i.e. health-related behavioral interventions. They report on how an expert panel reviewed the literature on control or comparison groups for behavioral trials and developed strategies for improving comparator choices and for resolving related controversies and disagreements. Based on this, a Pragmatic Model for Comparator Selection in Health-Related Behavioral Trials was developed. It was concluded that the optimal comparator is the one that best serves the primary purpose of the trial, taking the optimal comparator's limitations and barriers to its use into account. The authors report best practice recommendations for the selection of comparators for health-related behavioral trials and recommend use of the model to improve the comparator selection process and to resolve disagreements about comparator choices.

3. The ecosystem of evidence

Good comparison practice includes comparing investigators' promises and achievements, and how differences between them can be explained. This was studied by Koensgen and coworkers (9833), who compared non-Cochrane systematic reviews (SRs) and their published protocols. They reviewed published protocols of non-Cochrane SRs and their corresponding SRs, using the "Preferred Reporting Items for Systematic review and Meta-Analysis Protocols" (PRISMA-P) [8]. It was found that not less than 92.5% of the SRs differed from their protocols in at least one of the PRISMA-P items and subcategories, and that half the SRs had a major difference in at least one item. Of all differences, only 10% were reported in the SRs, two-thirds of which with an explanation. The authors conclude that reporting quality and transparency of non-Cochrane SRs needs improvement, and that all important changes made to the protocol and the SR publication should be reported and explained. They suggest that, for this process, guidance should be included in the updated PRISMA statement.

In planning and prioritizing updates of systematic reviews, given scarce resources, comparison is essential between potential reviews that are candidates for being updated. This was addressed by Bashir and co-authors (9836), who studied the estimation of risk of conclusion change in systematic review updates by learning from a

database of published examples. They applied classification tree methodology and modeled the risk of conclusion change using pairs of systematic reviews and their updates as samples. It was found that for estimating the risk of conclusion change, information about the presence and size of new and potentially relevant trials is most useful. According to the authors, future tools for signaling conclusion change risks would benefit from automated surveillance of relevant ongoing and completed trials. We suggest that the methodology developed by these authors may not only be relevant for decision making on review updates, but also for the research agenda: what priority should be given to which trials of what size, to be able to change the current state of evidence?

This reflection relates to 'the ecosystem of evidence' that, according to Cartabellotta and Tilson (9817), cannot thrive without efficiency of knowledge generation, synthesis, and translation. They discuss current challenges of Evidence-based medicine (EBM) in how to integrate best available research evidence with patient perspective and clinical expertise, and in how to address inefficiencies in the generation, synthesis, and translation of evidence. Interestingly, they define an ecosystem of evidence as being influenced by: living organisms, that is, stakeholders who compete and collaborate; their social, cultural, economic, and political environment; scientific evidence, influenced by rules, standards; and frameworks associated with evidence generation, synthesis, and translation. They also provide an analysis of the strengths and weaknesses of this ecosystem, and outline suggestions for building a stable and resilient ecosystem of evidence.

J. André Knottnerus
Peter Tugwell

References

- [1] O'Connor AB. Building comparative efficacy and tolerability into the FDA approval process. *JAMA* 2010;303:979–80.
- [2] Sox HC, Helfand M, Grimshaw J, Dickersin K, , PLoS Medicine Editors, Tovey D, Knottnerus JA, et al. Comparative effectiveness research: challenges for medical journals. *J Clin Epidemiol* 2010;63:862–4.
- [3] D'Hotman D, Pugh J, Douglas T. The case against forced methadone detox in the US prisons. *Public Health Ethics* 2016;12(1):89–93.
- [4] Neville RG, Crombie IK, McDevitt DG. A double-blind placebo-controlled trial of theophylline in general practice. *Br J Clin Pract* 1991;45(1):14–7.
- [5] De Jonge JW. Diuretic drug cessation in general practice: withdrawing diuretic drugs prescribed for ankle edema. ISBN 90-5170-221-3. PhD-thesis. Chapter 5: Methodological issues of withdrawal trials. Maastricht: Maastricht University; 1993.
- [6] Schneider A, Ay M, Faderl B, Linde K, Wagenpfeil S. Diagnostic accuracy of clinical symptoms in obstructive airway diseases varied within different health care sectors. *J Clin Epidemiol* 2012;65:846–54.
- [7] Knottnerus JA, Tugwell P. Effect modification by setting: how usual is usual care? *J Clin Epidemiol* 2012;65:815–6.
- [8] Shamseer L, Moher D, Clarke M, Ghersi D, Liberati A, Petticrew M, et al, PRISMA-P Group. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. *BMJ* 2016;354:i4086.