Redefining Representativeness of a Sample in Causal Terms[*]

Michał Sikorski · Alexander Gebharter · Barbara Osimani

Abstract

Despite its crucial role, sample representativeness remains a controversial topic in medical science methodology. There is an ongoing debate not only about how best to define and ensure the representativeness of a sample (e.g., Rudolph et al., 2023; Porta, 2016), but also about whether representativeness is worth pursuing at all (e.g., Rothman et al., 2013). We present a new definition of representativeness in terms of causal models and argue that it is more precise and more useful than existing alternatives. We use examples to demonstrate the types of evidence that can support assumptions of representativeness.

1 Introduction

Because in typical medical experiments it is not feasible to test all members of the target population, scientists rely on samples. This requires that the sample represents the target population at least to some degree. Surprisingly, despite its crucial role, representativeness remains a controversial topic within the methodology of medical science. It remains unclear not only how to best define and ensure representativeness, but also whether it is worth pursuing in the first place. In this paper we present a new definition of representativeness in terms of causally interpreted Bayesian networks (CBNs). We argue that

[*] This is a draft paper. Please do not cite or quote without permission. The final version of this paper forthcoming under the following bibliographical data: Sikorski, M., Gebharter, A., Osimani, B. (fothcoming). Redefining representativeness of a sample in causal terms. *Journal of Evaluation in Clinical Practice.*

this new definition is not only more precise, but also provides more guidance than existing alternatives.

We critically discuss the accepted definitions of representativeness in section 2. In section 3, we introduce CBNs and our new definition of representativeness. In section 4, we compare our proposal to already existing alternatives.

2 Existing Definitions and their Deficiencies

In a recent article Rudolph et al. (2023) discusses two definitions of representativeness used in medical science and proposes a new one. In this section, we describe all three proposals and highlight some of their shortcomings. We assess the proposals with respect to how adequate they are – i.e., how well they capture the concept of representativeness in the way it is typically used – and their potential for guiding scientists in their search for representative samples.

2.1 Representatives as Random Sampling

According to the first definition, a sample is representative if

RR “[...] the study sample is a simple random sample of the target population (i.e., the sample that arises through representative sampling)” (Rudolph et al., 2023 p. 1)

Randomization secures the representativeness of a sample with a high probability if each unit of the sampled population has the same probability of being selected. Formally, if it can be assumed that “sample draws” are independent and identically distributed or, at least, exchangeable (see Cox, 2006 or Bernardo, 2001). Despite that, the desirability of representativeness along the lines of **RR** is controversial. Rothman et al. (2013) argued that representative samples should be avoided. It is often more fruitful to study many non-representative samples (e.g., w.r.t. age categories) because this might show an effect that could be hidden when data are pooled into a single representative sample. The majority of the commentators agreed with the conclusions of Rothman et al. They argued that simple random sampling is often costly to the point of being impossible and that the results achieved in non-representative samples are often generalizable (e.g., Nøhr and Olsen, 2013, Elwood, 2013 or Richiardi et al., 2013). Surprisingly, none of these authors argued for revising the definition of representativeness rather than abandoning the requirement of representativeness. A lesson one may draw from this discussion is that randomization is not necessary for representativeness. Also note that **RR** is only applicable to cases in which the sample is a subset of the target population. This implies that **RR** does not apply to extrapolation where the study sample is by definition not drawn from the target population (Rothman et al., 2013). Samples composed of non-human model organisms, fragments of tissues, or computer simulations, for example are paradigmatic in this respect, since they cannot be immediately apply to a target population composed of humans. But also without crossing boundaires in levels of inference, results from studies performed on a given population do not necessarily apply to other populations (for socio-anagraphic factors or other relevant sets of conditions). This seems to be inconsistent with how extrapolation of results across species is understood in science. Thus, **RR** may at best be a good definition of a specific kind of representativeness. Finally, what lessons for obtaining a representative sample can be learned from **RR**? To obtain a representative sample, one needs to endorse randomization procedures which, as the extensive discussion in (Rothman et al., 2013) and commentators shows, is often expensive, time-consuming, and in some cases impossible. Thus, collecting a random sample might be the right approach in some cases, but seems to be unpractical as a general strategy.

2.2 Representativeness as Similarity

Another definition discussed by Rudolph et al. (2023) characterizes representativeness in terms of similarity:

SR1 "[...] 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." (Rudolph et al., 2023 p.1)

The source of this definition is a dictionary of epidemiology (Last, 1983). Thus, it seems likely that this or a similar definition is widely assumed by researchers. The main problem with this definition is that it is underspecified. It is not clear which personal characteristics should be considered. If all of the possessed properties should be considered, then **SR1** seems to collapse to **RR**, as using a randomized sample seems to be the best way to reliably reproduce the distribution of properties among individuals of the target population. Consequently, **SR1** would share all the deficiencies of **RR**. If, on the other side, only some of the characteristics should be considered, then which properties are the relevant ones?

What are the practical consequences of **SR1**? The definition seems to suggest that scientists should look for a sample that a) shears some of the characteristics of the target population and that b) delivers results similar to those that would be expected in the target population. This does not seem to be particularly helpful for at least two reasons. Firstly, as already mentioned, it is not clear what characteristics are relevant. Secondly, to satisfy b) scientists need to determine if the results reached in the sample population will resemble the effects that would be obtained in the target population. Since this assessment is difficult before having carried out the study, also **SR1** does not seem to give useful advice.

The newest edition of the *Dictionary of Epidemiology* (Porta, 2016) presents a slightly different definition which is worth citing in full:

SR2 “REPRESENTATIVE SAMPLE A sample that to a large extent resembles a population of interest. The term *representative* as it is commonly used is largely undefined in the statistical or mathematical sense. The use of probability sampling will not ensure that a sample will be representative of the population in all relevant aspects. It is unwarranted to assume that if the sample resembles the reference population on factors that have been checked, no differences exist in other relevant factors.” (Porta, 2016 p. 247)

The comparison between **SR2** and **SR1** as paraphrased in (Rudolph et al., 2023) reveals interesting differences. Firstly, the results are not mentioned in the definition and therefore subclause b) cannot be derived from the new version. Thus, the definition avoids the second problem of the earlier version. The definition is, though, still underspecified. Hence, also **SR2** does not provide significant guidance to scientists. Additionally, **SR2** mentions that the concept of representatives is not well defined and that using probability sampling does not warrant representatives. This does not seem to apply to the random simple samples described above but to other probabilistic sampling methods such as cluster sampling, or stratified sampling that employs some of the properties to group participants.

2.3 Rudolph et al.'s Definition

A new definition developed by Rudolph et al. (2023) is an improved version of a definition in terms of similarity:

GR "We define a study sample to be representative of a well-defined target population if the results estimated in that sample are generalizable to the target population."(Rudolph et al., 2023 p. 1).

The definition captures the purpose behind collecting representative samples. Because it is usually not possible to check for each member of the target population whether the effect of interest occurs, we need such a sample as a basis for projecting the actually observed effect to the target population. Therefore, it seems to be almost trivially true that all samples should be representative in this sense. **GR** was partly motivated by the discussion of **RR** and succeeded in providing a much less controversial definition.

However, **GR** does not provide much guidance to scientists in their search for a representative sample. Though it highlights the necessity of an explicit description of the target population (further developed in the paper), it remains unclear how to determine whether the results observed in the sample can be generalized before the experiment is conducted and to what degree they can be projected to the target population.

3 A Causal Analysis of Representativeness

In this section, we reconstruct the debate concerning the desirability of representativeness (understood as in **RR**) on the basis of causally interpreted Bayesian networks (CBNs). Graphical causal models such as CBNs and causal structural equation models (SEMs) systematically link complex causal structures to probabilistic dependence. They can be used for representing causal structures, predicting causal effects based on observation and intervention, and inferring causal structure based on observational and experimental data.[1] (For details on CBNs see, e.g., Pearl, 2000; Spirtes et al., 1993.) From a causal perspective, both proponents and critics of **RR** are partially right.

[1] This is an advantage over more traditional approaches such as Rubin's causal model. We chose CBNs rather than SEMs because they are more general, in that they can also work with nonparametric assumptions, i.e., without knowledge of functions and distributions

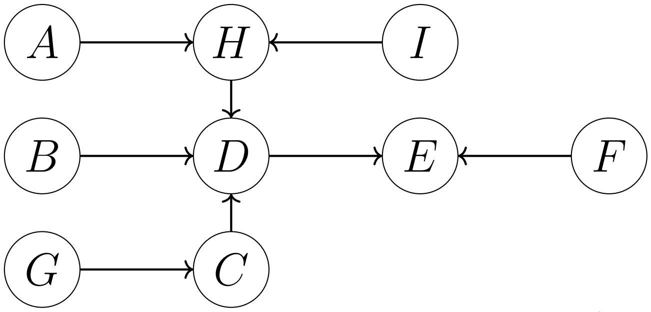


Figure 1: Simple exemplary causal model

To see why, let us construct a simple CBN $\langle \mathbf{V}, \mathbf{E}, Pr \rangle$ featuring the variables $A, B, C, D, E, F, G, H, I$. We assume that the model's graph is the one in Figure 1 and that all direct causal relations (represented by arrows) indicate positive causes, meaning that the higher the values of the cause variables, the higher the probabilities for higher values of the effect variables. In addition, we assume that A, B, C are interactive effect modifiers. In particular, we assume that they boost each others' influences on D, meaning that the higher the values of each of these variables are, the stronger their individual and joint influence is on D.

RR implies that all features of the individuals in a representative sample are distributed among these individuals as they are distributed among the individuals of the target population. Let us assume that our study's goal is to investigate the effect of A on D. Let us further assume that we have a sample s_1 in which the values of all nine variables are indeed distributed among the individuals as they are in the target population. Suppose that B- and C-values are skewed towards very low values. That is, most units in the sample appear to manifest very low values for B and C. Thus, carrying out our study based on s_1 will result in a weak positive effect of A on D. This effect can be expected to be the same as in the target population (random error excluded). Note that for this result it does not matter whether we measured any of the variables except A and D because due to randomization it can be assumed that, excluding sampling errors, the distributions of covariates are probabilistically equivalent in the two populations.

Let us now consider an alternative sample s_2. We assume that s_2 differs from s_1 insofar as B- and C-values are much more equally distributed among individuals than they are in s_1. As a consequence, we can expect that G- and

E-values are on average higher (since G and D are positive causes of C and E respectively) and that D-values are on average higher (because we assumed that A, B, C boost each other's influence on D). s_2 is clearly not representative of the target population according to **RR**. Now we can reconstruct Rothman et al.'s 2013 observation in terms of our simple model. In particular, we can specify a specific causal scenario in which a non-representative sample seems clearly better than a representative one. Based on s_2 we can learn more about A's potential effect on D and about how to optimally utilize A for controlling D than we can learn from the representative sample s_1. The reason is that, unless the size of the sample is large enough the rareness of individuals with other than very low B- and C-values in sample s_1 does not allow us to condition on these other B- and C-values because we do not have enough data to reliably infer how these B- and C-values would influence A's efficacy w.r.t. D. Not so in the non-representative sample s_2. Since in this sample we find all B- and C-values instantiated sufficiently often by individuals, we have access to this information and are able to obtain a richer understanding of the counterfactual dependence of D on A given different values of B and C. This speaks in favour of Rothman et al. Note that uncovering the patterns of how D counterfactually depends on A requires to having measured B and C.

Which sample is better, s_1 or s_2? We think that depends. Rothman et al.'s 2013 argument appears to conflate two distinct purposes a sample might have: (i) making a reliable inference about A's effect on D in the target population vs. (ii) understanding how D counterfactually depends on A when varying other factors such as B and C. Sample s_1 allows for (i), but not for (ii) due to B- and C-values, other than very low ones, being underrepresented in s_1. Sample s_2, on the other hand, allows for (ii) but, at least without further information about the target population, not for (i). While sample s_1 being representative (according to **RR**) guarantees that any observed effect of A on D in the sample will be the same in the target population, s_2 does not. Since in s_2 B- and C-values are distributed differently than in the target population (by assumption), it can – though allowing us to study how A would hypothetically influence D to the background of different B- and C-values – not be used to make reliable predictions about A's effect on D in the target population as long as we do not know in addition how exactly B- and C-values are distributed in the target population. Thus, which type of sample to prefer depends on one's specific goals: inference vs. understanding. Hence, the analysis shows that certain samples such as

s_1 are unsuitable for generating understanding in terms of counterfactual dependence, but does not provide evidence that a non-representative sample is generally preferable. Representative samples seem necessary to ensure generalizable results, at least as long as the distribution of other direct causes of the purported effect of interest is not known for the target population in which case a non-representative sample such as s_2 is always preferable. Since this information is usually not available for the target population and this is the context in which representativeness actually does matter, we will assume for the remainder of the paper that such knowledge about the distribution of other factors in the target population is not available to the scientist.

Moreover, the example demonstrates that **RR** is unnecessarily strong. One does not have to have all variables' values distributed in the sample as they are distributed in the target population. For example, in a study investigating the effect of A on D, other direct causes of D not lying on a path from A to D such as B and C should be distributed in the sample as they are in the target population, as we just saw. Also other variables that might distort A's impact on D due to influencing causes of D that are at the same time effects of A such as I need to be distributed equally. All other variables can in principle be distributed differently (see also recent graph theory contributions to the topic: Pearl and Bareinboim, 2011; Tennant et al., 2021; Webster-Clark and Breskin, 2021; Webster-Clark et al., 2024). More distant causes of direct causes of D such as G, for example, do not matter for A's impact on D in the target population as long as possible disturbing factors such as B, C, and I are equally distributed in the sample and the target population. The same goes for A, H, and D because we control for A in the experiment and H's probability is determined by A and I while D's is determined by its direct causes H, B, and C. Thus, bringing about A in the population will have the same effect on D as it has in the sample as long as B, C, and I are equally distributed. The same reasoning as for D applies to effects such as E. Finally, variables that are not causes of D such as F can also be distributed differently since their distribution will not play any role in evaluating A's impact on D.

Summarizing, if one wants to infer a variable's causal impact on another one in the target population and does not have additional knowledge about how the effect's direct causes are distributed in the target population, a representative sample is necessary. But not a fully representative sample in the sense of **RR**. We only require causes of the effect variable that might distort A's impact on D to be distributed equally in the sample as they are in the target population.

3.1 A New Definition of Representativeness

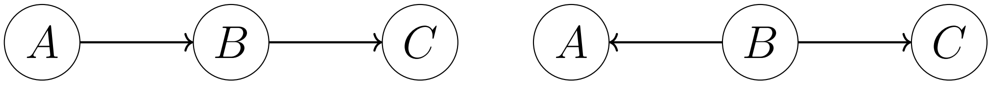
In this section, we present a new causally informed definition of representativeness based on the findings from section 3. The purpose of this definition is to specify as weak as possible conditions that are sufficient for generalizing a finding in a sample to the target population. So far, we always implicitly assumed that the causal structure underlying the sample is the same as the one underlying the target population. This is, however, not at all a trivial assumption (for evidence see, e.g., Cartwright and Hardie, 2012). To see why, let us discuss another simple example. Assume we are interested in A's impact on C. Assume further that B mediates between A and C. Suppose we have a sample s in which all values of causes of A, B, C are distributed as in the target population. Finally, let us assume that the causal structure underlying s is the one shown in Figure 2(a) while the one underlying the target population is the one in Figure 2(b).

Both structures are compatible with the exact same probability distributions. By assumption, the sample is representative in the sense of **RR**. However, results obtained from the sample s might get things horribly wrong about the target population. For example, assume we did a randomized controlled trial (RCT) based on our sample and found that A has a strong impact on C.[2] Naturally, we infer that the same holds for the target population. However, since A is not causally relevant for C in the target population, A will give us no control at all over C. To avoid situations like these, representativeness requires that the causal structure underlying the sample is the same as the one underlying the target population. If we also take into account our findings from section 3 about which variables need to be distributed equally, we arrive at the following definition:

CR A sample is representative of a target population w.r.t. C's causal effect on E if (i) both share the same causal structure (and parameters[3]) and (ii) the values of all variables in $\mathbf{Z}$ are distributed in the sample as they are in the target population, where $\mathbf{Z}$ is the set of all variables Z

[2] For details on the difference between observation and intervention (as in an RCT) see, for example, (Pearl, 2000).

[3] A variable's parameters are the probabilities of its values given all possible value combinations of its direct causes.



(a) (b)

Figure 2: Causal structures of the sample (a) and the target population (b)

different from C and E that are (ii.a) direct causes of E not lying on a directed path from C to E or (ii.b) direct causes of a variable X lying on a directed path from C to E not themselves lying on a directed path from C to E.

Our proposal pinpoints how the sample has to resemble the target population to guarantee that the obtained findings are generalizable to the latter. Consequently, **CR** can be seen as a natural development of both **GR** and **SR2**. Moreover, the definition is more precise and formal than its predecessors. Condition (i) reflects that the sample needs to have the same causal structure as the target population, while (ii) summarizes our findings from section 3: Only variables need to be distributed in the sample as in the target population that could distort C's impact on E. These are exactly the direct causes of E that are not effects of C (condition (ii.a)) as well as direct causes of variables on a path from C to E (condition (ii.b)).

Example 1: Negative Side-Effects of Efavirenz in the Population of Zimbabwe[a]

Efavirenz is a relatively cheap drug that prevents the replication of HIV and, therefore is effective in treating and preventing HIV/AIDS. The drug was tested in the U.S. and approved by the U.S. Food and Drug Administration in 1998. The degree to which its safety can be generalized beyond the population of the U.S. is crucial for the prospects of using the drug overseas (see Park et al., 2023).

Tested Effect: Negative Side-Effects of Efavirenz

Target population	Sample population
Population of Zimbabwe	Population of U.S.

Assessment of Causal Similarity:

The causal structure underlying the effects (including side effects) of Efavirenz is consistent across all humans. However, the prevalence of causally relevant factors varies among populations. For instance, a mutation in the CYP2B6 gene, which impedes the metabolization of Efavirenz, is much more common in Zimbabwe than in the U.S. (see Masimirembwa et al., 2016).

Assessment of Representativity:

Despite the shared underlying causal structure, differences in critical factors compromise the representativity of results obtained in the U.S. for the population of Zimbabwe. However, it has been argued that dosage adjustment may effectively counter these differences (Nyakutira et al., 2008), potentially making the drug as safe for use in Zimbabwe as it is in the U.S.

[a] The source of this example is (Park et al., 2023).

Example 2: Effects of Alcohol Use Disorder in Humans

Alcohol use disorder remains a prevalent and serious problem. Some of its aspects, such as underlying molecular mechanisms, remain poorly understood. Studies of the disorder in humans are changing due to practical and ethical reasons.

Tested Effect: Effects of Alcohol Use Disorder

Target population	Sample population
Humans	*Caenorhabditis elegans* worms

Assessment of Causal Similarity:

Studies have demonstrated that the causal structures underlying the physiological (e.g., Yu et al., 2011) and behavioral (e.g., Salim et al., 2021) effects of alcohol consumption, as well as Alcohol Use Disorder, are consistent in both humans and *Caenorhabditis elegans* worms. For example, corticotropin-releasing factor receptors contribute to the development of alcohol-seeking behavior in both humans and *Caenorhabditis elegans* (see Salim et al., 2021). Some causally relevant factors, particularly those related to the size of the organization, will need adjustment through tuning. However, it is likely that making these adjustments will not pose significant challenges.

Assessment of Representativity:

Given the current state of scientific knowledge regarding the causal structure underlying the effects of alcohol, it appears that results obtained in *Caenorhabditis elegans* will be representative of humans. For instance, research involving these worms will likely be valuable in studying the health consequences of Alcohol Use Disorder and testing pharmacological interventions aimed at treating it.

Example 3: Efficiency of Tumor necrosis factor blockers on Sepsis

Tumor necrosis factor (TNF) is believed to play a significant role in fatalities associated with sepsis. Consequently, medications that lower TNF are expected to improve the survival rates of patients suffering from sepsis. However, experimental findings have yielded surprisingly mixed results, revealing heterogeneity among sepsis cases (see e.g., Lorente and Marshall, 2005 or Marshall, 2014) .

Tested Effect: Efficiency of TNF neutralization.

Target population	**Sample population**
Patients suffering from sepsis caused by *Escherichia coli* infection	Patients suffering from sepsis caused by *Streptococcus pneumoniae* infection

Assessment of Causal Similarity:

Despite the substantial similarity in symptoms, different cases of sepsis are caused by different types of bacteria (e.g., Escherichia coli or Streptococcus pneumoniae). Due to differences in bacterial characteristics, such as Gram type, these pathogens respond differently to TNF-neutralizing medications. This, in turn, results in distinct causal structures for each type of sepsis.

Assessment of Representativity:

Due to significant differences in the causal structure of different types of sepsis, it cannot be assumed that patients suffering from one type will be representative in how they respond to TNF neutralization compared to patients with another type. This explains the observed mixed empirical results.

Example 4: Penicillin Toxicity in Bats

Scientists are striving to save the last population of critically endangered tropical species of bats dying due to a bacterial disease. They want to test if penicillin is a safe drug for bats.

Tested Effect: Penicillin Toxicity

Target population	**Sample population**
Tropical Bats	Guinea Pigs

Assessment of Causal Similarity:

The gut flora of guinea pigs is primarily composed of gram-positive organisms, which is atypical among mammals. Since the composition of intestinal flora causally influences the effects of penicillin consumption, it is reasonable to assume that the causal structure underlying the effects of penicillin in guinea pigs differs from that in bats (see Green, 1974).

Assessment of Representativity:

The causal structure underlying the effects of penicillin is likely different in guinea pigs than in bats. Consequently, studies using guinea pigs will not yield results that are representaive of bats.

4 Discussion

CR naturally translates into methodological instructions. To show that a sample is likely to be representative a scientist should: a) present evidence suggesting that the causal structure is the same in the sample and target population, and b) present evidence showing that causes that might distort the impact the variable intervened on has on the purported effect are distributed similarly in the sample and the target population. Evidence concerning a) can come in the form of scientific theories describing how the tested cause will affect the members of the sample and target population. For example, if the best available genetic and physiological theories (see Examples 2 and 3) predict that the causal structure is the same in the case of both populations, we have a strong reason to believe that a) is satisfied in a given case. Justification for b) should show that all (known) factors possibly distorting the intervention variable's purported effect are similarly distributed in both populations. This evidence can take the form of empirical results showing that the crucial properties are similarly distributed.

Checklist for justification of representativeness

No.	Item	
1.a	Present evidence demonstrating that the sample and the target population share the same underlying causal structure with respect to the tested variable C's effect on E.	□
1.b	Present evidence demonstrating that the causal structures of the sample and target populations exhibit the same parameters representing the strength of the connections between related variables. For example, general biological or physiological theories have demonstrated that the underlying processes (e.g., metabolism of the tested substance) are the same in both populations. Evidence showing the presence of functionally analogous relevant organs in members of both populations further supports this similarity etc.	□
2.a	Present evidence demonstrating that the values of direct causes of E that do not lie on a directed path from C to E – that is, causes of the tested effect E that are not caused by C – are the same in the sample as in the target population.	□
2.b	Present evidence demonstrating that the values of direct causes of a variable X that are not lying on a directed path from C to E are the same in the sample as in the target population. For example, empirical evidence showing that the values of causes of the tested effect (other than the tested cause) and the external causes of other variables on the path from C to E are comparable in both populations.	□
3.	Counterbalance any remaining significant differences in the parameters and variables mentioned above. For example, differences in body mass between the members of both populations can be accounted for by adjusting the dosage of the tested substance.	□

Another implication of **CR** is that there are two distinct ways in which a sample may fail to be representative. Firstly the causal structure that connects (or disconnects) studied variables may be different in the sample and target population. Secondly, the sample may be non-representative because some of the possible distorting variables are not distributed equally in both populations. It seems that the failure of the second kind may be amended, for example, by taking into account unbalances or counteracting them. As sketched in Example 2, reducing the dose of Efavirenz administered to the population in Zimbabwe may lower the prevalence of negative side effects to levels observed in the U.S. while preserving most of its antiviral efficacy (see Nyakutira et al., 2008 and Park et al., 2023 for discussion). At the same time, the failures of replicability of the first type seem to be harder to amend and perhaps require composing a new sample.

What are the limitations of **CR**? Having a sound and precise definition of representability is one thing, how to ensure representability in practice is another. It is a very difficult task to achieve certain knowledge about causal structure or to exclude confounding factors with certainty. However, **CR** provides more guidance for how to maximize certainty or track uncertainty. There is a vast literature on inferring causal structure based on observational data and there are ways to account for confounders, selection bias, and other disturbing factors (e.g., Richardson and Spirtes, 2002; Spirtes et al., 1999; Zhang, 2008). Though causal search algorithms will usually not output a single causal structure, but rather a set of structures compatible with the data, this might still be useful. Such results can be complemented by expert evaluations or background knowledge such as information about temporal order. Thus, even though the correct causal structure as well as disturbing factors can usually not be identified with certainty, **CR** might still guide scientists in evaluating the suitability of a sample for an inference task or choosing the most promising sample when possible. Among other things, **CR** provides awareness for the causal pitfalls that come with simply ignoring its implications.

What are the consequences of the new conceptualization for the relativity of representativeness? In line with Rudolph et al. (2023), **CR** predicts that a sample is representative only of a specific target population; therefore, this target population must be well described. Moreover, it implies that representativeness is relative to the variable intervened on and the tested effect(s). This is plausible: A sample may be representative of a target population w.r.t. one effect but not w.r.t. another. For example, the *Caenorhabditis el-*

egans worms appear to be representative of humans in terms of the effects of alcohol consumption but not with respect to the effects of adaptation to low temperatures. In contrast to Nøhr and Olsen, 2013 and Olsen, 2013, **CR** does not imply that representativeness is relative to time and location. A sample will remain representative as long as the relevant causal structure and causally relevant properties in the target population remain unchanged.

Finally, we would like to discuss the broader implications of the results of this paper. We demonstrated that knowledge of the causal structure can be used to guide and simplify the search for a representative sample. We do not need to replicate all features of the target population in the sample (as in **RR**), but only those that could causally distort the intervened on variable's impact on the purported effect. At the same time, we do not claim that the causal approach is the only way to go. Another plausible approach would be to redefine the notion of similarity (between the sample and the target population) in a mathematically precise non-causal manner, for example, along the lines of Douven et al. (2021). Regardless of the specific method chosen, we are confident that we have demonstrated the viability of a more formal approach.

References

Bernardo, J. M. (2001). The concept of exchangeability and its applications.

Cartwright, N., & Hardie, J. (2012). *Evidence-based policy: A practical guide to doing it better*. Oxford University Press.

Cox, D. R. (2006). Principles of statistical inference.

Douven, I., Elqayam, S., Gärdenfors, P., & Mirabile, P. (2021). Conceptual spaces and the strength of similarity-based arguments. *Cognition*, *218*.

Elwood, J. M. (2013). Commentary: On representativeness. *International journal of epidemiology*, *42 4*, 1014–5.

Green, R. H. (1974). The association of viral activation with penicillin toxicity in guinea pigs and hamsters 1. *The Yale Journal of Biology and Medicine*, *47*, 166–181.

Last, J. (1983). *A dictionary of epidemiology*. Oxford University Press.

Lorente, J. A., & Marshall, J. C. (2005). Neutralization of tumor necrosis factor in preclinical models of sepsis. *Shock*, *24*, 107–119.

Marshall, J. C. (2014). Why have clinical trials in sepsis failed? *Trends in molecular medicine*, *20 4*, 195–203.

Masimirembwa, C. M., Dandara, C., & Leutscher, P. D. C. (2016). Rolling out efavirenz for hiv precision medicine in africa: Are we ready for pharmacovigilance and tackling neuropsychiatric adverse effects? *Omics : a journal of integrative biology*, *20 10*, 575–580.

Nøhr, E. A., & Olsen, J. (2013). Commentary: Epidemiologists have debated representativeness for more than 40 years–has the time come to move on? *International journal of epidemiology*, *42 4*, 1016–7.

Nyakutira, C., Röshammar, D., Chigutsa, E., Chonzi, P., Ashton, M., Nhachi, C. F. B., & Masimirembwa, C. M. (2008). High prevalence of the cyp2b6 516g→t(*6) variant and effect on the population pharmacokinetics of efavirenz in hiv/aids outpatients in zimbabwe. *European Journal of Clinical Pharmacology*, *64*, 357–365.

Olsen, J. (2013). Random sampling - is it worth it? *Paediatric and perinatal epidemiology*, *27 1*, 27–8.

Park, A., Steel, D., & Maine, E. (2023). Evidence-based medicine and mechanistic evidence: The case of the failed rollout of efavirenz in zimbabwe. *The Journal of Medicine and Philosophy*, *48*, 348–358.

Pearl, J. (2000). *Causality* (1st ed.). Cambridge University Press.

Pearl, J., & Bareinboim, E. (2011). External validity and transportability: A formal approach. *JSM Proceedings*, 157–171.

Porta, M. (2016). *A dictionary of epidemiology*. Oxford University Press.

Richardson, T., & Spirtes, P. (2002). Ancestral graph Markov models. *Annals of Statistics*, *30*(4), 962–1030.

Richiardi, L., Pizzi, C., & Pearce, N. (2013). Commentary: Representativeness is usually not necessary and often should be avoided. *International journal of epidemiology*, *42 4*, 1018–22.

Rothman, K. J., Gallacher, J. E., & Hatch, E. E. (2013). Why representativeness should be avoided. *International journal of epidemiology*, *42 4*, 1012–4.

Rudolph, J. E., Zhong, Y., Duggal, P., Mehta, S. H., & Lau, B. (2023). Defining representativeness of study samples in medical and population health research. *BMJ Medicine*, *2*.

Salim, C., Kan, A. K., Batsaikhan, E., Patterson, E. C., & Jee, C. (2021). Neuropeptidergic regulation of compulsive ethanol seeking in c. elegans. *Scientific Reports*, *12*.

Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, prediction, and search* (1st ed.). Springer.

Spirtes, P., Meek, C., & Richardson, T. (1999). An algorithm for causal inference in the presence of latent variables and selection bias. In *Proceedings of the 11th conference on uncertainty in artificial intelligence* (pp. 499–506). Morgan Kaufman.

Tennant, P. W., Murray, E. J., Arnold, K. F., Berrie, L., Fox, M. P., Gadd, S. C., Harrison, W. J., Keeble, C., Ranker, L. R., Textor, J., et al. (2021). Use of directed acyclic graphs (dags) to identify confounders in applied health research: Review and recommendations. *International journal of epidemiology*, *50*(2), 620–632.

Webster-Clark, M., & Breskin, A. (2021). Directed acyclic graphs, effect measure modification, and generalizability. *American Journal of Epidemiology*, *190*(2), 322–327.

Webster-Clark, M., Ross, R. K., Keil, A. P., & Platt, R. W. (2024). Variable selection when estimating effects in external target populations. *American Journal of Epidemiology*, kwae048.

Yu, X., Zhao, W., Ma, J., Fu, X., & Zhao, Z. J. (2011). Beneficial and harmful effects of alcohol exposure on caenorhabditis elegans worms. *Biochemical and biophysical research communications*, *412 4*, 757–62.

Zhang, J. (2008). Reasoning with ancestral graphs. *Journal of Machine Learning Research*, *9*, 1437–1474.