Defining representativeness of study samples in medical and population health research

Jacqueline E Rudolph, Yongqi Zhong, Priya Duggal, Shruti H Mehta, Bryan Lau

ABSTRACT

Medical and population health science researchers frequently make ambiguous statements about whether they believe their study sample or results are representative of some (implicit or explicit) target population. This article provides a comprehensive definition of representativeness, with the goal of capturing the different ways in which a study can be representative of a target population. It is proposed that a study is representative if the estimate obtained in the study sample is generalisable to the target population (owing to representative sampling, estimation of stratum specific effects, or quantitative methods to generalise or transport estimates) or the interpretation of the results is generalisable to the target population (based on fundamental scientific premises and substantive background knowledge). This definition is explored in the context of four covid-19 studies, ranging from laboratory science to descriptive epidemiology. All statements regarding representativeness should make clear the way in which the study results generalise, the target population the results are being generalised to, and the assumptions that must hold for that generalisation to be scientifically or statistically justifiable.

Introduction

It is common if not a requirement for medical and population health science researchers to consider the inferences from a study beyond the context of their analysis. Accordingly, many papers mention whether their study sample is representative of, or study results generalise to, some implicit or explicit target population; others refer to a lack of generalisability or representativeness as a limitation. Despite being frequently discussed and debated, in common practice, the meaning of representativeness remains ambiguous. Here, we propose a comprehensive definition of representativeness and discuss it in the context of different study designs. We presume no bias in the study’s results; in any real world study, bias will need to be weighed alongside whether the sample is representative or its results are generalisable or applicable to a target population.

What is representativeness?

The ambiguity in meaning arises in part because the word “representative” has a broader meaning in English and a more technical definition, and the definition being used is not always clear. In a 2013 series of commentaries on representativeness, the concept was defined as occurring when the study sample is a simple random sample of the target population (ie, the sample that arises through representative sampling). A second definition is that the study sample and the results obtained merely resemble what would be expected in the target population, perhaps based on a similarity in personal characteristics. The first definition is more precise and implies a high standard for study design, while the second encompasses a variety of possible interpretations.

Here, we bridge these two uses of the word “representative” and attempt to concretise the second, broader definition. We define a study sample to be representative of a well defined target population if the results estimated in that sample are generalisable to the target population. We consider two ways in which study results can generalise to the target population: in estimate and in interpretation. Box 1 lists a summary of key terms used in this article and figures 1–2 show examples applying the definition of representativeness.

To help clarify these definitions, we use as an example a randomised controlled trial that was conducted to measure the efficacy of molnupiravir for treatment of covid-19. Among unvaccinated adults with mild to moderate covid-19 who were not in the hospital, researchers found that the risk of hospital admission or death at 29 days among participants randomised to molnupiravir was 6.8%, compared with 9.7% of participants randomised to placebo. They concluded that treatment with molnupiravir within five days of infection reduced the risk of hospital admission or death. In boxes 2–4, we explore the definitions in the context of other study designs. As with the trial, we use the question, study design, and sample description of these publications to construct a theoretical example. We do not delve into the specific details of the study or comment on whether the researchers fully achieved what we discuss.

KEY MESSAGES

- Researchers frequently refer to whether their sample is or is not representative without clarifying whether they mean that their sample is a simple random sample of their target population or that the results from their sample are merely reflective of what would be seen in the target population.
- This article provides a comprehensive definition of what it means for a study to be representative, and examines this definition in the context of examples with different study designs.
- When publishing research, researchers should critically assess whether a study sample is representative of a clearly defined target population, by carefully considering the manner in which they think the results generalise to the target population and the assumptions underlying that hypothesis.
Generalisable in estimate
A sample is representative if its results are generalisable in estimate. For a given estimand (eg, risk difference, odds ratio, population mean), the estimate obtained in the study sample is the same within a margin of error as what would be estimated in the target population. In the molnupiravir randomised controlled trial, we might hypothesise that the risk difference comparing molnupiravir with placebo estimated in the trial is the same risk difference as would be estimated in the target population of all adults with recent SARS-CoV-2 infection who were not in the hospital. Generalising the estimate obtained in a given study sample might be considered the primary goal when intending to quantitatively inform policy interventions or when obtaining effect estimates in the target population is impossible or infeasible.11

Generalisability in estimate can be achieved if the distributions of key covariates are the same as in the target population, as would occur in expectation with random sampling. Thus, generalising the estimate aligns closely with the definition of representativeness based on representative sampling. These key covariates are those that affect the variable under study (eg, hospital admission or death) and thus are potential effect measure modifiers of the effect of a treatment on that variable. By effect measure modifiers, we mean variables where the effect of the treatment differs by levels of that variable on some scale (eg, risk difference, risk ratio, odds ratio). In our example, age might be an effect measure modifier because the effect of molnupiravir on hospital admission or death (as quantified by the risk difference) might differ across ages.

More generally, even if the distribution of the key covariates differs between the sample and target population, the sample might still be representative within strata of the key covariates, such that the stratum specific estimates (eg, risk difference within age categories) can be generalised from the sample to the target population. While this generalisation requires that all the key covariates be measured, the proportion of the sample in the covariate strata need not exactly match the proportion who fall into that subgroup in the target population.

If we apply this definition of generalisability in estimate to the molnupiravir trial example, suppose our target population is all individuals recently infected with SARS-CoV-2 who were not in the hospital. In this case, we would not be able to generalise the trial’s estimate to the entire target population, even if we could control for post-randomisation factors such as non-adherence, because the trial sample did not include vaccinated individuals. However, the target population does include vaccinated individuals, and it is reasonable to assume that the effect of molnupiravir on disease progression to hospital admission or death would vary by vaccination status. On the other hand, we might be able to generalise our results within the stratum of unvaccinated individuals, provided all other effect measure modifiers were similar. In this case, we would say that the sample is representative within that stratum.

Generalisable in interpretation
A sample is representative if its results are generalisable in interpretation. While the estimates obtained in the sample are not quantitatively the same within a margin of error as those that would be estimated in the target population, we can hypothesise, based on background knowledge, that the interpretation (which could be the direction of effect, general inference from the results, or knowledge gained from an experiment) would remain the same.12 For example, we might hypothesise that molnupiravir is generally protective against hospital admission or death from covid-19, even in samples other than the study sample. Generalising the interpretation aligns with the broad definition of representative, which states that a study sample resembles what would be expected in the target population.

We often generalise the interpretation of our own results to external populations, and any study that generalises in estimate will also generalise in interpretation. The primary goal in studies examining fundamental laws of nature or asking research questions that are relatively independent of historical and
environmental context are to generalise the interpretation, rather than the estimate. Generalising the interpretation should be done cautiously, however, because it is based on hypotheses that the mechanisms or biological processes under investigation in the study sample are (at least approximately) identical to those that would be seen in the target population.

If we apply this definition of generalisability in interpretation to the molnupiravir trial example, it might be reasonable to hypothesise that molnupiravir would have a beneficial impact if given to those individuals infected with SARS-CoV-2, even beyond the enrolled sample of participants with moderate illness who were not in the hospital. We might base this hypothesis on our understanding of the drug’s biological mechanism and the validity of a properly conducted, double blind, placebo controlled trial. In this example, we generalise the interpretation despite the fact that the risk difference comparing molnupiravir with placebo estimated in the trial would differ from the risk difference estimated in the target population (again assuming that the target population is all individuals with recent SARS-CoV-2 infection who were not in the hospital).

Discussion
In summary, we consider a sample to be representative of a target population if its results can be generalised to that target population either in estimate or in interpretation. Any statements made regarding the representativeness of the study need to make this further qualification. Is it the estimate obtained or the interpretation of the results that are generalisable to the target population? Researchers should also do what they can to safeguard their results from being applied incorrectly. Even in studies with a strong scientific rationale for generalising the interpretation of results to the target population, researchers might need to mention that the estimate obtained in the sample should not be naively generalised to the target population.

Stating which form of representativeness was the goal of the study might also be useful. In the example of the molnupiravir randomised controlled trial, generalising the interpretation regarding drug administration
efficacy to the target population might have been the primary goal. Many trials have this same goal, because the investigators often over sample individuals at high risk for the outcome in order to increase the power of the study. (Even so, clinical trials have received some criticism that they rarely represent a more general target population.) If it was possible, generalising the estimate to the target population would be useful for predicting how molnupiravir would perform in practice but might not be immediately required for the study results to be meaningful. Further studies would likely need to be conducted to generalise the interpretation to other target populations, such as children recently infected with SARS-CoV2.

Several points relate to defining representativeness and are worth discussing. Firstly, irrespective of the way in which a sample is representative, the target population must be clearly defined. Stating that a sample is representative is meaningless unless researchers specify what population it represents or its results are being applied to. As an example, we showed how specifying different target populations (all individuals vs all unvaccinated individuals who were not in the hospital) for the molnupiravir randomised controlled trial had different implications for whether the results were generalisable in estimate.

Secondly, researchers must be clear about the assumptions required for generalising to the target population. When generalising the estimate, these assumptions might be made based on knowledge of whether the study was designed using a simple random sample or whether stratification by relevant key covariates is possible. When generalising the interpretation, the assumptions might be made based on a knowledge of basic scientific premises or the validity of a related animal model. If researchers attempted to generalise the interpretation but the scientific principles underlying that generalisation did not hold (eg, the validity of the animal model for describing human physiology), then the assumptions would be violated, and the inferences in the study would not be representative. In either case, the way to truly test whether the assumptions held would be to estimate the effect of interest in the target population. While we often generalise in estimate because designing a study in the target population would
not be feasible, we generally consider such a study necessary to prove hypotheses regarding generalisation of interpretation, especially when the sample is highly removed from the target population (eg, cell line vs human population).

Thirdly, a natural extension of generalising the (overall or stratum specific) estimate to a target population are methods to estimate the overall mean of an outcome or the average effect of a treatment on an outcome (rather than a stratum specific estimate).
Even further, any study that is representative in interpretation could theoretically be made representative in estimate if all relevant effect measure modifiers were measured and accounted for; however, that is not always possible when the study sample is distant from the target population (e.g., laboratory mice to humans).

Fourthly, the concepts of representativeness and generalisability discussed above also relate to the term “applicability” used in certain risk-of-bias tools, such as the PROBAST and QUADAS.16 17 All concepts centre on the idea that it is important to assess a study and its results in terms of how well they can be related to some target population. While we discussed causal and descriptive studies in this article, the two tools mentioned apply this concept to predictive and diagnostic studies.

Finally, one question that has been raised is whether generalising the interpretation or the estimate is intrinsically more important for health research and for science broadly. It could be argued that generalising the interpretation is the primary aim of scientific inference and thus should be our goal in most studies.1 The underlying premise is that the goal of science is the discovery of universal knowledge about nature that will hold true in most instances. If we view health research from this viewpoint, then generalisation of the interpretation is what matters. By contrast, generalisation of the estimate can never be universal. The estimate obtained in a particular study sample will always be tied to a specific scientific or public health question, and the study design and will vary based on the distribution of key covariates across time and populations. However, to inform policies and interventions in the real world, we must be able to predict health outcomes in human populations beyond those we studied. Therefore, generalisation of the estimate (whether obtained via study design or analytical methods) is an important goal. A further argument could be that these endeavours of statistical inference are just as informative for science as the inferences above. Science can be about discovering laws of nature; it can also be about discovering laws of nature; it can also seek to understand particular facets of nature. For some areas of health research, such as epidemiology and other population health sciences, the facet of nature under study is disease as it occurs in humans at a population level, and true understanding of the disease under study will be contextualised by time, place, history, and social environment. Consideration for how these factors have changed from the original setting to some new time or target population and how these changes might affect the estimate obtained is critical.

While such theoretical debates are important, our comprehensive definition of representativeness does not treat either generalisation of estimate or interpretation as inherently more relevant. That evaluation largely depends on the research question and study in the target population.5 13 14 While the study sample might not be representative of the target population as observed, it could be made representative by using methods for generalisability or transportability, such as weighting or standardization to control for the key covariates or effect measure modifiers that differ between the samples.15 These approaches require measuring and accounting for all relevant key covariates, meeting certain identifiability conditions, and often making model specification assumptions.13 14
design at hand. Health researchers both develop the universal knowledge related to the health of populations and investigate how that knowledge can be applied to improve the health of populations, and the two ends of the research spectrum are fundamentally linked. What is important, then, is that researchers are clear on the manner in which their results can be applied to the target population when they say their study is representative and the assumptions underlying that statement.

Conclusions

We have established the idea that a study sample can be representative of a target population if one of the following is true: the estimate obtained in the study sample is generalisable to the target population or the interpretation of the study results is generalisable to the target population. Whether a study sample can be representative of a target population through the first definition depends on the study design or whether the variables affecting the outcome (which could be effect measure modifiers of the effect of interest) have been measured. On the other hand, even in the absence of simple random sampling or measurement of all key covariates, we can say that the study is representative in terms of its interpretation, direction of effect, or inference, because this requires less stringent assumptions than generalisation of the study estimate. The example studies provided give guidance on how one might determine whether the study sample from different types of research is representative and whether, for the specific research question, generalising the estimate or the interpretation was the priority.

Twitter jacqueline E Rudolph @jerudolph13

Contributors All authors were responsible for the concept of the article. JER wrote the original draft. YZ, PD, SHM, and BL reviewed and edited the manuscript. JER is guarantor. The corresponding author was commissioned; externally peer reviewed.

Funding This work was supported in part by National Institutes of Health grants R01-CA250851, R01-Alz10240, U10-DA036977, and R01-DA057673. The funders had no role in the writing of the report or decision to submit the article for publication.

Competing interests All authors have completed the ICMJE uniform disclosure form at http://www.icmje.org/disclosure-of-interest/ and declare: support from the National Institutes of Health for the submitted work; all authors declare no financial relationships with any organisations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data sharing not applicable as no datasets generated and/or analysed for this study.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/4.0/.

ORCID iD
Jacqueline E Rudolph http://orcid.org/0000-0001-7177-1847

REFERENCES

2 Nuhr EA, Olsen I. Commentary: epidemiologists have debated representativeness for more than 40 years – has the time come to move on? Int J Epidemiol 2013;42:1016–7. 10.1093/iju/dyt102