

Appraisal

Research Note: Transporting causal effects from randomised trials to target populations to improve external validity

Randomised trials can reliably estimate average treatment effects from a sample that is drawn from a population.¹ When they are well conducted, trials can rigorously mitigate threats to internal validity and enable credible claims to be made about the causal effects of interventions. However, investigators and clinicians have noted differences between patients seen in practice and participants enrolled in trials.^{2,3} These concerns may cast doubt over the external validity of trials and limit the ability of trial findings to inform policy and practice with confidence.⁴ Recent discussions have suggested that for trials to reliably inform decisions about intervening on populations, they should not only generate internally valid estimates of treatment effects but also enable reliable inferences about relevant target populations.^{3,5} This research note discusses how average treatment effects from a randomised trial can be extended to clinically relevant target populations when trial samples may not represent target populations.

Most randomised trials cannot enrol simple random samples of a target population

Under most clinical settings it is unfeasible to enrol the entire target population into a randomised trial. Even the largest and most pragmatic physiotherapy trials typically recruit a fraction of the target population. Therefore, depending on the extent to which there is selective invitation and consent into the trial, the trial sample will be more or less resemblant of the target population. Selective recruitment is common: some hospitals do not participate in trials, clinicians may forget to extend invitations, participation may be limited to selected geographical regions, and language barriers could exclude certain populations. Rothwell claims that across medical disciplines, less than 10% of patients who would have been considered eligible for a relevant trial end up being recruited.³ In principle, comprehensive recruitment strategies can overcome selective invitation; however, selective consent is certainly inevitable. All individuals, guardians and attorneys have the right to refuse consent. This guarantees that the trial sample will never be a perfect random sample of the target population.^a Figure 1 illustrates key sampling processes that would typically occur for a randomised trial.

Randomised trials can make inferences about the population it represents, but that population may not reflect a clinically relevant target population

Most randomised trials aim to estimate treatment effects that apply to some population.^{1,6} More precisely, randomised trials aim to

estimate a contrast^b of the average potential outcome had all individuals in a population been treated versus the average potential outcome had all individuals in a population not been treated. Provided that there are minimal losses to follow-up and adequate adherence,⁷ the intention-to-treat effect can be interpreted as the expected average causal effect of the treatment in a population that is resembled by the trial sample (ie, the trial population). Unfortunately, for some trials, it is not guaranteed that the trial population reflects a target population the investigator wishes to make inferences about or the population where they wish to implement the findings. The relationships between the target population, trial population and trial sample are illustrated in Figure 2.

In practice, eligibility criteria are used to restrict the entire population to some *trial-eligible* population. It may appear that the eligibility criteria fully define the target population of interest. But that is not true. Trial eligibility criteria only serve to place boundaries around a relevant subset of the population, and it is expected that samples drawn from the trial-eligible population will vary with respect to individual characteristics. Also, applying reasonable eligibility criteria may still exclude individuals who would have been considered suitable for the treatment being tested in the trial. Therefore, eligibility criteria are just the investigators' intention to partially define a target population. Making inferences about populations requires further understanding about how individuals in a trial sample compare to those in the target population.

Selective trial participation alone rarely causes problems for making inferences about target populations, but when coupled with effect heterogeneity, external validity can be compromised

It may be easy to criticise the generalisability of a randomised trial just because the trial sample is not randomly drawn from the target population. However, non-random sampling would not compromise external validity^c if treatment effects are homogenous.^d Only when causes of the outcome in the target population are unequally represented in the trial sample, the trial may have limited capacity to make valid inferences about the target population. For example, if a falls prevention strategy for people with Parkinson's disease increases the rate of falling for those who have freezing of gait,⁹ and more people freeze in the target population compared to the trial sample, we might expect the average treatment effect from the trial to differ from the expected treatment effect in the target population.

^bTypically, a 'contrast' refers to a difference or ratio of potential outcome means, probabilities or hazards.

^cBroadly, any variation in treatment composition, treatment mechanisms, outcome measure, setting and individual characteristics between the trial sample and target population can influence external validity. To limit the scope of this Research Note, it is assumed that all factors are consistent across the trial sample and target population except for individual characteristics.

^dThe absence of evidence for treatment effect heterogeneity does not imply that treatment effects are homogenous.

^aThe population represented by the trial sample will always be different to the target population with respect to time. Because trial findings are typically applied to a target population after the trial has closed, the source population that produced the trial sample is historical and could be different to the target population where the findings are to be implemented.^{5,8}

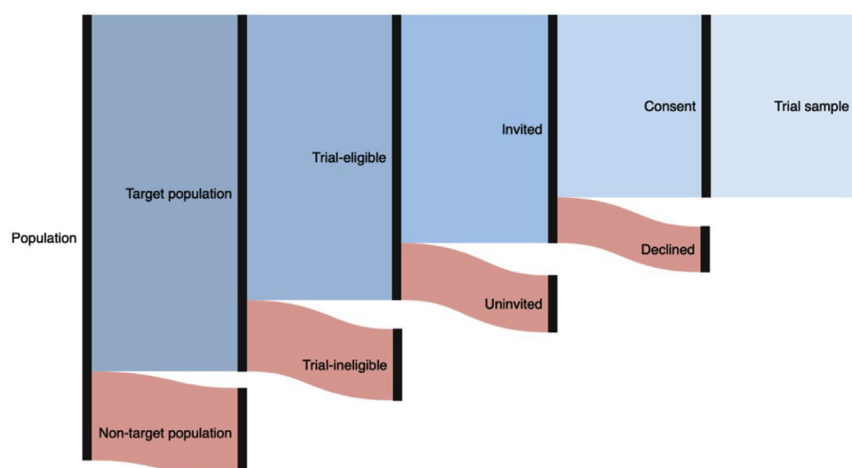


Figure 1. Selective sampling process from a population to a trial sample. In principle, there is a subset of the entire population that can be labelled the target population. From the target population investigators define a trial-eligible population. Trial-eligible individuals are then invited (or self-enrol), then a subset of the invited individuals provide consent to participate. Those who consent make up the trial sample. At each stage of selection there is a chance for the excluded subsets (in red) to differ from the included subsets (in blue), which would result in a trial sample that does not resemble the target population.

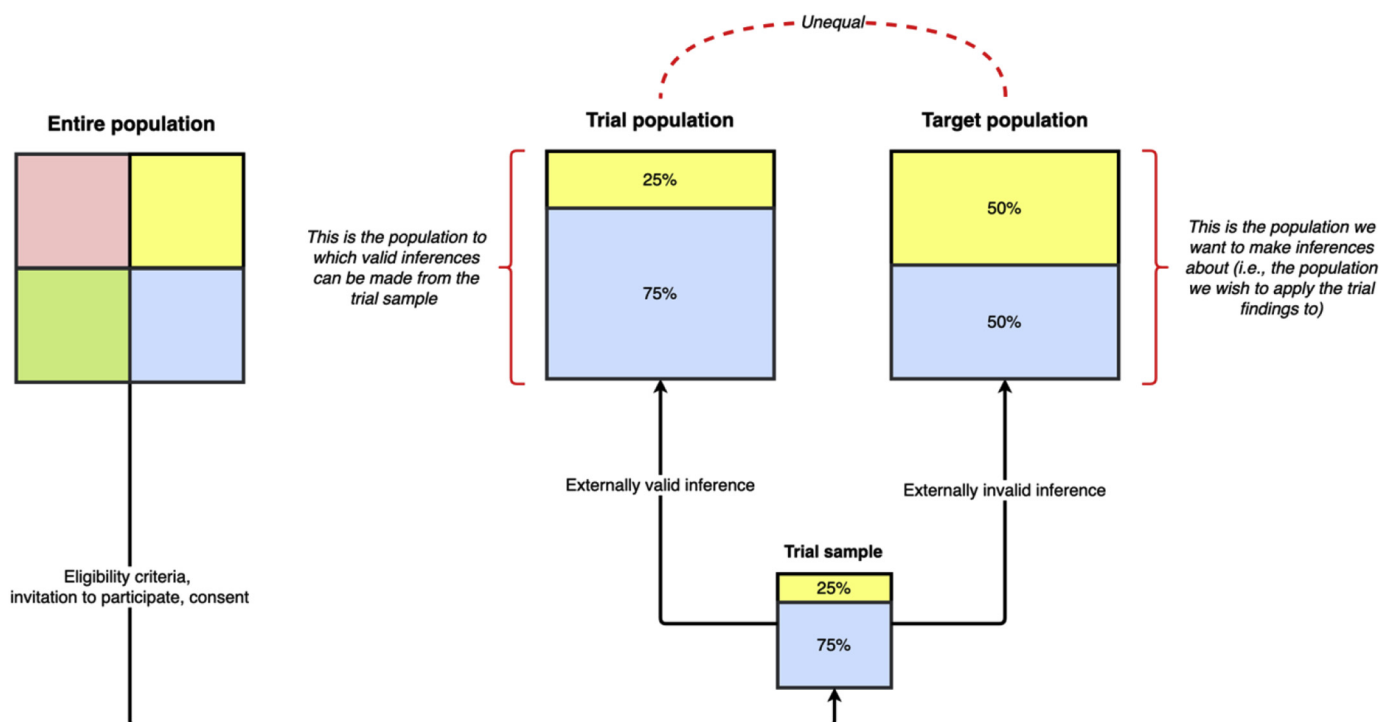


Figure 2. The relationship between the target population, trial sample and trial population. The trial sample that is drawn from the entire population can make valid inferences about a population it represents (the trial population). But when there is selective sampling (Figure 1), the trial population may not reflect the target population with respect to patient characteristics (indicated by the colours and percentages). This fictitious example only displays the distributions of one characteristic. In reality, there will be more characteristics and their combinations to consider.

Claims about the external validity of trials should always be made with reference to a target population

It is not meaningful to talk about the external validity of a trial without reference to a target population that is of scientific or practical interest.⁵ Because the fundamental goal of the randomised trial is to make inferences about a target population, a claim or assessment about the external validity of a trial should be made in reference to a target population. This means that a trial could be externally valid for one target population but not another. Take for example the Parkinson's falls prevention trial conducted in the UK's NHS⁹ and assume that it sampled participants at random so that all causes of the outcome were equally distributed in the trial and target populations. There is a good chance that this trial would allow valid inferences about the average treatment

effect in the UK population with Parkinson's. However, if there was interest in applying the findings to care home residents with Parkinson's in New Zealand where there is a higher prevalence of freezing, it is unlikely that the trial will provide reliable inferences about the New Zealand care home population. This shows that we cannot just claim that the Parkinson's trial is externally valid or invalid. These statements must be made with careful reference to the target population of interest. Arguably, where possible, trials should report baseline characteristics of a target population alongside the baseline characteristics of the trial sample (ie, Table 1 of the trial report) so that its external validity can be assessed descriptively.

The following sections aim to provide a broad overview of emerging methods that can extend the findings of a possibly non-representative trial to a target population.

Generalisability and transportability analyses

By combining data from a randomised trial and a registry or cohort that can be considered representative of a target population^e recent developments have produced rigorous frameworks^{8,10} and methods^{11–13} for estimating the average treatment effect in a target population. The intuition behind this approach is to utilise the rigour of randomised trials to ensure internal validity, and the breadth of registries and cohorts to improve external validity. Informally, these methods can be used to answer the question: ‘what would be the average treatment effect had we conducted the same trial in the entire target population?’.

In recent discussions, the term *generalisability* has been used to describe the objective of learning about a population of all trial-eligible individuals, and *transportability* as the objective of learning about a broader target population who are not necessarily part of the trial-eligible population.¹⁴ The methods outlined below generally apply to both generalisability and transportability, with a few exceptions that are noted as footnotes. Note there are also design adaptations to the randomised trial (eg, two-stage preference trials) that can also enhance generalisability,¹⁵ but they are not the focus of this paper. The following sections describe analytical methods that can extend trial findings when design adaptations are unfeasible, unethical or uneconomical.

To illustrate key concepts, we use an example of a generalisability analysis that extended the WHiTE-3 trial ($n = 958$) to a subset of the UK's National Hip Fracture Database who would have been considered eligible for the WHiTE-3 trial ($n = 190,894$).¹⁶ The WHiTE-3 trial compared the effects of two hip implants (modern Exeter hemiarthroplasty versus the traditional Thompson monoblock) on quality of life and length of hospital stay in a convenience sample of participants aged > 60 years with displaced intracapsular fractures of the hip.

Study design and data requirements

Dahabreh et al outlined two major study designs for extending trial findings:¹⁷ a nested trial design, where a randomised trial is embedded in a cohort or registry, and a non-nested trial design, where data from a randomised trial are combined with data from a separate cohort or registry. In both designs, it is assumed that the cohort or registry is either a census or a simple random sample of the target population. As most physiotherapy trials are not nested in cohorts or registries, the natural design choice would be the non-nested trial design.

In both designs, individual participant data on treatment, participant characteristics (plausible causes of outcome) and outcome are required from the trial, and individual participant data on participant characteristics are required from the cohort or registry. In most applications, particularly in the non-nested case, measures on participant characteristics will need to be harmonised across the trial and target population datasets.

Assumptions under which randomised trials can be extended to target populations

To obtain an unbiased estimate of the average treatment effect for a target population, a set of unverifiable assumptions must be met. The following key assumptions are required in addition to the standard assumptions for obtaining an unbiased average treatment effect from a randomised trial alone.^f

^eA registry or cohort may not be a random sample, unless it is a census of the target population. Therefore, we often make an extra assumption that the registry or cohort is a simple random sample of the target population the investigator is concerned about.

^fThese assumptions include consistency (well-defined interventions), positivity of treatment assignment and exchangeability of treatment assignment (no confounding).¹ In most cases these assumptions will be satisfied by design in a randomised trial. As noted in footnote ‘c’, we restrict our discussion to cases where the treatment mechanism is consistent between the trial and target populations.

The first assumption, *conditional exchangeability for trial selection*, requires that there are no unmeasured causes of outcomes that are unequally distributed between the trial and target population.^{10,11} For example, if the Exeter stem that was tested in WHiTE-3 offered better quality of life for people who took bone medication, and there were more bone medication users in the target population than the trial, this difference should be adjusted in the analysis. This assumption is often challenging to satisfy because the investigator will usually need to rely on subject matter knowledge to identify plausible causes of outcomes. Selection diagrams¹⁰ can be useful for identifying a sufficient set of variables that would relax this assumption (Figure 3).

The second assumption, *positivity of trial participation*, requires that all individuals in the target population have a positive probability of being selected into the trial. For example, say 2% of population represented by the UK National Hip Fracture Database had scoliosis and we believe that having scoliosis might cause the outcome (thus is required to achieve *conditional exchangeability for trial selection*). If nobody in the trial had scoliosis, positivity could be violated because the 2% of people in the target population with scoliosis are not represented in the trial (ie, have zero probability of being selected into the trial). In most cases, this type of violation can be avoided by carefully limiting the target population to individuals who would be considered eligible for the trial.^g

Estimating average causal effects in the target population

If all the above conditions are met, the average treatment effect in a clearly defined target population can be estimated. The analytical methods can be broadly classified into three types: weighting by trial participation,^{11,12} outcome modelling^{14,18} and their combination.¹⁴ As the weighting method is perhaps the most widely used approach, this section will focus on that approach, and readers are referred to the references for the other approaches.

The basic principle behind the weighting method is to weight individuals in the trial so that they closely resemble the target population with respect to variables that affect the outcome. The first step involves fitting a model for trial membership based on a set of variables collected in the trial and target population. This model is used to obtain predicted probabilities of trial membership that are used to calculate sampling weights^h for all individuals in the trial.¹¹ Intuitively, individuals in the trial who have characteristics that were over-sampled get assigned a small weight, and individuals who have characteristics that were under-sampled get assigned a large weight. Provided that the trial participation model is well specified, applying the sampling weights to all trial participants makes the trial sample more like the target population. Finally, the potential outcomes for the treatment and control groups can be estimated in the weighted trial sample to obtain the average treatment effect for the target population.^{11,12}

The generalisability analysis of the WHiTE-3 trial showed that there were differences between the trial sample and target population in plausible causes of the outcome such as the use of walking aids and pre-fracture living status. Despite these differences, the average treatment effect in the target population was comparable with the average treatment effect from the trial. For quality of life, the estimate for the target population was 0.05 (95% CI -0.01 to 0.11) and estimate from the trial was 0.06 (95% CI -0.01 to 0.12). Similarly, for length of hospital stay (in days), the estimate for the target population was -1.14 (95% CI -2.35 to 0.08), whereas the estimate from the trial was -0.70 (95% CI -1.90 to 0.51). Provided that the measured covariates were sufficient to achieve conditional exchangeability of

^gIf the goal is to estimate an average treatment effect in a broader population of individuals who are not necessarily trial-eligible (ie, transportability), investigators may intentionally breach the positivity assumption to allow extrapolation to a wider target population.

^hThe sampling weights are calculated as the inverse probability of trial membership or the odds of trial membership, depending on whether the inferential goal is generalisability or transportability, respectively.¹²

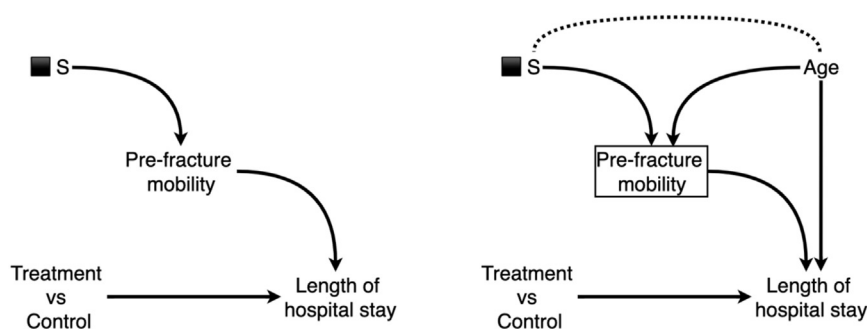


Figure 3. Selection diagrams. Selection diagrams are Directed Acyclic Graphs (DAGs) that represent the trial population, combined with selection nodes (■ S) that indicate differences between the trial and target populations. Please refer to a previous Research Note on causal inference (Herbert 2020) for a gentle introduction to DAGs.⁶ To estimate the effect of a treatment on an outcome in the target population, all selection nodes must be made independent of the outcome. On the left panel, pre-fracture mobility is a cause of the outcome (length of hospital stay) and could be differently distributed in the trial and target population, as indicated by the selection node. To make the selection node independent of the outcome, pre-fracture mobility would need to be adjusted. On the right panel, pre-fracture mobility is a cause of the outcome, and age is a cause of the outcome and pre-fracture mobility. As there is no selection node going into age, we assume age is equally distributed between the trial and target population. Here we want to adjust for pre-fracture mobility to separate the selection node and outcome. But doing so induces a correlation (dotted line) between selection and age because pre-fracture mobility is a collider of selection and age. This opens a back-door path from selection to the outcome via age. Therefore, the estimate of the treatment effect should adjust for mobility and age to achieve conditional exchangeability for trial selection.

trial selection, these results provide some assurance that the WHiTE-3 findings would apply to the wider UK population of individuals who would have been eligible for the trial.

The transported causal effect will often include the effect of treatment and the effects of trial participation

Lastly, it is important to acknowledge that participating in a randomised trial may have effects on the outcome that are not mediated through the actual treatment being tested. For example, participating in the WHiTE-3 trial may have increased the frequency and quality of postoperative exercises because participants in both arms were closely monitored by physiotherapists during the trial. Furthermore, especially for self-reported outcomes, there may be Hawthorne effects from trial participation.¹⁹ As it can be challenging to partition the effect of treatment from the effects of participating in a trial, when we generalise or transport trials to target populations, we are extending the effect of the treatment and also the possible effects of trial participation.²⁰

Concluding remarks

Well conducted randomised trials can provide robust estimates of average treatment effects. But they are more useful when results are externally valid and applicable to relevant target populations. The increasing availability of routinely collected data and recent methodological developments have strengthened the ability to improve the external validity of randomised trials. Generalisability and transportability analyses could help remove barriers to implementation by assuring clinicians and decision-makers about the external validity of selected trials to their populations of interest. They could also identify trial findings that are unsuitable for implementation, based on their lack of generalisability or transportability. Trials could provide greater return on investment if they can be transported to multiple settings to guide implementation and resource allocation.

These are exciting prospects, but they come with challenges. We should not forget that these methods rely on strong assumptions that must be reasonably satisfied to permit valid claims about generalisability and transportability. Relatedly, the quality of registry data and their overlap with trial data must be sufficient and well-aligned so that conditional exchangeability for selection assumption can be met. Finally, harmonisation of data across trials and registries through

common data models may create opportunities for robust generalisability and transportability analyses.

Competing interests: Nil.

Sources of Support: Nil.

Acknowledgements: We thank Dr Aidan Cashin for earlier comments and proofing of this paper.

Provenance: Invited. Peer reviewed.

Correspondence: Hopin Lee, Centre for Statistics in Medicine, Nuffield Department of Orthopaedics Rheumatology and Musculoskeletal Sciences (NDORMS), University of Oxford, Oxford, UK. Email: hopin.lee@ndorms.ox.ac.uk

Hopin Lee^{a,b} and Sarah E Lamb^{a,c}

^aCentre for Statistics in Medicine, Nuffield Department of Orthopaedics Rheumatology and Musculoskeletal Sciences (NDORMS), University of Oxford, Oxford, UK

^bSchool of Medicine and Public Health, University of Newcastle, Newcastle, Australia

^cCollege of Medicine and Health, University of Exeter Medical School, Exeter, UK

References

1. Rubin DB. *J Educ Psychol.* 1974;66:688–701.
2. Kennedy-Martin T, et al. *Trials.* 2015;16:495.
3. Rothwell PM. *Lancet.* 2005;365:82–93.
4. Huebschmann AG, et al. *Annu Rev Public Health.* 2019;40:45–63.
5. Westreich D, et al. *Am J Epidemiol.* 2019;188:438–443.
6. Herbert RD. *J Physiother.* 2020;66:273–277.
7. Kasza J. *J Physiother.* 2021;67:147–149.
8. Bareinboim E, Pearl J. *J Causal Inference.* 2013;1:107–134.
9. Ashburn A, et al. *Health Technol Assess.* 2019;23:1–150.
10. Pearl J, Bareinboim E. Transportability of Causal and Statistical Relations: A Formal Approach. In: *2011 IEEE 11th International Conference on Data Mining Workshops. IEEE;* 2011:540–547.
11. Dahabreh IJ, et al. *Biometrics.* 2019;75:685–694.
12. Westreich D, et al. *Am J Epidemiol.* 2017;186:1010–1014.
13. Stuart EA, et al. *Prev Sci.* 2015;16:475–485.
14. Dahabreh IJ, et al. *Stat Med.* 2020;39:1999–2014.
15. Marcus SM, et al. *Psychol Methods.* 2012;17:244–254.
16. Lee H, et al. *J Clin Epidemiol.* 2021;131:141–151.
17. Dahabreh IJ, et al. <http://arxiv.org/abs/1905.07764>. Accessed 9 March, 2021.
18. Kern HL, et al. *J Res Educ Eff.* 2016;9:103–127.
19. McCarney R, et al. *BMC Med Res Methodol.* 2007;7:30.
20. Dahabreh IJ, Hernán MA. *Eur J Epidemiol.* 2019;34:719–722.