Philosophy of Science (2022), **89**, 949–959
doi:10.1017/psa.2022.56

PSA PHILOSOPHY OF SCIENCE ASSOCIATION
PHILOSOPHY OF SCIENCE

SYMPOSIA PAPER

Conventional Choices in Outcome Measures Influence Meta-Analytic Results

Hamed Tabatabaei Ghomi* and Jacob Stegenga

Department of History and Philosophy of Science, University of Cambridge, Cambridge, UK
*Corresponding author. Email: ht396@cam.ac.uk

(Received 08 October 2021; revised 17 February 2022; accepted 02 May 2022; first published online 26 May 2022)

Abstract

It is a plausible speculation that conventional choices in outcome measures might influence the results of meta-analyses. We test that speculation by simulating data from trials on antidepressants. We vary real drug effectiveness while modulating conventional values for outcome measures. We had previously shown that one conventional choice used in meta-analyses of antidepressants falls in a narrow range of values that maximize estimates of effectiveness. Our present analysis investigates why this phenomenon occurs. Moreover, our results suggest the superiority of absolute outcome measures over relative measures. This research program can be extended to test numerous other aspects of clinical research.

1. Introduction

An outcome measure is used to analyze data from trials and is a quantitative assessment of the strength of a causal relation. The choice of outcome measure can influence one's inferences, such as how discordant data are between trials and how strong the tested causal relation is (Sprenger and Stegenga 2017). Here, we demonstrate that fine-grained conventional choices about a particular outcome measure can have a dramatic impact on estimates of the effectiveness of interventions.

This present article is part of a larger research program developing an innovative method that has not been widely used in the philosophy of science. We simulate patient-level data from trials on a particular class of drugs while varying features of the simulated research context, such as the real effectiveness of the drugs, the number of subjects in each trial, and in the present work, the conventional features of an outcome measure. We use simulations in the spirit of Mayo-Wilson and Zollman (2021), who argue that simulations are like thought experiments in philosophy and indeed are often superior to thought experiments.[1]

[1] Simulations have become an important tool in philosophy, testing a wide range of topics in epistemology, philosophy of science, and other domains; for a sample of this approach, see Zollman (2007), O'Connor (2015), Romero (2016), and Kummerfeld and Zollman (2015). However, we are not aware of

A widely used outcome measure in meta-analyses is the *responder odds ratio*, which is defined as the ratio of the odds of being a "responder" in the drug group divided by the odds of being a responder in the placebo group. A responder is a subject whose symptoms drop below a threshold *c*. In trials and meta-analyses of antidepressants, *c* is typically 50% of a subject's pretrial symptom severity. Symptom severity in antidepressant trials is measured with various scales; one common scale is the Hamilton Depression Rating Scale (we will refer to this as the *H scale*, with corresponding H scores), a 50-point scale in which higher scores represent greater symptom severity. So, if you are a subject in a trial and your H score at the start of the trial is 26, and it is 12 at the end of the trial, you would be deemed a responder. A responder odds ratio of greater than 1 implies that a drug is effective; conversely, a responder odds ratio of close to 1 suggests that the drug is relatively ineffective.

An inference from a particular value of a responder odds ratio to a claim that a drug is effective is dubious, and although establishing this point is not our main objective, we give an argument for this as well. Our main objective is to show that this commonly used value for *c* falls within a range that maximizes the responder odds ratio and to explain this phenomenon. If other values for *c* were used in meta-analyses of antidepressants, it is very likely that lower values of the responder odds ratio would be found.

A brief thought experiment is suggestive. Based on past trial data, we know that subjects start trials with an average H score of about 24, subjects in the drug group end trials with an average H score of about 11, and subjects in the placebo group end trials with an average H score of about 14. If *c* were very low—say, 5%—then few subjects in either the drug group or the placebo group would be deemed responders, and thus the responder odds ratio would be close to 1, and the drug would be deemed ineffective. Conversely, if *c* were very high—say, 95%—then many subjects in both groups would be deemed responders, and thus again the responder odds ratio would be close to 1, and thus again the drug would be deemed ineffective. A *c* value of 50% looks to be roughly in a "sweet range" to maximize the responder odds ratio.

One way to test these suggestions would be to get patient-level data from trials on antidepressants and then compute the responder odds ratio while varying *c*. However, save few exceptions, no one has access to these data. Even the latest and biggest meta-analyses get access only to group-level summaries of trial data (Cipriani et al. 2018).

Our approach gets around the problem of data access. Moreover, our approach allows us to test counterfactuals, such as scenarios in which the tested drugs are very effective or ineffective, and scenarios in which a different measurement scale is used. Tabatabaei Ghomi and Stegenga (forthcoming) show that setting *c* around 50% indeed nearly maximizes the responder odds ratio and thereby maximizes the estimate of drug effectiveness, and other values of *c* give lower values for the responder odds ratio. These results are consistent with previous work, such as Hadzi-Pavlovic (2009). In the present article, we explore the reasons for this phenomenon.

work in the philosophy of science in which data from trials are simulated to test meta-level hypotheses about research. Some articles in the field of statistics have simulated trial data of antidepressants (Chevance et al. 2019; Landin et al. 2000; Santen et al. 2009), but the questions posed by those researchers and, consequently, their methods are quite different from ours.

https://doi.org/10.1017/psa.2022.56 Published online by Cambridge University Press

Here is another way of thinking about our thought experiment mentioned earlier. For a causal relation evaluated in a trial with data generated by a particular measuring instrument, there is a sweet range for c which minimizes the proportion of subjects deemed responders in the placebo group and maximizes the proportion of subjects deemed responders in the drug group, thereby maximizing the responder odds ratio (for similar suggestions, see Ragland 1992).

Now consider another thought experiment. Suppose we modify the H scale by adding questions that are irrelevant to the causal relation under investigation and that add the same score to all subjects, such as "Are you human?" and "Is your body composed of at least five atoms?" This increases the total possible H score without changing the absolute difference in mean scores between the drug and placebo groups. The sweet range for c that maximizes the responder odds ratio will correspondingly increase. We evaluate this via our simulation approach, and we confirm our speculation.

Another objective of this article is to enter the recent debate between some philosophers, who argue that absolute outcome measures are informative while relative outcome measures are misleading, and other philosophers, who argue that both absolute and relative outcome measures can be informative. The responder odds ratio is an example of a relative outcome measure that is misleading and uninformative about the real effectiveness of interventions.

2. Methods

Our simulation approach was initially presented in another article (Tabatabaei Ghomi and Stegenga, forthcoming). We briefly reiterate it here, adding details particular to this article.

2.1 Modeling responders

OR is 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). The odds of being a responder in the placebo or the drug group equals the number of responders divided by the number of nonresponders in each 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 responders in the drug ($g = d$) or the placebo ($g = p$) group, and $|R_g|$ and $|g|$ are the sizes of R_g and g respectively.

A subject i in group g counts as a *responder* ($i \in R_g$) if their H score after treatment (H_i^a) is less than or equal to a fraction c of their H score before treatment (H_i^b):

$$R_g = \left\{ \forall i \in g \mid \frac{H_i^a}{H_i^b} \leq c \right\}.$$

A drug is deemed effective if the lower 95% confidence interval of *OR* is higher than 1 (Bland 2015).

https://doi.org/10.1017/psa.2022.56 Published online by Cambridge University Press

Table 1. Parameters Used in Simulations

Parameter	Calculated	Used for Simulation
Mean H score in placebo group before treatment, m_p^b	23.5 (unweighted: 23.3)	24
Mean H score in drug group before treatment, m_d^b	24.0 (unweighted: 23.8)	24
Mean H score in placebo group after treatment, m_p^a	14.2 (unweighted: 14.6)	14
Mean H score in drug group after treatment, m_d^a	11.3 (unweighted: 11.5)	11
Standard deviation (SD) of H score in placebo group before treatment, s_p^b	3.5	3.5
SD of H score in drug group before treatment, s_d^b	3.5	3.5
SD of H score in placebo group after treatment, s_p^a	7.9 (unweighted: 7.98)	7.9
SD of H score in drug group after treatment, s_d^a	7.4 (unweighted: 7.3)	7.4

2.2 Simulating patient data

We used a random generator to generate patient data before and after treatment with drug or placebo based on the parameters that we calculated from published meta-analyses of trials on antidepressants. Our parameter estimates were based on one of the most recent and most comprehensive meta-analyses of antidepressants (Cipriani et al. 2018). Because the absolute values of these parameters are central to this article, in this section we provide some details of how we estimated them.

The most frequently used scale was the 17-question H scale, used in 537 out of 1,199 parameters reported in the supplementary material of Cipriani et al. (2018). After excluding reports with missing data, we ended up with 105 parameter values for the placebo and 337 values for various drugs (drug values were combined, regardless of drug identity). We calculated both the simple and the weighted averages (by the number of participants in each study) of the parameter values (weighting had a negligible impact). Because Cipriani et al. (2018) report standard deviations only after treatment, we used Hieronymus et al. (2016) to obtain the standard deviations of the placebo and drug groups before treatment. The other distribution parameter values reported by Hieronymus et al. (2016) were close to the values we calculated and so cross-validated our results. The calculated values are reported in table 1.

We simulated trials of various sizes, from 100 to 500 subjects (equally distributed between the drug and placebo groups; the size of each group is denoted by n). In our previous study, we varied drug effectiveness, indicated by the after-treatment mean H score in the drug group (m_d^a). For each combination of values for m_d^a (seven values, from 9 to 15) and n (nine values, from 50 to 250), we repeated the simulation 5,000 times (thereby generating data for 315,000 trials in total). In the present study, we fixed the drug effectiveness to $m_d^a = 11$ (estimated from meta-analyses), but we varied the H scale as described in the following section, resulting in different absolute values for mean H scores in the drug and placebo groups before and after

https://doi.org/10.1017/psa.2022.56 Published online by Cambridge University Press

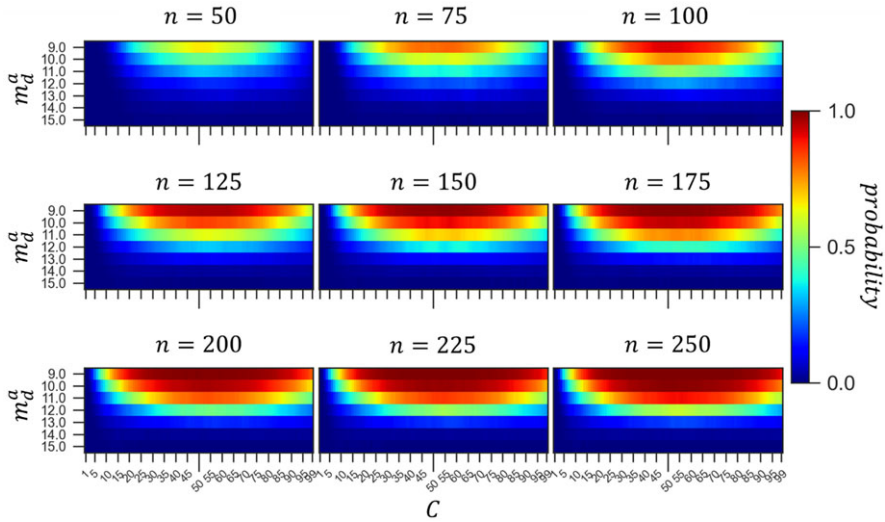

Figure 1. The probability of concluding that a drug is effective under various combinations of sample size (n), real drug effectiveness (m_d^a), and conventional threshold for definition of responder (c).

treatment.[2] We repeated the simulation 5,000 times for each combination of n and H scale (thereby generating data for 225,000 trials in total).

2.3 Adjusting measurement scale

We simulated data using multiple versions of the H scale. The base scale is H50, the 17-question scale commonly used in real trials, with a total of 50 points. The mean values reported in table 1 are based on this scale. We adjusted this scale by adding a number of irrelevant questions for which all subjects would answer affirmatively, generating a scale with a total of 57 points (H57), 67 points (H67), 88 points (H88), and 151 points (H151). Section 3 explains why we used these particular scales. These additional questions shift m_d^b, m_d^a, m_p^b and m_p^a up by a fixed value equivalent to the number of added questions without changing the absolute difference between the means of the placebo and drug groups. For each scale, we assessed the effect of varying c on the difference in the number of responders in the placebo and drug groups, the lower 95% confidence interval of *OR*, and the probability of concluding that a drug is effective.

3. Results and discussion

3.1 The sweet range

The probability of concluding that a drug is effective under various combinations of n, m_d^a and c is displayed in figure 1; this was one of the main results of Tabatabaei Ghomi

[2] There is a hard lower cutoff on the randomly generated data: H scores cannot be lower than zero. This can result in distributions with averages slightly higher than the m_d^a used for random generation, especially with lower m_d^a values. See Tabatabaei Ghomi and Stegenga (forthcoming) for more technical details.

https://doi.org/10.1017/psa.2022.56 Published online by Cambridge University Press

and Stegenga (forthcoming), and we reproduce it here to lay the grounds for the remainder of our results. We begin by noting three features of figure 1. First, as the effectiveness of the drug decreases from $m_d^a = 9$ to $m_d^a = 15$, the probability of concluding that a drug is effective decreases, as expected. Second, the probability of finding a drug to be effective increases by increasing the number of participants, which is a result of the increased power of larger trials. Third, and most relevant to our present interest, the influence of the (arbitrary and conventional) value of c on inferences of effectiveness is abundantly clear.

We observe three key features of the influence of c on the probability of concluding that a drug is effective. First, as mentioned earlier, $c \approx 50\%$ nearly maximizes the probability of concluding that a drug is effective. Second, the range of c with a high probability of concluding that a drug is effective narrows as we go toward less effective drugs (higher m_d^a). And third, the optimum range of c for concluding that a drug is effective shifts slightly to the right as we go toward less effective drugs. In Tabatabaei Ghomi and Stegenga (forthcoming), we noted these phenomena but did not explain them. Here, we provide an explanation for these phenomena, thereby articulating the source of a systematic bias in some trials and meta-analyses.

There is a "sweet range" for c that minimizes the proportion of subjects deemed responders in the placebo group and maximizes the proportion of subjects deemed responders in the drug group, thereby maximizing the lower confidence interval of the responder odds ratio; this can be formally represented as follows:

$$\frac{m_d^a}{m_d^b} < c < \frac{m_p^a}{m_p^b} \text{ (sweet range)}.$$

Here is why the sweet range is the case. Recall that an individual patient in the drug or placebo group is deemed a responder if the ratio of her H score after treatment over her H score before treatment (H_i^a/H_i^b) falls below c. A c value that on average satisfies this criterion for patients in the drug group but not for the patients in the placebo group increases the difference between the number of responders in the two groups and thus increases the *OR*. c values within the sweet range have this property.[3]

When using H17, on average the mean H score before treatment is 24, and the mean H score after treatment is 14 and 11 for the placebo and drug groups, respectively. Coincidently, these values are such that $c = 50\%$ falls within the sweet range:

$$\frac{11}{24} = 0.45 < c = 0.5 < \frac{14}{24} = 0.58.$$

This is why we see, in figure 1, the probability of deeming a drug to be effective clustered in a range bounding $c = 0.5$.[4]

As drugs become less effective, m_d^a increases, and consequently, the lower bound of the sweet range approaches the upper bound. This explains the dome shape of figure 1, in which a progressively smaller range of c values has a high chance of finding less effective drugs to be effective. Also, as the effectiveness of drugs

[3] The maximizing c value is within the sweet range, and other values within the range on either side of the maximizing c result in a relatively high, but less than maximum, *OR*.

[4] The lower bound of the sweet range found in the simulation may be slightly higher than 11/24 because of the technical point mentioned in footnote 2.

https://doi.org/10.1017/psa.2022.56 Published online by Cambridge University Press

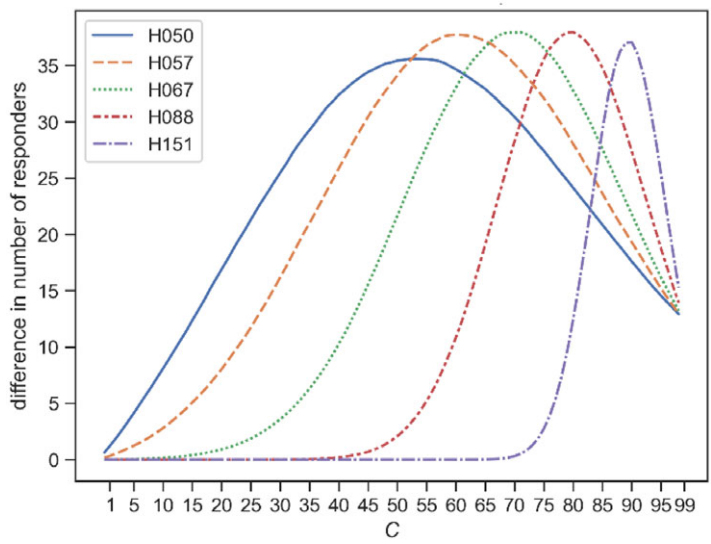

Figure 2. Difference in the number of responders between the drug and placebo groups.

decreases to that of placebo ($m_d^a \to m_p^a$), the lower bound of the sweet range increases so much that $c = 0.5$ falls out of the sweet range, and higher values of c are within the range. This explains the slight shift toward higher values of c at the peak of the inverse dome.

These are, in part, numerical consequences of the arbitrary range of scores of H17. Were the H scale designed differently, assigning different absolute values to patients, $c = 0.5$ would not have been within the sweet range. To show this, we repeated the simulations, this time with our hypothetical H scales with additional irrelevant questions. Although the drug effectiveness is fixed in these simulations (and consequently, the absolute mean difference between drug and placebo H scores is fixed), the absolute value that each H scale assigns to patients shifts, and consequently, the absolute values of mean H scores in both the drug and placebo groups differ between scales.

We chose the particular adjustments to the H scale such that the following values of c fall within the sweet range:

$$\text{H50} \Rightarrow c \approx 50\%$$
$$\text{H57} \Rightarrow c \approx 60\%$$
$$\text{H67} \Rightarrow c \approx 70\%$$
$$\text{H88} \Rightarrow c \approx 80\%$$
$$\text{H151} \Rightarrow c \approx 90\%$$

Figure 2 shows the average difference between the number of responders in the placebo and drug groups in the simulations of trials with 500 subjects. The results in other trial sizes are similar (save small stochastic fluctuations) and so are not shown. The average difference between the number of responders in the placebo and the drug groups maximizes approximately at 50%, 60%, 70%, 80%, and 90% for H17,

https://doi.org/10.1017/psa.2022.56 Published online by Cambridge University Press

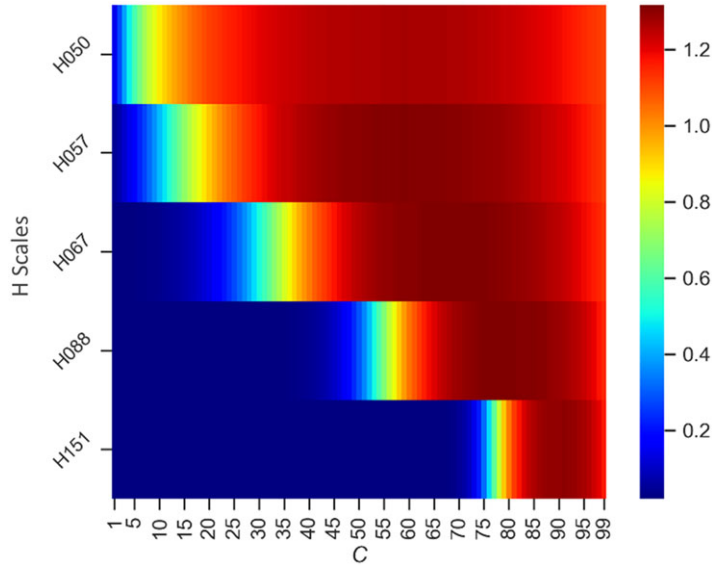

Figure 3. Average lower 95% confidence interval of *OR* under combinations of various H scales and *c* values.

H57, H67, H88, and H151, respectively. This maximum difference, in turn, maximizes the lower 95% confidence interval of *OR* (figure 3) and, consequently, the probability of concluding that the drug is effective (figure 4). Figures 3 and 4 also clearly visualize the concept of the sweet range (the spread of the darkest red tone). This result shows that were the H scales designed differently, some drugs that were deemed effective by the *OR* would have been deemed ineffective based on exactly the same data.

3.2 Superiority of absolute outcome measures

Our results emphasize the superiority of absolute measures, such as the difference in mean H score reduction between the drug and placebo groups, compared with relative measures, such as *OR*. The analyses with the newly created H scales show that the value of the *OR* is sensitive to arbitrary changes to the H scales, whereas the absolute mean H score reduction is not sensitive to such changes. Choose a particular *c* on the *x*-axis of figure 3, and move up the corresponding column; as the H scale changes, so does the probability of concluding that a drug is effective based on *OR*. Yet this is an undesirable feature of an outcome measure. Recall that the questions that were added to the newly formed H scales are, by stipulation, totally irrelevant to that which is being measured (namely, severity of depression symptoms).

A plausible desideratum for an outcome measure is that if an outcome measure is used to summarize the same data from a trial generated from two measurement scales that are identical in every respect relevant to the causal relation under investigation and differ only in ways causally unrelated to the causal relation, then the outcome measure should report the same value. For example, if you measure the temperature today at lunchtime and then at dinnertime using a blue Celsius thermometer

https://doi.org/10.1017/psa.2022.56 Published online by Cambridge University Press

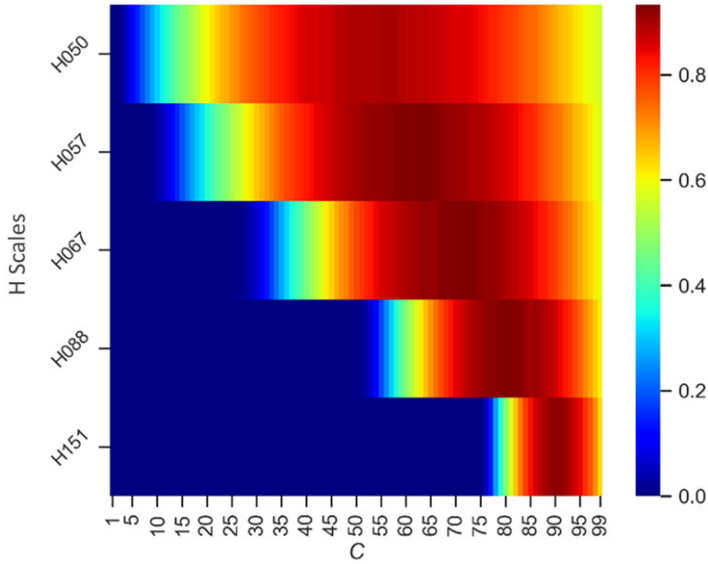

Figure 4. Probability of finding the same drug to be effective under combinations of various H scales and c values.

and a red Celsius thermometer, you should expect to compute the same lunch–dinner temperature difference with both thermometers because the color of the thermometers is causally irrelevant to measuring temperature. The difference in the mean reduction in symptom severity between the drug and placebo groups satisfies this desideratum (because the mean H scores for both the drug group and the placebo group scale up identically when changing the H scale, the mean difference between the groups remains the same). As we have seen in figure 3, the responder odds ratio does not satisfy this desideratum.

In general, the responder odds ratio involves a loss of information because the definition of responder does not include the magnitude of change in symptom severity for a subject. Consider a hypothetical weight-loss drug and a trial in which a responder is defined as anyone who loses at least a nonzero amount of weight. Suppose that in the trial of the drug, there were 100 subjects in each of the drug and placebo groups, and in the drug group, 75 subjects lost 10 grams, yet 25 subjects *gained* 5 kilograms, and in the placebo group, 50 subjects lost 10 grams, and 50 subjects gained 10 grams; this drug would have a responder odds ratio of 3 in favor of weight loss, even though it caused virtually no weight loss in most subjects and caused substantial weight gain in a large proportion of subjects. For reasons such as this, statistician Stephen Senn claims that measures such as the responder odds ratio are "liable to be extravagantly interpreted" (Senn 2003, 239).

Although some philosophers have argued that both absolute and relative measures should be reported on the presumed grounds that both are informative (e.g., Hoefer and Krauss 2021), our analysis here supports those who have argued that reporting relative measures can be misleading. Stegenga and Kenna (2017) and Sprenger and

https://doi.org/10.1017/psa.2022.56 Published online by Cambridge University Press

Stegenga (2017) argue for the superiority of absolute outcome measures for binary outcomes, and our argument here extends this to continuous outcome measures.

4. Conclusion

Our results show that the conventional choice of $c = 50\%$ in the responder odds ratio is in a range that maximizes estimates of the effectiveness of antidepressants. The dependence of the probability of concluding that a drug is effective on the choice of c indicates the problem with this arbitrary measure. We further offer an explanation for where this "sweet range" comes from, along with other nuances related to this sweet range.

In what is probably the most significant meta-analysis of antidepressants to date, the 95% lower confidence interval of the odds ratio for antidepressants was estimated to be between 1.37 and 2.13, with many dugs showing values around 1.6 (Cipriani et al. 2018). This might strike you as rather modest. Yet, we hope to have demonstrated how uninformative such a measure is about the real effectiveness of the drugs. Moreover, we have shown that this value depends on a conventional choice for c, and other choices for c would entail even lower values for the responder odds ratio. It is a happy coincidence for those who wish to substantiate the putative effectiveness of antidepressants that the commonly used value of c in meta-analyses of antidepressants is tuned to the commonly used measurement scale and actual facts about subjects in trials of antidepressants such that the responder odds ratio is nearly maximized.

Our method deployed here is an example of a possible way to evaluate speculative hypotheses about research practices and can be extended to other domains.

Acknowledgments. For commentary and discussion, we are grateful to Cristian Larroulet Philippi, Adria Segarra, Sophia Crüwell, Adrian Erasmus, Oliver Holdsworth, Ina Jantgen, Charlotte Zemmel, Zinhle Mncube, and two anonymous referees.

References

Bland, Martin. 2015. *An Introduction to Medical Statistics*. Oxford: Oxford University Press.

Chevance, Astrid, Florian Naudet, Raphaël Gaillard, Philippe Ravaud, and Raphaël 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, Andrea, Toshi A. Furukawa, Georgia Salanti, Anna Chaimani, Lauren Z. Atkinson, Yusuke Ogawa, Stefan Leucht, et al. 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 (10128):1357–66.

Hadzi-Pavlovic, Dusan. 2009. "Exploring Kirsch and Moncrieff's 'Response Rate Illusion.'" *Acta Neuropsychiatrica* 21 (1):38–40.

Hieronymus, Fredrik, Johan Fredrik Emilsson, Staffan Nilsson, and Elias 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, Carl, and Alexander Krauss. 2021. "Measures of Effectiveness in Medical Research: Reporting Both Absolute and Relative Measures." *Studies in History and Philosophy of Science Part A* 88 (9041):280–83.

Kummerfeld, Erich, and Kevin J. S. Zollman. 2015. "Conservatism and the Scientific State of Nature." *British Journal for the Philosophy of Science* 67 (4):1057–76.

https://doi.org/10.1017/psa.2022.56 Published online by Cambridge University Press

Landin, Richard, David J. DeBrota, T. A. DeVries, William Z. Potter, and M. A. Demitrack. 2000. "The Impact of Restrictive Entry Criterion during the Placebo Lead-In Period." *Biometrics* 56 (1):271–78.
Mayo-Wilson, Conor, and Kevin J. S. Zollman. 2021. "The Computational Philosophy: Simulation as a Core Philosophical Method." *Synthese* 199 (5):3647–73.
O'Connor, Cailin. 2015. "Ambiguity Is Kinda Good Sometimes." *Philosophy of Science* 82 (1):110–21.
Ragland, David R. 1992. "Dichotomizing Continuous Outcome Variables: Dependence of the Magnitude of Association and Statistical Power on the Cutpoint." *Epidemiology* 3 (5):434–40.
Romero, Felipe. 2016. "Can the Behavioral Sciences Self-Correct? A Social Epistemic Study." *Studies in History and Philosophy of Science Part A* 60:55–69.
Santen, Gijs W., Erik W. Van Zwet, Meindert Danhof, and Oscar 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.
Senn, Stephen. 2003. "Disappointing Dichotomies." *Pharmaceutical Statistics: The Journal of Applied Statistics in the Pharmaceutical Industry* 2 (4):239–40.
Sprenger, Jan, and Jacob Stegenga. 2017. "Three Arguments for Absolute Outcome Measures." *Philosophy of Science* 84 (5):840–52.
Stegenga, Jacob, and Aaron Kenna. 2017. "Absolute Measures of Effectiveness." In *Measurement in Medicine: Philosophical Essays on Assessment and Evaluation*, edited by Leah M. McClimans, 35–52. Lanham, MD: Rowman & Littlefield International.
Tabatabaei Ghomi, Hamed, and Jacob Stegenga. Forthcoming. "Simulation of Trial Data to Test Speculative Hypotheses about Research Methods." In *Experimental Philosophy of Medicine*, edited by Kristien Hens and Andreas De Block. London: Bloomsbury.
Zollman, Kevin J. S. 2007. "The Communication Structure of Epistemic Communities." *Philosophy of Science* 74 (5):574–87.

Cite this article: Tabatabaei Ghomi, Hamed and Jacob Stegenga. 2022. "Conventional Choices in Outcome Measures Influence Meta-Analytic Results." *Philosophy of Science* 89 (5):949–959. https://doi.org/10.1017/psa.2022.56