Social Science & Medicine xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Social Science & Medicine

journal homepage: www.elsevier.com/locate/socscimed

The "average" treatment effect: A construct ripe for retirement. A commentary on Deaton and Cartwright

S.V. Subramanian[a,b,*], Rockli Kim[a], Nicholas A. Christakis[c,d,e]

[a] Department of Social and Behavioral Sciences, Harvard T.H. Chan School of Public Health, Boston, MA, USA
[b] Harvard Center for Population & Department Studies, Cambridge, MA, USA
[c] Department of Sociology, Yale University, New Haven, CT, USA
[d] Department of Medicine, Yale University, New Haven, CT, USA
[e] Yale Institute for Network Science, Yale University, New Haven, CT, USA

"Don't cross a river if it is (on average) four feet deep".
-Nassim Nicholas Taleb, 2016 p.160

1. Introduction

When summarizing or analyzing a population, regardless of whether it consists of hundreds or millions of individuals, it is the norm in most social, medical, and health research to characterize it in terms of a single number: the *average*. The reliance on average is pervasive in descriptive, explanatory, or causal analyses. There is nothing inherently wrong with an "on average" view of the world. But whether such a view is actually meaningful, for populations or individuals, is another matter. The average can obscure as much as it illuminates. It is a lean summary of a distribution with no recognition of the rich variation between and within populations that is necessary to ascertain its relevance. And, on rare occasions, when summaries of variation are presented in analyses of populations in epidemiology or clinical trials, they are often simply and incorrectly labeled "error."

In this issue, Angus Deaton and Nancy Cartwright (hereafter, Deaton and Cartwright) provide a comprehensive assessment and critique of the use of Randomized Controlled Trials (RCTs) in the social sciences (Deaton and Cartwright, 2018). Their insights and critique are equally applicable to biomedical, public health, and epidemiologic research. Here, we elaborate on one aspect of the problem that Deaton and Cartwright mention in their essay, namely, that inference exclusively based on "Average Treatment Effect" (ATE) can be hazardous in the presence of excessive heterogeneity in responses. This inferential problem applies both for the study population – those with the same characteristics as the trial population, including even individuals within the trial itself – and the larger population of interest the intervention targets. While the latter (*i.e.*, the issue of external validity in RCTs) has received considerable attention, including by Deaton and Cartwright, the former remains sidelined even as it underscores the intrinsic importance of variation in any population.

Instead of expecting ATE from an RCT to work for any individual or population, Deaton and Cartwright argue that we can do better with "*judicious use of theory, reasoning by analogy, process tracing, identification of mechanisms, sub-group analysis, or recognizing various symptoms that a causal pathway is possible*" (Deaton and Cartwright, 2018). Their hypothetical example of an RCT based on a classroom innovation in two schools, St Joseph's and St Mary's, is most intuitive in this regard. Deaton and Cartwright argue that even if the innovation turns out to be successful on average, actual experiences in the school with comparable composition may be more informative when other schools decide to adopt and scale up the same innovation (Deaton and Cartwright, 2018).

Following a brief introduction to the problems of averages, we elaborate on why variation or heterogeneity matters from a substantive perspective and develop a generalized modeling framework to assessing "Treatment Effect" (TE) based on two constructs of a population distribution: the average *and* the variance. We show that existing, but woefully under-utilized, methodologies can be routinely applied to enhance the relevance and interpretation of TE in a population. We refer to treatment as a shorthand for any deliberate intervention and not just in the strict medical sense. We focus on RCT settings here because both the mean and the variance in the outcome of interest are expected to be equivalent at baseline due to randomization and any differential in the post-treatment variation clearly indicates something systematic. However, the points we raise in this commentary applies equally, and in fact more importantly, to analysis of observational data.

2. The fallacy of averages

There is nothing innately problematic about focusing only on the mean to summarize a distribution, provided it has some substantive meaning and application to the real world. The yawning gap between a statistical average and its application to the real world of individuals is well recognized (Christakis, 2014). For illustration, we present two examples from Todd Rose's thought-provoking book, "*The End of Averages*" (T. Rose, 2016).

* Corresponding author. Professor of Population Health and Geography, Harvard Center for Population & Development Studies, 9 Bow Street, Cambridge, MA 02138, USA.
E-mail address: svsubram@hsph.harvard.edu (S.V. Subramanian).

https://doi.org/10.1016/j.socscimed.2018.04.027

In 1942, in a quest to discover an "ideal" form of a woman, Dr. Robert L. Dickinson (an obstetrician) and Mr. Abram Belskie (a sculptor) decided to measure ~15,000 young adult women on 9 body dimensions (*e.g.*, height, bust, waist, hips, thigh, calf, ankle, foot, weight) and, based on the "average" across each, sculpted a female form called "Norma" (Creadick, 2010). They then decided to launch a contest, "*Are you Norma*?", encouraging women to submit their bodily dimensions. Of almost 4000 submissions received, how many resembled Norma on all 9 dimensions? Exactly zero. Indeed, Norma represented a misguided ideal that was both highly desirable yet impossible to observe. What was the impact of this exercise? Instead of confronting the individual variability around constructs of "normality", most doctors and scientists concluded that American women were physically unfit (T. Rose, 2016).

The second example illustrates an even more consequential case. During World War II, the United States Air Force aircrafts were crashing at a higher-than-expected rate even though no mechanical and human errors could be detected. After much probing, the Air Force commissioned a study in 1950 to design a better fitting cockpit based on the average of more than 4000 pilots on 140 body measurements. Yet, when Lieutenant Gilbert S. Daniels did an exercise to see how many pilots fit the so called "average pilot" on 10 dimensions (*i.e.*, height, sleeve length, crotch height and length, and circumferences for chest, vertical trunk, hip, neck, waist and thigh), the answer was, yet again, zero (Daniels, 1952; T. Rose, 2016). Yes, even in such an evidently homogeneous group of airmen, it was impossible to find even one individual who fit the average on all dimensions, even when the average was generously defined as falling within the middle 30 percent of the range of values for each of the dimensions. Essentially, by designing the cockpit to fit the average airman, it was ensured that it fit no one. Daniels concluded, "*It is virtually impossible to find an "average airman" in the Air Force population […] not because of any unique traits in this group of men, but because of the great variability of bodily dimensions which is characteristic of all men*" (Daniels, 1952 p. 1).

3. The reality of variation

The above illustrative examples point to an important limitation concerning ATE even in an ideal RCT. For the ATE to be truly meaningful even within the limited trial sample population, we argue, two dimensions need to be considered.

First, there should be a systematic and a statistically significant difference in the average outcome between the Treatment and the Control groups in the expected direction (*i.e.*, treatment, on average, had the intended effect). If this occurs, the trial is considered a success and, after few repeated demonstrations of a similar ATE, is usually followed by recommendations for scaling up intervention.

A second consideration of equal importance is: of the sample population that received the treatment, what percentage actually experienced the intended effect? Stated differently, what is the regularity or predictability with which individuals in the Treatment group experienced the desired effect? In the extant literature, this dimension is completely ignored. Consider two successful RCTs, both showing systematic differences in ATEs. However, in RCT 1, 90% of the individuals in the Treatment group experience the desired effect while in RCT 2 only 10% of the individuals in the Treatment group experience any therapeutic benefit. The remaining individuals in both groups are either unaffected or experience changes in the unintended direction. Assuming these are two types of treatments intended to have a similar effect, which one of these would we consider more successful overall? Arguably, the treatment from RCT 1! The substantially higher degree of regularity and predictability with which the treatment worked in RCT 1 not only is desirable because the ATE now is more meaningful as it applies to a majority, it also suggests a better understanding of who are more susceptible to the treatment, and potentially the mechanism of "why" it works, and the judiciousness in designing the treatment.

We consider a toaster to be working if it is able to toast the bread every time it is used. One does not take solace from the claim that the bread will pop up toasted, say, 2 out of every 10 times. In clinical settings, however, if a drug works 20% of the time in RCT, compared with 5–10% for a placebo, it is often accepted to be "effective" (Christakis, 2008). For instance, among the top 10 highest-grossing drugs in the United States, Humira, Enbrel, and Remicade each works for 1 in 4 people who take them, and Nexium only works for 1 in 25 people who take it for heartburn. Statins are effective in lowering cholesterol for as few as 1 in 50 individuals (Schork, 2015). The truth, therefore, is that, most people taking RCT-validated, effective treatments derive no benefit from them; even in the study population (let alone the larger real-world population) (Christakis, 2008). As clinicians struggle in their efforts to understand low adherence to several prescribed medication regimens, it is worth considering if the low adherence is because patients realize that the medication does *not* work for them. In fact, the growing recognition that the effectiveness of different treatments are vetted for the actual individual patient has motivated "precision medicine" and N-of-1 trials (Schork, 2015).

The case for recognizing individuals and the variability that is observed between individuals in matters of health was eloquently made by Stephen Jay Gould in his classic commentary, "*The median isn't the message* " (Gould, 1985). In this personal story of statistics written after Gould was diagnosed with abdominal mesothelioma, an incurable disease with a median mortality of only eight months, he noted two important aspects about statistical distributions. First, the distribution of experiencing adverse events is more likely to be heavily skewed than normally distributed. Second, the distribution may alter when circumstances change. Gould embodied these characteristics as he lived for 20 highly productive years after the initial diagnosis (and extremely competent surgery).

Another example concerns why doctors tend to offer "Do Not Resuscitate" orders to AIDS patients at much higher rates than to patients with advanced liver cirrhosis even though these two conditions might have equal average prognoses (Wachter et al., 1989). It might be tempting to conclude that doctors are more eager to avoid resuscitation in AIDS patients, perhaps for discriminatory reasons. But the real reason might be that the *variance* in survival in the AIDS group is much higher, and there may be many more patients in that group who will die imminently. It may be to this fact (*i.e.*, the greater variance) that the doctors are more oriented rather than to the average survival of the two groups; the doctors may reason that they can wait to offer DNR orders to the cirrhosis patients (Christakis, 2014).

Most "successful" (*i.e.*, a "statistically significant ATE" in the expected direction) social, health, and medical interventions, we speculate, will be characterized by such poor regularity and certainty with which the treatment works among those who have received the treatment. Closing the gap between a robustly estimated, but mythical, "average" and its ability to say anything meaningful about the constituents of both the trial population as well as the real-world population has to be an integral part of any scientific endeavor that claims to be "useful" in its motivation and inference.

4. Why this fixation with averages?

The origins of use of average to describe a characteristic or trait in a population appears to trace back to Adolphe Quetelet's 19th century notion of "*l'homme moyen*" or the "average man" (Krieger, 2012; Porter, 1985; Quetelet, 1842). This metaphor of "average man" was derived from the fields of astronomy and meteorology where the results of observations from multiple observatories were combined to determine a star's celestial coordinates. Quetelet argued that the distribution of a population's characteristics composed of "deviations" or "errors" resulting from the imperfect variations of individuals is analogous to the data produced by each observatory in astronomy, and hence can inform a population's true (inherent) value (Krieger, 2012).

This centuries-old concept of average as the most salient statistical summary as well as the target of inference for scientific research continues. Geoffrey Rose, whose insightful and succinct writings provided the foundation for a "population perspective" to health, argued explicitly that the basis for population health approach should focus primarily in explaining and modifying the mean/average values on a given variable of interest (*e.g.*, blood pressure, cholesterol, body weight, and alcohol intake) that differ across populations (G. Rose, 1989, 1991). The singular reliance on comparing means in two distributions (*e.g.*, Treatment and Control groups in RCTs) is based on the assumption that, as Rose put it, "*within each population the spread between the two extremes is rather similar*" (G. Rose, 1989 p. 411–2), *i.e.*, the assumption of "homoscedasticity" to use the regression jargon (Goldfeld and Quandt, 1965).

It is not an exaggeration to state that whether the assumption of homoscedasticity holds post-treatment (with variance in the outcome measure being the same in the Treated and the Control populations) is rarely tested and reported. But, this has consequential implications for the interpretation of the ATE. Indeed, when the expectation that variance within a population is the same across time or that variance is the same across different populations was systematically tested in observational settings with body mass index (BMI) as the focal variable of interest, it was found to be not true (S. Kim et al., 2017; Krishna et al., 2015; Razak et al., 2013; Razak et al., 2016; Vaezghasemi et al., 2016). Mean increases in BMI in a population were accompanied by increases in the variance over time (Block et al., 2013; S. Kim et al., 2017; Krishna et al., 2015; Stenholm et al., 2015); and, mean differences in BMI across populations were also accompanied by different variances within them (Collaboration, 2016; R. Kim et al., 2018; Razak et al., 2013; Vaezghasemi et al., 2016).

So why are researchers fixated with "ATE"? Because that is typically all we can statistically observe and estimate. In most research that involves "treating" individuals, what we are really interested is Individual Treatment Effect (ITE), the difference in expectations under treatment condition and control condition for a given individual (Wiedermann, 2016). Of course, that is impossible to observe and estimate in a causal sense; often, referred to as the "fundamental problem of causal inference" (Holland, 1986). In a typical trial, individual *i* receives either the treatment or a different exposure value (no treatment), and hence it is impossible to simultaneously observe the outcomes for both conditions on *i* (Holland, 1986). Therefore, identifying ITE constitutes a sort of "missing data" problem, with half of the potential outcomes missing (Subramanian et al., 2007).

It appears that our singular focus on ATE arises because that's all we can do and not because the average has some substantive primacy in hierarchy of inferential targets. It reminds one of the "Streetlight Effect" critique of scientific research, *i.e.*, when it is extremely difficult or even impossible to cleanly measure the object of real importance, scientists instead measure what they can, hoping it turns out to be relevant (Freedman, 2010). ATE is the statistical solution that researchers routinely accept in place of ITE, but the real answers are more likely to be hiding in the underlying distribution of ITEs.

In epidemiological research, however, there is a substantive argument that is presented to justify the sole focus on ATE. It has proved almost impossible to demonstrate any relation between an individual-level risk factor and outcome within a given population, whereas strong associations can be found between population mean values and outcome incidences (G. Rose, 2001). The neglect of individual heterogeneity or ITE is underpinned by the arguably widespread role of stochastic and random processes that influence morbidity and mortality (Davey Smith, 2011; G. Rose, 2001). For instance, Davey Smith critiquing the momentum in research focused on epigenetics and precision medicine writes, "*Several lines of evidence suggest that largely chance events, from the biographical down to the sub-cellular, contribute an important stochastic element to disease risk that is not epidemiologically tractable at the individual level.*" (Davey Smith, 2011 p.537).

Indeed, attesting to this view, in observational studies, less than 2% of the inter-individual variations in women's BMI was explained by basic socioeconomic factors (R. Kim et al., 2018) and less than 1% of the variability in child anthropometric status and growth failure was explained by conventional risk factors routinely conceptualized as representing successful interventions (Mejía-Guevara et al., 2018). Poor "discriminatory accuracy" (*i.e.*, the accuracy with which an intervention can discriminate who experiences the desired outcome) (Pepe et al., 2004; Ware, 2006) has been shown for well-known categories such as race/ethnicity for predicting influenza vaccine uptake (Mulinari et al., 2017) and traditional risk factors such as blood pressure, BMI, diabetes mellitus, cholesterol, and cigarette smoking for predicting coronary heart disease (Merlo et al., 2017).

At the same time, to dismiss ITE as simply down to luck (Davey Smith, 2011) is problematic as it is always individuals who experience disease and mortality, not populations.

5. Generalized framework for assessing treatment effect

If the heterogeneity between individuals is indeed all stochastic (and "natural") (Davey Smith, 2011; Davey Smith et al., 2016), then there should be no distinct patterning in the magnitude of variation between individuals within a defined population. In light of recent observational studies suggesting that the variation itself appears to be patterned for many health and well-being measures, we extend the following framework to better assess TE in RCT. In an RCT, the variance in the outcome measure in both the Treatment and Control groups should remain the same at the baseline (at least under the assumption that they are drawn from the same population). Indeed, one attractive feature of the RCT design is that not only will means in the outcome of interest and other covariates be the same across Treatment and Control groups, but also will the variances at the baseline (and any difference will be simply due to chance). With such a starting point, however, if the magnitude of variation in outcome differs across Treatment and Control groups at the follow-up (post-treatment), that would imply something more systematic happening.

The case to consider heterogeneity (between- and within-populations) can be routinely implemented within the well-known regression framework. Consider the following familiar linear regression model:

$$y_i = \beta_0 + \beta_T T_i + e_{0i} \quad (1)$$

where y_i is, say, BMI for individual *i*; β_0 represents the mean value of BMI for individuals in the Control group; β_T represents the mean BMI differential for individuals in the Treatment group T_i. The term e_{0i} represents the residual for each individual *i*, and assuming an Identical and Independent Distribution (IID) a variance $\sigma^2_{e0}$is estimated. The parameter σ^2_{e0} is the amount of variation that remains unexplained, after accounting for the TE or T_i. It is also assumed that σ^2_{e0} is the same in the Treatment and Control groups. Much of what is presented from an RCT is simply the β_T.

At most, heterogeneity in the TE is assessed in the following form:

$$y_i = \beta_0 + \beta_T T_i + \beta_1 x_{1i} + \beta_2 T_i * x_{1i} + e_{0i} \quad (2)$$

In Model (2), the ATE is now allowed to be different for different sub-groups of x_{1i}. The parameter β_2 is estimated for an interaction between T_i and x_{1i} specified in the fixed part of the regression model. While Model (2) certainly is more desirable, it still is rooted in the worldview of ATE, except now we have multiple ATEs for sub-groups instead of one ATE across all the population. Further, there are reasons why such an approach is often not advocated for RCT data (Dahabreh et al., 2016). A substantive equivalent of Model (2) would be to stratify the population by x_{1i} and specify separate Model (1) for each sub-group. To our knowledge, this is the current "state of the art" to incorporating heterogeneity into ATE assessments. While there is an increasing tendency to calculate standard errors allowing for the possibility that residual variances may be different in the Treatment and Control groups

(Deaton and Cartwright, 2018), σ^2_{e0} (the parameter summarizing the within-population variability) is still considered a "nuisance" parameter under this approach.

Treating heterogeneity as a characteristic of the population that is of substantive, intrinsic interest requires shifting our focus to the so-called "random" part of a regression model. Following Goldstein (2005), a generalized approach to incorporating the second moment of distribution into our regression would be as follows:

$$y_i = \beta_C C_i + \beta_T T_i + e_{Ci} C_i + e_{Ti} T_i \quad (3)$$

In Model (3), separate coding is used (*i.e.,* a dummy variable for both Control and Treatment groups) and hence there is no intercept (Goldstein, 2005). β_C now estimates the mean BMI for individuals in the Control group while β_T directly estimates the mean BMI for individuals in the Treatment group. Models (1) and (3) are substantively identical in being able to ascertain the ATE (*i.e.*, is β_T different from β_C in a meaningful way). However, what distinguishes Model (3) from the conventional Model (1) is the presence of two "residual" terms – one for observations in the Control group ($e_{Ci} C_i$) and one for observations in the Treatment group ($e_{Ti} T_i$). Model (3), thus, allows heteroscedasticity in unexplained variation by modeling two separate variances: σ^2_C : variance in BMI between subjects in the Control; and σ^2_T : variance in BMI between subjects in the Treatment. With separate mean and variance estimated for the outcome in the Treatment and Control groups, Model (3) allows us to conceptualize a 3 × 3 typology of TE based on an integration of all possible ATEs and variance in outcomes for the two groups (Table 1). A similar typology has been proposed previously to consider changes that occur both in the overall shift (average effect) and in the shape of the curve (standard deviation) as a result of polices and interventions (Benach et al., 2011, 2013). However, increase or decrease in variability does not necessarily translate to changes in social inequalities; we extend this framework in a more general sense and discuss its applicability for different units of analysis.

The great majority of RCTs typically assess the three ATE possibilities under the assumption of $\sigma^2_T = \sigma^2_C$ (Row 1, Table 1), with a goal often to show a desired effect of the ATE (*i.e.*, $\beta_T > \beta_C$). However, each ATE possibility has substantially different implication depending on the associated change in variance in the outcome. While a trial that has null effect on average ($\beta_T = \beta_C$) and same variance ($\sigma^2_T = \sigma^2_C$) (Type 1A) can be concluded as having truly no TE (neither harmful nor beneficial), increase in the Treatment group variance (1B) indicates that the treatment actually worked extremely well for some but made others considerably worse off. Conversely, even with the same null ATE, decrease in variance in the Treatment group (1C) implies that the treatment had fairly consistent effect on the outcome for most individuals.

Consider now the possibility (2B) where even though it is a "successful" trial ($\beta_T > \beta_C$), the Treatment group variance *increased* post-treatment ($\sigma^2_T > \sigma^2_C$), suggesting that the treatment had very heterogeneous effect. This could result from a differential TE by patient characteristics (*i.e.,* effect modification) and/or by different parts of the underlying risk distribution (*i.e.,* quantile regression) (Kent et al., 2010). Under such scenario, should the treatment still be "scaled" up? Perhaps the most desirable type of TE would be (2C) where ATE is positive (, $\beta_T > \beta_C$*i.e.*, a successful trial in a traditional sense) *and* variance in the outcome is also reduced in the Treatment group ($\sigma^2_T < \sigma^2_C$). This suggests that the treatment works regularly and predictably for a great majority of the participants, which is a key factor in scaling up decisions.

Similarly, even among "unsuccessful" trials where treatment has an ATE that is in opposite direction to what was intended ($\beta_T < \beta_C$), different lessons can be learned from those resulting in a larger variance ($\sigma^2_T > \sigma^2_C$) versus a smaller variance in the Treatment group ($\sigma^2_T < \sigma^2_C$). The former (3B) means that certain individuals experience therapeutic benefits even as others have considerably worse outcomes post-treatment, while the latter (3C) indicates that the treatment consistently had a harmful effect for almost everyone.

We should note that modeling heterogeneity in TE is not restricted to individual level, and Model (3) is equally applicable to populations when treatment is at the "cluster" level:

$$y_j = \beta_C C_j + \beta_T T_j + u_{Cj} C_j + u_{Tj} T_j \quad (4)$$

where, *j* represents the cluster. In Model (4), it is assumed that the unit of randomization and unit at which outcome (*y*) is measured is at the cluster/population/group level. Another variant of Model (4) would be Model (5):

$$y_{ij} = \beta_C C_j + \beta_T T_j + u_{Cj} C_j + u_{Tj} T_j + e_{0ij} \quad (5)$$

where, while unit of randomization is at the cluster/population/group level, the unit at which outcome (*y*) is measured is individual such that multiple individuals within *j* receive the *same* treatment giving the classic multilevel model (Goldstein, 2011; Subramanian et al., 2003).

Regardless, the substantive point made here can be easily applied to all these models; essentially, letting the TE vary at its own level. It is important to note that what we outline above is different from the typical "random slopes" in a multilevel model, where ATE associated with C_{ij} and T_{ij} (or any individual-level exposure in observational setting) is modeled to vary across *j* cluster/population/group units (Subramanian, 2004).

Indeed, the conditions that would lead to 1 of the 9 possibilities in Table 1 will depend on aspects such as the nature of the intervention, the population being studied, the outcome measured, and many of the issues that Deaton and Cartwright raise in their essay. Our goal here was to provide a framework to initiate a discussion that allows us to systematically and in a generalized manner anticipate and model heterogeneity both at the population and individual levels. In addition, clearly reporting the percentage of individuals in Treatment group who experience the desired effect as well as the relevant discriminatory accuracy statistics can substantially improve the interpretation of the regularity and predictability of TEs in RCTs.

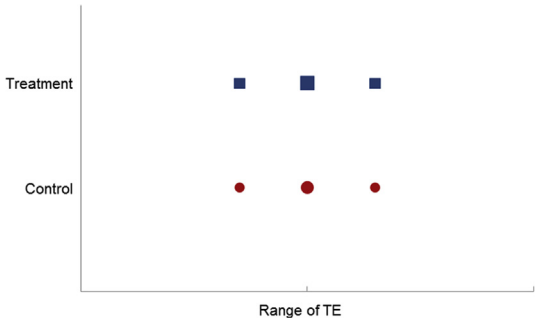
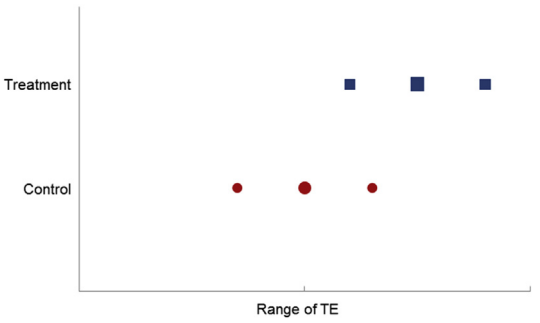
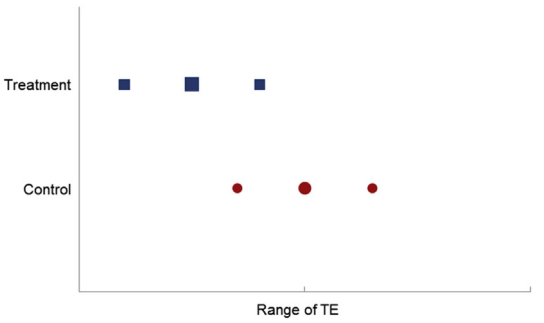
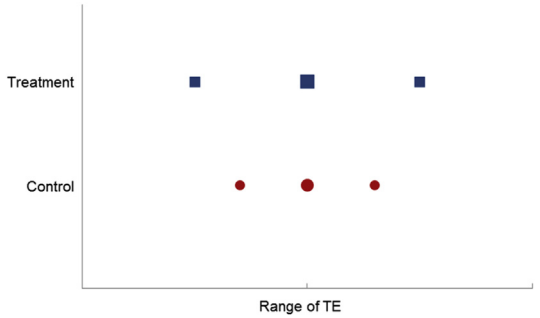
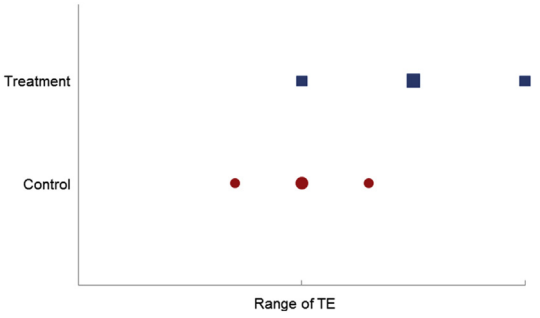
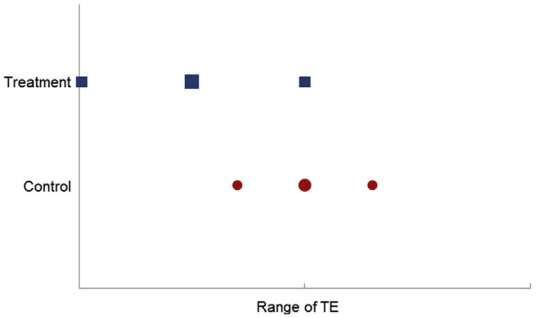
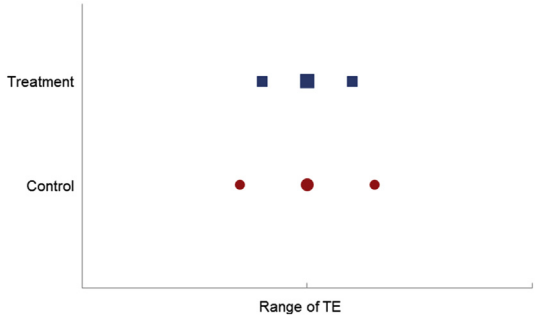
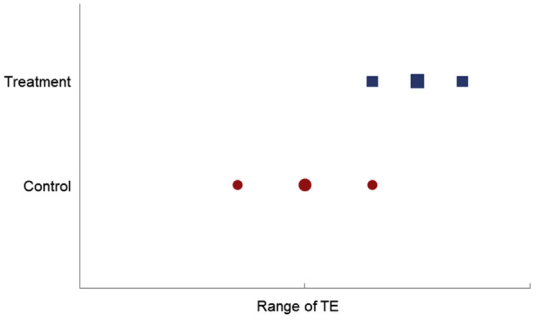
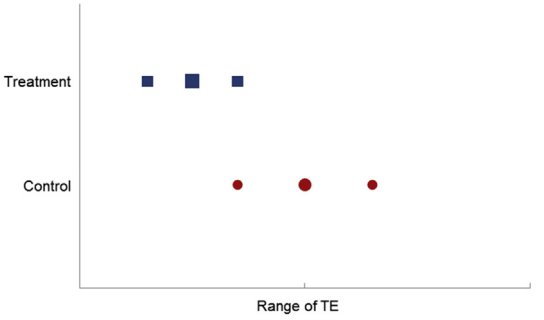
6. Concluding remarks

As Deaton and Cartwright rightly argue, to strip the challenging aspects of "*why*" something works from the question of "what works" inadvertently undermines the very challenge and enterprise of scientific inquiry. The thoughtful issues raised by Deaton and Cartwright on the scope and limits of RCTs in assessing "what works" provides a catalyst to substantively and methodologically incorporate ideas of heterogeneity at individual and population levels. While more complex trial designs attempt to better capture the differential TEs – conditional ATE marginalized over covariates, interaction tests, ATEs by subgroups – the fundamental focus still remains fixated in estimating ATEs. It is far less recognized that the relevance and interpretation of ATE depends on the definition of populations and heterogeneity in ITEs. As Gould (1985) reminded us: "*Variation is the hard reality, not a set of imperfect measures for a central tendency. Means and medians are the abstractions.*" Much of epidemiologic research and any research aimed at improving health and well-being at individual and population levels could do well to recognize this.

Acknowledgements

We are grateful to Etsuji Suzuki and Craig Duncan for their feedback on our manuscript. SVS is also grateful for the helpful comments and encouragements of participants at lectures given on this subject at University of Southampton, Pontificia Universidad Javeriana, Harvard University, Seoul National University, Princeton University, University of Turku, Stockholm University, City University New York, and Chinese University of Hong Kong.

Table 1
Typology of treatment effect (TE) based on combinations of possible average treatment effects (ATEs) and variance in outcome in the Treatment and Control groups.

	Type 1: $\beta_T = \beta_C$	Type 2: $\beta_T > \beta_C$	Type 3: $\beta_T < \beta_C$
Type A: $\sigma_T^2 = \sigma_C^2$	**1A.** 	**2A.** 	**3A.** 
Type B: $\sigma_T^2 > \sigma_C^2$	**1B.** 	**2B.** 	**3B.** 
Type C: $\sigma_T^2 < \sigma_C^2$	**1C.** 	**2C.** 	**3C.** 

Note. For simplicity we hold ATE and variance in outcome to be constant for the Control group, but they may also change post-treatment.

References

Benach, J., Malmusi, D., Yasui, Y., Martínez, J.M., 2013. A new typology of policies to tackle health inequalities and scenarios of impact based on Rose's population approach. J. Epidemiol. Community Health 67, 286–291.

Benach, J., Malmusi, D., Yasui, Y., Martínez, J.M., Muntaner, C., 2011. Beyond Rose's strategies: a typology of scenarios of policy impact on population health and health inequalities. Int. J. Health Serv. 41, 1–9.

Block, J.P., Subramanian, S., Christakis, N.A., O'Malley, A.J., 2013. Population trends and variation in body mass index from 1971 to 2008 in the framingham heart study offspring cohort. PLoS One 8 e63217.

Christakis, N.A., 2008. Does this work for you? Br. Med. J. 337, 1025.

Christakis, N.A., 2014. What Scientific Idea Is Ready for Retirement? the Average. online essay. www.edge.org.

Collaboration, N.R.F, 2016. Trends in adult body-mass index in 200 countries from 1975 to 2014: a pooled analysis of 1698 population-based measurement studies with 19· 2 million participants. Lancet 387, 1377–1396.

Creadick, A.G., 2010. Perfectly Average: the Pursuit of Normality in Postwar America. Univ of Massachusetts Press.

Dahabreh, I.J., Hayward, R., Kent, D.M., 2016. Using group data to treat individuals: understanding heterogeneous treatment effects in the age of precision medicine and patient-centred evidence. Int. J. Epidemiol. 45, 2184–2193.

Daniels, G.S., 1952. The "Average Man"? Technical Note Number WCRD 53–57. Wright Air Development Center: Air Force Aerospace Medical Research Lab Wright-Patterson AFB, OH.

Davey Smith, G., 2011. Epidemiology, epigenetics and the 'Gloomy Prospect': embracing randomness in population health research and practice. Int. J. Epidemiol. 40, 537–562.

Davey Smith, G., Relton, C.L., Brennan, P., 2016. Chance, Choice and Cause in Cancer Aetiology: Individual and Population Perspectives. Oxford University Press.

Deaton, A., Cartwright, N., 2018. Understanding and misunderstanding randomized controlled trials. Soc. Sci. Med in press.

Freedman, D.H., 2010. Why scientific studies are so often wrong: the streetlight effect. Discover Magazine 26.

Goldfeld, S.M., Quandt, R.E., 1965. Some tests for homoscedasticity. J. Am. Stat. Assoc. 60, 539–547.

Goldstein, H., 2005. Heteroscedasticity and complex variation. Encycl. Stat. Behav. Sci. 2, 790–795.

Goldstein, H., 2011. Multilevel Statistical Models. John Wiley & Sons.

Gould, S.J., 1985. The median isn't the message. Discover 6, 40–42.

Holland, P.W., 1986. Statistics and causal inference (with discussion and rejoinder). J. Am. Stat. Assoc. 81, 945–970.

Kent, D.M., Rothwell, P.M., Ioannidis, J.P., Altman, D.G., Hayward, R.A., 2010. Assessing and reporting heterogeneity in treatment effects in clinical trials: a proposal. Trials 11, 85.

Kim, R., Kawachi, I., Coull, B.A., Subramanian, S.V., 2018. Patterning of individual heterogeneity in body mass index: evidence from 57 low- and middle-income countries. Eur. J. Epidemiol. https://doi.org/10.1007/s10654-018-0355-2.

Kim, S., Subramanian, S., Oh, J., Razak, F., 2017. Trends in the distribution of body mass index and waist circumference among South Korean adults, 1998–2014. Eur. J. Clin. Nutr. 1.

Krieger, N., 2012. Who and what is a "population"? Historical debates, current controversies, and implications for understanding "population health" and rectifying health inequities. Milbank Q. 90, 634–681.

Krishna, A., Razak, F., Lebel, A., Smith, G.D., Subramanian, S., 2015. Trends in group inequalities and interindividual inequalities in BMI in the United States, 1993–2012. Am. J. Clin. Nutr. 101, 598–605.

Mejía-Guevara, I., Corsi, D.J., Perkins, J.M., Kim, R., Subramanian, S.V., 2018. Variation in anthropometric status and growth failure in low-and middle-income countries. Pediatrics 141 e20172183.

Merlo, J., Mulinari, S., Wemrell, M., Subramanian, S., Hedblad, B., 2017. The tyranny of the averages and the indiscriminate use of risk factors in public health: the case of coronary heart disease. SSM-Population Health 3, 684–698.

Mulinari, S., Wemrell, M., Rönnerstrand, B., Subramanian, S., Merlo, J., 2017. Categorical and anti-categorical approaches to US racial/ethnic groupings: revisiting the National 2009 H1N1 Flu Survey (NHFS). Crit. Publ. Health 1–13.

Pepe, M.S., Janes, H., Longton, G., Leisenring, W., Newcomb, P., 2004. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. Am. J. Epidemiol. 159, 882–890.

Porter, T.M., 1985. The mathematics of society: variation and error in Quetelet's Statistics. Br. J. Hist. Sci. 18, 51–69.

Quetelet, A., 1842. A Treatise on Man and the Development of His Faculties: Now First Translated into English. William and Robert Chambers.

Razak, F., Corsi, D.J., Subramanian, S., 2013. Change in the body mass index distribution for women: analysis of surveys from 37 low-and middle-income countries. PLoS Med. 10 e1001367.

Razak, F., Smith, G.D., Subramanian, S., 2016. The idea of uniform change: is it time to revisit a central tenet of Rose's "Strategy of Preventive Medicine"? Am. J. Clin. Nutr. 104, 1497–1507.

Rose, G., 1989. High-risk and population strategies of prevention: ethical considerations. Ann. Med. 21, 409–413.

Rose, G., 1991. Ancel keys lecture. Circulation 84, 1405–1409.

Rose, G., 2001. Sick individuals and sick populations. Int. J. Epidemiol. 30, 427–432.

Rose, T., 2016. The End of Average: How to Succeed in a World that Values Sameness. Penguin UK.

Schork, N.J., 2015. Personalized medicine: time for one-person trials. Nature 520, 609–611.

Stenholm, S., Vahtera, J., Kawachi, I., Pentti, J., Halonen, J.I., Westerlund, H., et al., 2015. Patterns of weight gain in middle-aged and older US adults, 1992–2010. Epidemiology 26, 165.

Subramanian, S., 2004. The relevance of multilevel statistical methods for identifying causal neighborhood effects. Soc. Sci. Med. 58, 1961–1967.

Subramanian, S., Glymour, M.M., Kawachi, I., 2007. Identifying Causal Ecologic Effects on Health: a Methodological Assessment. Macrosocial Determinants of Population Health. Springer, pp. 301–331.

Subramanian, S., Jones, K., Duncan, C., 2003. Multilevel Methods for Public Health Research: Neighborhoods and Health. Oxford University Press, New York.

Taleb, N.N., 2016. The Black Swan: the Impact of the Highly Improbable. Random house, US.

Vaezghasemi, M., Razak, F., Ng, N., Subramanian, S., 2016. Inter-individual inequality in BMI: an analysis of Indonesian family life surveys (1993–2007). SSM-Population Health 2, 876–888.

Wachter, R.M., Luce, J.M., Hearst, N., Lo, B., 1989. Decisions about resuscitation: inequities among patients with different diseases but similar prognoses. Ann. Intern. Med. 111, 525–532.

Ware, J.H., 2006. The limitations of risk factors as prognostic tools. N. Engl. J. Med. 355, 2615–2617.

Wiedermann, W., 2016. Statistics and Causality: Methods for Applied Empirical Research. John Wiley & Sons.