Trials

METHODOLOGY

Open Access

Evaluation of randomized controlled trials: a primer and tutorial for mental health researchers

Mathias Harrer[1,2*†], Pim Cuijpers[3,4†], Lea K. J. Schuurmans[1], Tim Kaiser[5], Claudia Buntrock[6], Annemieke van Straten[3] and David Ebert[1]

Abstract

Background Considered one of the highest levels of evidence, results of randomized controlled trials (RCTs) remain an essential building block in mental health research. They are frequently used to confirm that an intervention "works" and to guide treatment decisions. Given their importance in the field, it is concerning that the quality of many RCT evaluations in mental health research remains poor. Common errors range from inadequate missing data handling and inappropriate analyses (e.g., baseline randomization tests or analyses of within-group changes) to unduly interpretations of trial results and insufficient reporting. These deficiencies pose a threat to the robustness of mental health research and its impact on patient care. Many of these issues may be avoided in the future if mental health researchers are provided with a better understanding of what constitutes a high-quality RCT evaluation.

Methods In this primer article, we give an introduction to core concepts and caveats of clinical trial evaluations in mental health research. We also show how to implement current best practices using open-source statistical software.

Results Drawing on Rubin's potential outcome framework, we describe that RCTs put us in a privileged position to study causality by ensuring that the potential outcomes of the randomized groups become exchangeable. We discuss how missing data can threaten the validity of our results if dropouts systematically differ from non-dropouts, introduce trial estimands as a way to co-align analyses with the goals of the evaluation, and explain how to set up an appropriate analysis model to test the treatment effect at one or several assessment points. A novice-friendly tutorial is provided alongside this primer. It lays out concepts in greater detail and showcases how to implement techniques using the statistical software R, based on a real-world RCT dataset.

Discussion Many problems of RCTs already arise at the design stage, and we examine some avoidable and unavoidable "weak spots" of this design in mental health research. For instance, we discuss how lack of prospective registration can give way to issues like outcome switching and selective reporting, how allegiance biases can inflate effect estimates, review recommendations and challenges in blinding patients in mental health RCTs, and describe problems arising from underpowered trials. Lastly, we discuss why not all randomized trials necessarily have a limited external validity and examine how RCTs relate to ongoing efforts to personalize mental health care.

†Mathias Harrer and Pim Cuijpers contributed equally and share first authorship.

*Correspondence:
Mathias Harrer
mathias.harrer@tum.de
Full list of author information is available at the end of the article

Keywords Mental health, Randomized controlled trial, Data analysis, Tutorial

Background

Randomized controlled trials (RCTs) are widely considered the "gold standard" to determine if an intervention is effective or not [1]. RCTs form a crucial part of treatment policy decisions and are regarded as one of the highest levels of evidence [2, 3]. While their primacy is not uncontested [4, 5], hardly anyone would disagree that RCTs can be an exceptionally well-suited design to study the effectiveness of some treatment or intervention.

A more practical concern is that the methodological quality of RCT evaluations, in mental health research and elsewhere, leaves much room for improvement. Back in the 1990s, Altman [6] called the poor quality of health research a "scandal", and it has been argued that his assessment is still accurate today [7, 8]. In the past, methodological researchers have found fault with various aspects of RCT analyses; Table 1 provides an overview of commonly named problems in the literature.

As a remedy, it has been emphasized that researchers need to be equipped with a greater understanding of their methodology [7]. We agree with this assessment, and we believe all mental health researchers should care for and be helped to understand what makes a good RCT evaluation. In this spirit, we want to provide a non-technical introduction to some core ideas behind RCTs and how they relate to the larger topic of causal inference.

Table 1 Common problems with statistical analyses of RCTs in (mental) health research

Analysis steps	Common problems
Missing data handling	Many systematic reviews have found that, while standards have improved in recent years [9], the missing data handling in many RCTs remains poor [9–13]: • The amount of missing data is often insufficiently reported, as is the methodology used to handle missing values. • Assumptions of the missing data handling strategy remain undiscussed (are the data assumed to be missing completely at random, missing at random, missing not at random—and why?). • Methods that are inadequate (e.g., single imputation) or based on strong assumptions (e.g., complete case analysis) are used. • Although recommended by many regulatory guidelines [14], sensitivity analyses are still underused. If sensitivity analyses are conducted, they are often not suited to test the assumptions of the main missing data handling strategy. • While often plausible, methods that model the missing not at random (MNAR) assumption are employed very infrequently and are often poorly reported.
Baseline covariate tests	Methodologists have frequently commented that baseline covariate or "randomization tests" are superfluous and that they should not be conducted [15–19]. Nevertheless, these tests are frequently reported in RCT evaluations, and reviewers often demand them to show that the randomization "worked". Because *P* values of these tests are often included in the baseline characteristics table, some refer to this as the "Table-1 Fallacy" [20].
Analysis model	Even when data was derived form a parallel-group RCT, researchers often calculate change from baseline and pre-post effect sizes to assess intervention effects. While widespread and often requested by reviewers, this approach does not account for regression to the mean and can produce highly misleading results [21–23].
Interpretation of results	Null (viz., $p \geq 0.05$) results are often interpreted as showing the absence of an effect, while "absence of evidence does not imply evidence of absence" [24, 25]. This issue also pertains to negative effects, which may be uncommon but important to detect. This problem is exacerbated by the fact that (in mental health research), most trials are not even sufficiently powered to detect the main effect of the intervention [26]. In a similar vein, "post hoc" power analyses are often conducted (or requested), e.g., to calculate the power of a trial based on its final sample size and calculated effect size (often with the intention to check if there is a "true" effect that the trial was simply not powered to detect). This approach is circular and logically flawed, since the observed power is simply a function of the *P* value [27, 28].
Reporting	There is evidence that the quality of clinical trial reports has improved substantially since journals started adopting the Consolidated Standards of Reporting Trials (CONSORT [29, 30]). Nevertheless, the reporting of RCT results in mental health research remains suboptimal [31, 32]. In the abstract, for example, trialists often fail to report methods of randomization and/or allocation concealment, or do not disclose the funding source. Another concern is selective reporting. Still, many trials are not preregistered in a clinical trial registry; statistical analysis plans (SAPs) provided in these registrations are often vague. This makes it easier to conceal questionable research practices such as selective outcome reporting (i.e., only reporting outcomes that fit the researcher's objective) [33] or "outcome switching" [34] in clinical trial reports. Core outcome sets (COS [35]) are collections of outcomes that should be measured and reported in all clinical trials. They are a great way to ensure that endpoints are assessed consistently within a research field and using the appropriate instruments. A number of COS or related consensus papers has been developed for various mental and behavior disorders [36–41], but they remain underused. A comprehensive overview of available COS for mental health research and beyond is provided by the Core Outcome Measures in Effectiveness Trials (COMET [42]) initiative (www.comet-initiative.org).

In this primer and tutorial, we discuss fundamental concepts in trial evaluations and showcase their practical implementation using the free statistical programming framework R. We focus on issues that are particularly relevant in mental health research and describe some of the avoidable and unavoidable limitations of RCTs in this field. Naturally, it is out of scope for this article to give a comprehensive view of all the intricacies in RCT analyses; instead, we want to provide a starting point and show how to get the "basics" right.

Methods

This article consists of two parts: a conceptual primer on RCT methodology (presented here), as well as a practical tutorial in the supplementary material. The tutorial provides a practical guide on how to analyze RCTs using R, based on data of a real trial examining the effect of an Internet-based intervention for depression [43, 44]. The tutorial also presents more detailed background information on some of the concepts mentioned in the primer. Prior knowledge of R is not required to complete the tutorial.

Results

Potential outcomes

In mental health care, many of our research questions revolve around the *cause* and *effect* of different actions. If we administer a psychological intervention or prescribe some medication, we do so because we hope that this will *cause* our patient's mental health to improve. If a patient starts to feel better during our treatment, we may take this as a sign that the intervention was successful. Yet in reality, we will never know the *true* impact of our actions. This is because we cannot go back in time to see how our patient would have developed had we acted differently.

This inability to go back in time and directly observe the effect of different actions is encapsulated in the "fundamental problem of causal inference" [45]. It states that a causal effect can only be shown if we compare the outcome of some action A to the outcome had we not taken action A. The problem is that only one outcome will ever be realized; the other remains a *potential* outcome that cannot be observed.

This idea is formalized in the *potential outcome framework*, which is part of the so-called Neyman–Rubin causal model (NRCM; named after Jerzy Neyman and Donald Rubin; [46, 47]). This model has become a dominant approach on how to think about causality in biomedical contexts. It also allows us to understand how and why causal inferences can be drawn from clinical trials. We will now introduce some basic concepts of this model and the notation through which they are typically expressed.

Say that we, as mental health professionals, are approached by a person i who currently suffers from a depressive episode. Naturally, our goal is to help that person, and this means that we have to decide which course of action is most likely to make that person feel better. Let us, therefore, assume that our outcome of interest is the depression status of person i several weeks into the future. We call this outcome Y and write $Y_i = 1$ if person i still suffers from depression at this later stage, and $Y_i = 0$ if not. Naturally, assuming that patients simply "have" or do "not have" depression is quite a simplification. Diagnostic manuals have often been criticized for enforcing such a false dichotomy between health and disease [48]. Yet for now, it will be helpful to think of our outcome Y as binary: either patients still suffer from depression after some time ($Y_i = 1$) or they do not ($Y_i = 0$).

Imagine that we have just learned of a new treatment T that may be just right for person i, and now we have to decide if we want to provide it. This means that we have two courses of action: either the treatment is provided ($T = 1$) or we decide not to provide it ($T = 0$). For this brief moment, two *potential* outcomes exist: one value of Y_i that we would measure in a world in which T was provided; and the value of Y_i we would measure if T was not provided. These potential outcomes are typically denoted as $Y_i(T = 1)$ or $Y_i(T = 0)$, respectively. Note that, using this notation, $Y_i(\cdot)$ is a function; it works like a magical "crystal ball" that tells us the outcome Y_i depending on which action we plug into it.

For some mental health problems, it is common for patients to improve even without treatment [21]. For example, about one-third of untreated depression cases remit "spontaneously" within 6 months [49]. This means that patients' symptoms improve so much that they no longer meet the diagnostic criteria for depression, even though they have not received any intervention. Thus, if we decide to provide the treatment T and patient i improves, this in no way guarantees that T actually *caused* this improvement. Maybe the patient would have also improved had we not provided the treatment. To establish the true, *causal* effect of our treatment, the two potential outcomes have to be compared with each other. This true causal effect is denoted by τ_i:

$$\tau_i = Y_i(T = 1) - Y_i(T = 0) \quad (1)$$

Put differently, the causal treatment effect is the difference between the potential outcome if we provide the treatment, and the potential outcome if we do not provide it. If the potential outcomes do not differ, there is no "real" causal effect.

Usually, we are not only interested in an individual treatment effect (ITE) for one person i, but also in a "typical" or "overall" effect that can be expected in a patient

population. This quantity τ is also known as the average treatment effect (ATE [50]). Panel A in Fig. 1 illustrates this. Imagine that to test if the new treatment T really works, we recruit a sample of eight people for which it may be suited. For each person, there are two potential outcomes, one if we provide T, one if we do not. The ATE τ is the difference between the average outcome had we provided the treatment to all individuals, $\mathbb{E}[Y_i(T=1)]$, and the average outcome had we given it to no one, $\mathbb{E}[Y_i(T=0)]$.

There is something bittersweet about this definition of the ATE. On the one hand, it gives us a "recipe" for how we can obtain the true causal effect of an intervention in a population of interest. At the same time, it shows us that, as finite beings, this true effect will always be unknown to us because it is impossible to observe the two "ingredients" that define τ at the same time. We cannot provide *and* not provide some treatment to the same people at the same time. This is where RCTs try to provide a solution.

Randomization solves a missing data problem

We have now learned that if we want to confirm that some treatment is effective, we must show that there is a causal effect, defined by comparing the two potential outcomes. However, this is impossible since no two potential outcomes can ever be observed at the same time. Therefore, we need an instrument that lacking actual knowledge of τ at least allows us to approximate it as closely as possible.

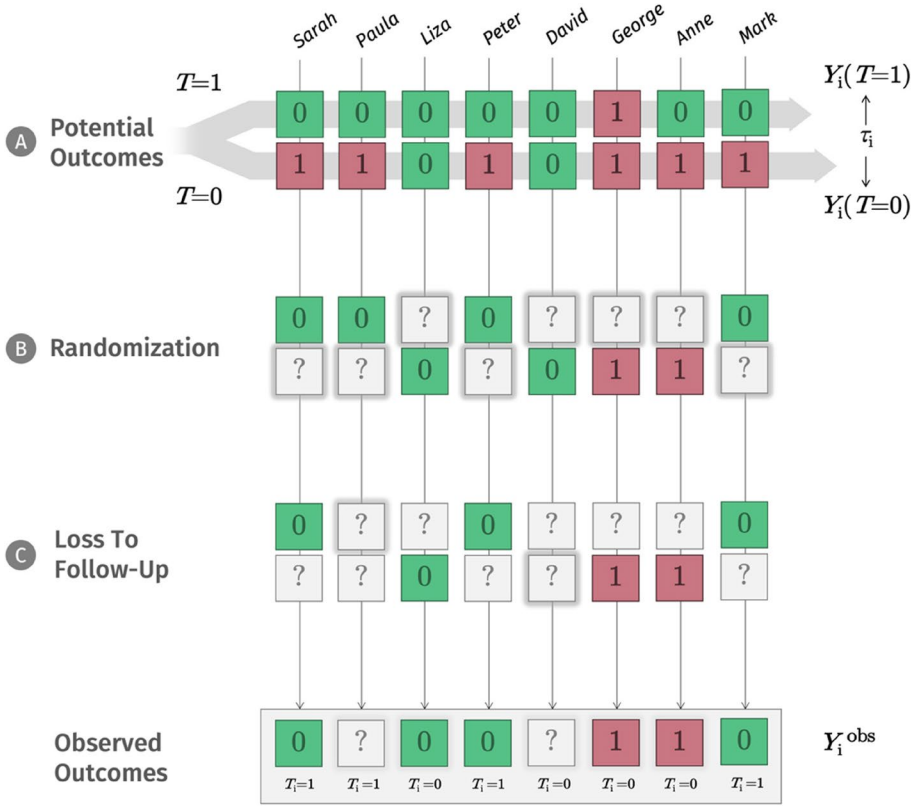

Fig. 1 Potential and observed outcomes in RCTs. *Note*: Going from top to bottom, this diagram illustrates the hidden "machinery" inside an RCT. The top panel (**A**) shows the potential outcomes for each person in our sample if we provide a new treatment ($T=1$) or not ($T=0$). In our example, "0" means that a person does not suffer from a depressive episode after several weeks, while "1" means that the person still suffers from depression. The potential outcomes are hypothetical; since they are based on counterfactuals, it is impossible to observe both at the same time, and so the true causal effect τ_i of our treatment also remains unobservable. Going down one step, panel **B** shows the process of randomization, which lets chance decide which potential outcome is realized, and which one is missing ("?"). Loss to follow-up (panel **C**) adds another layer of missingness. Here, it is much less plausible that the missings are added "completely at random". As analysts, all we end up having are the observed outcomes at the end of this process, which we need to use to estimate the unobservable causal effect τ on top as closely as possible. *Legend:* T_i = treatment allocation of patient i ($T_i=0$ for no treatment, $T_i=1$ for treatment); τ_i = causal treatment effect of patient i; Y_i = outcome of patient i: "1" (red box) if the patient still suffers from depression after several weeks, or "0" (green box) if the patient does not suffer from depression after several weeks; "?" (gray box) if the outcome was not recorded; $\mathbf{Y}_i^{obs}$ = observed outcomes of the trial

The NRCM tells us that to draw causal inferences (e.g., "treatment T causes patients' depression to improve"), we have to solve a missing data problem [51, 52]. As symbolized by panel B in Fig. 1, depending on our decision, only one potential outcome value Y will ever be observed. We use a question mark (?) to show that the other potential outcome will inevitably be missing from our records. In daily life, there may be countless reasons why one potential outcome is realized, while the other is "missing". In our depression example, we could imagine that people with greater symptom severity, or higher expectations, are more likely to opt for the new treatment; but it is probably an even more complex network of reasons that determines if $T=1$ or $T=0$.

In RCTs, through randomization, we suspend the influence of these countless and often unknown variables by replacing it with a single factor: chance. We cannot change that one potential outcome will always be missed; but successful randomization ensures that we at least know that potential outcomes are *missing completely at random* (MCAR) for each person in our population [46, 53, 54]. The fact that outcomes are "deleted" at random has a crucial implication: it means that the average potential outcome of those receiving treatment (let us call them "group A") and those without treatment ("group B") become *exchangeable* [55, 56]. We would have observed a comparable distribution of outcomes even if we had somehow made a mistake and had always given the treatment to group B instead of group A.

The concept of exchangeability can be difficult to understand at first. It essentially means that the treatment status ("does person i receive treatment or not?") is now completely independent of the potential outcomes of each person. Provided we have a sufficiently large sample, this allows us to get a representative cross-section of the potential outcomes we had observed if all patients had received the treatment; and it also provides us with an unbiased sample of all the potential outcomes we had observed if no patient had received the treatment.

Through randomization, we ideally achieve something in a sample that we learned was impossible to do for one person i: observing the outcome of some action, while at the same time also seeing the outcome had we not acted like this. The two groups have now become like real-life crystal balls for each other, where group B with $T=0$ indicates what would have happened to group A if it had not received the treatment, and group A suggests what would have happened to group B if we had in fact provided it with the treatment. This is possible because we know that, *in theory*, the other potential outcomes still exist for both randomized groups and that the average of these potential outcomes is exchangeable with the average of the other group. At any stage post-randomization, the difference in outcome means $\mu_1 - \mu_2$ between the two groups can therefore be used to approximate the true causal effect τ of our treatment T.

This is an important insight that is often neglected in practice. By definition, the ATE estimated in RCTs is a *between-group* effect. In many mental health trials, it is still common to see that patients' change from baseline is used to measure the "treatment effect" of an intervention. This is a flawed approach, even when the change scores over time in the intervention group are compared to the ones of a control group (see Table 1). There are various mathematical reasons that should discourage us from conducting change score analyses [57, 58], but they also often reveal a conceptual misunderstanding. In RCTs, we are not interested in *within-group* change over time: we want to estimate the true, causal effect of our treatment, and this causal effect is estimated by the difference between our randomized groups at some specific point in time. Naturally, many RCTs contain not only one but several follow-ups. It is also possible to include these multiple outcome assessments into a single statistical model and then estimate the overall effect of the intervention over the study period. However, even in such a longitudinal analysis, we are only interested in the average difference *between* the two groups that we observe across the different follow-up measurements, not how outcomes change from baseline *within* a single arm. A suitable method to conduct longitudinal analyses in RCTs known as mixed model repeated measures is demonstrated in the tutorial (S6.2).

Ideally, RCTs bring us into a privileged position: only by comparing the intervention and control group means at one or several specific time points after randomization, we generate a valid estimate of our treatment's average causal effect. Within-group changes from baseline are therefore typically neither relevant nor appropriate to estimate the ATE. It should also be noted that RCTs are not the only instrument to estimate τ. It simply becomes much more difficult, and requires much more untestable assumptions, once (potential) outcomes are not MCAR [55, 59]. We will see this in the following section.

The NRCM also reveals another crucial point about RCTs: they work, not necessarily because randomization makes the two randomized groups perfectly identical, but because they make the potential outcomes of both groups exchangeable. This point also allows to understand why baseline imbalance tests to show that randomization "worked" are misguided, even though they are commonly seen in practice (see Table 1). The goal of randomization is to create exchangeability in the potential *outcomes* because this allows us to draw causal inferences, not to ensure that the groups have identical *baseline values* [51]. Even successful randomization provides perfectly

balanced groups only *in the long run*; in our finite RCT sample, allocating treatment by chance is perfectly consistent with the idea that baseline means may also sometimes differ *by chance*. Random baseline imbalances can be relevant, but only if they occur in variables that are associated with differences in the (potential) outcomes [60, 61]. Below, we show that covariate adjustment in the analysis model can be used to address this concern.

Ignorability

Unsurprisingly, in practice, no such thing as an "ideal RCT" exists. In reality, we often have to deal with many additional problems, such as loss to follow-up. Panel C in Fig. 1 shows that we can think of loss to follow-up as a second layer of missingness added *after* randomization. These missings could occur for various reasons: maybe unobserved people moved to another city; they might have busier jobs; or they could have had negative side effects, leading them to discontinue treatment. In any way, on this "layer", it is often implausible that values are simply missing completely at random. Looking at the examples above, it is clear that loss to follow-up can distort our estimate of the ATE dramatically. Imagine that all treated individuals who had negative side effects were lost to follow-up. This would introduce *selection bias* [62]: our estimates are only based on individuals who tolerated the treatment well in the treatment group, and we would overestimate the ATE.

In this moment, it is helpful to go back to the NRCM. To us, missing values due to loss to follow-up are analogous to potential outcomes: they exist in theory, we just have not observed them. Yet, to approximate them as closely as possible, we now need an assumption that is more plausible than MCAR. Here, we resort to a trick. We now stipulate that values are missing randomly, but only in a subset of our data with identical covariates $\boldsymbol{X}$. Imagine that people working a full-time job were less likely to have time for the post-assessment, and that this fully explains our missing values. This implies that, once we only look at full-time workers, the outcomes of those who provide data, and those who do not, will not systematically differ (i.e., the observed and unobserved outcomes are *exchangeable* again). If we can identify some combination of covariates $\boldsymbol{X}$ conditional on which values are randomly missing again, we speak of *ignorable* missing data [63]. This is the core idea of the *missing at random* (MAR) assumption: facing missing follow-up data, we can "rescue" our estimate of the causal effect using prognostic information captured by our observed baseline covariates $\boldsymbol{X}$. This inherently untestable assumption is crucial since many relevant imputation methods depend on it.

Trial estimands

Arguably, the potential outcome framework we covered above is quite theoretical. It is still important to understand some of its core tenets because they define precisely when causal inferences can be drawn from data and what inherently unobservable effect RCTs actually try to get as close as possible to. In real-life RCTs, our research questions are much less abstract than that, and we now need a tool that links all this theory to the concrete analyses we should perform in our own evaluation. One way to align the theory and hands-on implementation of an RCT analysis is through so-called trial estimands.

Estimand means "what is to be estimated". Previously, we learned that RCTs allow us to estimate the ATE caused by the treatment. This sounds straightforward, but things get more complicated once we think of the intricacies of "real-life" RCTs. How, for example, should this effect be measured, and when? In what population is this effect to be expected? What do we do if some patients do not use the intervention as intended? Trial estimands allow us to answer these questions precisely and unambiguously. They are a list of statements that describe the (i) compared treatment conditions, (ii) targeted population, (iii) handling of "intercurrent events", (iv) measured endpoint, and (v) what population-level summary is used to quantify the treatment effect [64]. Table 2 shows a concrete example for a psychological intervention.

Trial estimands play an important role in regulatory affairs and have been adopted by both the European Medicines Agency [65] and the U.S. Food and Drug Administration (FDA; [66]). They are still much less common in mental health research, but there are very good reasons to make them a routine step in each RCT evaluation [67, 68].

Estimands also allow to us understand what is *not* estimated. Imagine that we conducted an RCT comparing a new type of psychotherapy to a waitlist. If the trial is successful, we could be inclined to say that our treatment has a true causal effect, meaning that therapists should now deliver it to help their patients. Looking at the estimand, we can immediately see that this reasoning is flawed because the trial only estimated the causal effect compared to a waitlist, not compared to what therapists usually do in their practice. In this particular case, we are confusing the "efficacy" of the treatment in a waitlist-controlled trial with its "effectiveness" as a routine-care treatment [69]. Many of such misinterpretations and sweeping overgeneralizations can be avoided by honestly stating what effect our RCT analysis actually estimates.

Detailing the handling of intercurrent events (such as treatment discontinuation, or use of other treatments) within the estimand is another important part since this can change the interpretation of the ATE. Typically, a

Table 2 Example of a trial estimand employing a treatment policy strategy

Attribute	Example
1. Treatment conditions	Internet-based, 8-week intervention for subthreshold depression with full access to care as usual, versus one-session psychoeducation with full access to care as usual.
2. Population	Elderly individuals (≥ 65 years) with mild depression (PHQ-9 ≥ 10) who do not fulfill diagnostic criteria of a major depressive episode according to the DSM-5.
3. Intercurrent events	Treatment policy strategy: effect regardless of treatment discontinuation (i.e., not completing all intervention sessions as intended) or use of other treatments (e.g., conventional psychotherapy or pharmacotherapy).
4. Endpoint	PHQ-9 depressive symptom severity score 8 weeks after randomization (post-test), measured regardless of intercurrent events.
5. Summary measure	Mean treatment group difference in the endpoint, expressed as a standardized mean difference (Cohen's *d*), with the pooled post-test *SD* used for standardization.

DSM-5 Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, *PHQ-9* Patient Health Questionnaire 9

so-called treatment policy strategy will come closest to the definition we elaborated earlier: we want to estimate the outcome of our treatment compared to what would have happened had we not provided it (or provided some alternative treatment instead), *no matter* if the treatment was actually used as intended [70]. This strategy largely overlaps with what is more commonly known as the "intention-to-treat" principle: once randomized, individuals are always analyzed. This strategy is most easily implemented if we were also able to obtain follow-up data of patients who did not follow the treatment protocol [71].

As part of their RCT evaluation, many mental health researchers also conduct a so-called per-protocol analysis, in which only participants who adhered to the treatment protocol are included in the analysis. This approach is often used as a sensitivity analysis to examine the treatment's efficacy under optimal conditions. However, this type of analysis is not supported by current guidelines because it extracts different subsets from both arms of our trial that may not be entirely comparable [72]. It is possible to estimate such an "optimal" effect of a treatment using the so-called principal stratum strategy [73], but this approach is based on many statistical assumptions and may be more difficult to implement in practice.

Climbing up the ladder: the analysis model

To appreciate what the goal of the analysis model in RCT evaluations is, we must first go back to Fig. 1. In its totality, this visualization illustrates what is sometimes called the *data-generating process*, the "hidden machinery" inside an RCT. On top, we have the lofty realm of the potential outcomes, which exist in theory, but are never fully observable. As we proceed downwards, many of these outcomes "go missing": first through randomization, where we know that potential outcomes are "deleted" completely at random and later through loss-to-follow-up, where the missingness mechanism is much less clear. All we RCT analysts end up with are the observed data at the very bottom of this process. The statistics we use in RCT evaluations are an attempt to climb this ladder back up.

Previously, we described that missing data, and how they are handled, play a critical role in RCTs. There is some degree of loss-to-follow-up in virtually all clinical trials, and this means that some unknown missingness mechanism will be lurking behind our observed data. Multiple imputation (MI [74]), in which several imputed datasets are created, and parameters of interest (e.g., the ATE) estimated in each of them, has now become a common approach to handle missing data in clinical trials. Through what is known as "Rubin's rules" [75], MI allows to calculate pooled estimates which properly reflect that we can estimate the true value of an imputed variable but will never know it *for certain.*

MI is a very useful and highly flexible method, and the tutorial shows how to apply it in practice (see S4.2). Nevertheless, it is important to keep in mind that MI, like all imputation procedures, is based on untestable assumptions. Typically, MI approaches assume that data are missing at random and will only be valid if our imputation model is a plausible approximation of the true missingness mechanism [76]. In the tutorial, we also cover alternative approaches which assume that the data are missing not at random (MNAR; S4.2.3) and which are often sensible for sensitivity analyses.

When conducting an RCT evaluation, it is helpful to understand the imputation and analysis model as two sides of the same coin. Both try to approximate the true data-generating mechanism within our trial, and they should therefore be *compatible* with each other [77]. In the literature, this is also known as the *congeniality* of the imputation and analysis model [78]. We describe this concept in greater detail in the tutorial (S4.2.3).

Various statistical models can be used to estimate the ATE within RCTs [61, 79, 80], and more options are

discussed in the tutorial. For now, we focus on analysis of variance (ANOVA), which remains the one of the most widely used methods to test the treatment effect in clinical trials. First published in 1935, Ronald A. Fisher's "The Design of Experiments" laid the foundations for randomized designs, while popularizing ANOVA as a suitable analysis method [81]. This historical legacy could explain why ANOVA is still often thought of as a "specific" method for randomized experiments and as unrelated to other statistical models. In reality, ANOVA models are simply a special type of linear regression [79]. In a "vanilla" two-armed single-center RCT with one clearly defined primary outcome, the ATE can be estimated (and tested) using the following regression equation:

$$\widehat{Y} = \alpha + \beta_T T + \beta_1 x. \tag{2}$$

where Y is the primary outcome of our trial (which we for now assume is continuous), α (the intercept) is the outcome mean in the control group, and β_T is the coefficient estimating our treatment effect: the amount by which the expected value of Y is shifted up or down for the treatment group ($T=1$; where $T=0$ for the control group). Since we only have two groups, a t-test of our group coefficient β_T will be mathematically equivalent to the F-test in an ANOVA. In the tutorial (S5.1.3) we show that, once the linear regression model in Eq. 2 has been fitted, we can also easily produce the F-test results we would have obtained from an ANOVA. If (2) were to only include the aforementioned terms, β_T would be the simple difference in means between the two groups in our trial; this quantity is also known as the "marginal" treatment effect. Here, however, we *adjust* the estimate by including the baseline covariate x as a predictor (in ANOVA language, this results in an analysis of covariance, ANCOVA). Typically, x is a baseline measurement of the outcome variable, and it is also possible to adjust for multiple covariates, provided they were specified before the analysis (see also S5.2.1 in the tutorial).

Intuitively, one may think that, thanks to randomization, covariate adjustments are unnecessary; but there are several reasons to include them. First, "good" covariates explain variation of our outcome Y *within* treatment groups, so adjusting for them increases our statistical power (i.e., the confidence interval around β_T shrinks [82]). Second, adjustment for prognostic covariates automatically controls for potential baseline imbalances *if they matter*: that is, when there is a disbalance in baseline covariates that are strongly predictive of the outcome [83]. Third, it is sometimes argued that covariate adjustment is helpful because it provides a *personalized* interpretation of β_T as the predicted difference in outcomes between two patients with identical covariate values x, but different treatment ($T=0$ vs. $T=1$ [84]).

In practice, treatment decisions are made for individuals, so it is tempting to follow this interpretation. Yet, this is not the reason why we adjust for covariates. In RCTs, our goal is to estimate the mean difference between the intervention and control group from our sample, and the covariate-adjusted model in Eq. 2 just happens to allow to estimate this marginal effect more precisely, at least for continuous outcomes. In logistic regression models, which are commonly used for binary outcomes, covariate adjustment has a different effect than what we described above: the confidence intervals do not tighten up, but the value of β_T increases instead [85]. This behavior is associated with the "non-collapsibility" of the odds ratio [86, 87], a numerical averaging failure that causes the average of conditional odds ratios (e.g., odds ratios calculated in subgroups of our trial sample) to not necessarily equal the unadjusted odds ratio that we observe in the entire sample. We explain this statistical "oddity" in greater detail in S5.2.2 in the tutorial, but the main point is that this behavior inadvertently changes our estimand. The odds ratio measured by β_T does not estimate the ATE anymore; instead, we obtain a conditional treatment effect that is typically higher than the "original" ATE, and which depends on the covariate distribution in our trial, as well as the covariates we decide to adjust for [88, 89]. One way to deal with this problem is to fit a logistic regression model in the first step, and then use a method known as *regression standardization* to obtain the desired estimate of the effect size (e.g., an odds or risk ratio). Briefly put, this method first uses the logistic regression model with covariates to predict the outcome Y while "pretending" that all participants had been allocated to the treatment group. Then, it does the same assuming that everyone had been assigned to control. Comparing the means of these two counterfactual predictions, we obtain a valid estimate of the marginal effect size in our trial sample, while taking into account the covariates in our model. In the tutorial, we explain this method in further detail and show to apply it in practice (S5.2.2).

In contrast, when Y is continuous, the main analysis model can be used directly to calculate effect size measures. If we divide ("standardize") the estimate of β_T and its confidence interval in our linear regression model by the pooled standard deviation of Y, we obtain an estimate of the between-group standardized mean difference (the well-known Cohen's d), as well as its confidence interval (see S5.1.4 in the tutorial).

There are also cases in which effect sizes and their significance play less of an important role, for example in pilot trials. Large-scale RCTs are time-consuming and costly, so external pilot trials are a helpful way to examine on a smaller scale if a new treatment can be successfully administered in the desired setting, how well the

recruitment strategy works, or if patients adhere to the treatment [90]. As emphasized by the 2016 CONSORT extension [91], pilot trials should focus on a pre-defined set of feasibility objectives (e.g., "at least 25% of eligible patients can be recruited for the trial" or "at least 85% of patients are retained by follow-up"). These feasibility objectives can also serve as progression criteria to determine if a confirmatory trial can safely be rolled out [92]. Although tempting, the primary goal of pilot trials is not to estimate an effect size on clinical outcomes, or its significance, because they typically do not have the power to detect such effects.

Overall, effect sizes are an apt conclusion for our tour because they are often used to "summarize" the results of a trial. Effect sizes are also what meta-analysts use to synthesize to results of multiple studies and often an entire research field. There is a risk to "reify" effect sizes derived from RCTs, to take them as "proof" that some treatment has a true, real effect. Looking back, we can now see how fragile effect sizes really are. They are only *estimates* of an inherently unobservable quantity that will only be valid if the assumptions of our statistical analyses are correct; many of which (like the underlying missingness mechanism) are untestable. RCTs can be easily "tweaked" to show an intervention effect even when there is none [93], and for many of these design flaws, there is no direct statistical remedy at all. This is sobering, but it underlines that effect sizes and *P* values alone are not sufficient to draw valid causal inferences from RCTs.

Discussion

It is crucial to keep in mind that clinical trial evaluations are an intricate topic and that this article barely scratches the surface. There is much more to learn about good practice in RCT analyses; a more in-depth and "hands-on" look at trial evaluations is provided in the tutorial in the supplement.

Furthermore, many problems of RCTs arise at the design stage, well before the actual analysis. Therefore, before concluding this primer, we also want to summarize a few more general shortcomings of RCTs as they are frequently observed in mental health research. Some of these problems are "human-made" and can be avoided by improving research practices and trial designs. Others are inherent limitations of RCTs in the mental health field that we have to keep in mind to draw valid inferences from them.

Avoidable limitations

One limitation of RCTs that is both widespread and easy to avoid is the lack of prospective registration. There is a broad consensus that the protocol of a trial, including its design, planned sample size, inclusion criteria, and primary outcome should be published *before* the first patient is recruited. The International Committee of Medical Journal Editors (ICMJE) has made prospective registration a condition for publication in one of their journals, and this mandate has been in effect since 2005 [94]. Nevertheless, many mental health researchers still fail to prospectively register their trial. For example, two meta-epidemiological studies found that only 15–40% of recent psychotherapy trials were prospectively registered [95–97], and similar numbers are also found in pharmacotherapy trials [98].

Without prospective registration, analytic details can easily be tweaked to make the results of a trial appear better than they really are. One of these "methods" is known as *outcome switching*: if our original primary outcome does not show the desired effect, one can simply switch to another assessed endpoint with better results to show that the intervention "worked". There is evidence that this and other post hoc discrepancies are widespread in mental health RCTs [33, 96, 99–101]. Naturally, the pressure of producing positive results this way may be most pronounced among trialists with financial interests in the treatment. Antidepressants are a commonly named example here [102], but similar conflicts of interest may also pertain to, e.g., the blooming "digital therapeutics" industry [103], who also need to show that their treatment is effective to sell it. The best way to circumvent these issues is to register a detailed protocol before the beginning of the trial in one of the primary registries listed by the WHO International Clinical Trials Registry Platform (ICTRP) and to analyze and report results in accordance with the original registration. The trial registration may also be supplemented with a statistical analysis plan [104, 105], which should define the trial estimand as well as the exact statistical procedures employed to estimate it. Core outcome sets (COS; see Table 1) should also be included at this stage to ensure that psychometrically valid instruments are used and to make it easier to compare the results to other trials.

A related problem is *allegiance bias*. Even without any obvious financial interests, some trialists may feel a strong sense of commitment to the treatment under study, for example because they have contributed to its development. There is a substantial body of research, especially for psychological treatments, that this type of allegiance can lead to inflated effect estimates in RCTs [106–109]. Allegiance biases can occur through different pathways. Trialist may, for example, provide better training or supervision for the personnel administering their "preferred" treatment, or they may simply have more experience in implementing it [93]. In psychotherapy research, for instance, non-directive counseling is often used as a control condition to which new interventions

are compared to. Since the researchers "favor" the new treatment, the non-directive intervention is frequently implemented as an "intent-to-fail" condition [110]. This is in contrast to empirical findings which show that non-directive supportive therapy is an effective treatment in its own right and that its purported inferiority to other psychotherapies may be caused by allegiance bias [111]. One way to prevent allegiance bias is to conduct an independent evaluation by researchers who have not been involved in the development of any of the studied treatments. Guidelines such as the Template for Intervention Description and Replication (TIDieR [112]) also remain underutilized in the mental health field, but can help to clarify the components of interventions or active control conditions, and how well they may compare to other trials.

Another common weak spot are the control groups used in mental health trials. For psychological interventions, waitlists are still one of the most frequently used comparators to determine the effectiveness of the treatment. An advantage of waitlist controls is that they allow to provide the intervention to all recruited participants; patients in the control group just have to wait for it until the end of the trial. However, there is evidence that waitlists may function as a type of "nocebo" in clinical trials: since patients know that they will receive an intervention soon anyway, they may be less inclined to solve their problems in the meantime [113–115]. For treatments of depression, for example, we know that patients on the waitlist typically fare worse than under care as usual and even worse than patients who receive no treatment at all [116, 117]. In this primer, we learned that the causal effect of a treatment can only be defined in reference to some other condition. Thus, to find out if our intervention is really beneficial to patients, we must choose a *plausible* comparator for our research question. This could be, for example, a care as usual group, or another established treatment for the mental health problem under study.

A last avoidable issue of mental health RCTs concerns their sample size. It is commonly understood that the number of participants to be included our trial should be determined in advance using a power analysis [118]. Nonetheless, there is evidence that most mental health trials are woefully underpowered. A recent meta-review found that RCTs in mood, anxiety, and psychotic disorder patients had a median power of 23%, which is far below the accepted level of 80% [26]. This finding is concerning because it implies that most trials in the field are unable to attain statistically significant results for the effect they try to measure. This opens the door for a host of systemic issues that we already discussed above: selective publication, reporting bias, and data dredging, which can all be read as attempts to squeeze out significant effects from studies that do not have the sample size to detect them in the first place. Obviously, clinical trials are costly and most trialists recruit such small samples for logistic reasons, not on purpose. Yet, in this case, it may be helpful to shift to trial designs that make more efficient use of the available sample, such as adaptive [119, 120], fractional factorial [121], stepped wedge [122], or pragmatic trial [123] designs. Failure to recruit the targeted sample size remains a widespread issue in both pharmacotherapy [124, 125] and psychotherapy trials [126–129]. Sometimes, it may still be impossible for trialists to reach the required sample size established via power analysis, but the resulting lack in statistical power should then be clearly named as a limitation of the trial. Naturally, a much better approach is to identify uptake barriers beforehand. Most reasons for recruitment failures have been found to be preventable, and pilot trials are a good way to assess how difficult it will be to enroll patients in practice [130].

Inherent limitations

There are also limitations of RCTs that are intrinsic to this research design or at least difficult to avoid in mental health research. One example is blinding. Pharmacological trials are typically designed in such a way that patients remain unaware of whether they are receiving the medication or a pill placebo. Similarly, clinicians who evaluate the outcomes are also kept blinded to the treatment assignments. Such double-blinded placebo-controlled trials are considered one of the strongest clinical trial designs, but even they can fail, for example if patients and raters recognize the side-effects of the tested medication [131]. Conducting a double-blinded trial of a psychological intervention is even more challenging and seldom attempted in practice [132]. This is because patients will typically be aware if they were assigned to a placebo condition designed to have no therapeutic effect or a "bona fide" psychological treatment.

Often, it is not easy to define for what exact placebo effects we should control for in RCTs of psychological interventions and what control groups can achieve this without breaking the blinding. For medical devices (e.g., digital mental health interventions), the U.S. FDA recommends using "sham" interventions to control for placebo effects (i.e., interventions that appear like the tested treatment, but deliver no actual therapy); but also acknowledges that constructing such control groups can be difficult [133]. Some authors have argued that the idea of a blinded placebo control does not apply to psychological interventions altogether, since both can be regarded as treatments that work solely through psychological means [134–138]; and that placebo controls should therefore be abandoned in psychotherapy research [136].

This conclusion is not uncontroversial, and others contend that some types of placebo controls can be helpful to test the "true" effect of psychological interventions [139].

Another limitation of RCTs relates to the ongoing efforts to "personalize" mental health care and to explore heterogeneous treatment effects (HTE [140–142]). Randomized trials are an excellent tool to determine the average effect of a treatment. Yet, in practice, we do not treat an "average" patient, but individuals. For mental disorders such as depression, various treatments with proven effects are available, but none of them work sufficiently well in all patients, and many people have to undergo several rounds of treatment until an effective therapy is found [143]. Thus, many researchers want to better understand which treatment "works for whom" and reliably predict which person will benefit most from which type of treatment.

In this primer, we already described that we will never know the true effect that a treatment had on an individual, since this effect is defined by counterfactual information. As described by Senn [144], this has crucial implications once we are interested in determining "personalized" treatment effects. Sometimes, the fact that some patients' symptoms improve strongly while on treatment, whereas others do not improve, is taken as an indication that the treatment effect must vary among individuals. If we develop a model that predicts the "response" to treatment based on pre-post data alone, we implicitly follow the same rationale. Looking at panel A in Fig. 1, we see that this is a deeply flawed logic, because we do not consider how individuals would have developed without treatment. Contrary to common belief, meta-analyses of variance ratios suggest that the improvements caused by antidepressants are mostly uniform across patients [145], while a higher degree of HTE may exist in psychological treatments [146].

At first glance, RCTs may seem like a more natural starting point to examine HTE because they are based on a counterfactual reasoning. Indeed, subgroup and moderator analyses are often used in RCT evaluations to assess if treatment effects differ for patient groups. The problem is that RCTs are typically not designed for these types of analyses. Most RCTs are barely sufficiently powered to detect the average treatment effect, let alone to provide a usable estimate of the effect in specific patient subgroups [88, 147]. This problem becomes even more severe if we make these subgroups even smaller by prognosticating "individualized" treatment benefits for patients with an identical combination of baseline characteristics, as is typically done in clinical prediction modeling [148, 149]. Several methods have been proposed to deal with this limitation; for example to examine effect modification in individual participant data meta-analysis (IPD-MA), in which many RCTs are combined into one big data set [150] or to develop causally interpretable models in large-scale observational data [151].

A similar problem is that RCTs can show *that* a treatment works, but they typically cannot reveal the mechanisms that make it effective. Unfortunately, more than a single RCT is usually needed to understand which components generate the effects of an intervention. For example, although numerous RCTs have demonstrated that psychotherapy is effective, there is a decades-long debate about its working mechanisms that has not been resolved until today [152].

A last inherent limitation of RCTs concerns their "generalizability" and "transportability" [153]. By generalizability, we mean that results within our study sample can be extended to the broader population from which our study sample was drawn. Transportability means that results of our trial also translate to another (partly) different context. A frequently voiced criticism of RCTs is that, although they have high internal validity, they often lack external validity [154–156]. Treatments in RCTs are often delivered in a highly standardized way and under tightly controlled conditions. This may not accurately reflect routine care, where healthcare providers may have much less time and resources at their disposal.

If this is valid criticism depends on the goals of trial. For a newly developed treatment, it may be more important to first show that it has a causal effect under optimized conditions, while tightly controlling for potential biases. On the other hand, pragmatic trials can be conducted to examine the effects of more established treatments under conditions that are closer to routine care and therefore also have a greater external validity [123]. Please to prioritize "real-world evidence" over RCTs have also been criticized on the grounds that high internal validity of randomized evidence is a prerequisite of external validity [157, 158] and that regulators should rather focus on reducing the bureaucratical burden associated with conducting RCTs [158].

A related concern involves the fact that RCTs often have very strict inclusion criteria, and therefore their findings may not be transportable to the actual target population [159]. Many trials exclude, e.g., individuals with subthreshold symptoms or comorbidities, or they may fail to recruit from minority groups and hard-to-reach populations [160–163]. Furthermore, RCTs only include patients who actively decide to participate in such a study, which means that trial samples can often be very selective. This means that a treatment may be effective in a homogenous trial sample, but less so in the more diverse target population in which it should be implemented.

Representativeness is an important consideration in clinical trial design, but it is not automatically true that randomized evidence from more restricted samples is not transportable to another population. This may only be the case if there is heterogeneity of treatment effects [153]. We mentioned that, even though it might look different at first, the true effect of some treatments can be quite homogeneous. If there are no strong effect modifiers, there is also no strong reason to believe that the treatment will have a substantially different effect in populations with varying patient characteristics. In this case, the ATE of our RCTs also provides an internally valid estimate of the effect we can expect in the target population. Of course, there are many scenarios in which HTE are indeed plausible. Then, we have to use methods that extend beyond the trial data to accurately estimate the causal effect of our treatment in a different target population of interest [164–166]. Mueller and Pearl [167] make the compelling case that RCT and routine care data, when combined, can allow to make more informed decisions about the individual risks or benefits of a treatment that may remain undetected when looking at RCTs alone. This underlines that experimental and observational studies both have their place in mental health research and that we obtain better inferences if various sources of data are considered—a practice that some refer to as "data fusion" [168].

This concludes our journey through some of the pitfalls that we should keep in mind when we evaluate and interpret the results of an RCT. We hope that this primer and tutorial delivered some helpful insights for mental health researchers. More importantly, we also hope our introduction illustrates that trial methodology is a fascinating topic worthy of further exploration.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13063-023-07596-3.

Additional file 1.

Acknowledgements
We would like to thank Stella Wernicke for her helpful feedback on this article.

Authors' contributions
MH and PC had the idea for this article. MH and LKJS and drafted a first version of the manuscript and supplementary tutorial under supervision of PC. All authors contributed to the further development of the manuscript. All authors read and approved the final manuscript.

Funding
Open Access funding enabled and organized by Projekt DEAL. MH is supported by a fellowship of the Bavarian Research Institute for Digital Transformation (BIDT), an institute of the Bavarian Academy of Sciences and Humanities.

Availability of data and materials
All material used to compile the tutorial presented in the supplement is openly available on Github (github.com/mathiasharrer/rct-tutorial).

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
DDE is a stakeholder of the Institute for Health Trainings Online (GET.ON), which aims to implement scientific findings related to digital health interventions into routine care.

Author details
[1]Psychology and Digital Mental Health Care, Technical University Munich, Georg-Brauchle-Ring 60-62, Munich 80992, Germany. [2]Clinical Psychology and Psychotherapy, Institute for Psychology, Friedrich-Alexander-University Erlangen-Nuremberg, Erlangen, Germany. [3]Department of Clinical, Neuro and Developmental Psychology, Amsterdam Public Health Research Institute, Vrije Universiteit Amsterdam, Amsterdam, the Netherlands. [4]WHO Collaborating Centre for Research and Dissemination of Psychological Interventions, Vrije Universiteit Amsterdam, Amsterdam, the Netherlands. [5]Methods and Evaluation/Quality Assurance, Freie Universität Berlin, Berlin, Germany. [6]Institute of Social Medicine and Health Systems Research (ISMHSR), Medical Faculty, Otto Von Guericke University Magdeburg, Magdeburg, Germany.

Received: 16 May 2023 Accepted: 18 August 2023
Published online: 30 August 2023

References

1. Jones DS, Podolsky SH. The history and fate of the gold standard. Lancet. 2015;385(9977):1502–3. Elsevier.
2. Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, Norris S, Falck-Ytter Y, Glasziou P, DeBeer H. Introduction evidence profiles and summary of findings tables. J Clin Epidemiol. 2011;64(4):383–94. Elsevier.
3. Howick J, Chalmers I, Glasziou P, Greenhalgh T, Heneghan C, Liberati A, Moschetti I, Phillips B, Thornton H. The 2011 Oxford CEBM evidence levels of evidence (introductory document). Oxf Cent Evid Based Med. 2011. http://www.cebm.net/index.aspx?o=5653.
4. Cartwright N. Predicting what will happen when we act. What counts for warrant? Prev Med. 2011;53(4–5):221–4. Elsevier.
5. Deaton A, Cartwright N. Understanding and misunderstanding randomized controlled trials. Soc Sci Med. 2018;210:2–21. Elsevier.
6. Altman DG. The scandal of poor medical research. BMJ. 1994;308(6924):283–4. British Medical Journal Publishing Group.
7. Van Calster B, Wynants L, Riley RD, van Smeden M, Collins GS. Methodology over metrics: current scientific standards are a disservice to patients and society. J Clin Epidemiol. 2021;138:219–26. Elsevier.
8. Pirosca S, Shiely F, Clarke M, Treweek S. Tolerating bad health research: the continuing scandal. Trials. 2022;23(1):1–8. BioMed Central.
9. Bell ML, Fiero M, Horton NJ, Hsu C-H. Handling missing data in RCTs; a review of the top medical journals. BMC Med Res Methodol. 2014;14(1):118. https://doi.org/10.1186/1471-2288-14-118.
10. Akl EA, Briel M, You JJ, Sun X, Johnston BC, Busse JW, Mulla S, Lamontagne F, Bassler D, Vera C, et al. Potential impact on estimated treatment effects of information lost to follow-up in randomised controlled trials (LOST-IT): systematic review. BMJ. 2012;344:e2809. British Medical Journal Publishing Group.
11. Akl EA, Shawwa K, Kahale LA, Agoritsas T, Brignardello-Petersen R, Busse JW, Carrasco-Labra A, Ebrahim S, Johnston BC, Neumann I, et al. Reporting missing participant data in randomised trials: systematic survey

of the methodological literature and a proposed guide. BMJ Open. 2015;5(12):e008431. British Medical Journal Publishing Group.
12. Powney M, Williamson P, Kirkham J, Kolamunnage-Dona R. A review of the handling of missing longitudinal outcome data in clinical trials. Trials. 2014;15(1):1–11. BioMed Central.
13. Rabe BA, Day S, Fiero MH, Bell ML. Missing data handling in non-inferiority and equivalence trials: a systematic review. Pharm Stat. 2018;17(5):477–88. Online Library.
14. Cro S, Morris TP, Kenward MG, Carpenter JR. Sensitivity analysis for clinical trials with missing continuous outcome data using controlled multiple imputation: a practical guide. Stat Med. 2020;39(21):2815–42. Wiley Online Library.
15. Moher D, Hopewell S, Schulz KF, Montori V, Gøtzsche PC, Devereaux PJ, Elbourne D, Egger M, Altman DG. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. Int J Surg. 2012;10(1):28–55. Elsevier.
16. Senn S. Testing for baseline balance in clinical trials. Stat Med. 1994;13(17):1715–26. Wiley Online Library.
17. Altman DG, Dore C. Randomisation and baseline comparisons in clinical trials. Lancet. 1990;335(8682):149–53. Elsevier.
18. Begg CB. Significance tests of covariate imbalance in clinical trials. Control Clin Trials. 1990;11(4):223–5. Elsevier.
19. De Boer MR, Waterlander WE, Kuijper LD, Steenhuis IH, Twisk JW. Testing for baseline differences in randomized controlled trials: an unhealthy research behavior that is hard to eradicate. Int J Behav Nutr Phys Act. 2015;12(1):1–8. BioMed Central.
20. Pijls BG. The table I fallacy: p values in baseline tables of randomized controlled trials. J Bone Joint Surg. 2022;104(16):e71. Lippincott Williams & Wilkins.
21. Cuijpers P, Weitz E, Cristea I, Twisk J. Pre-post effect sizes should be avoided in meta-analyses. Epidemiol Psychiatr Sci. 2017;26(4):364–8. Cambridge University Press.
22. Bland JM, Altman DG. Comparisons against baseline within randomised groups are often used and can be highly misleading. Trials. 2011;12(1):1–7. BioMed Central.
23. Bland JM, Altman DG. Best (but oft forgotten) practices: testing for treatment effects in randomized trials by separate analyses of changes from baseline in each group is a misleading approach. Am J Clin Nutr. 2015;102(5):991–4. Oxford University Press.
24. Altman DG, Bland JM. Statistics notes: absence of evidence is not evidence of absence. BMJ. 1995;311(7003):485. British Medical Journal Publishing Group.
25. Alderson P. Absence of evidence is not evidence of absence. BMJ. 2004;328:476–7. British Medical Journal Publishing Group.
26. de Vries YA, Schoevers RA, Higgins JP, Munafò MR, Bastiaansen JA. Statistical power in clinical trials of interventions for mood, anxiety, and psychotic disorders. Psychol Med. 2022;53(10):4499–4506. Cambridge University Press.
27. Hoenig JM, Heisey DM. The abuse of power: the pervasive fallacy of power calculations for data analysis. Am Stat. 2001;55(1):19–24. Taylor & Francis.
28. Althouse AD. Post hoc power: not empowering, just misleading. J Surg Res. 2021;259:A3–6. Elsevier.
29. Kane RL, Wang J, Garrard J. Reporting in randomized clinical trials improved after adoption of the CONSORT statement. J Clin Epidemiol. 2007;60(3):241–9. Elsevier.
30. Plint AC, Moher D, Morrison A, Schulz K, Altman DG, Hill C, Gaboury I. Does the CONSORT checklist improve the quality of reports of randomised controlled trials? A systematic review. Med J Aust. 2006;185(5):263–7. Wiley Online Library.
31. Dal-Ré R, Bobes J, Cuijpers P. Why prudence is needed when interpreting articles reporting clinical trial results in mental health. Trials. 2017;18(1):1–4. BioMed Central.
32. Song SY, Kim B, Kim I, Kim S, Kwon M, Han C, Kim E. Assessing reporting quality of randomized controlled trial abstracts in psychiatry: adherence to CONSORT for abstracts: a systematic review. PLoS One. 2017;12(11):e0187807. Public Library of Science San Francisco, CA USA.
33. Miguel C, Karyotaki E, Cuijpers P, Cristea IA. Selective outcome reporting and the effectiveness of psychotherapies for depression. World Psychiatry. 2021;20(3):444. World Psychiatric Association.
34. Altman DG, Moher D, Schulz KF. Harms of outcome switching in reports of randomised trials: CONSORT perspective. BMJ. 2017;356:j396. British Medical Journal Publishing Group.
35. Williamson PR, Altman DG, Blazeby JM, Clarke M, Devane D, Gargon E, Tugwell P. Developing core outcome sets for clinical trials: issues to consider. Trials. 2012;13(1):1–8. BioMed Central.
36. Chevance A, Ravaud P, Tomlinson A, Berre CL, Teufer B, Touboul S, Fried EI, Gartlehner G, Cipriani A, Tran VT. Identifying outcomes for depression that matter to patients, informal caregivers, and health-care professionals: qualitative content analysis of a large international online survey. Lancet Psychiatry. 2020;7(8):692–702. Elsevier. PMID:32711710.
37. Prevolnik Rupel V, Jagger B, Fialho LS, Chadderton L-M, Gintner T, Arntz A, Baltzersen Å-L, Blazdell J, van Busschbach J, Cencelli M, Chanen A, Delvaux C, van Gorp F, Langford L, McKenna B, Moran P, Pacheco K, Sharp C, Wang W, Wright K, Crawford MJ. Standard set of patient-reported outcomes for personality disorder. Qual Life Res. 2021;30(12):3485–500. https://doi.org/10.1007/s11136-021-02870-w.
38. Krause KR, Chung S, Adewuya AO, Albano AM, Babins-Wagner R, Birkinshaw L, Brann P, Creswell C, Delaney K, Falissard B, Forrest CB, Hudson JL, Ishikawa S, Khatwani M, Kieling C, Krause J, Malik K, Martínez V, Mughal F, Ollendick TH, Ong SH, Patton GC, Ravens-Sieberer U, Szatmari P, Thomas E, Walters L, Young B, Zhao Y, Wolpert M. International consensus on a standard set of outcome measures for child and youth anxiety, depression, obsessive-compulsive disorder, and post-traumatic stress disorder. Lancet Psychiatry. 2021;8(1):76–86. Elsevier. PMID:33341172.
39. Retzer A, Sayers R, Pinfold V, Gibson J, Keeley T, Taylor G, Plappert H, Gibbons B, Huxley P, Mathers J, Birchwood M, Calvert M. Development of a core outcome set for use in community-based bipolar trials—a qualitative study and modified Delphi. PLoS One. 2020;15(10):e0240518. https://doi.org/10.1371/journal.pone.0240518. Public Library of Science.
40. Karnik NS, Marsden J, McCluskey C, Boley RA, Bradley KA, Campbell CI, Curtis ME, Fiellin D, Ghitza U, Hefner K, Hser Y-I, McHugh RK, McPherson SM, Mooney LJ, Moran LM, Murphy SM, Schwartz RP, Shmueli-Blumberg D, Shulman M, Stephens KA, Watkins KE, Weiss RD, Wu L-T. The opioid use disorder core outcomes set (OUD–COS) for treatment research: findings from a Delphi consensus study. Addiction. 2022;117(9):2438–47. https://doi.org/10.1111/add.15875.
41. McKenzie E, Matkin L, Sousa Fialho L, Emelurumonye IN, Gintner T, Ilesanmi C, Jagger B, Quinney S, Anderson E, Baandrup L, Bakhshy AK, Brabban A, Coombs T, Correll CU, Cupitt C, Keetharuth AD, Lima DN, McCrone P, Moller M, Mulder CL, Roe D, Sara G, Shokraneh F, Sin J, Woodberry KA, Addington D. Developing an international standard set of patient-reported outcome measures for psychotic disorders. Psychiatr Serv. 2022;73(3):249–58. https://doi.org/10.1176/appi.ps.202000888. American Psychiatric Publishing.
42. Williamson PR, Altman DG, Bagley H, Barnes KL, Blazeby JM, Brookes ST, Clarke M, Gargon E, Gorst S, Harman N, Kirkham JJ, McNair A, Prinsen CAC, Schmitt J, Terwee CB, Young B. The COMET Handbook: version 1.0. Trials. 2017;18(3):280. https://doi.org/10.1186/s13063-017-1978-4.
43. Buntrock C, Ebert DD, Lehr D, Smit F, Riper H, Berking M, Cuijpers P. Effect of a web-based guided self-help intervention for prevention of major depression in adults with subthreshold depression: a randomized clinical trial. JAMA. 2016;315(17):1854–63. American Medical Association.
44. Ebert DD, Buntrock C, Lehr D, Smit F, Riper H, Baumeister H, Cuijpers L, Berking M. Effectiveness of web-and mobile-based treatment of subthreshold depression with adherence-focused guidance: a single-blind randomized controlled trial. Behav Ther. 2018;49(1):71–83. Elsevier.
45. Holland PW. Statistics and causal inference. J Am Stat Assoc. 1986;81(396):945–60. Taylor & Francis.
46. Aronow PM, Miller BT. Identification with potential outcomes. In: Found Agnostic Stat. 1st ed. New York: Cambridge University Press; 2019.
47. Imbens GW, Rubin DB. Causal inference for statistics, social, and biomedical sciences. 1st ed. New York: Cambridge University Press; 2015. ISBN:978-0-521-88588-1.
48. Roefs A, Fried EI, Kindt M, Martijn C, Elzinga B, Evers AW, Wiers RW, Borsboom D, Jansen A. A new science of mental disorders: using personalised, transdiagnostic, dynamical systems to understand, model, diagnose and treat psychopathology. Behav Res Ther. 2022;153:104096. Elsevier.

49. Whiteford HA, Harris M, McKeon G, Baxter A, Pennell C, Barendregt J, Wang J. Estimating remission from untreated major depression: a systematic review and meta-analysis. Psychol Med. 2013;43(8):1569–85. Cambridge University Press.
50. Vegetabile BG. On the distinction between "conditional average treatment effects" (CATE) and "individual treatment effects" (ITE) under ignorability assumptions. ArXiv210804939 Cs Stat 2021 Aug 10. Available from: http://arxiv.org/abs/2108.04939. Accessed 25 Feb 2022.
51. Greenland S, Robins JM. Identifiability, exchangeability and confounding revisited. Epidemiol Perspect Innov. 2009;6(1):1–9. BioMed Central.
52. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol. 1974;66(5):688–701. https://doi.org/10.1037/h0037350. US: American Psychological Association.
53. Rubin DB. Inference and missing data. Biometrika. 1976;63(3):581–92. Oxford University Press.
54. Schouten RM, Vink G. The dance of the mechanisms: how observed information influences the validity of missingness assumptions. Sociol Methods Res. 2021;50(3):1243–58. SAGE Publications Sage CA: Los Angeles, CA.
55. Greenland S, Morgenstern H. Confounding in health research. Annu Rev Public Health. 2001;22:189. Annual Reviews, Inc.
56. Hernán MA. Beyond exchangeability: the other conditions for causal inference in medical research. Stat Methods Med Res. 2012;21(1):3–5. Sage Publications Sage UK: London, England.
57. Clifton L, Clifton DA. The correlation between baseline score and post-intervention score, and its implications for statistical analysis. Trials. 2019;20(1):43. https://doi.org/10.1186/s13063-018-3108-3.
58. Harrell Jr FE. Statistical errors in the medical literature. Available from: https://web.archive.org/web/20230319005251/https://www.fharrell.com/post/errmed/. Accessed date 2023-03-19.
59. Aronow P, Robins JM, Saarinen T, Sävje F, Sekhon J. Nonparametric identification is not enough, but randomized controlled trials are. ArXiv Prepr ArXiv210811342. 2021.
60. Senn S. Seven myths of randomisation in clinical trials. Stat Med. 2013;32(9):1439–50. Wiley Online Library.
61. Tackney MS, Morris T, White I, Leyrat C, Diaz-Ordaz K, Williamson E. A comparison of covariate adjustment approaches under model misspecification in individually randomized trials. Trials. 2023;24(1):1–18. BioMed Central.
62. Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. Epidemiology. 2004;15(5):615–25. JSTOR.
63. van Buuren S. Implications of ignorability. In: Flex Imput Missing Data. Boca Raton: Chapman and Hall/CRC; 2018.
64. Clark TP, Kahan BC, Phillips A, White I, Carpenter JR. Estimands: bringing clarity and focus to research questions in clinical trials. BMJ Open. 2022;12(1):e052953. British Medical Journal Publishing Group.
65. EMA. ICH E9 (R1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials. 2020. Available from: https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e9-r1-addendum-estimands-sensitivity-analysis-clinical-trials-guideline-statistical-principles_en.pdf. Accessed 12 Jan 2023.
66. FDA. E9(R1) Statistical principles for clinical trials: addendum: estimands and sensitivity analysis in clinical trials. 2021. Available from: https://www.fda.gov/media/148473/download. Accessed 12 Jan 2023.
67. Cro S, Kahan BC, Rehal S, Ster AC, Carpenter JR, White IR, Cornelius VR. Evaluating how clear the questions being investigated in randomised trials are: systematic review of estimands. BMJ. 2022;378:e070146. British Medical Journal Publishing Group.
68. Kahan BC, Morris TP, White IR, Carpenter J, Cro S. Estimands in published protocols of randomised trials: urgent improvement needed. Trials. 2021;22(1):1–10. BioMed Central.
69. Singal AG, Higgins PDR, Waljee AK. A primer on effectiveness and efficacy trials. Clin Transl Gastroenterol. 2014;5(1):e45. PMID:24384867.
70. Han S, Zhou X-H. Defining estimands in clinical trials: a unified procedure. Stat Med. 2023;42(12):1869–87. Wiley Online Library.
71. Pétavy F, Guizzaro L, Antunes dos Reis I, Teerenstra S, Roes KCB. Beyond "intent-to-treat" and "per protocol": improving assessment of treatment effects in clinical trials through the specification of an estimand. Br J Clin Pharmacol. 2020;86(7):1235–9. PMID:31883123.
72. Fletcher C, Hefting N, Wright M, Bell J, Anzures-Cabrera J, Wright D, Lynggaard H, Schueler A. Marking 2-years of new thinking in clinical trials: the estimand journey. Ther Innov Regul Sci. 2022;56(4):637–50. Springer.
73. Bornkamp B, Rufibach K, Lin J, Liu Y, Mehrotra DV, Roychoudhury S, Schmidli H, Shentu Y, Wolbers M. Principal stratum strategy: potential role in drug development. Pharm Stat. 2021;20(4):737–51. Wiley Online Library.
74. Rubin DB. Multiple imputation after 18+ years. J Am Stat Assoc. 1996;91(434):473–89. Taylor & Francis.
75. Barnard J, Rubin DB. Small-sample degrees of freedom with multiple imputation. Biometrika. 1999;86(4):948–55. [Oxford University Press, Biometrika Trust].
76. Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, Wood AM, Carpenter JR. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009;338:b2393. British Medical Journal Publishing Group.
77. Bartlett JW, Seaman SR, White IR, Carpenter JR, Alzheimer's disease initiative*. Multiple imputation of covariates by fully conditional specification: accommodating the substantive model. Stat Methods Med Res. 2015;24(4):462–87. Sage Publications Sage UK: London, England.
78. Meng X-L. Multiple-imputation inferences with uncongenial sources of input. Stat Sci. 1994;9(4):538–58. JSTOR.
79. Twisk JW. Analysis of RCT data with one follow-up measurement. In: anal data randomized control trials pract guide. 1st ed. Cham (CH): Springer; 2021.
80. Twisk JW. Analysis of RCT data with more than one follow-up measurement. In: anal data randomized control trials pract guide. 1st ed. Cham (CH): Springer; 2021.
81. Lehman EL. The design of experiments and sample surveys. In: fish Neyman Creat class Stat. 1st ed. New York: Springer; 2011.
82. Kahan BC, Jairath V, Doré CJ, Morris TP. The risks and rewards of covariate adjustment in randomized trials: an assessment of 12 outcomes from 8 studies. Trials. 2014;15(1):1–7. BioMed Central.
83. Johansson P, Nordin M. Inference in experiments conditional on observed imbalances in covariates. Am Stat. 2022;76(4):394–404. Taylor & Francis.
84. Hauck WW, Anderson S, Marcus SM. Should we adjust for covariates in nonlinear regression analyses of randomized trials? Control Clin Trials. 1998;19(3):249–56. Elsevier.
85. Jiang H, Kulkarni PM, Mallinckrodt CH, Shurzinske L, Molenberghs G, Lipkovich I. Covariate adjustment for logistic regression analysis of binary clinical trial data. Stat Biopharm Res. 2017;9(1):126–34. Taylor & Francis.
86. Cummings P. The relative merits of risk ratios and odds ratios. Arch Pediatr Adolesc Med. 2009;163(5):438–45. American Medical Association.
87. Daniel R, Zhang J, Farewell D. Making apples from oranges: comparing noncollapsible effect estimators and their standard errors after adjustment for different covariate sets. Biom J. 2021;63(3):528–57. Wiley Online Library.
88. Permutt T. Do covariates change the estimand? Stat Biopharm Res. 2020;12(1):45–53. Taylor & Francis.
89. Xiao M, Chu H, Cole SR, Chen Y, MacLehose RF, Richardson DB, Greenland S. Controversy and debate: questionable utility of the relative risk in clinical research: paper 4: odds ratios are far from "portable"—a call to use realistic models for effect variation in meta-analysis. J Clin Epidemiol. 2022;142:294–304. Elsevier.
90. Kistin C, Silverstein M. Pilot studies: a critical but potentially misused component of interventional research. JAMA. 2015;314(15):1561–2. https://doi.org/10.1001/jama.2015.10962.
91. Eldridge SM, Chan CL, Campbell MJ, Bond CM, Hopewell S, Thabane L, Lancaster GA. CONSORT 2010 statement: extension to randomised pilot and feasibility trials. BMJ. 2016;355:i5239. British Medical Journal Publishing Group.
92. Avery KN, Williamson PR, Gamble C, Francischetto EO, Metcalfe C, Davidson P, Williams H, Blazeby JM. Informing efficient randomised controlled trials: exploration of challenges in developing progression criteria for internal pilot studies. BMJ Open. 2017;7(2):e013537. British Medical Journal Publishing Group.
93. Cuijpers P, Cristea I. How to prove that your therapy is effective, even when it is not: a guideline. Epidemiol Psychiatr Sci. 2016;25(5):428–35. Cambridge University Press.

94. De Angelis C, Drazen JM, Frizelle FA, Haug C, Hoey J, Horton R, Kotzin S, Laine C, Marusic A, Overbeke AJP. Clinical trial registration: a statement from the International Committee of Medical Journal Editors. Lancet. 2004;364(9438):911–2. Elsevier.
95. Bradley HA, Rucklidge JJ, Mulder RT. A systematic review of trial registration and selective outcome reporting in psychotherapy randomized controlled trials. Acta Psychiatr Scand. 2017;135(1):65–77. https://doi.org/10.1111/acps.12647.
96. Stoll M, Mancini A, Hubenschmid L, Dreimüller N, König J, Cuijpers P, Barth J, Lieb K. Discrepancies from registered protocols and spin occurred frequently in randomized psychotherapy trials—a meta-epidemiologic study. J Clin Epidemiol. 2020;1(128):49–56. https://doi.org/10.1016/j.jclinepi.2020.08.013.
97. Cybulski L, Mayo-Wilson E, Grant S. Improving transparency and reproducibility through registration: the status of intervention trials published in clinical psychology journals. J Consult Clin Psychol. 2016;84(9):753–67. PMID:27281372.
98. Shinohara K, Tajika A, Imai H, Takeshima N, Hayasaka Y, Furukawa TA. Protocol registration and selective outcome reporting in recent psychiatry trials: new antidepressants and cognitive behavioural therapies. Acta Psychiatr Scand. 2015;132(6):489–98. https://doi.org/10.1111/acps.12502.
99. Roest AM, de Jonge P, Williams CD, de Vries YA, Schoevers RA, Turner EH. Reporting bias in clinical trials investigating the efficacy of second-generation antidepressants in the treatment of anxiety disorders: a report of 2 meta-analyses. JAMA Psychiat. 2015;72(5):500–10. https://doi.org/10.1001/jamapsychiatry.2015.15.
100. Turner EH, Cipriani A, Furukawa TA, Salanti G, de Vries YA. Selective publication of antidepressant trials and its influence on apparent efficacy: updated comparisons and meta-analyses of newer versus older trials. PLoS Med. 2022;19(1):e1003886. https://doi.org/10.1371/journal.pmed.1003886. Public Library of Science.
101. Turner EH, Matthews AM, Linardatos E, Tell RA, Rosenthal R. Selective publication of antidepressant trials and its influence on apparent efficacy. N Engl J Med. 2008;358(3):252–60. Massachusetts Medical Society. PMID:18199864.
102. Ioannidis JP. Effectiveness of antidepressants: an evidence myth constructed from a thousand randomized trials? Philos Ethics Humanit Med. 2008;3(1):14. https://doi.org/10.1186/1747-5341-3-14.
103. Wang C, Lee C, Shin H. Digital therapeutics from bench to bedside. Npj Digit Med. 2023;6(1):1–10. https://doi.org/10.1038/s41746-023-00777-z. Nature Publishing Group.
104. Gamble C, Krishan A, Stocken D, Lewis S, Juszczak E, Doré C, Williamson PR, Altman DG, Montgomery A, Lim P, Berlin J, Senn S, Day S, Barbachano Y, Loder E. Guidelines for the content of statistical analysis plans in clinical trials. JAMA. 2017;318(23):2337–43. https://doi.org/10.1001/jama.2017.18556.
105. Kahan BC, Forbes G, Cro S. How to design a pre-specified statistical analysis approach to limit p-hacking in clinical trials: the Pre-SPEC framework. BMC Med. 2020;18(1):253. https://doi.org/10.1186/s12916-020-01706-7.
106. Leykin Y, DeRubeis RJ. Allegiance in psychotherapy outcome research: separating association from bias. Clin Psychol Sci Pract. 2009;16(1):54. Wiley-Blackwell Publishing Ltd.
107. Dragioti E, Dimoliatis I, Fountoulakis KN, Evangelou E. A systematic appraisal of allegiance effect in randomized controlled trials of psychotherapy. Ann Gen Psychiatry. 2015;14:1–9. Springer.
108. Munder T, Brütsch O, Leonhart R, Gerger H, Barth J. Researcher allegiance in psychotherapy outcome research: an overview of reviews. Clin Psychol Rev. 2013;33(4):501–11. Elsevier.
109. Munder T, Gerger H, Trelle S, Barth J. Testing the allegiance bias hypothesis: a meta-analysis. Psychother Res. 2011;21(6):670–84. Taylor & Francis.
110. Baskin TW, Tierney SC, Minami T, Wampold BE. Establishing specificity in psychotherapy: a meta-analysis of structural equivalence of placebo controls. J Consult Clin Psychol. 2003;71(6):973. American Psychological Association.
111. Cuijpers P, Driessen E, Hollon SD, van Oppen P, Barth J, Andersson G. The efficacy of non-directive supportive therapy for adult depression: a meta-analysis. Clin Psychol Rev. 2012;32(4):280–91. Elsevier.
112. Hoffmann TC, Glasziou PP, Boutron I, Milne R, Perera R, Moher D, Altman DG, Barbour V, Macdonald H, Johnston M, Lamb SE, Dixon-Woods M, McCulloch P, Wyatt JC, Chan A-W, Michie S. Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. BMJ. 2014;348:g1687. British Medical Journal Publishing Group. PMID:24609605.
113. Furukawa TA, Noma H, Caldwell DM, Honyashiki M, Shinohara K, Imai H, Churchill R. Waiting list may be a nocebo condition in psychotherapy trials: a contribution from network meta-analysis. Acta Psychiatr Scand. 2014;130(3):181–92.
114. Mohr DC, Spring B, Freedland KE, Beckner V, Arean P, Hollon SD, Ockene J, Kaplan R. The selection and design of control conditions for randomized controlled trials of psychological interventions. Psychother Psychosom. 2009;78(5):275–84. Karger Publishers.
115. Mohr DC, Ho J, Hart TL, Baron KG, Berendsen M, Beckner V, Cai X, Cuijpers P, Spring B, Kinsinger SW. Control condition design and implementation features in controlled trials: a meta-analysis of trials evaluating psychotherapy for depression. Transl Behav Med. 2014;4(4):407–23. Oxford University Press.
116. Michopoulos I, Furukawa TA, Noma H, Kishimoto S, Onishi A, Ostinelli EG, Ciharova M, Miguel C, Karyotaki E, Cuijpers P. Different control conditions can produce different effect estimates in psychotherapy trials for depression. J Clin Epidemiol. 2021;132:59–70. Elsevier.
117. Cuijpers P, Miguel C, Harrer M, Plessen CY, Ciharova M, Ebert D, Karyotaki E. Cognitive behavior therapy vs. control conditions, other psychotherapies, pharmacotherapies and combined treatment for depression: a comprehensive meta-analysis including 409 trials with 52,702 patients. World Psychiatry. 2023;22(1):105–15. Wiley Online Library.
118. Schulz KF, Grimes DA. Sample size calculations in randomised trials: mandatory and mystical. Lancet. 2005;365(9467):1348–53. Elsevier.
119. Pallmann P, Bedding AW, Choodari-Oskooei B, Dimairo M, Flight L, Hampson LV, Holmes J, Mander AP, Odondi L, Sydes MR, Villar SS, Wason JMS, Weir CJ, Wheeler GM, Yap C, Jaki T. Adaptive designs in clinical trials: why use them, and how to run and report them. BMC Med. 2018;16(1):29. https://doi.org/10.1186/s12916-018-1017-7.
120. Blackwell SE, Schönbrodt FD, Woud ML, Wannemüller A, Bektas B, Braun Rodrigues M, Hirdes J, Stumpp M, Margraf J. Demonstration of a "leapfrog" randomized controlled trial as a method to accelerate the development and optimization of psychological interventions. Psychol Med. 2022;4:1–11. PMID:36330836.
121. Chakraborty B, Collins LM, Strecher VJ, Murphy SA. Developing multicomponent interventions using fractional factorial designs. Stat Med. 2009;28(21):2687–708. https://doi.org/10.1002/sim.3643.
122. Hussey MA, Hughes JP. Design and analysis of stepped wedge cluster randomized trials. Contemp Clin Trials. 2007;28(2):182–91. https://doi.org/10.1016/j.cct.2006.05.007.
123. Ford I, Norrie J. Pragmatic trials. N Engl J Med. 2016;375(5):454–63. Massachusetts Medical Society. PMID:27518663.
124. Haberfellner EM. Recruitment of depressive patients for a controlled clinical trial in a psychiatric practice. Pharmacopsychiatry. 2000;33(4):142–4. PMID:10958263.
125. Rendell J, Licht R. Under-recruitment of patients for clinical trials: an illustrative example of a failed study. Acta Psychiatr Scand. 2007;115:337–9. Blackwell Publishing.
126. Brown JSL, Murphy C, Kelly J, Goldsmith K. How can we successfully recruit depressed people? Lessons learned in recruiting depressed participants to a multi-site trial of a brief depression intervention (the 'CLASSIC' trial). Trials. 2019;20(1):131. https://doi.org/10.1186/s13063-018-3033-5.
127. Bolinski F, Kleiboer A, Neijenhuijs K, Karyotaki E, Wiers R, de Koning L, Jacobi C, Zarski A-C, Weisel KK, Cuijpers P. Challenges in recruiting university students for web-based indicated prevention of depression and anxiety: results from a randomized controlled trial (ICare Prevent). J Med Internet Res. 2022;24(12):e40892. JMIR Publications Toronto, Canada.
128. Woodford J, Farrand P, Bessant M, Williams C. Recruitment into a guided internet based CBT (iCBT) intervention for depression: lesson learnt from the failure of a prevalence recruitment strategy. Contemp Clin Trials. 2011;32(5):641–8. https://doi.org/10.1016/j.cct.2011.04.013.
129. Fairhurst K, Dowrick C. Problems with recruitment in a randomized controlled trial of counselling in general practice: causes and implications.

J Health Serv Res Policy. 1996;1(2):77–80. SAGE Publications Sage UK: London, England.
130. Briel M, Olu KK, von Elm E, Kasenda B, Alturki R, Agarwal A, Bhatnagar N, Schandelmaier S. A systematic review of discontinued trials suggested that most reasons for recruitment failure were preventable. J Clin Epidemiol. 2016;1(80):8–15. https://doi.org/10.1016/j.jclinepi.2016.07.016.
131. Schulz KF, Grimes DA. Blinding in randomised trials: hiding who got what. Lancet. 2002;359(9307):696–700. Elsevier.
132. Juul S, Gluud C, Simonsen S, Frandsen FW, Kirsch I, Jakobsen JC. Blinding in randomised clinical trials of psychological interventions: a retrospective study of published trial reports. BMJ Evid-Based Med. 2021;26(3):109–109. Royal Society of Medicine. PMID:32998993.
133. Food And Drugs Administration. Design considerations for pivotal clinical investigations for medical devices: guidance for industry, clinical investigators, institutional review boards and FDA staff (FDA-2011-D-0567). 2013. Available from: http://web.archive.org/web/20221122041708/https://www.fda.gov/media/87363/download.
134. Rosenthal D, Frank JD. Psychotherapy and the placebo effect. Psychol Bull. 1956;53(4):294. American Psychological Association.
135. Wampold BE, Frost ND, Yulish NE. Placebo effects in psychotherapy: a flawed concept and a contorted history. Psychol Conscious Theory Res Pract. 2016;3:108–20. https://doi.org/10.1037/cns0000045. US: Educational Publishing Foundation.
136. Kirsch I, Wampold B, Kelley JM. Controlling for the placebo effect in psychotherapy: noble quest or tilting at windmills? Psychol Conscious Theory Res Pract. 2016;3(2):121. Educational Publishing Foundation.
137. Kirsch I. Placebo psychotherapy: synonym or oxymoron? J Clin Psychol. 2005;61(7):791–803. https://doi.org/10.1002/jclp.20126.
138. Justman S. From medicine to psychotherapy: the placebo effect. Hist Hum Sci. 2011;24(1):95–107. https://doi.org/10.1177/0952695110386655. SAGE Publications Ltd.
139. Gaab J, Locher C, Blease C. Chapter thirteen - placebo and psychotherapy: differences, similarities, and implications. In: Colloca L, editor. Int Rev Neurobiol. Academic; 2018. p. 241–255. https://doi.org/10.1016/bs.irn.2018.01.013.
140. Fernandes BS, Williams LM, Steiner J, Leboyer M, Carvalho AF, Berk M. The new field of 'precision psychiatry.' BMC Med. 2017;15(1):80. https://doi.org/10.1186/s12916-017-0849-x.
141. Arns M, van Dijk H, Luykx JJ, van Wingen G, Olbrich S. Stratified psychiatry: tomorrow's precision psychiatry? Eur Neuropsychopharmacol. 2022;1(55):14–9. https://doi.org/10.1016/j.euroneuro.2021.10.863.
142. Hsin H, Fromer M, Peterson B, Walter C, Fleck M, Campbell A, Varghese P, Califf R. Transforming psychiatry into data-driven medicine with digital measurement tools. Npj Digit Med. 2018;1(1):1–4. https://doi.org/10.1038/s41746-018-0046-0. Nature Publishing Group.
143. Kessler RC. The potential of predictive analytics to provide clinical decision support in depression treatment planning. Curr Opin Psychiatry. 2018;31(1):32–9. Wolters Kluwer.
144. Senn S. Mastering variation: variance components and personalised medicine. Stat Med. 2016;35(7):966–77. Wiley Online Library.
145. Plöderl M, Hengartner MP. What are the chances for personalised treatment with antidepressants? Detection of patient-by-treatment interaction with a variance ratio meta-analysis. BMJ Open. 2019;9(12):e034816. British Medical Journal Publishing Group. MID:31874900.
146. Kaiser T, Volkmann C, Volkmann A, Karyotaki E, Cuijpers P, Brakemeier E-L. Heterogeneity of treatment effects in trials on psychotherapy of depression. Clin Psychol Sci Pract. 2022;29(3):294. Educational Publishing Foundation.
147. Unger EF. Subgroup analyses and pre-specification. Clin Trials. 2023:20(4):338–40. SAGE Publications Sage UK: London, England.
148. Kent DM, Paulus JK, Van Klaveren D, D'Agostino R, Goodman S, Hayward R, Ioannidis JP, Patrick-Lake B, Morton S, Pencina M. The predictive approaches to treatment effect heterogeneity (PATH) statement. Ann Intern Med. 2020;172(1):35–45. American College of Physicians.
149. Kent DM, Van Klaveren D, Paulus JK, D'Agostino R, Goodman S, Hayward R, Ioannidis JP, Patrick-Lake B, Morton S, Pencina M. The predictive approaches to treatment effect heterogeneity (PATH) statement: explanation and elaboration. Ann Intern Med. 2020;172(1):W1–25. American College of Physicians.
150. Cuijpers P, Ciharova M, Quero S, Miguel C, Driessen E, Harrer M, Purgato M, Ebert D, Karyotaki E. The contribution of "individual participant data" meta-analyses of psychotherapies for depression to the development of personalized treatments: a systematic review. J Pers Med. 2022;12(1):93. https://doi.org/10.3390/jpm12010093. Multidisciplinary Digital Publishing Institute.
151. Kessler RC, Bossarte RM, Luedtke A, Zaslavsky AM, Zubizarreta JR. Machine learning methods for developing precision treatment rules with observational data. Behav Res Ther. 2019;120:103412. Elsevier.
152. Cuijpers P, Reijnders M, Huibers MJH. The role of common factors in psychotherapy outcomes. Annu Rev Clin Psychol. 2019;7(15):207–31. PMID:30550721.
153. Degtiar I, Rose S. A review of generalizability and transportability. Annu Rev Stat Its Appl. 2023;10:501–24. Annual Reviews.
154. Franklin JM, Schneeweiss S. When and how can real world data analyses substitute for randomized controlled trials? Clin Pharmacol Ther. 2017;102(6):924–33. PMID:28836267.
155. Carey TA, Stiles WB. Some problems with randomized controlled trials and some viable alternatives. Clin Psychol Psychother. 2016;23(1):87–95. Wiley Online Library.
156. Van Poucke S, Thomeer M, Heath J, Vukicevic M. Are randomized controlled trials the (g) old standard? From clinical intelligence to prescriptive analytics. J Med Internet Res. 2016;18(7):e185. JMIR Publications Toronto, Canada.
157. Lilienfeld SO, McKay D, Hollon SD. Why randomised controlled trials of psychological treatments are still essential. Lancet Psychiatry. 2018;5(7):536–8. Elsevier.
158. Collins R, Bowman L, Landray M, Peto R. The magic of randomization versus the myth of real-world evidence. N Engl J Med. 2020;382(7):674–8.
159. Kennedy-Martin T, Curtis S, Faries D, Robinson S, Johnston J. A literature review on the representativeness of randomized controlled trial samples and implications for the external validity of trial results. Trials. 2015;16(1):495. https://doi.org/10.1186/s13063-015-1023-4.
160. Stirman SW, DeRubeis RJ, Crits-Christoph P, Brody PE. Are samples in randomized controlled trials of psychotherapy representative of community outpatients? A new methodology and initial findings. J Consult Clin Psychol. 2003;71(6):963–72. https://doi.org/10.1037/0022-006X.71.6.963.
161. Zimmerman M, Mattia JI, Posternak MA. Are subjects in pharmacological treatment trials of depression representative of patients in routine clinical practice? Am J Psychiatry. 2002;159(3):469–73. https://doi.org/10.1176/appi.ajp.159.3.469. American Psychiatric Publishing.
162. Lorenzo-Luaces L, Zimmerman M, Cuijpers P. Are studies of psychotherapies for depression more or less generalizable than studies of antidepressants? J Affect Disord. 2018;1(234):8–13. https://doi.org/10.1016/j.jad.2018.02.066.
163. Polo AJ, Makol BA, Castro AS, Colón-Quintana N, Wagstaff AE, Guo S. Diversity in randomized clinical trials of depression: a 36-year review. Clin Psychol Rev. 2019;1(67):22–35. https://doi.org/10.1016/j.cpr.2018.09.004.
164. Dahabreh IJ, Robertson SE, Steingrimsson JA, Stuart EA, Hernán MA. Extending inferences from a randomized trial to a new target population. Stat Med. 2020;39(14):1999–2014. https://doi.org/10.1002/sim.8426.
165. Dahabreh IJ, Robertson SE, Tchetgen EJ, Stuart EA, Hernán MA. Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals. Biometrics. 2019;75(2):685–94. https://doi.org/10.1111/biom.13009.
166. Jackson D, Rhodes K, Ouwens M. Alternative weighting schemes when performing matching-adjusted indirect comparisons. Res Synth Methods. 2021;12(3):333–46. https://doi.org/10.1002/jrsm.1466.
167. Mueller S, Pearl J. Personalized decision making – a conceptual introduction. J Causal Inference. 2023;11(1). https://doi.org/10.1515/jci-2022-0050. De Gruyter.
168. Bareinboim E, Pearl J. Causal inference and the data-fusion problem. Proc Natl Acad Sci. 2016;113(27):7345–52. https://doi.org/10.1073/pnas.1510507113. Proceedings of the National Academy of Sciences.

Publisher's Note