Randomized experiments and causal inference: Randomization balances the impact of confounders in the statistical sense

Mariusz Maziarz (corresponding author)

mariusz.maziarz@uj.edu.pl

ORCiD: 0000-0003-1979-0746

Jagiellonian University, Faculty of Philosophy, Institute of Philosophy

Grodzka 52, Kraków, Poland

Abstract

In their recent defense of randomization, Martinez and Teira (2022) endorsed Worrall's (2002; 2007) arguments that randomization does not assert the balance of confounding factors and delivered two other epistemic virtues of random assignment (efficiency balance and Fisherian balance). Worrall's criticism claiming that randomization does not assert Millean balance shape the philosophical debates concerned with the role of randomization in causal inference and evidence hierarchies in medicine. We take issue with Worrall's claim that randomization does not assert the balance of confounders. First, we argue that randomization balances the influence of confounders on an outcome in the statistical sense. Second, we analyze the potential outcome approach to causal inference and show that the average treatment effect (ATE) is an unbiased estimator of the average causal effect and observe that actual causal inferences rely on randomization balancing the impact of confounders.

Key-words: Randomized controlled trials, RCT, causal inference, randomization, potential outcomes approach, POA

Funding:

The work of Mariusz Maziarz on the first draft of the manuscript has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 805498). The additional research required by the Reviewers for *History and Philosophy of the Life Sciences* was funded by The National Science Centre, Poland, under grant number 2022/45/B/HS1/00183. For the purpose of Open Access, the author has applied a CC-BY public copyright license to any Author Accepted Manuscript (AAM) version arising from this submission.

1. Introductory Remarks

Recently, Martinez and Teira (2022) investigated the purpose of randomization. They reconstructed the notion of balance underlying Worrall's (2002) criticism of the view that randomization controls for both known and unknown confounders and argued that even though randomization does not achieve Millean balance (i.e., "identifying and equalizing ex ante confounding factors"), it is nevertheless beneficial for causal inference because it asserts other types of balance – efficiency balance ("equalizing the value of the antecedent conditions ex ante according to the efficiency they yield in the estimation of the treatment outcome") and Fisherian balance ("ex post influence of uncontrolled conditions"). As Martinez and Teira (2022) put it, "Worrall is right in showing that randomization does not provide a good warrant of experimental balance in Mill's sense. But for both frequentist and Bayesian statisticians, such understanding of balance is not necessary for causal inference, while randomization is not so easy to dispense with." (Martinez and Teira 2022, p. 5). However, Worrall's criticism of randomization is problematic because it involves the pre-statistical concept of Millean balance and applies it to the domain of randomized controlled trials in medicine that relies heavily on statistics. We show that still another goal of the experimenters (and possibly the most significant one) is to obtain an unbiased estimate of the average treatment effects (ATEs) and argue that randomization balances the summary influence of all confounders (a'la Mill) in the statistical sense, what contradicts Worrall's (2002) claim.

This claim concerned with whether randomization equalizes the influence of confounding factors on an outcome of interest is at the heart of the debate about the role of randomization in causal inference. This discussion has started in response to the insufficiently examined view of the early proponents of evidence-based medicine that the results of randomized trials are superior to those reported by observational studies, at least for the assessment of treatment efficacy. Disagreeing with Papineau (1994), Worrall (2002) notably argued that "[e]ven if there is only a small probability that an individual factor is unbalanced, given that there are indefinitely many possible confounding factors, then it would seem to follow that the probability that there is some factor on which the two groups are unbalanced (…) might for all anyone knows be high." (p. 324). Several fellow philosophers have endorsed his argument. For instance, Thomson (2011) criticized the use of RCTs for causal inference, repeating after Worrall that in actual clinical studies, treatment and control groups are heterogeneous because there are many confounders in medicine. He further added that "[i]n medicine, randomization is almost always gerrymandered (sampling is not from the entire relevant population, some individuals assigned to a sample are removed after the fact, samples are adjusted to eliminate relevant differences observed after sampling or known to be likely from past experience (…). In addition, the assumption of homogeneity that is reasonably robust in Fisher's agricultural work is absent in medicine (…)." Borgerson (2009) argued against the privileged position of RCTs by pointing out that actual RCTs have only finite samples and may not reflect the average treatment effect of an ideal RCT with an infinite sample. Finally, Worrall (2007) himself further supported the criticism of randomization by arguing that "[t]here is no reason to think that [(…) average treatment effects (ATE) observed in individual studies agree with] the 'limiting average'" (p. 465) observable only if one re-randomized infinitely.

Others disagree with Worrall's objections to randomization. Cartwright (2010) distinguished between ideal RCTs that, by definition, assert the equal distribution of confounding factors between the treatment and control groups and the actual studies. This argument was further extended by Deaton and Cartwright (2018), who criticized RCTs for their problems with extrapolation and analyzed the misunderstandings concerned with balance present in the literature but contended that randomization asserts the balance of confounders "in expectation". They also convincingly argued against Worrall's (2002; 2007) assumption that the equal distribution of each confounder is needed for sound inferences instead of balancing the average influence of confounders.

Similarly, Philippi (2022) responded to the objection of Worrall (2002; 2007) by pointing out that estimating accurate ATEs depends on the summary effect of confounders not differing significantly from zero instead of balancing each confounder. The most substantial criticism comes from the statistician discussing seven myths of randomization. In his rebuttal of Myth 2, Senn (2013a) differentiated between "a probability statement regarding the possible effects of possible imbalances (which is what with the usual statistical calculations provide) with a requirement for perfect balance (which does not exist)" and argued that any baseline imbalances do not undermine results due to the application of statistical testing. He further contended that Worrall (2002) confused a probability statement regarding the expected distribution of confounders with the requirement for perfect balance. He also rebuked Borgerson's (2009) claim by pointing out that the in-principle possibility of existing infinitely many confounders undermines the reason for randomization: even if there were an infinite number of confounders, it is their effect size that matters (see also Senn 2013b).

Still others agreed with Worrall's claim that balance in confounding factors is unattainable but defended randomization on other grounds. In *Philosophy of Evidence-Based Medicine,* Howick (2011) argued that randomization might not assert an equal balance of confounders due to limited sample sizes. Still, randomized studies are nevertheless better at this task than non-randomized research designs and hence deserve to be prioritized by the evidence hierarchies. La Caze et al. (2012) took issue with Worrall's (2002; 2007) criticism of the epistemic virtues of randomization, but they endorsed the purported falsity of the view that "random allocation controls for known and unknown confounders". Backmann (2017) pointed out that the potentially infinite number of confounding factors can be substantially limited on the basis of mechanistic evidence regarding treatment action. In his attempt to explain the evidence-based medicine (EBM) evidence hierarchy, La Caze (2009) argued that randomization asserts higher internal validity than non-randomized interventional studies and observational epidemiology. Later, La Caze (2013) delivered a Bayesian justification for the higher trustworthiness of results stemming from randomized interventional studies.

We argue that randomization balances the impact of confounders between the treatment and control groups (a'la Mill) in the statistical sense. That is, randomization asserts that the most likely division of participants into the treatment and control groups is such that the overall impact of confounders is equal and large imbalances unlikely (what overlaps with the 'in expectancy' balance claim defended by Deaton and Cartwright (2018), but it also allows for assessing the probability of deviations from the balance and certainty of estimates. The paper proceeds as follows.

First, we analyze the notion of 'Millean' balance and distinguish among the 'applicable in each case', 'in expectancy', and 'in the statistical sense' versions of the claim. Herein, we justify our defense of the latter and contrast it with Martinez and Teira's (2022) notion of quasi-Millean balance. Second, we use a toy example of a coin-flipping exercise to show that randomization balances the influence of confounding factors in the statistical sense. Third, we analyze how the variability of the impact of confounders reduces with sample size, discuss some quantitative approaches to measuring this variability, and analyze pre-study power analysis to show that researchers can assess the risk of being 'unlucky'. Fourth, we discuss the potential outcome approach to show that assessing treatment efficacy based on average treatment effect crucially relies on the balance of the effect of confounders between the treatment and control groups. Our analysis shows that the epistemic goal of randomization is to make the average treatment effect an unbiased estimator for the average causal effect of the intervention under test and allow for calculating the accuracy of the estimates. In concluding remarks, we analyze the implications of randomization balancing the impact of confounders 'in the statistical sense' for causal inference and the trustworthiness of conclusions supported with RCTs.

2. Explicating the balance claim

Let us begin with a more detailed look at Martinez and Teira's (2022) view on the claim that randomization does not control for the 'Millean' balance of confounding factors. They endorsed this position and the argumentation presented by Worrall (2002; 2007) (e.g., in the last paragraph of p. 5). But they also contended that biomedical researchers are committed to the view that randomization asserts 'Millean balance' (pp. 3-4). This raises the question of whether the researchers are wrong about randomization's vices or whether Worrall's arguments support a factually inaccurate position. The answer, in our view, depends on the exact meaning of the claim concerning balancing confounders.

On the one hand, the balance of confounders can be interpreted deterministically, as a claim that each randomization asserts that the distribution of confounders between treatment and control groups is strictly equal. Such a claim is blatantly false or, as Cartwright (2010) and others argued, holds only for the ideal RCT with an infinite sample. Furthermore, if the deterministic interpretation of the balance claim were true, there would be no need to conduct statistical hypothesis testing and calculate confidence intervals (CIs) since the observed $\widehat{ATE}$ would reflect the actual treatment efficacy.

On the other, the Millean balance can be taken as a probabilistic claim about the outcome of randomization, i.e., balancing confounders in expectancy-CONFUSES MODE WITH THE MEAN???. Both Worrall (2002; 2007) and Martinez and Teira (2022) seem to argue that randomization does not assert even the latter notion of balance, but they remain vague to some extent in this regard. Our reading results from Martinez and Teira's (2022, p. 5) endorsement of Worrall's arguments and their claim that randomization does not balance "ex ante value of each possible confounder" (p. 8) as we read *ex ante* in a way synonymous to *in expectancy*. The two clauses indicate that the balance claim is concerned with what randomization achieves in the long run (the mathematical expectancy) since it is the only meaningful way in which confounding can be studied before (*ex ante*) a trial and can be contrasted with 'ex post' analysis of whether a particular random allocation has achieved the equal distribution of a confounder or the overall impact of all confounders. Worrall is more apparent in his claim that RCTs do not control for the equal balance of confounders even probabilistically (see Fuller 2019). He contended that the defenders of randomization hold the weaker stance that randomization asserts the equal distribution of confounders 'in some probabilistic sense' (Worrall 2002, 322). Similarly, he claimed that "[t]he idea that randomization controls all at once for known and unknown factors (or even that it "tends" to do so) is a will-o'-the-wisp" (Worrall 2002, 328).

However, we should notice that the other reading also receives some support. For example, Martinez and Teira (2022, 4) wrote: "[i]f the confounders are unknown, there is no way for the experimenters to realize they have been unlucky [in a particular instance]. Hence randomization does not control for the lack of balance.

But this alternative reading shows a misunderstanding of the role of randomization in balancing confounders since each statistical inference is susceptible to being wrong. This misunderstanding can also be found in other voices in the debate (e.g., Borgerson 2009). It dates back to Worrall's (2002) misguided argumentation and motivates our defense and explication of the claim that randomization balances the influence of confounders in the statistical sense. The claim we defend differs from the notion of 'Millean Balance' as understood by Martinez and Teira in being neither about balancing each confounder in every randomization nor in expectancy, but about balancing the overall influence of all confounders, asserting that large deviations are unlikely and allowing for calculating their probability. It is also distinct from the notion of the balance claim defended by Deaton and Cartwright (2018), because statistical inferences are not only about having an unbiased estimator of the first moment (here: average treatment effect) but also calculating its accuracy and hence knowing the certainty of one's estimate.

Quasi-Millean balance, denoting the situation when "the relevant covariates in the two groups [are (…)] not too different on average (Martinez and Teira 2022, 14) seems to align with the statistical notion of balance defended in the paper. However, quasi-Millean balance requires relevant covariates (i.e., those potential confounders that have an impact on the outcome of interest) to be balanced. The emphasis on each confounder being balanced dates back to Worrall's (2002, p. 324) requirement that each 'individual factor' should be balanced for sound causal inferences. The phrase 'not too different on average' could be interpreted in terms of no statistically significant difference between the average values of a confounder in the treatment and control groups (e.g., the average value of age across the trial arms). Martinez and Teira (2022, 17) are right in observing that "Millean balance is a pre-statistical notion targeting the single run experiment in which all confounding factors are kept at the same value", but their notion of quasi-Millean balance remains too restrictive as it focuses on the balance of each confounder and hence unattainable by randomization. Instead, we argue that randomization balancing the summary influence of all confounders (balance in the statistical sense) is sufficient for the causal interpretation of average treatment effect (ATE) and that randomization achieves this notion of experimental balance. To reiterate, the notion of balance in the statistical sense is less restrictive than the quasi-Millean balance as it allows for some confounders to have different average values across the treatment and control groups because it defines balance in terms of the overall (summary) influence of all confounders.

The statistical notion of the balance claim is needed to shed light on the role of randomization in causal inference. Randomized controlled trials in medicine, experiments in psychology, and randomized field experiments in economics are primarily conducted to estimate the effects of tested interventions. To do so, the average treatment effects (ATEs) are estimated, which, as we argue below, are unbiased estimates of the causal effects only if the balance claim is accurate. If randomization did not balance the overall impact of confounders in the statistical sense, ATEs would be systematically off the true effect sizes. It is utterly convincing that "[t]he choice between these three competing notions of balance depends, crucially, on how the experimenter understands causal and statistical inference" (Martinez and Teira 2022, p. 17), but the potential outcomes approach (POA) seems to be the main framework for causal inference, and the statistical notion of the balance claim underlies the causal inferences aligning with POA (see section 5).

3. What randomization achieves

Now, let us consider a simple coin-flipping exercise to illustrate how randomization balances the impact of confounders in the statistical sense. Bernoulli process is the simplest random process and hence well shows what randomization achieves and explains the probabilistic nature of statistical inferences. Assessing the efficacy of an intervention in an RCT can be compared to the (simplified) case of checking if a coin is fair. If you encounter a coin and wonder if it is fair, you may want to proceed with an empirical test. If you flip it several times and get heads all the time, you are likely to conclude that the coin is biased despite a small probability of using a fair coin and being unlucky. Such a situation resembles the one mentioned by Martinez and Teira in the above quote: you are unable to conclude (with certainty) if you are unlucky and the coin is fair or whether it is biased. However, probability calculus and statistics allow for assessing the probabilities of these possibilities. Your best guess of the coin's chance of falling on heads (an unbiased estimator) is as follows:

$$\hat{p} = \frac{h}{h+t}$$

Where:

$\hat{p}$– the estimated probability of getting heads;

h - the number of heads in your coin-flipping exercise;

t – the number of tails in your coin-flipping exercise.

Obviously, due to the inherent randomness of the Bernoulli process, the estimated probability of getting heads ($\hat{p}$) may differ from the actual probability of a particular coin (p). For instance, if you flipped a fair coin four times, your chances are only 6/16 to get an accurate estimate. In fact, in 2 out of 16 possible outcomes, you will get four heads or tails in a row. However, in stark contrast with Martinez and Teira's assertion quoted above, experimenters expect the variability of the estimator and, in this case, would not draw any conclusions due to sample size making this coin-flipping exercise underpowered. Precisely the same can be said about an estimate of the $\widehat{ATE}$ (which is also an unbiased estimator of the average causal effect) in an RCT: due to random imbalances between the treatment and control groups, $\widehat{ATE}$ may diverge from the true efficacy. However, these divergences do not undermine the claim that randomization balances the influence of confounding factors in the statistical sense for the exact same reason that getting four heads in a row should not jeopardize your trust in a fair coin, at least if you endorse the threshold for statistical significance at the level of 0.05.

It is so for the reason that the use of statistics allows for quantifying the certainty of the estimate and what divergences can be expected with different probabilities. These calculations can also be used to design the Bernoulli trial (and RCTs) in such a way as to assert a required level of certainty. Calculating how many times you need to flip a coin to accurately estimate the probability of getting heads requires the following steps. First, you need to estimate the standard deviation of a random variable being a realization of the Bernoulli process as follows:

$$\sigma = \sqrt{p(1-p)}$$

Considering that the actual probability of getting heads is: $0 \leq p \leq 1$, and the function's maximum is for $p = 0.5$, this value can be endorsed for calculating sample size to stay on the safe side and not underestimate the variance (σ^2) of the random variable. The next step is to assess the variability of the estimator ($\hat{p}$), denoted by standard error SE (i.e., the standard deviation of the estimator of the first moment, which measures how far off the estimate is likely to be):

$$SE = \sigma/\sqrt{n}$$

Where:

σ – standard deviation;

n – the number of observations.

The standard error of the estimator depends positively on the standard deviation of an estimated variable and negatively on the number of observations in a sample used for estimation: the larger the sample size, the smaller the average deviation from the mean (p). If the sample size were infinite, SE would equal zero because the variability of the confounders' summary impact diminishes with the sample size. SE of the estimator of the true probability of getting heads ($\hat{p}$) is given by the following:

$$SE = \sqrt{p(1-p)/n}$$

For $p = 0.5$:

$$SE = \frac{1}{(2\sqrt{n})}$$

Assuming you are interested in estimating the 95% confidence interval with the maximum error of (0.1), you need to solve the following equation for n:

$$0.1 = \frac{1.9599^2}{(2\sqrt{n})}$$

Solving for n allows for concluding that you need to flip your coin at least 369 times to be 95% sure that the true chance of getting heads p does not differ from the value of your estimate $\hat{p}$ by more than 0.1. That is, the 95% CI for p includes the following values: $(\hat{p} - 0.1;\ \hat{p} + 0.1)$. This coin-flipping exercise shows that randomization does indeed balance the impact of confounders (here: physical forces determining the result of a flip) in expectancy. Still, it is not the only sense in which the 'balance' claim is valid. This is so because randomization additionally allows for estimating how large deviations are to be expected and estimating their probabilities. This is the sense of the claim that randomization controls for the influence of confounders in the statistical sense. The coin-flipping exercise shows that the Millean balance and Fisherian balance considered by Martinez and Teira (2022) are intertwined. If randomization controlled for Fisherian balance only but not for the influence of confounders (e.g., by miraculously leading to the biased division of participants in expectancy), then the estimator of the coin's probability of getting heads ($\hat{p}$) and $\widehat{ATE}$ in clinical trials would be systematically off from its actual value (p), which undermines inferences. Furthermore, this analysis shows that, in contrast to

Worrall's (2002; 2007) claims, slight imbalances between the treatment and control groups do not undermine inferences because of statistical hypothesis testing. For this reason, Worrall's (2002) claim seems to confuse a probability statement regarding the expected average impact and distribution of confounders with the requirement for perfect balance. To reiterate, if randomization asserted perfect balance, then there would be no need for statistical hypothesis testing.

4. Designing clinical trials

RCTs are designed in precisely the same way regarding sample size, which depends on the level of statistical significance, expected effect size, and outcome variability. However, the implications of how clinical trials are designed to minimize the risk of false inferences and maximize their epistemic value. Let us consider Greenland's (1990, 421) thought experiment:

"[s]uppose I wish to study whether lidocaine prophylaxis prevents death within the 72 hours following hospital admission for acute myocardial infarction. I will enroll two patients for this study, two successive admissions to a hospital emergency room. When the first patient is admitted, I will toss a fair coin: If heads, the first patient will receive lidocaine and the second will not; if tails, the second admission will receive lidocaine and the first will not. Suppose now that the first admission is massively compromised and is certain to die within 72 hours of admission, whereas the second is a mild case and is certain to survive, whether or not either of them receives lidocaine therapy."

The story shows that if the overall impact of confounders (here: the severity of heart attack) is not distributed equally across the treatment and control groups, then the inferences drawn from observing the difference in treatment effects between the two groups cannot be ascribed to the intervention such as lidocaine treatment. RCTs are planned to estimate the chances of being 'unlucky' (i.e., reporting a false-positive or false-negative result). The primary purpose of precision analysis (also known as pre-study power analysis) is to choose an appropriate sample size, given the expected power to detect the clinically meaningful difference and the cost of increasing the number of participants. The most common approach to data analysis is testing for equality in means, i.e., addressing the question of whether there is any statistically significant difference between the treatment and control groups. Despite some problems (see Lawler and Zimmermann 2021), the usual practice is to test the null hypothesis stating that the outcome of the treatment (μ_T) and control groups (μ_C) are equal ($H_0: \mu_T = \mu_C$) vs. the alternative ($H_1: \mu_T \neq \mu_C$). Under the i.i.d assumption (independent and identically distributed effects), when the variance of the primary outcome is known, the null is rejected at the significance level of α if (Cook & DeMets 2008, 115-139):

$$\left|\frac{\widehat{\mu_T} - \widehat{\mu_C}}{\sigma\sqrt{\frac{1}{n_T} + \frac{1}{n_C}}}\right| > z_{\alpha/2}$$

Where:

$\widehat{\mu_T}$ – estimated outcome in the treatment group;

$\widehat{\mu_C}$ – estimated outcome in the control group;

n_T – the number of participants in the treatment group;

n_C – the number of participants in the control group;

$z_{\alpha/2}$ – the upper α-th quantile of the standard normal distribution.

In case when H_0 is false, the power of the above test can be approximated as follows (see, for details, Chow et al. 2018, 47-49):

$$z_\beta = \frac{|\widehat{\mu_T} - \widehat{\mu_C}|}{\sigma\sqrt{\frac{1}{n_T} + \frac{1}{n_C}}} - z_{\frac{\alpha}{2}}$$

Where:

z_β – the β-th quantile of the standard normal distribution.

Assuming an equal number of participants randomized into the treatment and control groups, $n_T = n_C$, the solution is as follows:

$$n_T = n_C = \frac{\left(z_{\alpha/2} + z_\beta\right)^2 2\sigma^2\,|\widehat{\mu_T} - \widehat{\mu_C}|}{(\widehat{\mu_T} - \widehat{\mu_C})^2}$$

What follows, the minimal sample size needed for a study to obtain the power (β) with statistical significance threshold α depends on the absolute value of effect size ($|\widehat{\mu_T} - \widehat{\mu_C}|$) and the variance of the primary outcome σ^2. It is so because small samples lead to large standard deviations and undermine making statistically significant inferences. However, if the treatment effect is substantial compared to the overall impact of confounding factors, then even a small-size study may have power sufficient for establishing efficacy. Both power (β) and statistical significance threshold (α) are chosen conventionally, considering the costs (financial and ethical) of running a clinical trial with larger samples and reporting a statistically insignificant result despite the treatment's efficacy. One popular choice is the 80% and 5% threshold, respectively.

The effect size ($\widehat{\mu_T} - \widehat{\mu_C}$) and variance ($\sigma^2$) needs to be chosen considering a particular treatment under test. One method of choosing a value for expected effect size is to use the minimal clinically important difference, which denotes the weakest improvement relevant to patients (McGlothlin & Lewis 2014). Another approach is to base the estimate on other existing treatments or the results obtained in trials of an earlier phase.

Instead, the standard deviation of the variable chosen for the primary outcome can be known at the planning stage, which has significant implications for the discussion regarding the role of randomization. For example, if an RCT tests a drug for hyperglycemia, then the variability of plasma glucose level (measured in mmol/l) among diabetes patients should be estimated. This can be done by measuring glucose levels in a sample of patients. Considering that such an observational study does not test any intervention, the variance of glucose levels for the sample σ^2 results from the overall summary impact of the randomly distributed factors that are actual confounders in the planned trial. Therefore, the variability in the overall average effect of confounders (balanced by randomization) can be measured during the design phase. Furthermore, the variance of the primary outcome does not depend on the unbalanced values of single confounders but on the overall influence of all confounders. This is so for the reason that some confounders may influence the variable of interest positively (e.g., exercise) and others negatively (eating processed food), and hence one may suspect them to average out. Let C_i denote the overall impact of n confounders on i-th patient. If there were indeed an infinite number of actual confounders, then their overall effect could be represented as follows:

$$C_i = \sum_{n=1}^{n} X_n$$

Assuming that each confounder has a normal distribution with a standard deviation σ and mean $\mu = 0$, the overall influence of confounders will average out in the limit (understood here as an infinite number of actual confounders):

$$E(C) = 0$$

Considering the implausibility of there being an infinite number of actual confounders, the actual overall impact of confounders on the treatment effect experienced by individual patients is likely to vary, but the more confounders there are, the more likely they are to average out, and the smaller is the standard deviation of their sum (C). For example, if there were just two confounding factors, then the standard deviation of the variable C being a sum of two confounders X and Y:

$$C = X + Y$$

is lower than the sum of standard deviations of two variables:

$$\sigma_C = \sqrt{\sigma_X^2 + \sigma_Y^2} < \sigma_X + \sigma_Y$$

For this reason, Worrall's consideration of actual and possible confounders misses the point as an infinite number of confounders would average out. In particular, Worrall's (2007, 483-484) claim that "even if this [the balance claim] was convincing for the case of a single confounder, it is not at all clear that the argument works even on its own terms when we take into account the fact that there are indefinitely many possible confounders" is at odds with what statistical theory says about a sum of random variables.

If researchers interpreted any difference between the ATE observed in the treatment and control groups as an indication of the treatment's efficacy, such a perfect balance would be needed, and the criticism voiced by Worrall and others would be convincing. However, statistical hypothesis testing and the use of confidence intervals allow for accounting for the random imbalances in the overall impact of confounders. Furthermore, pre-study power analysis is a standard approach used to assess the variability of primary outcome that allows for concluding, with a high degree of certainty, what divergences from the true effect size can be expected. For this reason, in contrast to Worrall's (2002; 2007) assertions endorsed by Martinez and Teira (2022, 4), experimenters are able to realize how unlucky they may be even when confounders remain unknown.

This being said, one remark needs to be made. By chance, randomization can lead to unbalanced arms of a trial. Still, such random imbalances are not at odds with the claim that randomization balances the impact of confounders *in the statistical sense*. Imbalanced outcomes of simple randomizations (such as coin-flipping) that suggest treatment efficacy despite no actual effect of the intervention on the outcome (false-positive result) are expected to happen with frequency depending on the threshold of statistical significance (e.g., 5%). For this reason, if other forms of bias (such as questionable research practices and fraud) can be excluded, a positive result (rejecting the null hypothesis) can be interpreted in two ways: either as a sign of treatment efficacy or as an indication that an unlikely event has happened and the trial arms are imbalanced (Rubin 1980). A few solutions to this problem can be divided into those that can be used when an imbalance is discovered after assigning participants to the treatment and control groups (statistical control for confounders and rerandomization) and at the planning stage (such as stratified randomization or covariate-adaptive randomization).

Suppose researchers discovered that randomization has led to imbalanced trial arms by comparing differences in average values of covariates between the treatment and control or pre-exposure outcome measures. In that case, they could statistically control for the balance by, for example, adding the unbalanced confounder to the regression. Using regression analysis instead of testing for a difference in means might be problematic as this approach is based on stricter assumptions (e.g., linearity). Furthermore, the results reported by studies involving regression analysis to control for covariates might be considered less trustworthy when this step was not preregistered as "it is almost always possible to obtain a false positive by adding covariates indiscriminately, requiring a pool of only a few dozens." (Mutz et al. 2019, 51). This approach is primarily useful when small sample sizes undermine making statistically significant inferences because adding explanatory variables reduces the unexplained variability of the outcome (ε).

Another solution is repeating the randomization procedure (re-randomize) when some variables known to determine the outcome or pre-exposure outcome measures are unbalanced (in the Millean, non-statistical sense) (Morgan and Rubin 2012; 2015). Re-randomization is popular in economics (Bruhn and McKenzie 2009) but rarely used in clinical trials, where baseline (im)balance is reported only in trial reports. The pragmatic explanation of this situation is that, in economics, covariates of individuals or groups (e.g., neighborhoods assigned to the treatment and control groups) are known before exposing them to the intervention. Instead, in clinical trials, patients are often randomized subsequently after being recruited and before the value of confounding factors is measured. Another reason for the limited use of rerandomization procedures in medicine is that CONsolidated Standards of Reporting Trials (CONSORT) guidelines (Moher et al. 2012), which are followed by most significant medical journals, discourage balance testing (and hence rerandomization) and advise reporting the values of covariates only. As the guidelines remain silent on why rerandomization is not recommended, we can only mention some arguments against this procedure without indicating which of them had been convincing for the CONSORT Group. First, the decision regarding rerandomization can depend on either testing the balance of known confounders or the average baseline outcome measure. While the latter can show an imbalance in significant (having a large impact) but unknown confounders, the former analysis is only informative about known confounders. Considering that those variables that have not been recognized as confounders so far may be statistically independent of known confounders, assuring the quasi-Millean balance in known confounders does not assert balance in unknown confounders. Second, using rerandomization makes statistical inference more complicated as alternative statistical tests should be used for the reason that "[r]erandomization changes the distribution of the test statistic, most notably by decreasing the true standard error, thus traditional methods of analysis that do not take this into account will result in overly "conservative" inferences in the sense that tests will reject true null hypotheses less often than the nominal level and confidence interval will cover the true value more often than the nominal level (Morgan and Rubin 2012, pp. 1264-1265).

However, the primary reason seems to be that the balance in known confounders can be controlled for by designing randomization in a specific way. One way to do so is to use stratified randomization, where separate samples were chosen based on known confounders (e.g., different age groups) are randomized separately. In principle, researchers could control for all known confounders before randomization, so there is no need to test for the balance in known confounders post randomization. When the number of confounders is large, covariate-adaptive randomization procedures come in handy and allow for obtaining the (approximately) equal balance of known confounders (in the quasi-Millean sense) and the equal balance of unknown confounders (in the statistical sense) (Greevy et al. 2004; Proschan & Dodd 2019). Mutz et al. (2019) used simulation to show that these balanced randomization procedures perform similarly to rerandomization in large datasets, outperform rerandomization in small samples, and have lower computational requirements. For these reasons, it might be both more convenient and more warranted methodologically to control for the equal balance of known confounders (in the quasi-Millean sense) by choosing randomization procedures other than simple randomization.

5. Randomization and causal inference

Furthermore, the claim that randomization balances the average impact of confounders in the statistical sense can be supported with considerations regarding the presupposed notion of cause. Such an analysis sheds some light on the goal of experimenters that use randomization to achieve the balance understood in the statistical sense. The method of difference inspiring the notion of 'Millean balance' emerged in the context when the regularity view on causality had been the mainstream position and dated back to the Humean (constant conjunction of events) tradition (Psillos 2014). This explains why John Stuart Mill (2017) has only considered the deterministic case of two events equal in every factor but the one causally contributing to the outcome. However, the regularity account of causality had been considered an unattainable ideal for quantum physics, social sciences, and medicine, which has led to the development of probabilistic accounts of causality that identify probability-raising with causality. For example, in Cartwright's (2010, p. 60) probabilistic theory of causality:

$$C\ causes\ E \text{ iff } P\big(E\,\big|\,C\&K\big) > P\big(E\,\big|\,{\sim}C\&K\big)$$

Where:

C – a causal factor;

${\sim}C$ – the lack of a causal factor;

E – an effect;

K – the overall impact of all other causes.

That is, C causes E iff C raises the probability of E when all other factors K are held constant. If researchers could observe an outcome (Y) of treating the same n-th patient and at the same time with either a new therapy under test ($Y_T(n)$) or a control ($Y_C(n)$), there would be no need for running clinical trials since:

$$TE(n) = Y_T(n) - Y_C(n)$$

Where:

This preprint has not undergone peer review (when applicable) or any post-submission improvements or corrections. The Version of Record of this article is published in *History and Philosophy of the Life Sciences*, and is available online at https://doi.org/10.1007/s40656-024-00655-1.

$TE(n)$ – the treatment effect experienced by n-th patient;

$Y_T(n)$ – the outcome of n-th patient receiving treatment;

$Y_C(n)$ – the outcome of n-th patient receiving control.

Obviously, observing the treatment effect of the same patient and at the same time with treatment and control is impossible. This is the fundamental problem of causal inference (see Rubin 1974). The workaround for this problem is to focus on the population-wide average treatment effects. The estimator ($\widehat{ATE}$) for ATE measures the difference in average outcomes between the treatment and control groups (see Rubin 2005):

$$\widehat{ATE} = \frac{1}{N}(\sum_{i=1}^{n} Y_T(n) - \sum_{j=1}^{m} Y_C(m)$$

This approach to assessing treatment efficacy is only true if the treatment and control groups are sufficiently similar:

$$P(E\,\big|\,C\&K') > P(E|{\sim}C\&K'')$$

$$ATE_{true} = \widehat{ATE} \leftrightarrow\ K' \approx\ K''$$

Otherwise, the $\widehat{ATE}$ would measure both the impact of the intervention (C) and the difference between K' and K'' (see Rubin 1974, 692). But, as we argued above, randomization asserts that $\widehat{ATE}$ is an unbiased (in the statistical sense) estimator of the average causal effect (see also Hernan & Robins 2018, chapter 2). For this reason, Martinez and Teira's (2022, 8) claim emerging from Worrall's misguided argumentation that "Causal inference in well-designed experiments does not depend on Millean balance, but on the proper statistical interpretation of the outcome" is false since pre-study power analysis and statistical hypothesis testing are only helpful for distinguishing random divergences of $\widehat{ATE}$ from the null hypothesis, but the effect size estimate relies crucially on there being a balance in the overall effect of confounders between the treatment and control groups. The potential outcomes approach (POA) to causal inference has been criticized for being over-restrictive regarding the concept of cause and identifying it with manipulability by humans despite the long tradition of considering such unmodifiable features as race, sex, and genetic constitution to cause (Marcellesi 2013; Broadbent 2015; Vandenbroucke et al. 2016). But regardless of these counterarguments, POA remains the mainstream view on causality underlying inferences from randomized controlled trials in medicine and other disciplines using similar designs (such as randomized field trials utilized in economics) (Rubin 2005; Hernan & Robins 2018).

Considering the POA is helpful for investigating what randomization achieves and what is the epistemic goal of the randomized assignment. Above, we argued that randomized assignment to the treatment and control groups balances the summary impact of both known and unknown (or unobservable) confounders in the statistical sense, i.e., randomization makes large deviations in the overall impact of confounders unlikely and allows for assessing the probability of obtaining those unlikely draws that generate an imbalance in trial arms. Randomization asserts that the only difference (in the statistical sense) between the treatment and control groups is the intervention delivered to the treatment group and hence $\widehat{ATE}$ in a randomized trial is an unbiased estimator for the average causal effect (Neyman

1990 [1923]; Greenland 1990). This is so because the randomization of study participants asserts that each patient (with their own distinct characteristics) is equally likely to enter the treatment and control groups, similarly to a fair coin that is equally likely to land up heads or tails.

Neither the balance claim understood in this way nor $\widehat{ATE}$ being an unbiased estimate for the average causal effect excludes the possibility that minor deviations from the balance can emerge by pure chance and, rarely, large deviations lead to false-positive results (Bird 2021). As we argued above, the larger the variance produced by the overall influence of all confounders, the larger the sample size is needed to keep the risk of false positives at a prespecified level. To reduce the required sample size, experimenters often control for the impact of the known or suspected confounders (the choice of which will be considered below) by either using balanced randomization (e.g., stratified randomization) or deterministic assignment aimed at minimizing differences across trial arms (such as minimization assignment). As Rubin (2007, p. 27) put it, such randomization methods, “by creating treatment and control groups within which the distributions of observed covariates are more similar than would be expected if we simply assigned treatments to units completely at random, eliminates conditional (on these covariates) bias, which when averaged over in a completely randomized design becomes variance.” For this reason, these methods lead to a lower variance of the $\widehat{ATE}$ than simple randomization and allow for reducing sample size while keeping power constant.

However, the use of balanced randomization poses the question of what are the grounds for deciding what baseline characteristics should be balanced. In general, this is a rules-of-thumb decision process as no strict algorithms are in use. The decision is concerned with the trade-off between using too few balancing variables and omitting important confounders, which lowers the trial’s statistical power, and controlling for variables unrelated to the outcome, which may inflate type-I error rates (Greenberg et al. 2018). Additionally, since balancing variables create a correlation between trial arms (Raab et al. 2000), they should be adjusted for at the analysis stage, which makes controlling for too many variables infeasible as it complicates statistical analysis. Otherwise, larger p-values and wider confidence intervals are reported, which negatively impacts the accuracy of decisions (Ciolino et al. 2014).

The general rule is to balance those baseline characteristics that are highly predictive of outcomes. Balancing such variables increases the power of the study (Kahan 2014) and allows for reducing the sample size. Usually, experimenters control for the impact of center (in the case of multi-center trials) and such prognostic factors as age and disease stage. Prognostic factors (confounders of primary outcomes) are usually identified by previous epidemiological studies or other clinical trials. In some cases, standard measures of disease stage are developed. They can be used as balancing variables, e.g., pneumonia-specific severity index (PSI) based on several prognostic factors is correlated with the risk of hospital admission and mortality of pneumonia patients. In other cases, limited empirical grounds for choosing balancing variables are available. Such a situation troubles multi-center trials, as researchers must choose if the center effects should be controlled for at the design stage despite limited empirical evidence available. Parzen et al. (1998) argued that center adjustment leads to a more complicated analysis and as long as the effect size of centers is not expected to be large, using simple randomization may be better. Center effects are rarely known beforehand as they depend on context and vary significantly across trials: they are close to zero in the case of clinical trials conducted in the primary care settings, but bear substantial effect on the outcomes of the patients of surgical trials. (Cook et al. 2012; Adams et al. 2013). Kahan (2014) advised considering if center characteristics are likely to influence patient outcomes and whether the baseline risk of individuals treated at different centers is likely to vary. Most RCTs control for only one or two most significant prognostic factors at the design stage and virtually no trial uses more than eight balancing variables (Kahan et al., 2012). Therefore even those studies that use balanced randomization rely on randomization to balance the overall impact of all other confounders.

In sum, causal inferences rely on the assumption regarding the equal distribution of the overall impact of confounders across the treatment and control groups. Randomization equalizes the impact of confounders on the outcome across the treatment and control groups in the statistical sense. This asserts that $\widehat{ATE}$ is an unbiased estimate of the average causal effect of the intervention under the test. In other words, the epistemic goal of randomization is to obtain an unbiased estimate of the effect of the intervention under test. Controlling for known confounders (with balanced randomization) can lead to a lower variance of $\widehat{ATE}$, and such designs can be preferred in some situations.

6. Concluding Remarks

Herein, we have argued that randomization balances the overall impact of confounders across the treatment and control groups in the statistical sense. This balance makes average treatment effect ($\widehat{ATE}$) an unbiased estimator of the average causal effect of an intervention in comparison to its control. In light of our analysis, Worrall's (2007, 465-466) conclusion that "the results from the actual random allocation made in some particular trial (as opposed to the results from an indefinite series of such random allocations) can give no extra reason at all for thinking that the division between an experimental and control group is not biased in some significant way" is not accurate. Randomization balances the distribution of outcomes in the statistical sense and, jointly with appropriate research design (considering the variability of outcome and expected effect size), asserts that positive results are unlikely to emerge by chance.

However, the unbiasedness of the average treatment effect ($\widehat{ATE}$) does not exclude the possibility that each randomization leads to a balanced distribution of confounders. The chance for a well-designed RCT to report a false-positive result is equal to the threshold for statistical significance (α). Still, this risk of spurious findings does not undermine the balance claim as long as the statistical sense of the claim is endorsed. It also does not undermine inferences made by medical researchers despite sometimes leading to empirical controversies and failed replications (Bird 2021) because several measures can be applied to counteract making false efficacy claims. Researchers routinely check for baseline imbalances among known confounders such as age, ethnicity, and comorbidities or use balanced randomization to control for factors known to have a large effect on an outcome of interest. This asserts that well-conducted randomized trials are at least as trustworthy as non-randomized (observational) studies. Still, randomization controls (in the statistical sense) for both known and unknown confounders, while observational studies can only do so for known confounders.

For this reason, "the evidence provided by an observational study (or a historically controlled trial) regarding an intervention's efficacy is not equivalent to the evidence provided by a well-conducted randomized intervention study" (La Caze 2013 p. 353). Moreover, this small probability of false-positive results is further diminished by not relying on evidence from single RCTs. For example, Food and Drug Administration requires two positive results stemming from phase-III trials for drug approval, while the

Movement of Evidence-Based Medicine prioritizes systematic reviews of RCTs that can be very helpful in discovering which clinical trials report false-positive results.

All in all, we believe that the defense of randomization relying on the control for the overall impact of both known and unknown confounders (in the statistical sense) is not only more straightforward than other arguments present in the literature, but also aligns with what medical and social science researchers believe to be true, and statistical theory. Our analysis accords with the frequentist paradigm. It may suggest that 'statistical theory' is more homogenous than it actually is. On the contrary, statistics is a heterogenous field that experiences heated debates between the frequentist, Bayesian, and likelihood approaches. Our argumentation is indeed limited to one of the approaches to statistical analysis. Further research is needed to analyze the role of randomization from the perspectives of the other approaches to data analysis (see Berchialla et al. 2019 for a defense of randomization from the Bayesian perspective).

Finally, we feel obliged to admit that the strong views of the early opponents of randomization in the philosophical literature are understandable as they emerged as a response to equally strong opinions of the EBM proponents claiming that: "If the study wasn't randomized, we'd suggest that you stop reading it and go on to the next article in your search. […] Only if you can't find any randomized trials should you go back to it" (Straus et al. 2018 [2005], p. 118). However, the current EBM position is less randomization-oriented (Oxford Centre for Evidence-Based Medicine 2009). Hence, a more nuanced debate among philosophers is needed as it will align with statistical theory and contribute to contemporary debates concerning causal inference.

References:

Adams, G., Gulliford, M. C., Ukoumunne, O. C., Eldridge, S., Chinn, S., & Campbell, M. J. (2004). Patterns of intra-cluster correlation from primary care research to inform study design and analysis. *Journal of Clinical Epidemiology*, *57*(8), 785-794.

Backmann, M. (2017). What's in a gold standard? In defence of randomised controlled trials. *Medicine, Health Care and Philosophy*, *20*(4), 513-523.

Berchialla, P., Gregori, D., & Baldi, I. (2019). The role of randomization in Bayesian and frequentist design of clinical trial. *Topoi*, *38*(2), 469-475.

Bird, A. (2021). Understanding the replication crisis as a base rate fallacy. *The British Journal for the Philosophy of Science 72*(4), 965-993.

Borgerson, K. (2009). Valuing evidence: bias and the evidence hierarchy of evidence-based medicine. *Perspectives in Biology and Medicine*, *52*(2), 218-233.

Broadbent, A. (2015). Causation and prediction in epidemiology: a guide to the "Methodological Revolution". *Studies in history and philosophy of science part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 54, 72-80.

Bruhn, M., & McKenzie, D. (2009). In pursuit of balance: Randomization in practice in development field experiments. *American Economic Journal: Applied Economics*, *1*(4), 200-232.

Cartwright, N. (2010). What are randomised controlled trials good for?. *Philosophical Studies*, *147*(1), 59-70.

Chow, S.-C., Shao, J., Wang, H., & Lokhnygina, Y. (2018). *Sample Size Calculations in Clinical Research*. Boca Raton: CRC Press.

Ciolino, J. D., Renee'H, M., Zhao, W., Jauch, E. C., Hill, M. D., & Palesch, Y. Y. (2014). Continuous covariate imbalance and conditional power for clinical trial interim analyses. *Contemporary Clinical Trials*, *38*(1), 9-18.

Cook, J. A., Bruckner, T., MacLennan, G. S., & Seiler, C. M. (2012). Clustering in surgical trials-database of intracluster correlations. *Trials*, *13*(1), 1-8.

Cook, Th. & DeMets, D. (2008) *Introduction to Statistical Methods for Clinical Trials.* London&New York: CRC Press.

Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*, *210*, 2-21.

Fuller, J. (2019). The confounding question of confounding causes in randomized trials. *The British Journal for the Philosophy of Science 3*(70), 1-26.

Greenberg, L., Jairath, V., Pearse, R., & Kahan, B. C. (2018). Pre-specification of statistical analysis approaches in published clinical trial protocols was inadequate. *Journal of Clinical Epidemiology*, *101*, 53-60.

Greenland, S. (1990). Randomization, statistics, and causal inference. *Epidemiology*, 421-429.

Greevy, R., Lu, B., Silber, J. H., & Rosenbaum, P. (2004). Optimal multivariate matching before randomization. *Biostatistics*, *5*(2), 263-275.

Hernan, M. A., & Robins, J. M. (2018). *Causal Inference: What If*. New York: CRC Press.

Howick, J. H. (2011). *The philosophy of evidence-based medicine*. John Wiley & Sons.

Kahan, B. C. (2014). Accounting for centre-effects in multicentre trials with a binary outcome–when, why, and how?. *BMC medical research methodology*, *14*(1), 1-11.

Kahan, B. C., & Morris, T. P. (2012). Reporting and analysis of trials using stratified randomisation in leading medical journals: review and reanalysis. *British Medical Journal*, *345*.

Kahan, B. C., Jairath, V., Doré, C. J., & Morris, T. P. (2014). The risks and rewards of covariate adjustment in randomized trials: an assessment of 12 outcomes from 8 studies. *Trials*, *15*(1), 1-7.

Kernan, W. N., Viscoli, C. M., Makuch, R. W., Brass, L. M., & Horwitz, R. I. (1999). Stratified randomization for clinical trials. *Journal of Clinical Epidemiology*, *52*(1), 19-26.

La Caze, A. (2009). Evidence-based medicine must be…. *Journal of Medicine and Philosophy*, *34*(5), 509-527.

La Caze, A. (2013). Why randomized interventional studies. *Journal of Medicine and Philosophy*, *38*(4), 352-368.

La Caze, A., Djulbegovic, B., & Senn, S. (2012). What does randomisation achieve?. *BMJ Evidence-Based Medicine*, *17*(1), 1-2.

Lawler, I., & Zimmermann, G. (2021). Misalignment between research hypotheses and statistical hypotheses: A threat to evidence-based medicine?. *Topoi*, 40(2), 307-318.

Marcellesi, A. (2013). Is race a cause?. *Philosophy of Science*, 80(5), 650-659.

McGlothlin, A. E., & Lewis, R. J. (2014). Minimal clinically important difference: defining what really matters to patients. *JAMA*, *312*(13), 1342-1343.

Moher, D., Hopewell, S., Schulz, K. F., Montori, V., Gøtzsche, P. C., Devereaux, P. J., ... & Altman, D. G. (2012). CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *International journal of surgery*, *10*(1), 28-55.

Morgan, K. L., & Rubin, D. B. (2012). Rerandomization to improve covariate balance in experiments. *The Annals of Statistics*, *40*(2), 1263-1282.

Morgan, K. L., & Rubin, D. B. (2015). Rerandomization to balance tiers of covariates. *Journal of the American Statistical Association*, *110*(512), 1412-1421.

Kahan, B. C., & Morris, T. P. (2013). Adjusting for multiple prognostic factors in the analysis of randomised trials. *BMC Medical Research Methodology*, *13*(1), 1-11.

Mutz, D. C., Pemantle, R., & Pham, P. (2019). The perils of balance testing in experimental design: Messy analyses of clean data. *The American Statistician*, *73*(1), 32-42.

Neyman, J., (1990) [1923]. On the application of probability theory to agricultural experiments. Essay on principles Dabrowska, D. M., & Speed, T. P. (trans.). *Statistical Science*, *4*(5), 465-472.

Oxford Centre for Evidence-based Medicine. (2009). Levels of evidence. *BJU Int*, *104*(11).

Papineau, D. (1994). The virtues of randomization. *The British Journal for the Philosophy of Science*, *45*(2), 437-450.

Philippi, C.L. (2022). There is no Cause to Randomize. *Philosophy of Science.* 89, 152-170.

Proschan, M. A., & Dodd, L. E. (2019). Rerandomization tests in clinical trials. *Statistics in Medicine*, *38*(12), 2292-2302.

Psillos, S. (2014). *Causation and explanation*. New York and London: Routledge.

Raab, G. M., Day, S., & Sales, J. (2000). How to select covariates to include in the analysis of a clinical trial. *Controlled Clinical Trials*, *21*(4), 330-342.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and non-randomized studies. Journal of Educational Psychology, 66(5), 688.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, 34-58.

Rubin, D. B. (1980). Randomization analysis of experimental data: The Fisher randomization test comment. *Journal of the American Statistical Association*, *75*(371), 591-593.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469), 322-331.

Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Statistics in Medicine*, *26*(1), 20-36.

Senn, S. (2013a). A brief note regarding randomization. *Perspectives in Biology and Medicine*, *56*(3), 452-453.

Senn, S. (2013b). Seven myths of randomisation in clinical trials. *Statistics in Medicine*, *32*(9), 1439-1450.

Straus, S. E., Glasziou, P., Richardson, W. S., & Haynes, R. B. (2018 [2005]). *Evidence-based medicine E-book: How to practice and teach EBM*. Elsevier Health Sciences.

Thompson, R. P. (2011). Causality, theories, and medicine. Illari, Ph., F. Russo, and J. Williamon (eds.) *Causality in the Sciences*, 25-44. Oxford: Oxford University Press.

Vandenbroucke, J. P., Broadbent, A., & Pearce, N. (2016). Causality and causal inference in epidemiology: the need for a pluralistic approach. International Journal of Epidemiology, 45(6), 1776-1786.

VanderWeele, T. J. (2021). Can sophisticated study designs with regression analyses of observational data provide causal inferences?. *JAMA Psychiatry*, *78*(3), 244-246.

Wicherts, J. M., Veldkamp, C. L., Augusteijn, H. E., Bakker, M., Van Aert, R., & Van Assen, M. A. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, 1832.

Worrall, J. (2002). What evidence in evidence-based medicine?. *Philosophy of science*, *69*(S3), S316-S330.

Worrall, J. (2007). Why there's no cause to randomize. *The British Journal for the Philosophy of Science*, *58*(3), 451-488.