Tabatabaei Ghomi, Hamed and Stegenga, Jacob. 2023. "Simulation of Trial Data to Test Speculative Hypotheses about Research Methods." In Advances in Experimental Philosophy of Medicine, Eds. Hens and De Block. London:Bloomsbury

6

Simulation of Trial Data to Test Speculative Hypotheses about Research Methods

Hamed Tabatabaei Ghomi and Jacob Stegenga

Introduction

Commentators claim that various features of clinical research – such as conventional choices in analytic measures, placebo run-in periods, and publication bias – contribute to overestimating the effectiveness of medical interventions. These suggestions are plausible, yet such speculative empirical claims could and should be tested. One way to test such hypotheses would be to have access to all patient-level data for particular interventions from all relevant trials, and then analyse the influence of these features of trials on estimates of intervention effectiveness. However, save some limited examples, access to such data is practically impossible; although some steps have been made toward improving data access, nobody has access to proprietary and confidential patient-level data from all trials in a domain. Moreover, without knowing the real effectiveness of interventions, there is no way to estimate the effects of various features of trials on estimates of effectiveness. In real trials, scientists do not know the real effectiveness of the interventions they are studying – that is, of course, the entire point of performing the trials. And finally, real data does not allow considering counterfactual scenarios that can be valuable for investigating these speculative hypotheses. Therefore, real trial data is limited in its utility for evaluating various hypotheses about medical research.

To get around these problems, we simulate patient-level data from trials, based on higher-order characteristics of real trial data, such as published means and standard deviations of measured outcomes. This approach solves the problem of data access. Moreover, our approach has a major advantage over real trial data: we can specify the true effectiveness of an intervention, and then vary features of the clinical research to quantify the influence of those features on estimates of effectiveness. For example, we can set the real effectiveness of a drug to zero, set the degree of publication bias to whatever we want, then estimate the effectiveness of the drug based on the published data and quantify the impact of publication bias on the estimate of effectiveness. And conversely, given an observed measure of effectiveness under certain biases we can estimate the real effectiveness. In the present paper we focus on the specific example of

research on antidepressants, though our research programme is extendable to other domains.

Experimental philosophy has allowed philosophers to test intuitions pertaining to a range of philosophical questions. The approach has been experimental in the usual sense of the term. Yet some philosophers have used simulations to test intuitions about scientific research, such as the optimal degree of communication between scientists (Zollman 2007), or the optimal division of labour in a scientific community (Alexander, Himmelreich and Thompson 2015; Thoma 2015; Weisberg and Muldoon 2009), or the conditions under which science can self-correct (Bruner and Holman 2019; Romero 2016). In our approach we simulate scientific data to evaluate meta-scientific intuitions and offer insights into the nature of scientific and medical research.

Some recent work in philosophy of medicine has shown that much medical research is sensitive to arbitrary methodological choices. For example, Stegenga (2016) argues that randomized trials involve numerous methodological choices that render trials more sensitive to detecting putative benefits of medical interventions yet less sensitive to detecting harms of those interventions. We extend this concern about the influence of arbitrary methodological choices on estimates of effectiveness. Our first target is the effect of arbitrary conventional choices in the definition of a particular outcome measure, the responder odds ratio (OR), which is defined as the ratio of the odds of being a 'responder' in the drug group of a trial (O_d) divided by the odds of being a responder in the placebo group of the trial (O_p). A responder is a trial subject (in either the drug or placebo group) whose symptom severity improves by more than a certain threshold c. In trials and meta-analyses of antidepressants, c is conventionally set to 50 percent. This commonly used value for c, however, is arbitrary. Some methodologists have criticized responder analyses in general because they involve a loss of information (Altman and Royston 2006; Collister et al. 2021). Our results extend this criticism by demonstrating that the choice of $c = 0.5$ can maximize estimates of drug effectiveness, and other values of c would give lower values for the responder OR – in short, the choice of c is an arbitrary methodological decision which has dramatic impact on estimates of effectiveness. This is consistent with previous work which suggested the phenomenon of a 'response rate illusion', in which estimates of effectiveness are exaggerated when continuous data are transformed into responder odds ratios (Hadzi-Pavlovic 2009; Kirsch and Moncrieff 2007). Our work demonstrates this phenomenon. Kirsch and Moncrieff (2007) provide an analytic argument which is further expanded by (Hadzi-Pavlovic 2009). Their approach provides a limited view of this effect. Our extensive simulation approach, however, shows the patterns of the effect of c on estimates of effectiveness under various conditions, and illustrates the details of the maximizing effect of $c \approx 0.5$.

Our results here are relevant to a recent debate among philosophers of medicine about the merits of different families of outcome measures. Some have argued that so-called 'absolute' outcome measures are superior to so-called 'relative' outcome measures (see, e.g., Stegenga (2015), Stegenga and Kenna (2017) and Sprenger and Stegenga (2017). Conversely, Hoefer and Krauss (2021) claim that relative measures are also informative and ought to be reported alongside absolute measures. This debate has been based on outcome measures for binary properties (e.g. whether a patient is dead

or alive at the end of a trial). Our simulations involve more fine-grained measures on a presumably continuous property (severity of depression). A responder odds ratio is a relative measure, while a simple and informative absolute measure of continuous data is just the mean difference between groups. Our demonstration that the responder odds ratio is dramatically sensitive to the arbitrary choice of *c* adds to the extant arguments for the superiority of absolute over relative measures.

Our second target is the impact of publication bias on estimates of effectiveness. Publication bias is a ubiquitous phenomenon in medical research (Wieseler et al. 2013). It is plausible that publication bias exaggerates estimates of effectiveness, since publication bias is usually directional: evidence which suggests that interventions are effective is published more often than evidence which suggests that interventions are ineffective (for demonstration see Turner et al. (2008), and for discussion of the phenomenon in philosophical contexts see Biddle (2007), Jukola (2017) and Stegenga (2018)). We offer some confirmation of this view about publication bias. More subtly, however, we show that publication bias has less impact on the apparent effectiveness of drugs that are truly effective, while it has more impact on the apparent effectiveness of drugs that have lower real effectiveness. This confirms similar findings in Friese and Frankenbach (2020) and Nuijten et al. (2015), while illustrating the details of the pattern and the magnitude of the effect of publication bias in the particular case of estimates of antidepressants' effectiveness.

Our third target is the impact of placebo run-in periods on estimates of effectiveness. A run-in period of a trial involves giving subjects a placebo prior to the formal data-gathering phase of a trial, eliminating subjects who appear very responsive to placebo, and then distributing the remaining subjects between the placebo and the drug groups. Such run-in periods are performed in many RCTs of antidepressants (Posternak et al. 2002). The assumption behind this method is that the subjects that are placebo-responsive during the run-in period would be the same subjects that would be placebo responsive in the main part of the study. In other words, placebo-responsiveness is assumed to be a constant character of subjects. We call this the *assumption of constancy*. If this assumption holds, then run-in periods should eliminate highly placebo-responsive subjects and thus reduce the placebo effect observed in the main study. Reduced placebo effect allows the detection of smaller real drug effects.

Despite its wide use, analysis of published RCTs shows that run-in periods do not seem to reduce the placebo effect (Greenberg, Fisher and Riter 1995; Lee et al. 2004; Posternak et al. 2002). We investigate the reasons behind this empirical observation and demonstrate that run-in periods require three conditions to reduce the placebo effect: high number of placebo-responsive subjects, constancy, and high real drug effectiveness. We conclude that the failure of run-in periods in practice indicates that one or more of these conditions fail to hold.

Our ambition in this paper is twofold. First, we aim to offer insights into the hypotheses about medical research noted above, in a manner that goes beyond mere intuition or limited empirical evidence. Second, and more generally, we aim to expand the relatively young literature that shows computer simulations can be fruitfully employed to address a wide range of second-order questions about scientific and medical research.

Methods

Here we present an overview of our methods. The appendix presents more technical details.

Simulating patient data

Symptom severity in antidepressant trials is measured with a scale called the Hamilton Depression Rating Scale (HAMD), a 50-point scale in which higher scores are said to represent greater severity of depression. We generated random HAMD values for the drug and the placebo groups, before and after intervention, by a Gaussian random generator. We applied a lower value of HAMD = 19 as an inclusion criterion (Santen, Horrigan, et al. 2009), so that all patients had HAMD scores above this value before treatment. Randomly generated values were limited between 0 and 50. The HAMD score of an individual subject after treatment could be lower (better), higher (worse), or equal to the subject's HAMD before treatment.

We set the distribution parameters for the random generator based on actual patient data. We calculated the parameters based on the most recent and probably the most comprehensive meta-analysis of antidepressant RCTs so far published (Cipriani et al. 2018). Table 6.1 lists all the parameters used for simulations.

We repeated the simulations for a wide range of stipulated drug effectiveness, indicated by after-treatment mean HAMD score in the drug group (m_d^a). We also repeated the simulations for various sizes of the placebo and drug groups (n). For each combination of values for m_d^a and n, we repeated the simulation 5000 times (thereby generating simulated data for 5000 trials per combination, resulting in 315,000 trials total).

One measure of effectiveness for a trial was the responder *OR*: a drug was declared effective in a trial with some level of statistical significance if the lower confidence interval of *OR* at that level of significance was greater than one (in this study, we used 95 per cent confidence intervals). The probability of finding a drug to be effective (P_E)

Table 6.1 Parameters used in simulations. © Jacob Stegenga and Hamed Tabatabaei Ghomi.

Parameter	Used for simulation
mean HAMD in placebo group before treatment, $\boldsymbol{m_p^b}$	24
mean HAMD in drug group before treatment, $\boldsymbol{m_d^b}$	24
mean HAMD in placebo group after treatment, $\boldsymbol{m_p^a}$	14
mean HAMD in drug group after treatment, $\boldsymbol{m_d^a}$	9,10,11,12,13,14,15
SD of HAMD in placebo group before treatment, $\boldsymbol{s_p^b}$	3.5
SD HAMD in drug group before treatment, $\boldsymbol{s_d^b}$	3.5
SD HAMD in placebo group after treatment, $\boldsymbol{s_p^a}$	7.9
SD HAMD in drug group after treatment, $\boldsymbol{s_d^a}$	7.4
Size of the placebo and drug groups, $\boldsymbol{n}$	50, 75, 100, 125, 150, 175, 200, 225, 250

under a certain combination of parameters was calculated by dividing the number of trials demonstrating effectiveness with that particular parameter set by the total number of trials with that parameter set. Another measure of effectiveness was the average difference in mean HAMD reduction between placebo and drug groups under a certain combination of parameters, which was calculated by averaging over all the trials with that combination of parameters.

Modelling responders: Odds ratio

Responder *OR* is defined as the ratio of the odds of being a 'responder' in the drug group (O_d) divided by the odds of being a responder in the placebo group (O_p). Odds of being a responder in each group is calculated by dividing the number of responders by the number of non-responders in that group.

$$OR = \frac{O_d}{O_p}$$

$$O_g = \frac{|R_g|}{|g| - |R_g|}, g \in \{d, p\}$$

where R_g is the set of all responders in group g (either the drug d, or the placebo p groups), and $|R_g|$ and $|g|$ are the sizes of R_g and g respectively.

A subject i in any of the drug or placebo groups ($g = d$ (for drug group) or $g = p$ (for placebo group)) is declared a *responder* ($i \in R_g$) if their HAMD score after treatment ($HAMD_i^a$) was less than or equal to a certain fraction c of their HAMD before treatment ($HAMD_i^b$):

$$R_g = \{\forall i \in g | \frac{HAMD_i^a}{HAMD_i^b} \leq c\}$$

To investigate the effect of c on the probability of finding a drug to be effective, we re-analysed data by varying c from 1% to 99%, with 1% increments

$$c \in \{1\%, 2\%, \ldots, 99\%\}$$

Modelling publication bias

To model the effect of publication bias, we applied a probabilistic filter. A trial that found a drug to be effective (based on statistically significant (t-test, $p \leq 0.05$) difference between the average HAMD score of the placebo and drug groups) had a 100% chance of being published, but a trial that found a drug to be ineffective had a β chance of being published (1 = no publication bias, 0.1 = severe publication bias).

$$\beta \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$$

Only published trials were used in the subsequent calculations of the average difference in mean HAMD reduction between placebo and drug groups.

Modelling run-in period

One way to model run-in periods would be to have a two-phase simulation, the first for the run-in period and the second for the main trial, and follow the patients' placebo responses through the two phases. This method, however, faces some methodological problems. First, it would require assigning a value of placebo responsiveness for each individual patient, yet there are no empirical grounds for such values or its distribution in a population (to determine this would require patient-level data from the placebo groups of trials). Second, it is hard to model the way the property of placebo responsiveness changes over time and between the run-in period and the main study. To bypass these methodological problems, we mimicked the effect of a run-in period on a trial's outcome in a one-step simulation as described below.

In short, we eliminate a number of patients from the main trial as would have been eliminated by a run-in period, re-analyse the data, and compare the outcome with the results of the full data set. The number and the attributes of the eliminated patients are set in a way such that the elimination mimics the elimination that would have resulted from a real a run-in period. The elimination is controlled by two variables, ρ and χ, that are associated with two main assumptions behind run-in periods. The first assumption is that would-be placebo-responsive patients in fact exhibit enough placebo response during the run-in period such that they are eliminated. The assumption is controlled by χ, the fraction of excluded patients. Usually, when a run-in period is used in trials of antidepressants, a reduction of ≥20% in a subject's HAMD score during a run-in period results in excluding the subject from the main study (Landin et al., 2000; Lee et al., 2004). We set the fraction of the excluded patients (χ) based on reports in the literature: Quitkin et al. (1998) reported 9%, Heiligenstein et al. (1993) reported 5.5%, and Lydiard et al. (1997) reported 17%. We thus set:

$$\chi \in \{0.05, 0.1, 0.17\}$$

The second assumption is the assumption of constancy, which holds if the patients who show high placebo response during the run-in period are the same patients that would have shown high placebo response had they entered the main trial. This assumption can be controlled by ρ, the fraction of eliminated patients that necessarily show ≥20% reduction in HAMD score. Varying ρ allows investigating the effect of the assumption of constancy on the effectiveness of run-in periods. The higher is ρ, the stronger the assumption of constancy holds. If the assumption holds strongly, then a higher fraction of excluded patients are necessarily among those with a HAMD reduction ≥20% in the main phase of trials, and vice versa.

We aimed to exclude a total of $\chi \times |g|$ (where $|g|$ is the size of the placebo or drug group) patients from each group of a trial. The *constancy* values (ρ) shows the proportion of excluded subjects that necessarily had ≥20% reduction in HAMD. We randomly picked $\rho \times \chi \times |g|$ subjects from those with ≥20% reduction in HAMD, and excluded them from the study. Then we picked the rest of the subjects for exclusion

$((1 - \rho) \times \chi \times |g|)$ without any constraints on their HAMD reduction. We tested three strengths of the assumption of constancy:

$$\rho \in \{0, 0.5, 1\}$$

Software

We ran the simulations and the following analysis by developing a Python 3.8 code. The computations were distributed over many CPUs to be able to complete the large number of trial simulations in a reasonable time.

Results and discussion

Probability of 'effective' conclusion by varying values of *c*

Figure 6.1 shows the probability of concluding that a drug is effective under various combinations of n, m_d^a, and c. Naturally, as the effectiveness of the drug decreases from $m_d^a = 9$ to $m_d^a = 15$, the probability of finding a drug to be effective decreases. Also, the probability of finding a drug to be effective increases by increasing the number of participants because the power of RCTs increases by their size. But there are patterns in Figure 6.1 that depend on c and signify the importance of the choice of c that is conventionally set to 50% with no obvious reason.

Figure 6.1 shows a dome-shaped higher probability of deeming drugs to be effective, centring around $c = 50\%$. This pattern shows that the conventional practice of setting

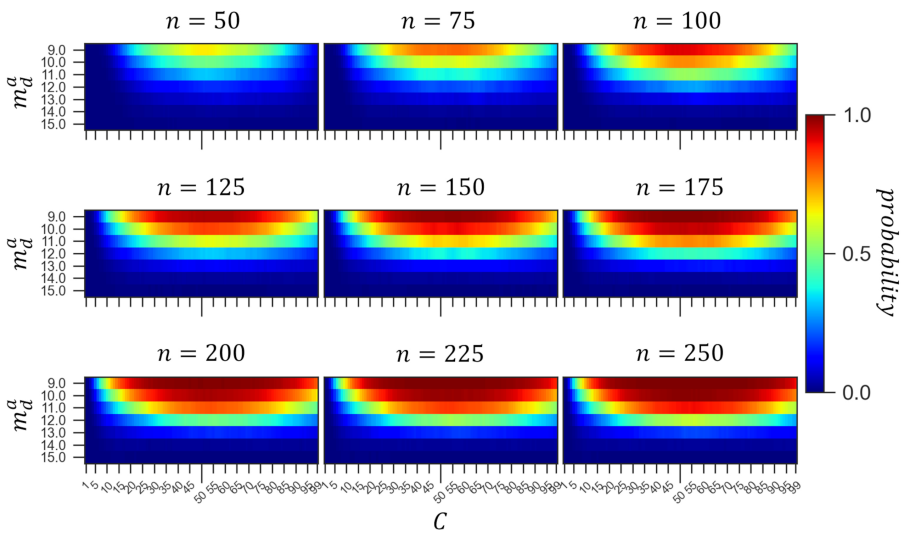

Figure 6.1 The probability of concluding that a drug is effective under various combinations of sample size (n), real drug effectiveness (m_d^a), and conventional threshold for definition of responder (c). © Jacob Stegenga and Hamed Tabatabaei Ghomi. This figure is reproduced in colour in Tabatabaei Ghomi and Stegenga (2022).

$c = 50\%$ appears to almost maximize the probability that a drug will be deemed effective according to the responder odds ratio. In underpowered studies such as $n = 50$, it increases the chance of identifying a truly effective drug (e.g. $m_d^a = 9$). In larger studies (e.g. $n = 200, 225$), however, it increases the chance of declaring a weak drug ($m_d^a = 12$) to be effective. In a study that is large enough, this will result in a statistically significant difference between placebo and drug where the real difference between the two is negligible (e.g. $m_d^a = 13$). This statistical effectiveness can be clinically misleading as some have argued that a HAMD reduction of 3 points is undetectable in the clinic (Moncrieff and Kirsch 2015).

Apparent effectiveness given publication bias

Figure 6.2 shows the difference in the mean HAMD reduction between the placebo and the drug group according to the published data with different levels of publication bias (varying β, $\beta = 0.1$ being the most severe bias, and $\beta = 1$ being equivalent to no bias) for drugs of various strengths (m_d^a). As expected, the observed difference in mean HAMD reduction increases by stronger publication bias. With a strong publication bias (e.g. $\beta = 0.1$), the difference in mean HAMD reduction between the drug and the placebo groups may increase by 1 to 2 points, especially for smaller RCTs. This increase is enough to suggest apparent effectiveness even for completely ineffective drugs, and even for drugs that are less effective than placebo. This is particularly worrying because our results show that the effect of publication bias is larger for weaker drugs. This is in line with Friese and Frankenbach (2020) and Nuijten et al. (2015), who similarly show that publication bias is most impactful when the true effect is small. As we simulated patient-level data with realistic parameters, our results further show the magnitude of this effect in the case of antidepressants, under various conditions.

The results also show that the size of RCTs is another factor affecting the impact of publication bias. Comparing the left side of Figure 6.2 with the right side of Figure 6.2

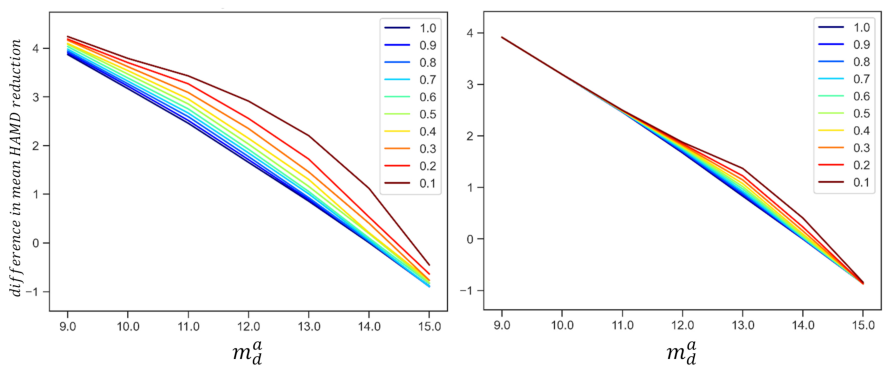

Figure 6.2 The difference in mean HAMD reduction between the drug and the placebo groups, for various intensities of publication bias, at sample size n=50 (left) and n=250 (right). © Jacob Stegenga and Hamed Tabatabaei Ghomi.

suggests that the impact of publication bias on estimates of effectiveness is mitigated with larger sample sizes.

Although a publication bias of $\beta = 0.1$ might seem to be extreme, we believe that it is representative of actual publication practices in some domains of medical research. Turner et al. (2008) show that while almost all RCTs with positive results get published, RCTs with negative outcomes either do not get published, or get published in a way that coveys a positive outcome. For instance, according to FDA reports about 50% of RCTs of antidepressants have positive outcomes, yet 94% of the published RCTs report positive results.[1] In our method of modelling publication bias, going from a 50-50 success-to-failure ratio to a 6-94 ratio corresponds to $\beta \simeq 0.06$ (lower than even our most severe publication bias parameter of $\beta = 0.1$, calculations included in the appendix). Another study reported a publication bias of 23% ($\beta \simeq 0.2$) (Wieseler et al. 2013).

These realistic estimates of β allow us to compute a correction factor for analysing the results of real published RCTs. For example, suppose that we observe a reduction of two HAMD points in the drug group compared to the placebo group. As we see in Figure 6.2, in the absence of any publication bias, that value corresponds to an average HAMD score of 12 after treatment with drugs. However, assuming we have publication bias of the amounts $\beta = 0.1$ or $\beta = 0.2$ we can correct the estimate of effectiveness in the following way: draw a horizontal line running across the graph from 2 on the y-axis and let that line intersect with the curve representing the publication bias parameter 0.1 (the rightmost curve), then draw a vertical line down to intersect the x-axis; the value of that x-axis intersection is the inferred real effectiveness of the drug. The conclusion would be that the estimated real HAMD score after treatment is approximately 13.5, and since the observed HAMD score after placebo was 14, this would entail that the estimated difference in HAMD reduction between drug and placebo groups was 0.5. Thus, this method would amount to correcting the reduction in HAMD scores observed in trials from 2 to 0.5.

Run-in period

Figure 6.3 shows the effect of run-in periods by depicting the difference in the mean HAMD reduction between the drug and the placebo groups, under various *constancy* values (ρ), and exclusion fractions (χ). The small change in the difference in the mean HAMD reduction is in line with the previous metanalyses of run-in periods that show the method is ineffective at increasing the observed difference between drug and placebo groups (Greenberg, Fisher and Riter 1995; Lee et al. 2004; Posternak et al. 2002).

Figure 6.3 suggests three potential reasons for the ineffectiveness of run-in periods. First, only a small fraction of subjects shows enough placebo effect during the short run-in periods to be excluded from the main phase of the RCT. As Figure 6.3 shows, when excluded subjects constitute only 5% or 10% of the subjects, the change in the difference in the mean HAMD reduction is very small for all constancy values and drug strength. Second, for low constancy values, run-in periods show no conspicuous effect, and this shows that run-in periods change the difference in mean HAMD

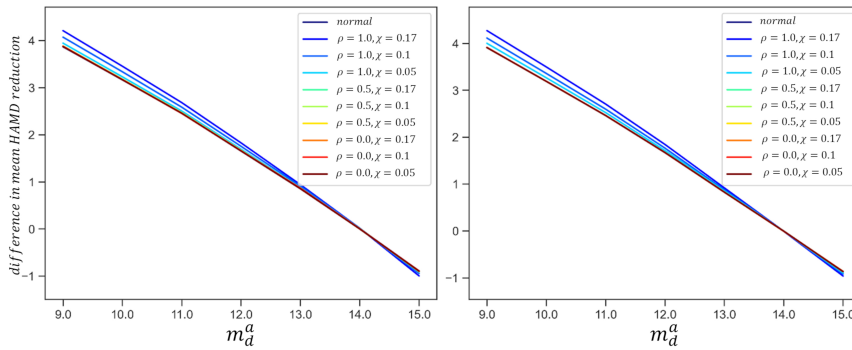

Figure 6.3 The difference in mean HAMD reduction between the drug and the placebo groups, under various constancy values (ρ) and exclusion proportions (χ), for n=50 (left), and n=250 (right). © Jacob Stegenga and Hamed Tabatabaei Ghomi.

reduction only if the assumption of constancy strongly holds ($\rho = 1$). Third, even when 17% of the subjects get excluded, and the assumption of constancy strongly holds, run-in periods increase the difference in mean HAMD reduction only for drugs with strong real effects. Less effective drugs do not appear much more effective with run-in periods.

Discussion

Our results show that the conventional choice of $c = 50\%$ in the responder odds ratio almost maximizes estimates of effectiveness. This can be beneficial in underpowered small RCTs by increasing the chance of detecting effective drugs. On the other hand, in studies with large number of participants, this bias increases the chance of drugs with negligible real effects to appear effective. And in general, the strong dependence of the chance of finding a drug to be effective on the choice of c indicates the problem with an arbitrary choice for c, and the need for substantiating the use of any value for this variable.

These results about responder analyses address two other recent debates in philosophy of medicine. First, some philosophers have argued that meta-analysis is not as reliable as it is often made out to be (Stegenga 2011), while others defend meta-analysis (Holman 2019). Responder analyses, though not intrinsically part of the methodology of meta-analysis, are often used in meta-analyses, because a responder analysis allows the pooling of trials that use different measurement scales. In a domain of research in which multiple incommensurable measurement scales are used (such as research on antidepressants, in which there are multiple versions of the HAMD scale and other scales used in various trials), analysts must choose between either excluding many trials from a meta-analysis or including all trials and

pooling their results using a standardized outcome measure such as responder *OR*. Our results here show a problem with the latter choice, and since the former choice amounts to not including relevant evidence, this amounts to a dilemma for many meta-analyses: exclude evidence or use suspect outcome measures. Second, some philosophers have recently criticized so-called relative outcome measures, which include the responder *OR* (Sprenger and Stegenga 2017), while others defend the use and reporting of relative measures (Hoefer and Krauss 2021). We have identified a salient problem with the responder *OR*, lending some support to those critics of relative outcome measures.

Our results show that publication bias has less effect on the apparent efficacy of drugs with strong real effects. This could be taken as pushback against the appeal to publication bias as one of several arguments for general scepticism about medical research and medical interventions in, for example, Stegenga (2018). On the other hand, we show that publication bias can spuriously increase the apparent efficacy of interventions with negligible real effects.

It is interesting to compare our results with Romero (2016). Although Romero used a different approach, his simulations included some scenarios which are broadly comparable to ours. Romero modulated the real effect size, the sample size of experiments, and whether negative findings were published, and then assessed the extent to which accumulated data tracked the real effect size (he also varied what he called 'direction bias', which is not relevant to our findings here). An important difference between Romero's method and ours is that each of his parameters could only take one of two values: effect size could be zero or medium, sample size could be 'sufficient' (e.g. large) or not, and publication bias either existed (negative findings not published) or not. In contrast, our corresponding parameters can be tuned to a variety of levels. Nevertheless, comparing some of his findings with our is illuminating.

In Romero's medium effect size simulations with large sample sizes, publication bias had little impact on estimates (compare his S1 with S9), which is consistent with our results displayed in the right side of Figure 6.2 (a medium effect size would be a final HAMD score of about 12, in the middle of the x-axis). In Romero's medium effect size simulations with small sample size, publication bias contributed to a large overestimate of effectiveness (compare his S2 with S10), which is consistent with our results displayed on the left side of Figure 6.2. In Romero's zero effect size simulations with large sample size, publication bias has no impact (compare his S5 with his S13), which is consistent with our results displayed in the right side of Figure 6.2, in the bottom-right region of the graph. Thus, some of Romero's findings are replicated by our simulations.

However, other findings in Romero (2016) are not replicated by our simulations. In Romero's zero effect size simulations with small sample size, publication bias has little impact (compare his S6 with his S14), yet in our simulations publication bias had a large impact in similar conditions, as displayed by our results on the left side of Figure 6.2 in the bottom-right region of the graph. In Romero's medium effect size simulations with small sample size, the estimated effect size was exaggerated by publication bias,

while in the zero effect size simulations it was not exaggerated by publication bias (compare Romero's S10 with S14). This finding contradicts our simulations (and other work cited in the section titled, 'Apparent Effectiveness Gven Publication Bias'), as seen in the left side of Figure 6.2, in which publication bias has a greater exaggerating impact on smaller real effect sizes.

We showed that run-in periods are not very effective in decreasing the placebo effect. We also showed three conditions that are necessary for run-in periods to be effective. The assumption of constancy should strongly hold, a drug should have strong real effect, and a large fraction of subjects should show significant placebo effect during the run-in period. The empirical observation that run-in periods do not increase estimates of drug efficacy, therefore, shows that one, two, or all three of these conditions do not generally hold.

There have been previous simulations of antidepressant RCTs (Chevance et al. 2019; Landin et al. 2000; Santen, Horrigan et al. 2009; Santen, Van Zwet et al. 2009), though the target questions of these studies and methods used to simulate trial data are different from ours. These studies had scientific rather than meta-scientific concerns. Also, they did not investigate the conventional choices in analytic measures that we investigate, and do not model run-in periods. Although they touched on the issue of publication bias, they examined it from a different angle. For example, one article suggested that publication bias increases the apparent effectiveness of antidepressants and this group modelled publication bias indirectly by correcting the effect sizes in their simulations (Chevance et al. 2019). From a methodological perspective, all of these studies simulated the time course of response to antidepressants and therefore needed to train a model on existing patient data. We do not need the exact time course of each subject for our purposes and thus can avoid model fitting and its associated caveats, such as the bias imposed on the results by limited training data.

We used computational simulations to put some speculative hypotheses about clinical research to test. We not only verified some speculations, but also showed the nuances and conditions of the effects of these biases. The results of our study inform the specific case of research on antidepressants, yet they are extendable to RCTs in general. Our methods can serve as examples of possible ways to evaluate speculative hypotheses about research practices.

Appendix

Here we offer further technical details about our methods.

Random data generation

The overall process

We generated random HAMD values for the drug and the placebo groups, before and after intervention, by a Gaussian random generator. By using a Gaussian random

generator we assume that distribution of HAMD values within each group is normal. This is an assumption that can be further investigated in future work. The parameters fed into the random generator were calculated based on real patient data so that the resulting distributions resembled actual patient data.

Calculating distribution parameters

We used one of the most recent and probably the most comprehensive meta-analysis of antidepressant RCTs so far published to calculate the parameter estimates (Cipriani et al. 2018). The supplementary material of Cipriani et al. (ibid.) reports the mean and standard deviations of the drug and placebo groups before and after treatment for several RCTs. The parameters reported are for various drugs and different versions of HAMD. We aggregated these values regardless of the drug. However, we limited our calculations only to values reported for HAMD-17 so that the scores used in the calculations were all on the same scale. HAMD-17 was the most frequently used scale in Cipriani et al. (ibid.), and 537 out of 1199 parameter values were included in the analysis. Out of these 537 values, 137 were for placebo and 400 were for various drugs. After excluding reports with missing data, we ended up with 105 parameter values for placebo, and 337 values for various drugs. For cases where the mean HAMD after treatment was reported as the change from the initial mean HAMD, the standard deviation was excluded from the calculations of standard deviation because the standard deviation in these cases was for the change and not the mean HAMD after treatment. The averages were weighted by the number of participants in each study, although this weighting had negligible impact.

Cipriani et al. (2018) report standard deviation only after treatment. We obtained the standard deviations of the placebo and the drug groups before treatment from (Hieronymus et al., 2016). The other parameters reported in Hieronymus et al. (2016) are close to values we calculated, thereby providing some cross-validation.

The only parameter that is different in our calculations from other references is the size of the placebo and the drug groups. We calculated sizes of 44 (placebo) and 62 (drug), while Cipriani et al. (2009) reports average sizes of about 100 participants per group and Chevance et al. (2019) say that usually the group sizes are between 100–300 participants and occasionally even more. The difference between our calculation and previous reports is not as drastic as it appears because our estimates are from trial reports after dropouts are deducted from the number of participants, while others report the sizes at the beginning of the study counting the dropouts in group sizes. Nevertheless, we repeated our simulation for 9 different sample sizes ranging from 50 to 250 to cover both the parameters obtained in our calculations and those in the previous reports.

On smaller values of m_d^a the unavoidable truncation of the normal distribution on the lower side (we cannot have HAMD scores less than zero) results in averages slightly higher than the assigned m_d^a. As a result, on these values of m_d^a, the difference in the mean HAMD reduction between the drug and placebo groups becomes slightly less than what is expected from the difference of m_d^a and m_p^a.

Responders based on odds ratio

Responder *OR* is defined as the ratio of the odds of being a 'responder' in the drug group (O_d) divided by the odds of being a responder in the placebo group (O_p). Odds of being a responder in each group is calculated by dividing the number of responders by the number of non-responders in that group.

$$OR = \frac{O_d}{O_p}$$

$$O_g = \frac{|R_g|}{|g| - |R_g|}, g \in \{d, p\}$$

where R_g is the set of all responders in group g (either the drug d, or the placebo p groups), and $|R_g|$ and $|g|$ are the sizes of R_g and g respectively. The 95% confidence interval of *OR* is calculated by the following formula:

$$\exp(\ln(OR) \pm 1.96 \sqrt{\frac{1}{|R_p|} + \frac{1}{|R_d|} + \frac{1}{|p| - |R_p|} + \frac{1}{|d| - |R_d|}})$$

To avoid zero division in effectiveness calculations, Haldane-Anscombe correction was applied by adding 0.5 to $|R_p|$, $|p| - |R_p|$, $|R_d|$, and $|d| - |R_d|$.

Average difference in mean HAMD reduction between placebo and drug groups

For each simulated RCT, we calculated the difference in the average reduction of mean HAMD score between the placebo and drug groups as follows:

Average HAMD reduction in the group g (placebo or drug):

$$\overline{HAMD}_g^{Red} = \frac{1}{|g|} \sum_{i \in g} |HAMD_i^a - HAMD_i^b|$$

The difference in the average reduction of HAMD score between the placebo and drug groups of an RCT was calculated as:

$$\Delta \overline{HAMD}_{RCT}^{Red} = \overline{HAMD}_d^{Red} - \overline{HAMD}_p^{Red}$$

We calculated the average difference in the reduction of HAMD score between the placebo and drug groups for a particular combination of parameter values by averaging [eqn_163b] of all RCT repeats with that specific parameter combination (number of repeats = 1000):

$$\Delta \overline{HAMD}^{Red} = \frac{1}{1000} \sum \Delta \overline{HAMD}_{RCT}^{Red}$$

For studying publication bias, only the trials that pass the publication bias filter enter the calculation of average difference in the mean HAMD reduction between the placebo and the drug groups (|*published*|: number of published studies).

$$\Delta\overline{HAMD}^{Red} = \frac{1}{|published|}\sum_{RCT\in published}\Delta\overline{HAMD}^{Red}_{RCT}$$

We test a range of values for β. But to have an estimate of the realistic value for β, we can use the reports that show publication bias converts a 50-50 success-failure ratio to a 94-6 ratio. This means:

$$\begin{cases} \dfrac{Success}{failure} = \dfrac{50}{50} \\ \dfrac{Success}{\beta \times failure} = \dfrac{94}{6} \end{cases} \Rightarrow \beta \simeq 0.06$$

Notes

1 A limitation of our discussion here is that the move from the FDA dataset to the published dataset may not be merely a matter of publication bias but may also be affected by p-hacking, multiple publication of positive studies, etc.

References

Alexander, J. M., J. Himmelreich and C. Thompson (2015), 'Epistemic Landscapes, Optimal Search, and the Division of Cognitive Labor', *Philosophy of Science*, 82 (3): 424–53.

Altman, D. G. and P. Royston (2006), 'The Cost of Dichotomising Continuous Variables', *British Medical Journal*, 332 (7549): 1080.

Biddle, J. (2007), 'Lessons from the Vioxx Debacle: What the Privatization of Science Can Teach Us about Social Epistemology', *Social Epistemology*, 21 (1): 21–39.

Bruner, J. P. and B. Holman (2019), 'Self-Correction in Science: Meta-Analysis, Bias and Social Structure', *Studies in History and Philosophy of Science Part A*, 78: 93–7.

Chevance, A., F. Naudet, R. Gaillard, P. Ravau and R. Porcher (2019), 'Power behind the Throne: A Clinical Trial Simulation Study Evaluating the Impact of Controllable Design Factors on the Power of Antidepressant Trials', *International Journal of Methods in Psychiatric Research*, 28 (3): e1779.

Cipriani, A., T. A. Furukawa, G. Salanti, J. R. Geddes, J. P. T. Higgins, R, Churchill, N. Watanabe, A. Nakagawa, I. M. Omori, H. McGuire, M. Tansella and C. Barbui (2009), 'Comparative Efficacy and Acceptability of 12 New-Generation Antidepressants: A Multiple-Treatments Meta-Analysis', *Lancet*, 373 (9665): 746–58.

Cipriani, A., T. A. Furukawa, G. Salanti, A. Chaimani, L. Z. Atkinson, Y. Ogawa, S. Leucht, H. G. Ruhe, E. H. Turner, J. P. T. Higgins, M. Egger, N. Takeshima, Y. Hayasaka, H. Imai, K. Shinohara, A. Tajika, J. P. Ioannidis and J. R. Geddes (2018), 'Comparative Efficacy and Acceptability of 21 Antidepressant Drugs for the Acute Treatment of Adults with

Major Depressive Disorder: A Systematic Review and Network Meta-Analysis', *Lancet*, 391: 1357–66.
Collister, D., S. Bangdiwala, M. Walsh, R. Mian, S. F. Lee, T. A. Furukawa and G. Guyatt (2021), 'Patient Reported Outcome Measures in Clinical Trials Should be Initially Analyzed as Continuous Outcomes for Statistical Significance and Responder Analyses Should be Reserved as Secondary Analyses', *Journal of Clinical Epidemiology*, 134: 95–102.
Friese, M. and J. Frankenbach (2020), 'p-Hacking and Publication Bias Interact to Distort Meta-Analytic Effect Size Estimates', *Psychological Methods*, 25 (4): 456.
Greenberg, R. P., S. Fisher and J. A. Riter (1995), 'Placebo Washout is Not aMeaningful Part of Antidepressant Drug Trials', *Perceptual and Motor Skills*, 81 (2): 688–90.
Hadzi-Pavlovic, D. (2009), 'Exploring Kirsch and Moncrieff's "Response Rate Illusion"', *Acta Neuropsychiatrica*, 21 (1): 38–40.
Heiligenstein, J. H., G. D. Tollefson and D. E. Faries (1993), 'A Double-Blind Trial of Fluoxetine, 20 mg, and Placebo in Out-Patients with DSM-III—R Major Depression and Melancholia', *International Clinical Psychopharmacology*, 8 (4): 247–51.
Hieronymus, F., J. F. Emilsson, S. Nilsson and E. Eriksson (2016), 'Consistent Superiority of Selective Serotonin Reuptake Inhibitors over Placebo in Reducing Depressed Mood in Patients with Major Depression', *Molecular Psychiatry*, 21 (4): 523–30.
Hoefer, C. A. Krauss (2021), 'Measures of Effectiveness in Medical Research: Reporting Both Aabsolute and Relative Measures', *Studies in History and Philosophy of Science Part A*, 88: 280–3.
Holman, B. (2019), 'In Defense of Meta-Analysis', *Synthese*, 196 (8): 3189–211.
Jukola, S. (2017), 'A Social Epistemological Inquiry into Biases in Journal Peer Review', *Perspectives on Science*, 25 (1): 124–48.
Kirsch, I. and J. Moncrieff (2007), 'Clinical Trials and the Response Rate Illusion', *Contemporary Clinical Trials*, 28 (4): 348–51.
Landin, R., D. J. DeBrota, T. A. DeVries, W. Z. Potter and M. A. Demitrack (2000), 'The Impact of Restrictive Entry criterion during the Placebo Lead-in Period', *Biometrics*, 56 (1): 271–8.
Lee, S., J. R. Walker, L. Jakul and K. Sexton (2004), 'Does Elimination of Placebo Responders in a Placebo Run-in Increase the Treatment Effect in Randomized Clinical Trials? A Meta-Analytic Evaluation', *Depression and Anxiety*, 19 (1): 10–19.
Lydiard, R. B., S. M. Stahl, M. Hertzman and W. M. Harrison (1997), 'A Double-Blind, Placebo-Controlled Study Comparing the Effects of Sertraline versus Amitriptyline in the Treatment of Major Depression', *Journal of Clinical Psychiatry*, 58 (11): 484–91.
Moncrieff, J. and I. Kirsch (2015), 'Empirically Derived Criteria Dast doubt on the Clinical Significance of Antidepressant-Placebo Differences', *Contemporary Clinical Trials*, 43: 60–2.
Nuijten, M. B., M. A. L. M. Van Assen, C. L. S. Veldkamp and J. M. Wicherts (2015), 'The Replication Paradox: Combining Studies Can Decrease Accuracy of Effect Size Estimates', *Review of General Psychology*, 19 (2): 172–82.
Posternak, M. A., M. Zimmerman, G. I. Keitner and I. W. Miller (2002), 'A Reevaluation of the Exclusion Criteria Ued in Antidepressant Efficacy Trials', *American Journal of Psychiatry*, 159 (2): 191–200.
Quitkin, F. M., P. J. McGrath, J. W. Stewart, K. Ocepek-Welikson, B. P. Taylor, E. Nunes, D. Delivannides, V. Agosti, S. J. Donovan, D. Ross, E. Peetkova and D. Klein (1998), 'Placebo Run-in Period in Sudies of Depressive Disorders', *British Journal of Psychiatry*, 173 (3): 242–8.

Romero, F. (2016), 'Can the Behavioral Sciences Self-Correct? A Social Epistemic Study', *Studies in History and Philosophy of Science Part A*, 60: 55–69.

Santen, G., J. Horrigan, M. Danhof and O. Della Pasqua (2009), 'From Trial and Error to Trial Simulation. Part 2: An Appraisal of Current Beliefs in the Design and Analysis of Clinical Trials for Antidepressant Drugs', *Clinical Pharmacology & Therapeutics*, 86 (3): 255–62.

Santen, G., E. Van Zwet, M. Danhof and O. Della Pasqua (2009), 'From Trial and Error to Trial Simulation, Part 1: The Importance of Model-Based Drug Development for Antidepressant Drugs', *Clinical Pharmacology & Therapeutics*, 86 (3): 248–54.

Sprenger, J. and J. Stegenga (2017), 'Three Arguments for Absolute Outcome Measures', *Philosophy of Science*, 84 (5): 840–52.

Stegenga, J. (2011), 'Is Meta-Analysis the Platinum Standard of Evidence?', *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 42 (4): 497–507.

Stegenga, J. (2015), 'Measuring Effectiveness', *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 54: 62–71.

Stegenga, J. (2016), 'Hollow Hunt for Harms', *Perspectives on Science*, 24 (5): 481–504.

Stegenga, J. (2018), *Medical Nihilism*. Oxford: Oxford University Press.

Stegenga, J. and A. Kenna (2017), 'Absolute Measures of Effectiveness', in McClimans (ed.), *Measurement in Medicine: Philosophical Essays on Assessment and Evaluation*, 35–52, London and New York: Rowman & Littlefield International.

Tabatabaei Ghomi, H. and J. Stegenga (2022), 'Conventional Choices in Outcome Measures Influence Meta-Analytic Results', *Philosophy of Science*, 89 (5): 949–959.

Thoma, J. (2015), 'The Epistemic Division of Labor Revisited', *Philosophy of Science*, 82 (3): 454–72.

Turner, E. H., A. M. Matthews, E. Linardatos, R. A. Tell and R. Rosenthal (2008), 'Selective Publication of Antidepressant Trials and Is Influence on Apparent Efficacy', *New England Journal of Medicine*, 358 (3): 252–60.

Weisberg, M. and R. Muldoon (2009), 'Epistemic Landscapes and the Division of Cognitive Labor', *Philosophy of Science*, 76 (2): 225–52.

Wieseler, B., N. Wolfram, N. McGauran, M. F. Kerekes, V. Vervölgyi, P. Kohlepp, M. Kamphuis and U. Grouven (2013), 'Completeness of Reporting of Patient-Relevant Clinical Trial Outcomes: Comparison of Unpublished Clinical Study Reports with Publicly Available Data', *PLoS Med*, 10 (10): e1001526.

Zollman, K. J. S. (2007), 'The Communication Structure of Epistemic Communities', *Philosophy of Science*, 74 (5): 574–87.