Nonrandomized studies of interventions - complementary or just convenient?

Darren L Dahly[1], Jack Wilkinson[2]

[1] HRB Clinical Research Facility UCC, School of Public Health, University College Cork, Cork, Ireland.

[2] Centre for Biostatistics, Manchester Academic Health Science Centre, Division of Population Health, Health Services Research, and Primary Care, University of Manchester, Manchester, UK.

* email: jack.wilkinson@manchester.ac.uk

Schwarze and colleagues present recommendations to improve the quality of non-randomised studies of interventions (NRSIs) (1). Given the methodological errors we regularly encounter when reading and reviewing NRSIs in ART, we support the need for such recommendations to help researchers better design, analyse and interpret these studies. That said, we are concerned that those who require such guidance will, almost by definition, not understand *when* the use of NRSIs is warranted, or to mistake them as viable *alternatives* to RCTs, despite reassurances from Schwarze et al that NRSIs should be viewed as *complementary*. Our concerns are particularly relevant for ART, where regulation is limited, commercial conflicts of interest are common, and interventions are regularly adopted before they are evaluated using RCTs. To be clear, we don't see these as points of disagreement with Schwarze et al, but rather as additional points for consideration.

Examining the case for NRSIs

Schwarze et al follow a common narrative that casts NRSIs as a means of filling evidence gaps when RCTs are unethical, too expensive, and lacking generalizability. While there are hypothetical situations where randomizing patients to particular interventions would be unethical (smoking and empty-parachutes come to mind), we fail to see the relevance for RCTs in ART, since they are about evaluating any benefits of an assignable intervention over existing, accepted treatment options, where ethics is largely underpinned by the principle of equipoise.

With respect to costs, we don't deny they are an important consideration, but any concern must consider the costs *for whom.* For example, our view of costs might be one thing in context of research on a rare disease without commercial incentive for development of treatments, but another thing entirely for the development of treatments likely to generate considerable profits for a commercial company. It is also important to consider *where* costs are going. Any experienced trialist will agree that trial conduct can be needlessly burdensome, and thus costly from an administrative perspective. However, we must recognize that many RCT costs are for things we'd hardly want to do without. These include the capture of costly but highly informative data that wouldn't otherwise be collected, high-quality data collection processes more generally, and efforts made to ensure participant safety.

In terms of generalizability of treatment effects, it is certainly true that clinical RCTs are rarely (if ever) conducted in random samples of target populations. Thus, at best, they can help us understand sample average treatment effects. However, the degree to which this inherent, long-understood limitation might matter for medical decision-making is situation specific, and even the general importance of

representativeness has been questioned. For example, surely the generalizability gap from animal to human is larger than that between RCT and NSRI, yet preclinical animal studies remain relevant for drug development. Similarly, there are many medical interventions that have been brought to market on the back of RCT evidence that indeed panned out to be widely effective treatments. One could even argue that non-representative trials producing widely generalizable results are a key driver of that medical progress. Further, some RCTs are designed to be widely representative (e.g. pragmatic and registry trials), and while some NRSIs are not (e.g. UKBiobank). And while it is true that RCTs don't always reflect so called "real world" practice, often enrolling patients most likely to benefit from the treatment and supporting treatment compliance, these studies are often the most informative. This is because a null result from a high-powered trial in this "best case" scenario can efficiently rule out treatments that aren't likely to work, potentially stopping expensive ghost-chases in their tracks. Finally, Schwarze et al rightly highlight concerns about inclusivity. Efforts to remove barriers to health research participation and improve inclusivity are crucial for both RCTs and NRSIs. In the UK, the National Institute for Health research has launched a requirement to demonstrate how research will address inequalities as a condition of funding (https://www.nihr.ac.uk/about-us/who-we-are/research-inclusion).

Potential risks of target trial terminology

There are risks that the term target trial emulation (TTE) might mislead readers without methodological training. We believe some readers may assume TTE represents a distinct form of observational study design, rather than a framework to help researchers identify and avoid self-inflicted errors in observational studies. We are particularly concerned that claims to have *emulated* an RCT will be mistaken for a claim to have *successfully emulated* an RCT. The latter statement implies that the inference is of a comparable standard to a well-conducted RCT. To be certain, the TTE framework (2), used correctly, will help investigators specify estimands, and side-step avoidable selection bias, but we are probably more pessimistic than Schwarze et al about the ability of "advanced methods" to address the threat of other sources of bias. However, the difference in how randomization accounts for confounding vs "advanced methods" like covariate adjustment, weighting, etc is a difference in kind, not in degree. This is because randomization is guaranteed to eliminate confounding for some key estimands, and to do so in a matter that predominately rests on our trust in a potentially verifiable randomization process. In contrast, acting on the effect estimate returned from an NRSI requires us to trust in the decisions and competence of the investigator, and even then no guarantee is possible, even in relatively simple sets of relationships among a small set of potential covariates. This distinction will be obvious to methodological experts, but they only constitute a small proportion of researchers and research-consumers.

Care is therefore needed in how TTEs are reported, to prevent any unintended obfuscation of the study limitations. We have already seen some misleading language in TTEs. While peer-reviewing an emulated trial in ART, we saw that authors had included the subheading "Randomisation". In this section of the manuscript, they described how participants "were randomly assigned" to treatment or control groups. This was a retrospective, observational study -- no one had been randomly assigned, but a reader could not be blamed for thinking otherwise.

We prefer an approach where the TTE framework is used not to emphasise study strength but rather as a way to highlight limitations of an NRSI. The approach adopted by the ROBINS-I (Risk Of Bias In Nonrandomised Studies of Interventions) tool is to define a hypothetical target RCT as a point of

comparison to the NRSI (3). This can be used to reflect on potential deficiencies in the study that could cause bias.

Closing remarks

The recommendations of Schwartz et al., help draw our attention to the importance of rigorous methods in ART where critical errors in the design and analysis of NRSIs are common. In an evaluation of PGT-A, investigators may include any participant with no embryos in a control group, on the grounds that they did not undergo PGT-A, introducing a form of immortal time bias (4). In an evaluation of an embryo culture medium, investigators might adjust for the number of embryos transferred, introducing bias due to overadjustment (e.g. 5). The recommendations, if followed, will prevent instances such as these. Nonetheless, we believe it is important to hold the line. NRSIs should be used when necessary, but the impracticalities and apparent ethical limitations of conducting RCTs should not be overstated.

References

1. Schwarze J-E, Tennant PWG, Barnhart K, Platt RW, Gupta S, Venetis C, et al. Recommendation to improve the rigor and impact of nonrandomized studies of interventions in fertility treatment research. Fertility and Sterility. 2025.
2. Hernan MA, Wang W, Leaf DE. Target Trial Emulation: A Framework for Causal Inference From Observational Data. JAMA. 2022;328(24):2446-7.
3. Sterne JAC, Hernan MA, Reeves BC, Savovic J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. Bmj-British Medical Journal. 2016;355.
4. Roberts SA, Wilkinson J, Vail A, Brison DR. Does PGT-A improve assisted reproduction treatment success rates: what can the UK Register data tell us? J Assist Reprod Genet. 2022;39(11):2547-54.
5. De Vos A, Janssens R, Van de Velde H, Haentjens P, Bonduelle M, Tournaye H, et al. The type of culture medium and the duration of in vitro culture do not influence birthweight of ART singletons. Hum Reprod. 2015;30(1):20-7.