MISALIGNMENT BETWEEN RESEARCH HYPOTHESES AND STATISTICAL HYPOTHESES – A THREAT TO EVIDENCE-BASED MEDICINE?

INSA LAWLER[1] & GEORG ZIMMERMANN[2]

[1]University of North Carolina at Greensboro, irlawler@uncg.edu
[2]Paracelsus Medical University & Paris Lodron University of Salzburg, georg.zimmermann@sbg.ac.at

This is a pre-print of an article published in *Topoi*. The final authenticated version is available online at: https://doi.org/10.1007/s11245-019-09667-0

Abstract

Evidence-based medicine frequently uses statistical hypothesis testing. In this paradigm, data can only disconfirm a research hypothesis' competitors: One tests the negation of a statistical hypothesis that is supposed to correspond to the research hypothesis. In practice, these hypotheses are often misaligned. For instance, directional research hypotheses are often paired with non-directional statistical hypotheses. Prima facie, one cannot gain proper evidence for one's research hypothesis employing a misaligned statistical hypothesis. This paper sheds lights on the nature of and the reasons for such misalignments and it provides a thorough analysis of whether they pose a threat to evidence-based medicine. The upshots are that the misalignments are often hidden for clinicians and that although some cases of misalignments can be partially counterbalanced, the overall threat is non-negligible. The counterbalances either lead to methodological inadequacy (in addition to the misalignment), loss of statistical power, or involve a (potential) lack of information that could be crucial for decision making. This result casts doubt on various findings of medical studies in addition to issues associated with under-powered studies or the replication crisis.

Acknowledgements

We thank Arne Bathke, Robyn Bluhm, Charlotte Werndl, the audiences in Genoa, Paris, and Munich, the editors of the special issue Fabrizio Macagno and Carlo Martini, as well as two anonymous reviewers for their constructive criticisms and suggestions.

1 INTRODUCTION

Evidence-based medicine (EBM) involves the statistical analyses of data. Such analyses play a crucial role for evaluating whether a treatment is promising and for a clinician's recommendation regarding a patient's therapy. 'Evidence' is a relative term; evidence is evidence for or against something. A common approach to statistical analysis of

data is *statistical hypothesis testing*. For instance, a clinician's research hypothesis may be that transcranial magnetic stimulation (TMS) has a positive effect on spinal cord injury (SCI) patients. In this paradigm, data cannot directly support a hypothesis, but only disconfirm its competitors. So, one examines the *negation* of the research hypothesis, namely that TMS has no or a negative effect on SCI patients. If the negation is disconfirmed, one can reason that the research hypothesis is indirectly supported. One thus tests the negation of a statistical hypothesis that corresponds to the research hypothesis (cf. Fig. 1).

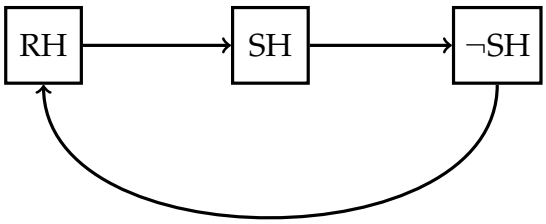

Figure 1: Indirect testing of a research hypothesis

It goes without saying that the reasoning involved in statistical hypothesis testing is risky. The conclusion does not follow from the data. The amount of risk depends on how well the sample reflects the population, on whether the samples have been drawn independently of each other, on the statistical power of the study, etc. Moreover, several researchers have argued that the results of hypothesis testing alone should not be a decisive guide (e.g., Cohen 1994; Ioannidis 2005; Cumming 2012, ch. 1), and some researchers reject dichotomous hypothesis testing and focus on other statistical means, such as the so-called confidence intervals (e.g., Meehl 1978; Cumming 2012, ch. 1). However, our main focus is not on these topics here. Instead, we are concerned with a prevailing issue that *prima facie* poses a threat even to rather low-risk cases of hypothesis testing or confidence interval calculations: Intuitively, it is essential that the statistical hypothesis closely *corresponds* to the research hypothesis; it should align with its claim. Take the previous example: The hypothesis that TMS has a positive effect on SCI patients is *directional*.[1] So, its corresponding statistical hypothesis should feature the same direction. Yet, in EBM, often a *non-directional* hypothesis is used, such as the negation of a so-called nil-null hypothesis.[2] Nil-null hypotheses reflect the claim that, say, there is no effect whatsoever. Accordingly, its negation reflects the claim that the treatment of interest has *some* effect – of whatever direction. This misalignment of pairing a directional research hypothesis with a non-directional statistical hypothesis is not uncommon. The authors examined 30 papers in paraplegia research, which

[1] We briefly discuss the normative issue of whether directional research hypotheses should be used in evidence-based medicine *at all* in sect. 4.4.

[2] Cho and Abe (2013) claim that this issue also prevails in business research, and given the reasons provided in sect. 4, it is likely to be also found in other disciplines, e.g., psychology.

were part of a systematic review on SCI trials (Zimmermann et al. 2019). 20 of the 30 papers feature such a misalignment and nonetheless involve conclusions regarding the research hypotheses. *Prima facie*, such conclusions are illegitimate, as Casella and Berger emphasize (1987, p. 106). Arguably, one cannot gain proper evidence for one's research hypothesis without an appropriately corresponding statistical hypothesis. Yet, misalignments prevail in EBM. What are the reasons for this practice? Does it pose a threat or is the misalignment benign upon closer examination?

This paper sheds light on these questions with an interdisciplinary approach, combining insights from applied medical statistics and philosophical considerations. We proceed as follows: In section 2, we briefly describe the basics of statistical hypothesis testing in EBM. In section 3, we distinguish between different forms of a misalignment between research hypotheses and statistical hypotheses. In section 4, we provide basic reasons for the misalignment practice in EBM. Our upshots are (i) that these reasons do not justify the misalignments and (ii) that the misalignments are often hidden for clinicians. In section 5, we evaluate whether the misalignments pose a threat to evidence-based medicine, with an eye on informed decision making by clinicians, physicians, and patients. Our upshot is that some cases of misalignments can be partially counterbalanced. However, the counterbalances either lead to methodological inadequacy (in addition to the misalignment), loss of statistical power, or involve a (potential) lack of information that could be crucial for informed decision making. The threat is thus non-negligible. This result of our analysis casts doubt on various findings of medical studies in addition to issues associated with under-powered studies (cf., e.g., Ioannidis 2005) or the fact that a substantial number of findings of medical studies cannot be replicated (cf., e.g., Davis, R. 2014). In section 6, we suggest some remedies to the misalignment practice.

2 STATISTICAL HYPOTHESIS TESTING IN CLINICAL STUDIES

Statistical hypothesis testing is a form of *inferential* statistics. The aim is to infer properties of a population (e.g., all SCI patients) by investigating a sample (e.g., 50 SCI patients). There have been extensive discussions about the vices and virtues of this testing, but we do not want to go into detail here.[3] The relation between research hypotheses and statistical hypotheses is also crucial to statistical analyses where estimating the effect size or confidence intervals is the primary goal.

Statistical hypothesis testing involves *four* central steps. *First*, a research hypothesis is established that contains measurable variables for its evaluation, such as measures of a patient's electrical perceptual threshold (EPT). This step is also called *operationalization*.

[3] For an overview see, e.g., Nickerson 2000; Gigerenzer 2004; Lecoutre and Poitevineau 2014.

Second, the research hypothesis needs to be paired with a corresponding statistical hypothesis. This requires setting up a specific statistical model for the collected data. For instance, a statistical hypothesis could be concerned with the difference $\Delta = \mu_T - \mu_P$ of mean electric perceptual thresholds under treatment and placebo in a TMS study with SCI patients. To assess the research hypothesis that, say, the treatment leads to an increased EPT compared to a placebo, one constructs a statistical hypothesis that is supposed to align with the research hypothesis. This is the so-called *alternative* hypothesis (H_1), e.g., $H_1 : \Delta > 0$. However, as mentioned before, its *negation* is tested, i.e., the so-called *null hypothesis* (H_0), e.g., $H_0 : \Delta \leq 0$. Importantly, H_0s can also be concerned with specific differences, e.g., $H_0 : \Delta \leq 2$ is also a fine H_0.[4]

Third, the evidence against H_0 can be quantified by calculating the so-called p-value. This value is defined as the probability of observing data which is at least as 'extreme' as the data at hand, when H_0 is true. For illustration, assume that the test statistic (i.e., basically the empirical mean difference, multiplied with some scaling factor) in the treatment-versus-placebo example is 1. Although this is a value greater than 0, it could also be a result of by-chance differences. So, it is reasonable to calculate the probability of getting a test statistic of 1 or even a larger value, given that the treatment is not superior. This probability is illustrated in the first plot in Figure 2.

Fourth, the p-value is taken as the basis for a formal decision rule: H_0 is rejected if and only if the p-value is less than a pre-specified cutoff α, where often $\alpha = 0.05$. So, loosely speaking, if it is very unlikely to observe data like the one at hand under H_0, you do not trust in H_0, where 'very unlikely' is specified by α. If the p-value is smaller than α, the result is said to be statistically significant.

A decision based on the p-value can be wrong. There could be a *false positive* result or type I error: the test is significant although H_0 is true. Conversely, the test might not be significant although H_1 is true, which is a *false negative* result or a type II error (β). 1 - β is also called the *statistical power*. In regulatory guidelines, emphasis is placed on keeping the type I error rate below a pre-specified threshold α, but also on determining the number of samples or subjects that have to be included in the study to achieve sufficient power (cf., e.g., ICH E9). The former can be accomplished by following the aforementioned decision rule (i.e., rejecting the null hypothesis if $p < \alpha$). For sample size calculation, one has to specify α, the hypothesized effect size, the desired power (e.g., 80 or 90%), and some nuisance parameters (e.g., the hypothesized variability of data, other variances, correlations, depending on the particular hypothesis test). In clinical studies, the hypothesized effect size needs to be at least as large as the so-called *minimal clinically important difference* (MCID).

[4] There are also cases where the null hypothesis is not the negation of the alternative hypothesis (e.g., when considering fixed point alternatives). However, since these cases are the exception rather than the rule, they are not our main focus of this article, except for cases where they may serve as a potential remedy for avoiding what we call 'magnitude misalignment' (see below).

When calculating the p-value, one has to distinguish between one- and two-sided testing problems. The EPT example above represents an instance of a so-called 'one-sided hypothesis testing problem' because its H_1 has *one* direction. In other settings, for instance, when comparing two treatments in the development phase, a researcher might only hypothesize that there is *some* difference in terms of efficacy. In this case, the null hypothesis will be non-directional or two-sided, that is, $H_0 : \mu_T - \mu_P = 0$ vs. $H_1 : \mu_T - \mu_P \neq 0$. The two cases are illustrated in Figure 2. In the one-sided case (corresponding to the EPT example), the p-value is the right-tail probability, whereas the probabilities in both tails are considered in the two-sided setting. The p-value for the two-sided case in Figure 2 is twice the p-value for the one-sided case.

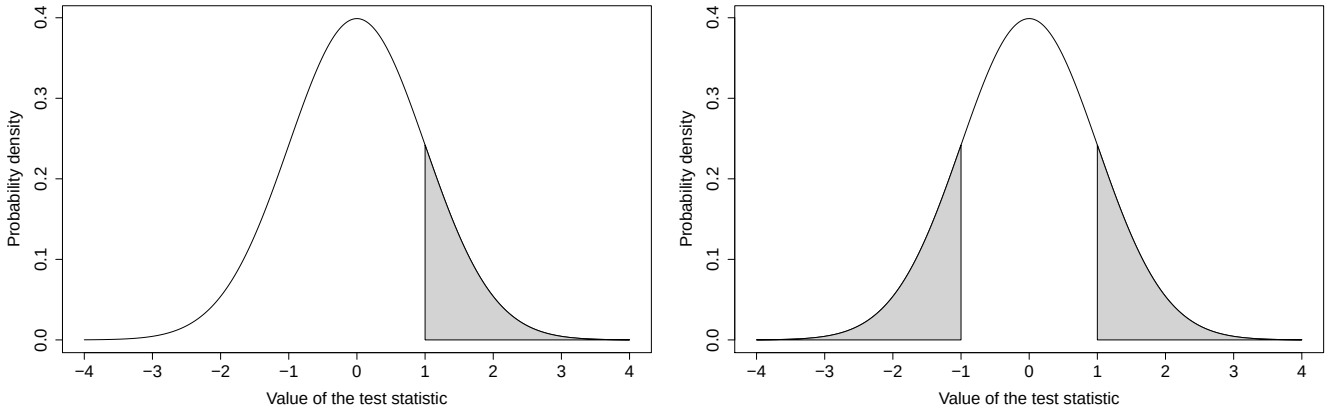

Figure 2: One-sided vs. two-sided tests; Grey areas represent the p-values.

Hypothesis testing is centered around the p-value. But, as we have mentioned, the hypothesized size of an effect plays an important role when it comes to the power analysis and the sample size considerations for a given study. Nevertheless, considering the *actual* effect size in a data set is not part of the formal decision rule about H_0 as such.[5] It goes without saying that it is nonetheless important for informed clinical decision making. We go back to this point in section 5.

In medical research, hypotheses are also evaluated by calculating so-called *confidence intervals*. For instance, in order to test $H_0 : \Delta \leq 0$, one might check whether the corresponding one-sided confidence interval contains 0 or not. In what follows, we focus on p-values. But, as we indicate, the problems discussed may also arise when using confidence intervals instead of p-values.

[5] Note that it would be methodologically *inadequate* to exclusively consider effect sizes because by-chance variations cannot be accounted for.

3 PREVAILING MISALIGNMENTS BETWEEN RESEARCH HYPOTHESES AND STATISTICAL HYPOTHESES

There is a variety of possible misalignments between research hypotheses and their statistical hypotheses. We focus on two common ones, namely on what we call DIRECTION MISALIGNMENT and MAGNITUDE MISALIGNMENT. A research hypothesis and its statistical hypothesis are misaligned when the statistical hypothesis does not reflect the *direction* of the research hypothesis or the *magnitude* of the effect stated in the research hypothesis. The latter case is meant as follows: A directional research hypothesis is typically not about an effect of whatever magnitude but about an effect of a particular size. For instance, in clinical studies, the effect size needs to be at least as large as the *minimal clinically important difference* (MCID). Ideally, the desired effect size should be reflected in the corresponding statistical alternative hypothesis, too: For example, one should consider setting $H_1 : \mu_T - \mu_P = \delta$ or $H_1 : \mu_T - \mu_P \geq \delta$, where δ denotes the desired effect size, e.g., the MCID.[6]

Take as a paradigmatic example for DIRECTION MISALIGNMENT a study by Derakhshanrad et al. 2015. Their research hypothesis is clearly directional: "The objective of this study was to determine whether an integrated and an intensive outpatient program would result in functional *improvement* of spinal cord injury (SCI) patients [...]." (p. 860, our italics) However, they explicitly state that only two-sided tests were used (p. 861), which implies that they used *non-directional* statistical hypotheses. 20 of the 30 studies we reviewed feature such a misalignment. DIRECTION MISALIGNMENT might also occur in the other direction, e.g., when a non-directional research hypothesis is tested with a one-sided test (presumably due to data-driven considerations). But since such cases typically involve a conclusion to a matching *directional* research hypothesis (rather than to the initial non-directional one), we do not focus on this variant of DIRECTION MISALIGNMENT.

A frequently occurring instance of misalignment in clinical studies features both DIRECTION MISALIGNMENT and MAGNITUDE MISALIGNMENT: A directional research hypothesis is paired with the negation of a statistical nil-null hypothesis, i.e., with a thesis that reflects that there is *no* effect. The statistical hypothesis is then tested with a two-sided test. This is a twofold misalignment: Although the research hypothesis is about a directional effect, the statistical hypotheses are just concerned with whether there is any effect (DIRECTION MISALIGNMENT) of any magnitude (MAGNITUDE MISALIGNMENT). 17 of the 30 studies we reviewed feature such a twofold misalignment.

A less common form of a misaligned statistical hypothesis is a direction-matching statistical hypothesis that does not reflect the magnitude claim (MAGNITUDE MISALIGNMENT). As Bigirumurame and Kasim emphasize (2017), such a misalignment is quite

[6] As we indicated in footnote 4, in such cases H_1 and H_0 do not exhaust the parameter space.

common in medical research. For instance, Jeong and Yoo conclude that air stacking leads to significant improvement for cervical SCI patient (2015, p. 1952), although the statistical hypotheses used in their study did not specify any effect magnitude.[7]

There is also one form of misalignment, which occurs at an earlier stage, and one form of misalignment, which occurs at a later stage. The former occurs at the first step of statistical hypothesis testing. We call it OPERATIONALIZATION MISALIGMENT.[8] It can occur when the measurable variables are chosen for the statistical evaluation. For instance, positive differences in the mean electric perceptual threshold might not adequately capture the hypothesis that the treatment leads to an improvement for the SCI patients. The other form of misalignment occurs at the testing stage. We call it TESTING MISALIGNMENT. Recall the indirect testing method (cf. Fig. 1): One tests the negated corresponding statistical hypothesis. There are cases, where the statistical evaluation method tests a negated statistical hypothesis other than the initially constructed one. For instance, *non-parametric* methods are often recommended as a standard remedy or rule-of-thumb when the validity of the assumptions underlying classical parametric tests (such as the *t*-test) are suspected (cf., e.g., Field 2000, p. 49). However, non-parametric methods are based on statistical hypotheses *other* than the ones used for parametric tests. Briefly, the effect measure that underlies frequently used non-parametric tests like, for example, the Wilcoxon-Mann-Whitney test, may also detect effects other than a shift of the group means (i.e., the parametric effect measure). So, if a non-parametric method is used *post hoc* and one draws conclusions regarding one's research hypothesis, one reasons based on a *non-matching* negated statistical hypothesis. Importantly, OPERATIONALIZATION MISALIGMENT and TESTING MISALIGNMENT can occur independently of whether the initial statistical hypothesis is aligned.

Analyzing OPERATIONALIZATION MISALIGMENT and TESTING MISALIGNMENT is a project on its own. In what follows, we focus on DIRECTION and MAGNITUDE MISALIGNMENT.

4 REASONS FOR THE PREVAILING MISALIGNMENTS

We identified *four* basic reasons for the misalignments between research hypotheses and their statistical hypotheses in medical research: THE TAUGHT ERROR, THE PRACTICAL ENFORCEMENT, THE AIM OF CONSERVATIVE TESTING, and THE AIM OF RESEARCH OPENMINDEDNESS. Our upshots are that these reasons do not *justify* the misalignment practice and that the misalignments are often *hidden* for clinicians.

[7] In fact, Jeong and Yoo's study also features DIRECTION MISALIGNMENT. They conclude from *two-sided* tests that there is a significant improvement (cf., e.g., Jeong and Yoo 2015, p. 1952).

[8] We owe this suggestion to Gerit Pfuhl.

4.1 *The taught error*

The first reason is that misalignments occur in teaching of statistical testing. This issue comes in at least *three* flavors: The *first* is that directional research hypotheses are not uncommonly paired with non-directional statistical hypotheses, when illustrating hypothesis testing. Consider the following case (Cumming 2012, p. 21):

> Suppose we want to know whether the new treatment for insomnia is better than the old. To use NHST [Null hypothesis significance testing] we test the null hypothesis that there's no difference between the two treatments in the population.

So, Cumming effectively suggests that to test a directional research hypothesis one should use a statistical nil-null hypothesis. To test whether the treatment is *better* one tests a statistical hypothesis that reflects the claim that the treatment does not make any difference whatsoever (e.g., $H_0 : \Delta = 0$). So, H_1 is $\Delta \neq 0$, which does *not* correspond to the directional research hypothesis. Here, a misalignment case is given as a *paradigmatic* example. A less extreme case is stating a statistical nil-null hypothesis along with a directional research hypothesis (cf., e.g., Hacking 2001, ch. 18). Although this does not imply that these hypotheses are paired, it is at least suggested. A related issue is that some textbooks suggest that one can conclude from a statistically significant p-value for the nil-null hypothesis that a *directional* statistical hypothesis is (indirectly) supported (and thus a corresponding directional research hypothesis). For instance, Machin, Cambell, and Walters illustrate hypothesis testing with the following example: The blood pressure of patients is examined before and after exercise. The p-value calculated for a nil-null hypothesis is statistically significant. Machin et al. then claim that "[...] there is sufficient evidence to [...] accept the [statistical] alternative hypothesis that there is a difference (a rise) in the mean blood pressure of middle-aged men before and after exercise" (2007, p. 108). Yet, the difference need not be a *rise*. Without additional considerations, one cannot conclude anything directional from rejecting a non-directional statistical hypothesis.

Second, the fact that H_0 is the negation of a thesis which is supposed to reflect the research hypothesis is often not emphasized or not even mentioned. For instance, Cumming writes (2012, p. 21):

> Many textbooks describe NHST as a series of steps, something like this:
>
> 1. Choose a null hypothesis [i.e., H_0] [...] Sometimes, in addition to specifying H_0, an alternative hypothesis, H_1 is also specified.

This idea of first choosing some H_0 and then perhaps additionally specifying H_1 neglects the fact that H_1 is supposed to correspond to the research hypothesis and that H_0 is the negation of H_1. A less extreme case is exemplified by Machin,

Cambell, and Walters' claim that "[...] the null hypothesis [i.e., H_0] is *often* the negation of the research hypothesis [...]" (2007, p. 106, our italics). This is misleading. H_0 cannot negate the research hypothesis because the latter is not a statistical hypothesis. Moreover, H_0 should negate the hypothesis that reflects the research hypothesis, i.e., H_1. It is no surprise that neither the direction of nor the magnitude involved in the research hypothesis is of great concern throughout these textbooks. We found this neglect in other textbooks, too (cf., e.g., Bland 2000; Kirkwood and Sterne 2003; Everitt 2006).

The *third* issue affects MAGNITUDE MISALIGNMENT. In some textbooks, it is suggested that H_0s simply are nil-null hypotheses (cf. Altman 1991, p. 165; Bland 2000, ch. 9.1; Kirkwood and Sterne 2003, p. 59; Everitt 2006, p. 165; Machin et al. 2007, p. 106). Authors like Altman thus unsurprisingly state that "[...] there is no direct reference in this [testing] method to the *magnitude* of the effect of interest [...]" (cf. Altman 1991, p. 166, his italics). Yet, as we have seen, point hypotheses can be used as H_1.

THE TAUGHT ERROR is linked to THE PRACTICAL ENFORCEMENT.

4.2 *The practical enforcement*

THE PRACTICAL ENFORCEMENT is the fact that two-sided tests seem to be the norm in practically relevant settings. There are at least *three* reasons for this. The *first* is that some popular statistical methods only involve two-sided tests. For instance, for comparing the means of several groups – without any incorporation of co-variates – the so-called Analysis of Variance (ANOVA) F-test is often used. ANOVAs only involve two-sided tests. The null hypothesis is here $H_0 : \mu_1 = \ldots = \mu_a$, where μ_1 denotes the mean of group 1, and so forth. However, the H_0 is rejected if and only if the test statistic exceeds a certain cutoff value, which is determined based on the so-called (central) F-distribution. This structure of the rejection criterion for certain two-sided tests like the ANOVA might give the false impression of being a *one*-sided test procedure, although in fact, a two-sided hypothesis is tested, as emphasized by Kaiser (1960) and Cho and Abe (2013, p. 1261). Another example is the Chi-square test, which is frequently used for the analysis of contingency tables. A related example stems from a vivid debate about how to specify the previously mentioned type I error bound α. A large group of researchers from renowned institutions recently opted for a much lower standard, namely 0.005 (Benjamin, D. et al. 2017).[9] In their paper, they also argue for a statistical method to overcome certain problems. Yet, this proposal is concerned with two-sided testing.

The *second* reason is concerned with *statistical tools*; for some frequently encountered settings, it is more straightforward to access two-sided tests in statistical software

[9] Similar demands are voiced by, e.g., Berger and Sellke 1987; Nickerson 2000; Colquhoun 2014.

packages (such as SPSS and R) than one-sided ones. For instance, when using multiple regression models, the regression coefficients are usually tested for significant differences from 0 by two-sided test procedures. Although this might be appropriate when using regression techniques for *exploratory* purposes (e.g., variable selection), the risk of misalignment is increased in *confirmatory* settings. For instance, regression models are frequently used for comparing means between several groups (e.g., treatment groups) while adjusting for several co-variates (e.g., baseline measurements of the outcome variable of interest). In this case, the significance of the coefficient corresponding to the group indicator variable is of primary interest. However, a two-sided p-value provided by the statistical software does not correspond to a directional research hypothesis.

The *third* reason is that *guidelines* for good scientific (statistical) practice might (perhaps involuntarily) discourage researchers from using one-sided tests. For instance, according to the New England Journal of Medicine guidelines, all reported p-values should be two-sided, except when one-sided tests are required by study design, such as in non-inferiority trials (cf. https://www.nejm.org/author-center/new-manuscripts). The ICH statistical principles for clinical trials E9 state that researchers must provide a clear justification if they decide to use a one-sided test instead of a two-sided one (ICH E9, section 5.5). Thus, it seems that two-sided tests are the default and one-sided tests are the more difficult cases.

In light of THE TAUGHT ERROR and THE PRACTICAL ENFORCEMENT, it seems plausible to hypothesize that the misalignments we pointed out in section 3 are often *hidden* for clinicians who conduct studies. They apply what they have learned and what popular statistical tests, common statistical software packages, and guidelines seem to (involuntarily) encourage. Additionally, our review suggests that clinicians do not always pay special attention to the directionality of the theses involved. In the 30 papers we examined, the directions of all hypotheses involved are clearly stated in 3 only. Although these findings do not mean that the respective authors do not care about directionality – the papers might only cover a part of their considerations – they indicate that there is some lack of awareness concerning the alignment of research and statistical hypotheses. Yet, none of the reasons discussed so far *justify* the misalignment practice. For instance, using ANOVAs does not justify testing a non-directional statistical hypothesis to reason about a directional research hypothesis.

THE TAUGHT ERROR and THE PRACTICAL REINFORCEMENT are no accident. They are at least partially motivated by two *theoretical* reasons, namely THE AIM OF CONSERVATIVE TESTING and THE AIM OF RESEARCH OPEN-MINDEDNESS.

4.3 *The aim of conservative testing*

As Cho and Abe highlight (2013), there are theoretical reasons for using two-sided tests instead of one-sided tests. The first reason is the worry that one-sided tests are not *rigorous* enough. It has been argued that statistical significance can typically be obtained more easily by using one-sided tests instead of two-sided tests (cf. Kimmer 1957; Braver 1975; Altman 1991, p. 171; Howell 2007, pp. 98-100). In the one-sided setting, the p-value is smaller than for the corresponding two-sided test since the latter is the accumulation of the probabilities in both tails (cf. Fig. 2). So, as a safeguard against potential criticism regarding 'fishing for significance', researchers might prefer conducting two-sided tests. When it comes to the calculation of one-sided confidence intervals, there might be an additional issue with respect to interpretation, since one of the confidence interval limits is infinite.

Note, however, that the rigor achieved with two-sided tests with respect to the type I error rate comes at the price of losing power (i.e., with a reduced probability of detecting a statistically significant effect, given that the intervention is indeed efficacious). Moreover, using two-sided tests might also be sub-optimal due to the fact that usually, in contrast to the directional setting, uniformly most powerful tests can only be obtained under additional restrictions on the classes of test statistics.[10]

4.4 *The aim of research open-mindedness*

The second theoretical reason in favor of two-sided tests is that one-sided tests violate the aim of research open-mindedness. One-sided tests carry the risk of not catching effects in the direction opposite to the initial assumption (cf., e.g., Altman 1991, p. 171; Dubey 1991; Bland 2000, ch. 9.5; Ruxton and Neuhäuser 2010; Cho and Abe 2013). For instance, consider a study where the research hypothesis is that a new treatment yields an average quality of life score that is *larger* than under the current gold standard treatment. But the study data show that the average quality of life score was substantially *lower* in the group receiving the new treatment. Using a one-sided test, the statistical significance of this adverse result cannot be assessed, and this potentially important unexpected finding might thus not be adequately considered and reported.

But, importantly, THE AIM OF RESEARCH OPEN-MINDEDNESS does not *justify* misalignments. It simply recommends to use non-directional *research* hypotheses.

This leads us to a fundamental issue, namely the normative question of whether directional research hypotheses should be used in evidence-based medicine *at all*. For instance, according to the *principle of clinical equipoise*, clinicians should not predict what they will discover in clinical trials. They should not hypothesize that treatment A is better than treatment B or than a placebo, etc. (for more on this principle see,

[10] There are proposals for solving this problem, see, e.g., Lehmann and Romano 2005, pp. 229 ff.

e.g., Freedman 1987; Djulbegovic 2009). Accordingly, clinicians should only use non-directional research hypotheses (and only two-sided statistical tests). However, to discuss this principle or the normative question in general would lead us too far afield here. Either way, such normative principles do not justify the misalignments we are interested in. They only favor non-directional *research* hypotheses.

To sum up, misalignments between a research hypothesis and its corresponding statistical hypothesis seem to be favored by taught examples and common practices, which are connected to theoretical reasons for using two-sided tests. Yet, none of theses reasons justifies the misalignment practice. So, we proceed with considering whether this practice poses a threat to evidence-based medicine.

5 DO MISALIGNMENTS POSE A THREAT TO EVIDENCE-BASED MEDICINE?

Employing a misaligned statistical hypothesis is methodologically flawed. It also seems clear that one cannot gain proper evidence regarding the hypothesis that, say, some medication is *better* than a placebo by examining whether there is any difference whatsoever between the parameters of interest. A close alignment between a research hypothesis and its statistical hypothesis seems to be methodologically essential. As we have seen, there are also no justificatory reasons for the misalignments. One could thus reason that the misalignments pose a threat to evidence-based medicine simply by virtue of being a fundamental methodological flaw. But although we do not want to diminish the issue of such a flaw, we follow a different approach. We want to examine whether the misalignments are *benign* upon closer examination; they could be *counterbalanced* by other factors.

In what follows, we first discuss promising counterbalancing factors for DIRECTION MISALIGNMENT and then promising counterbalancing factors for MAGNITUDE MISALIGNMENT. Our discussions are not only guided by methodological concerns but also by considerations about medical research practice and considerations about informed decision making by clinicians, physicians, and patients. After all, one main function of hypothesis testing is to aid good clinical decision making.

5.1 *Counterbalances for direction misalignment*

DIRECTION MISALIGNMENT occurs when the direction of the research hypothesis and the direction of its 'corresponding' statistical hypothesis do not match. At first glance, the toolbox of statistics offers a counterbalancing instrument for the discussed cases where a direction research hypothesis is combined with a non-directional statistical hypothesis. According to the so-called *closure testing principle* (Marcus et al. 1976), if a two-sided p-value is statistically significant, one can subsequently conduct two

one-sided tests using the *same* level α. Thereby, one could obtain the p-value that is relevant for the directional research hypothesis.

But one should not celebrate too soon for at least *three* reasons. (i) Such a method does not counterbalance *all* cases: Post hoc one-sided tests are only legitimate if the two-sided test is statistically significant. If the two-sided test is non-significant, the procedure stops without any further testing. So, although the procedure allows for control of type I errors, the price to pay is a loss in power (cf. sect. 4.3). (ii) Methodologically, it does not seem appropriate to change one's statistical hypothesis *post hoc*; one should decide whether to test one-sided or two-sided *before* analyzing the data, as Altman urges (1991, p. 171) and as is required for preregistrations of trials and studies. (iii) The post hoc move conflicts with classical principles of statistical hypothesis testing, as outlined in section 2. For instance, if initially the effect is supposed to be positive, but, in fact, it turns out to be negative, the closure testing principle effectively leads to the rejection of the initial alternative hypothesis. This *contradicts* the principle that evidence should only be used to disconfirm the null hypothesis.

Regarding (i), one might object that in the case of non-significant p-values, the data are not conclusive *independent* of the research hypothesis' direction. It should thus be no issue that the closure testing principle does not cover all cases. Yet, this is not the case. A one-sided test result could be statistically significant although a two-sided test result based on the *same data* is not.

One might also object that (ii) is not an issue. Instead of characterizing the closure testing principle as a post hoc change of statistical hypotheses, one should construe it as a *new* testing situation. One re-employs the data to conduct two one-sided tests, where each corresponds to a different research hypothesis. If so, no alternative hypothesis is being rejected. However, this change of research hypotheses is not part of the closure testing principle. So, from the point of a research hypothesis *evaluation*, the apparent remedy involves a dubious ad hoc move.

Up to this point, we have only considered p-values. In practice, as recommended by guidelines for the conduct and reporting of clinical trials (e.g., ICH E9; Kirkwood and Sterne 2003, ch. 8), not only the test decision, but also the precise p-values and, most importantly, the *effect size(s)* (e.g., mean difference between treatment groups) or confidence intervals should be considered. The final conclusion regarding the research hypothesis is then the result of taking a number of different aspects of the evidence into account. Given the fact that the p-value and the respective test decision only play one role among others, one might wonder whether misalignments are more benign in such scenarios. Can the dubious ad hoc move criticized in (ii) be counterbalanced by considering effect sizes or the like? We think that it can be – at least to a substantial degree. If one considers the effect size, one might have a good justification for doing a

one-sided test that matches the effect size's direction and for modifying one's research hypothesis. If so, our objection that the closure testing principle is methodologically unmotivated is weakened. Yet, it is not rebutted, because changing one's research hypothesis is not part of a *confirmatory* testing setting.

Moreover, the other problematic issues are not mitigated; there is still no counterbalance for non-statistically significant cases (and thus a loss of power), and there is still a conflict with classical principles of statistical hypothesis testing. Analogous problems of one- vs. two-sided testing arise in the case of using confidence intervals instead of p-values, as well. For instance, using two-sided instead of one-sided confidence intervals also means a loss in power (in some sense).

In addition to these methodological considerations, we would like to note that in medical research practice the closure testing principle is not commonly used (at least in our experience). One might think that this does not have any bad *practical impact*. If one tests two-sided but observes a clearly positive effect, it might seem benign that no closure testing principle was used. It just seems obvious that the treatment has *some* positive effect. While this might be true when effect sizes are in fact considered, the issue in medical research practice is that effect sizes have been neglected and are still neglected (also for non-significant results). It is to be expected that there is a substantial amount of published findings where the conclusions were exclusively drawn based on the (two-sided) p-values.[11] And if the effect sizes are not specified in the study report, there is room for claiming that the test results (indirectly) 'support' a directional research hypothesis about an *improvement* with a p-value that is statistically significant due to *negative* effects. In light of various known issues with questionable research practices, we thus doubt that effect size considerations have substantially counterbalanced DIRECTION MISALIGNMENT in practice.

To sum up: The result of our analysis is that DIRECTION MISALIGNMENT cannot be fully counterbalanced by the closure testing principle (combined with effect size considerations), especially from a methodological, but also from an applied point of view. This casts doubt on findings of studies that feature this kind of misalignment.

5.2 *Counterbalances for magnitude misalignment*

MAGNITUDE MISALIGNMENT occurs when the statistical hypothesis does not reflect the *magnitude* of the effect stated in the research hypothesis. At first glance, the toolbox of statistics also offers a counterbalancing instrument for this kind of misalignment: As we have mentioned, in the planning phase of a clinical study, it is mandatory to calculate the minimum number of subjects that have to be included to detect a certain

[11] Statistical significance still seems to play a dominant role in medical research insofar as that statistically insignificant results are less frequently published (cf., e.g., Altman 1991, ch. 8.5.4, 15.5.2; Dwan et al. 2008) – this phenomenon is called 'publication bias'.

effect: the minimal clinically important difference (MCID), with the statistical power (1 - β) usually set to 80% or 90%. So, if the research hypothesis is merely concerned with a clinically important difference and the sample has the appropriate size, a MAGNITUDE MISALIGNMENT might be be regarded as benign.

Yet, there are also at least *three* worries regarding this counterbalancing mean. (i) Arguably, not all cases are covered by the MCID consideration; not all cases are just about noticeable effects. Hypotheses about 'significant improvements' or 'substantial improvements' go beyond a *minimal* clinically important difference.

(ii) A crucial question is whether MCID satisfyingly captures the use of *qualitatively evaluative* terms like 'better' that are frequently used in research hypotheses. MCID is aimed at accommodating for qualitative changes, but at least conceptually and in practice MCID and qualitative differences could come apart. MCID is calculated based on estimates. Typically, these are not based on patient-centered considerations, but drawn from previous studies, subject-matter experts, or a given standard. If available, meta-analyses may serve as a convenient means of obtaining estimates not only of the effects, but also of nuisance parameters (e.g., variances). But for a particular group of patients, fulfilling MCID might not lead to a qualitative change. Consider a case where a three point difference on some quality-of-life scale is used for determining the required sample size because this difference was the MCID in a comparable study. This does not ensure that the three point difference is a qualitative difference for the patients in question. In other words, MCID might be necessary but not sufficient for capturing qualitative improvements. For capturing the latter, a patient-centered approach is to be preferred, i.e., one would need to consider in more detail the particular patients' needs and expectations. These considerations might not only apply to assessment of efficacy, but also to safety aspects: Some patients might be willing to tolerate side effects the medical doctors would call severe, and *vice versa*. In other words, there might be differences in the harm-benefit assessments. This aspect is of particular importance for patients with diseases that have potentially life-threatening consequences.

(iii) Not specifying the desired effect size *in* the statistical hypothesis leads to the (potential) *lack of crucial information*. On the one hand, clinicians might disagree about the MCID estimates. If such a disagreement is hidden in the sample size planning, disagreements might easily be overlooked. On the other hand, not incorporating the MCID estimate into the statistical hypotheses leads to a more difficult comparison of two treatments. This loss of information can have bad practical consequences. A physician or patient might decide for a risky treatment without realizing that the potential benefits are minuscule from the patient's point of view, etc. If the estimates were reflected in the hypothesis, differences could be visible to everyone, including patients who want to make an empirically informed decision about their treatments.

So far, we neglected effect size considerations. What if we add them? Based on the effect size, one could determine whether the MCID threshold has been reached. Moreover, the variability of the effect sizes (in comparison to the risks and disadvantages of the treatment) is also valuable information for clinician, physicians, and especially patients for informed decision making. Providing information about effect sizes pointing in opposite directions, or about their corresponding variability plays an essential role when it comes to formulating evidence-based recommendations (cf., e.g., Guyatt et al. 2008). Although we agree that considering effect sizes softens the blow of MAGNITUDE MISALIGNMENT, it does not solve the issue of the potential lack of crucial information. And that this issue is non-benign for medical *practice* should be evident from the fact that systematic reviews indicate that sample size calculation issues are still poorly reported in a considerable number of publications (cf., e.g., Bariani et al. 2015). In our review, we also examined whether the required sample sizes, in particular with regard to the hypothesized effect size, are provided. Only 3 papers report the MCID that was used for determining the required sample size. Even in these rare cases, however, the MCID estimates were inferred from previous studies, and not based on patient-centered considerations. The cited reviews also give reason to believe that a substantial amount of studies might not be adequately powered. In addition, and as discussed before, although effect sizes are strongly encouraged or even required to be stated in study reports and publication, the interpretations and conclusions of medical studies often focus on the (non-)significance of the testing results instead of effect size considerations.

To sump up: The result of our analysis is that MAGNITUDE MISALIGNMENT cannot be fully counterbalanced by MCID considerations or sample size planning. On the one hand, not all cases of MAGNITUDE MISALIGNMENT are covered. On the other hand, both methodologically and for informed clinical decision making, there is a danger of not capturing the qualitative differences in the research hypotheses and of not displaying important information. Our considerations cast doubt on findings of medical studies that feature MAGNITUDE MISALIGNMENT.

All in all, we conclude that misalignments pose a non-negligible threat to evidence-based medicine. Our results especially cast doubt on findings of medical studies that involve both DIRECTION MISALIGNMENT and MAGNITUDE MISALIGNMENT.

6 CONCLUDING REMARKS: CONSEQUENCES FOR CLINICIANS' THEORETICAL AND PRACTICAL REASONING

In this paper, we have identified four forms of misalignment between research hypotheses and their statistical hypotheses, namely DIRECTION MISALIGNMENT, MAGNITUDE MISALIGNMENT, OPERATIONALIZATION MISALIGNMENT, and TESTING MISALIGNMENT. We have focused on the first two, and we have identified two main reasons for the

occurrence of such misalignments, namely that they are favored by common practices (e.g., popular statistical tests, involuntarily misleading guidelines, or taught errors), which are based on theoretical considerations in favor of two-sided tests. Our upshot regarding whether such misalignments pose a threat to evidence-based medicine is mixed. Some cases of misalignments can be partially counterbalanced and rendered more benign. Yet, not all misalignments can be counterbalanced and the counterbalancing instruments have severe disadvantages: They either lead to methodological inadequacy (in addition to the misalignment) or involve a (potential) lack of information that could be crucial for medical research, informed clinical decision making, and patient-centered considerations. In addition, using two-sided tests for testing *directional* hypotheses results in a loss of power. This means that potentially beneficial treatments might be missed. Given the dominance of statistical significance in the interpretation of medical studies, a non-significant result will most likely lead to stopping the marketing application of the particular treatment under consideration, and any further investigations might be considered unnecessary. Apart from that, flaws in calculating the required sample size can still be found in a considerable number of studies, which means that MAGNITUDE ALIGNMENT might be a more serious problem than it seems at first glance. This further influences statistical power and, thus, might pose an additional threat to EBM in practice.

In view of these issues that misalignments face in addition to the flaw of a misalignment on its own, we strongly recommend avoiding them. We propose *five* remedies that can be applied to clinicians' theoretical and practical thinking.

The *first* remedy is to urge clinicians to pay more attention to the relation between their research hypotheses and the chosen statistical hypotheses – in teaching, in regulatory guidelines, and in practice. Using two-sided tests should involve a non- or bi-directional research hypothesis. If one wants to use directional research hypotheses, one needs to test one-sided. If one does not want to use one-sided tests, one should not specify directional research hypotheses.

The *second* remedy is to demand the preregistration of trials and studies. This might also draw more attention to the adequacy of the chosen statistical hypothesis, and it hampers what one could call 'hypothesis hacking': *after* statistical testing, one specifies a research hypothesis that matches the testing results.[12] In our experience, this does not rarely occur.

The *third* remedy is concerned with MAGNITUDE MISALIGNMENT. We suggest incorporating the MCID estimate into the statistical hypotheses by using an alternative hypothesis that is related to the effect size, e.g., $H_1 : \mu_T - \mu_P \geq \delta$, or a point hypothesis (cf. sect. 3). The results of such testings would be also a better guide for patients.

[12] We owe this suggestion to a researcher in the audience of our talk at the 8th Philosophy of Medicine Roundtable.

Having good reason to believe that TMS has an effect of a size that is apt for the patient's needs is more informative than just having good reason to believe that TMS has some unspecified effect. However, it might be more difficult to explain such hypotheses to practitioners (since the alternative is restricted to one single value).

The third remedy might lead to more one-sided testing situations. Our *fourth* remedy is indeed to re-consider using one-sided tests. Two-sided tests are the right choice for many cases but presumably not for all. An important benefit of one-sided tests is that the loss of power that two-sided tests face would be diminished. Although their reasons for employing directional hypotheses (and one-side tests) must be clearly stated and justified, researchers should not be afraid of using one-sided tests. They should also keep in mind that sacrificing power for little reason might also be unethical because potentially promising new treatment approaches are likely to be overlooked. However, we think that two important conditions for using one-sided tests are as follows: (i) The test needs to be *rigorous* enough. As mentioned in section 4.3, the use of one-sided tests is criticized for its 'fishing for significance' potential. But one could simply test with half of α to obtain appropriate rigor (cf. ICH E9, sect. 5.5). (ii) The test needs to fit the confirmatory testing setting. For instance, Ruxton and Neuhäuser (2010) propose a decision criterion for using a one-sided test: If a significant effect in the opposite direction yields the same consequences as no effect, a one-sided test can be used. This could be sensible in many frequently encountered situations in medical research. For instance, when evaluating safety outcomes, it is important to detect effects in both directions. However, when the main interest is on efficacy, the consequences of a non-significant one-sided test result might be the same, regardless whether the effect is 0 or opposite to the initial assumption: the marketing application of the treatment would be stopped. Such a decision criterion could also used be in preregistration to justify the use of one-sided tests.

Last but not least, we suggest as a *fifth* remedy that clinicians should better distinguish *confirmatory* statistical settings from *exploratory* ones. Statistical testing is not limited to disconfirming hypotheses. It can be also used for exploring one's data set. Moreover, in some cases, clinicians can also rely more on *descriptive* statistics. For instance, when it comes to safety and tolerability, guidelines suggest "[...] applying descriptive statistical methods to the data, supplemented by calculation of confidence intervals wherever this aids interpretation" (ICH E9, section 6.4). Yet, descriptive statistics cannot replace inferential statistics. Despite all the potential problems associated with hypothesis testing, its merits must be taken into account (e.g., accommodating by-chance variability).

In our opinion, all stakeholders in medical research should move the 'scientific paradigm' toward a multi-factorial decision model, with an eye on the interpretation of the effect size and its clinical relevance. This does not mean that hypothesis testing

should be banned from research. *p*-values or confidence intervals and effect size estimation are two aspects of the available evidence, which are complementary to each other. With respect to patient-physician communication and involvement of patients into clinical decisions, placing more emphasis on effect sizes is crucial: A SCI patient is primarily interested in the improvement of motor function that can be expected 'on average' for a particular treatment. Her or his decision will be mostly based on this aspect – and so should the treating physician's decision be. Clinicians should thus be encouraged to provide detailed and reliable information about how one therapy compares to another with respect to the clinical outcome (e.g., improvement in gait performance, bladder function, quality of life). Especially in rare disease settings such as SCI, however, the number of subjects are usually small, which leads to a considerable impact of by-chance variation. Therefore, it is especially important to consider the variability of the effect size. This can be communicated and explained to the patients by referring to analogous paradigmatic examples from daily life (e.g., uncertainty in the results of opinion polls). Moreover, especially in rare diseases, case reports or series represent a substantial amount of the evidence that is used for clinical decision-making. These data might also facilitate emphasizing the inter-individual differences in the communication with patients. Case series provide useful additional evidence, complementing the results from clinical trials, which naturally yield summary statements about the treatment efficacy 'on average' rather than being focused on evaluations at the single-subject level.

We have deliberately considered only one problem, namely misalignments of research and statistical hypotheses, within the broad topic of medical data analysis. It is needless to say that the paper thus could not serve as a critical appraisal of the methodological and statistical quality of evidence in SCI or medical research in general. Moreover, we did not take into account that often not a single research hypothesis is considered. In such cases, the corresponding statistical hypotheses would be more complex. We have also only considered a frequentist framework. In future research, it is worth examining if analogous problems arise for Bayesian accounts.

REFERENCES

Altman, D. (1991). *Statistic for Medical Research*. Chapan & Hall, first edition.

Bariani, G., de Celis Ferrari, A., Precivale, M., Arai, R., Saad, E., and Riechelmann, R. (2015). Sample Size Calculation in Oncology Trials: Quality of Reporting and Implications for Clinical Cancer Research. *American Journal for Clinical Oncology*, 38(6):570–574.

Benjamin, D. et al. (2017). Redefine Statistical Significance. https://psyarxiv.com/mky9j/.

Berger, J. and Sellke, T. (1987). Testing a Point Null Hypothesis: The Irreconcilability of P Values and Evidence. *Journal of the American Statistical Association*, 82(397):62–71.

Bigirumurame, T. and Kasim, A. S. (2017). Can Testing Clinical Significance Reduce False Positive Rates in Randomized Controlled Trials? A Snap Review. *BMC Research Notes*, 10:775.

Bland, M. (2000). *Introduction to Medical Statistics*. Oxford University Press, third edition.

Braver, S. (1975). On Splitting the Tails Unequally: A New Perspective on One- versus Two-Tailed Tests. *Educational and Psychological Measurement*, 32:283–301.

Casella, G. and Berger, R. (1987). Reconcilling Bayesian and Frequentist Evidence in the One-Sided Testing Problem. *Journal of the American Statistical Association*, 82(397):106–111.

Cho, H.-C. and Abe, S. (2013). Is Two-Tailed Testing for Directional Research Hypotheses Tests Legitimate? *Journal of Business Research*, 66:1261–1266.

Cohen, J. (1994). The Earth is Round ($P < .05$). *American Psychologist*, 49(12):997–1003.

Colquhoun, D. (2014). An Investigation of the False Discovery Rate and the Misinterpretation of *P*-Values. *Royal Society Open Science*, 1: 140216:1–16.

Cumming, G. (2012). *Understanding The New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis*. Routledge.

Davis, R. (2014). *Reproducibility Project: Cancer Biology*. https://elifesciences.org/collections/9b1e83d1/reproducibility-project-cancer-biology.

Derakhshanrad, N., Vosoughi, F., Yekaninejad, M., Moshayedi, P., and Saberi, H. (2015). Functional Impact of Multidisciplinary Outpatient Program on Patients with Chronic Complete Spinal Cord Injury. *Spinal Cord*, 53:850–865.

Djulbegovic, B. (2009). The Paradox of Equipoise: The Principle that Drives and Limits Therapeutic Discoveries in Clinical Research. *Cancer Control*, 16(4):342–347.

Dubey, S. (1991). Some Thoughts on the One-sided and Two-sided Tests. *Journal of Biopharmaceutical Statistics*, 1(1):139–150.

Dwan, K., Altman, D., Arnaiz, J. A., Bloom, J., Chan, A.-W., Cronin, E., Decullier, E., Easterbrook, P., Von Elm, E., Gamble, C., Ghersi, D., Ioannidis, J., Simesa, J., and Williamson, P. R. (2008). Systematic Review of the Empirical Evidence of Study Publication Bias and Outcome Reporting Bias. *PLoS ONE*, 3(8):e3081.

Everitt, B. (2006). *Medical Statistics from A to Z – A Guide for Clinicians and Medical Students*. Cambridge University Press, second edition.

Field, A. (2000). *Discovering Statistics Using SPSS for Windows*. Sage Publications, London.

Freedman, B. (1987). Equipoise and the Ethics of Clinical Research. *The New England Journal of Medicine*, 317(3):141–145.

Gigerenzer, G. (2004). Mindless Statistics. *The Journal of Socio-Economics*, 33:587–606.

Guyatt, G. H., Oxman, A. D., Vist, G. E., Kunz, R., Falck-Ytter, Y., Alonso-Coello, P., and Schünemann, H. J. (2008). GRADE: An Emerging Consensus on Rating Quality of Evidence and Strength of Recommendations. *BMJ*, 336(7650):924–926.

Hacking, I. (2001). *An Introduction to Probability and Inductive Logic*. Cambridge University Press.

Howell, D. (2007). *Statistical Methods for Psychology. 6th Edition*. Thomson Wadsworth.

International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (1998). ICH Harmonized Tripartite Guideline: Statistical Principles for Clinical Trials E9.

Ioannidis, J. (2005). Why Most Published Research Findings Are False. *PLOS Medicine*, 2.

Jeong, J. and Yoo, W. (2015). Effects of Air Stacking on Pulmonary Function and Peak Cough Flow in Patients with Cervical Spinal Cord Injury. *Journal of Physical Therapy Science*, 27(6):1951–1952.

Kaiser, H. (1960). Directional Statistical Decisions. *Psychological Review*, 67:160–167.

Kimmer, H. (1957). Three Criteria for the Use of One-Tailed Tests. *Psychological Bulletin*, 54:351–353.

Kirkwood, B. and Sterne, J. (2003). *Essential Medical Statistics*. Blackwell, second edition.

Lecoutre, B. and Poitevineau, J. (2014). *The Significance Test Controversy Revisited. The Fiducial Bayesian Alternative*. Springer Verlag.

Lehmann, E. and Romano, J. (2005). *Testing Statistical Hypotheses*. Springer, New York, third edition.

Machin, D., Campbell, M., and Walters, S. (2007). *Medical Statistics. A Textbook for the Health Sciences*. John Wiley & Sons, fourth edition.

Marcus, R., Peritz, E., and Gabriel, K. (1976). On Closed Testing Procedures with Special Reference to Ordered Analysis of Variance. *Biometrika*, 63(3):655–660.

Meehl, P. (1978). Theoretical Risks and Tabular Asterisks: Sir Karl, Sir Ronald, and The Slow Progress of Soft Psychology. *Journal of Consulting and Clinical Psychology*, 46:806–834.

Nickerson, R. (2000). Null Hypothesis Significance Testing: A Review of an Old and Continuing Controversy. *Psychological Methods*, 5(2):241–301.

Ruxton, G. D. and Neuhäuser, M. (2010). When Should We Use One-Tailed Hypothesis Testing? *Methods in Ecology and Evolution*, 1:114–117.

Zimmermann, G., Bolter, L.-M., Sluka, R., Höller, Y., Bathke, A. C., Thomschewski, A., Leis, S., Lattanzi, S., Brigo, F., and Trinka, E. (2019). Sample Sizes and Statistica Methods in Interventional Studies on Individuals with Spinal Cord Injury: A Systematic Review. *Journal of Evidence-Based Medicine*, published online ahead of print.