On the use of one-sided statistical tests in biomedical research

Ricardo Murphy

University of Oslo, Department of Molecular Medicine,

Physiology Section, Pb. 1103, Blindern, 0317 Oslo, Norway

Short title: One-sided tests in biomedical research

Ricardo Murphy,

University of Oslo,

Department of Molecular Medicine,

Physiology Section,

Pb. 1103, Blindern,

0317 Oslo,

Norway

Tel: +47 95333157

Fax: +47 22851249

Email: ricardo.murphy@medisin.uio.no

SUMMARY

There is a tendency to automatically use two-sided tests to assess the statistical significance of experimental results. Yet if a theory predicts the direction of an experimental outcome, or if for some practical (e.g. clinical) reason an outcome in that direction is the only one of interest, then it makes sense to use a one-sided test. The use of a two-sided test in these situations will lead to too many false negatives. Consequently treatment effects that corroborate a theory or that are of practical importance may be missed. This problem becomes particularly acute in the case of borderline results. Following a nonsignificant one-sided test, the possibility of an effect in the direction opposite to that predicted or required can be assessed in an exploratory fashion by computing the odds in favour of such an effect. Anyone is then at liberty to pursue this possibility as they see fit. The question of whether to use a one-sided or two-sided statistical test should always be decided on logical grounds not statistical ones, and suspicions regarding the motives of the investigator(s) should be disregarded. On the other hand, this choice can be avoided altogether by assuming that a treatment always has some effect (however small) and then computing the strength of the evidence in favour of the observed or predicted/required effect (i.e. 1–*P*, where *P* is the one-sided significance level of the test). With this approach one-sided and two-sided tests yield identical results, and so there is effectively only one type of test.

Keywords*:* Clinical trials; drug testing; hypothesis testing; odds; one-sided tests; philosophy of science; scientific method; statistical significance; statistical tests.

INTRODUCTION

In one widely used approach to statistical testing, an experimental result is declared "significant" or "not significant" according to whether the reported level of significance (P) of the test is less than or greater than some 'critical' value, α (the probability of a Type I Error or false positive). If $P < \alpha$ the null hypothesis (H_0) of no treatment effect is rejected in favour of some alternative hypothesis (H_1); otherwise one concludes that there are insufficient grounds for rejecting H_0 (with, however, no implication that H_0 is true). Hurlbert & Lombardi[1] and others cited by them have criticized this approach, and proposed instead the use of P values on a continuous scale and a more nuanced description of the strength of the evidence against H_0. Walpole *et al.*[2] call this the "P-value" approach, and it forms the core of the "neoFisherian" paradigm of statistical inference discussed by Hurlbert & Lombardi[1]. Today there is no excuse for not reporting P values – rather than just $P < \alpha$ or $P > \alpha$ - regardless of whether they are significant or not. Such values do indeed give a sense of the strength of the evidence against H_0. Yet to some it may seem natural to want a maximum value (α) of P to aim for, such that for $P < \alpha$ one feels comfortable with rejecting H_0. And at least in my fields of research (animal and plant physiology) $\alpha = 0.05$ seems to be the *de facto* 'industry standard'. However, it can be argued[3,4] that α should be set *post hoc* so as to minimize the quantity $(\alpha+\beta)/2$ (or a cost-weighted version thereof), where β is the probability of a Type II Error (false negative). On the other hand, if α is to be set *a priori* at the planning stage, with a view to estimating the required sample size (n), then I would argue that α and β should be set to the same value (e.g. $\alpha = \beta = 0.1$) unless different values can be justified on the grounds of different costs of Type I and Type II Errors. As pointed out by Baker and Mudge[3], there are situations in which a Type II Error may be much more costly than a Type I

Error. But however α is set, there is the problem of borderline results, i.e. nonsignificant results with *P* values close to α[5]. This is especially so if α is set to what might be considered a 'low' value such as 0.05 or 0.01. To an extent, recognizing the existence of borderline results amounts to a discretized version of the *P*-value/neoFisherian paradigm, with *P* values classified as "significant", "borderline" and "nonsignificant". Indeed these terms might just as well be replaced with, for example, "satisfactory", "suggestive" and "unsatisfactory", and the evidence against H_0 classified as "strong", "moderate" and "weak". But note that α is retained. Of course what is to be considered 'close' to α is arbitrary, but then so is α. In their discussion of clinical trials, Hackshaw & Kirkwood[5] assumed $\alpha = 0.05$ and considered *P* values in the range 0.05-0.10 to be borderline. This seems reasonable to me, and it brings me to the main topic of the present paper. Statistical tests of significance can generally be one- or two-tailed (also called one- or two-sided), i.e. they can generally use one or both tails of the sampling distribution of the test statistic. But for a symmetric distribution such as the *t* distribution, a significance level of *P* for a two-sided test becomes *P*/2 for a one-sided test (< *P* for an asymmetric distribution). Thus the borderline result $P = 0.08$ obtained with a two-sided *t* test becomes the significant result $P = 0.04$ with a one-sided *t* test. While the concept of a borderline result has lessened the impact of the choice of test (one- or two-sided) there is still a dramatic effect on how we view and present the result. So when – if ever – should a one-sided test of significance be used?

LOGIC

In attempting to answer this question I take as my starting point the Popperian view of the scientific method (or at least the logic of that method). One has some theory, invented to solve some problem or explain some observation. Here the term "theory" implies an

unjustified, unjustifiable conjecture about the causes of some phenomenon (or set of phenomena) as envisaged by Popper[6,7]. One then deduces potentially falsifiable predictions from this conjecture, together with certain other premises (Popper's "background knowledge" and "initial conditions") and subjects these predictions to experimental tests. The analysis of such experiments often involves statistical tests. In statistics a prediction deduced from a theory is called a 'hypothesis' (the "alternative hypothesis", H_1). This is an example of a statistical hypothesis, and it should not be confused with a scientific hypothesis, which for the purposes of this paper is the same thing as a theory. A scientific hypothesis/theory offers a causal explanation of some phenomenon; a statistical hypothesis is a prediction deduced from that conjectured explanation. To avoid confusion I will generally avoid the use of the term "scientific hypothesis", preferring instead the term "theory', and reserving the term "hypothesis" for statistical hypotheses.

If a prediction turns out to be false then, provided it is the conclusion of a valid argument, we can conclude that one or more of the premises (possibly the theory) is mistaken. On the other hand, if it turns out to be true we can conclude nothing, because an argument's validity does not exclude the possibility that the premises are false but the prediction is true. So the question that needs to be answered is this: is the prediction true or false? But as we all know in the real world only a probabilistic answer can be given to this question because of uncontrolled, unaccountable and apparently random variation (at least we generally model it as random). Thus an experiment may produce a 'positive' outcome (i.e. one that confirms the prediction) or a 'negative' outcome (one that contradicts the prediction). But in either case there is a possibility that the observed outcome occurred because of random variation. One way to address this problem is with a test of statistical significance. That is, for the nonzero predictions $\mu \neq 0$, $\mu > 0$ and $\mu < 0$, we negate the prediction to generate its logical complement ($\mu = 0$, $\mu \leq 0$ and $\mu \geq 0$, respectively). This is the so-called statistical null

'hypothesis', H_0 (also not to be confused with a scientific hypothesis). [Zero predictions, or more generally precise predictions ($\mu = \mu_0$, where μ_0 is some specified constant) are outside the scope of this paper.] Here μ is the true mean value of the sampling distribution of some parameter. For the purposes of this article we will take that parameter to be a difference in populations means resulting from some treatment effect. Then, under the assumption that H_0 is correct, we estimate the probability (P) of obtaining a value of the test statistic as large or larger than the one actually observed. If we choose to accept our prediction as true, then P is the probability that we are making a mistake (a Type I Error or false positive). We may also call P the rate of false positives or the Type I Error rate, with the understanding that the word "rate" implies the expected relative frequency if the study in question were to be repeated many times under identical conditions. Interpreted in this way, the Type I Error rate (= P) is a hypothetical quantity, although its maximum value (α) is real enough. It should be emphasized that P, which is also known as the "significance level" of the test, is *not* the probability that H_0 is true (e.g. [1]). Rather P should be interpreted as the strength of the evidence against H_0, with low values (say $P < \alpha$) leading us to reject it in favour of H_1, while accepting the risk (P) of committing a Type I Error.

Because of the logic of significance testing, it is generally said that we test H_0 rather than H_1 ("null hypothesis significance testing"). But this is simply because we require a statistical hypothesis specified by an equality in order to compute P (since the location of the sampling distribution of the test statistic must be specified in order to calculate these values). And under H_0 that hypothesis is $\mu = 0$ (specifying an upper or lower bound in the one-sided case). Yet it should be remembered that from a philosophical (i.e. Popperian) point of view it is the prediction (H_1) of a scientific theory that is under scrutiny, not H_0 (which, after all, is merely the negation of H_1; it is not a genuine hypothesis).

For a two-sided prediction of the form $\mu \neq 0$ it makes sense to consider the test statistic without regard to sign, so that *P* is apportioned to the two tails of its sampling distribution (equally for a symmetric distribution such as the *t* distribution). This is simply because the negated prediction (H_0) is $\mu = 0$, and so any nonzero outcome – in either direction (i.e. $\bar{x} < 0$ or $\bar{x} > 0$, where $\bar{x}$ is the observed mean treatment effect) - can be considered positive, confirming the prediction (H_1). [In which case it can be argued that a two-sided prediction is not scientific in the Popperian sense because measurements of a sufficiently high precision and/or a sufficiently large *n* will always yield a statistically significant nonzero outcome. This points to the need to specify a minimum effect size that is worth knowing about.] But what if H_1 is one-sided, i.e. of the form $\mu > 0$ (or $\mu < 0$)? In this situation H_0 is $\mu \leq 0$ (or $\mu \geq 0$), and so any experimental outcome $\bar{x} \leq 0$ (or $\bar{x} \geq 0$) must be considered negative, i.e. contradicting H_1. *It is my contention that it does not make sense to count negative results when computing the rate of false positives.* Therefore in common with Ludbrook[8,9], Peace[10,11], and Curran-Everett[12], I believe that only the side of the *t* distribution consistent with the direction of the predicted effect should be considered when computing *P*. Or in other words a one-sided prediction should be assessed with a one-sided statistical test. However I differ with Ludbrook in some respects, as will become clear. Also I do agree with Hurlbert & Lombardi[13] – although for different reasons - that it is hard to see a role for two-sided tests with unequal tail probabilities. Either your theory allows you to predict the direction of a treatment effect (in which case you do a one-sided test with tail probability *P*) or it doesn't (in which case you do a two-sided test with tail probabilities < *P*; *P*/2 for a symmetric distribution). Note also that for the one-sided test *P* actually gives an upper bound on the Type I Error rate. This follows from the fact that *P* is determined under the assumption $\mu = 0$. But if in reality $\mu < 0$ when the prediction is $\mu > 0$ (or *vice versa*), then the true Type I Error

rate will be less than *P*. Lombardi & Hurlbert[14] make the same point with regard to α. It therefore follows that:

> "…type I errors are no more likely to occur for one-sided tests at any significance level α than for two-sided tests with the same significance level α…a one-sided test is not a weaker standard than a two-sided test…"[15].

The use of a two-sided test regardless of whether the prediction being tested is one-sided or two-sided implies that any outcome $\bar{x} \neq 0$ should be considered positive. But I insist that it is a nonsense to define a positive outcome in such a way that it is capable of contradicting the prediction of the theory under test. Indeed this practice seems to me merely an example of "data dredging"[16], i.e. performing statistical tests in the hope that something significant will show up, rather than being guided by theoretical predictions. Moreover it leads to an overestimate of the Type I Error rate for one-sided predictions. For a two-sided test significant at a level *P* amounts to two one-sided tests, one of which is significant at a level *P*/2 (assuming symmetry for simplicity)[17]. Assuming $\mu = 0$, either of these one-sided tests may be significant by chance, but in the long run only one half of such tests will lead to acceptance of a one-sided prediction. Therefore the true Type I Error rate for that prediction is *P*/2, not *P*. We can also approach this problem from the point of view of multiple testing. In testing a two-sided prediction we effectively test the same prediction twice, once in the right hand tail of the test-statistic distribution and once in the left hand tail (either will do). Therefore a Bonferroni adjustment for multiple testing is required[18]: the critical tail probability is $\alpha/2$, not α. But if we test two theories which make opposing one-sided predictions ($\mu > 0$ and $\mu < 0$) each with its own H_0 ($\mu \leq 0$ and $\mu \geq 0$), then each prediction is tested only once (i.e. in the appropriate tail). Therefore, according to the arguments of Perneger[19], a Bonferroni adjustment is not required: each critical tail probability is α. Exactly

this approach is described by Schuirmann[20] in the context of drug bioequivalence testing. Schuirmann notes that his method is equivalent to the use of a two-sided $(1-2\alpha)\times 100\%$ confidence interval (note 2α, not α) as advocated by Westlake[21] and Peace[10]. Of course in bioequivalence testing the limits placed on the difference in drug efficacies are not theoretical predictions but *requirements*. Drug testing is an example of applied research.

APPLIED RESEARCH

A number of other authors have emphasized that a statistical hypothesis does not correspond to a scientific hypothesis (theory) but to one of its predictions (e.g. [1]). But it should be recognized that in practice at least one other scenario arises. Thus, H_1 may represent a decision criterion. i.e. a requirement that must be fulfilled for some practical purpose (it may or may not also be a theoretical prediction).. Lombardi & Hurlbert[14] acknowledge that one-tailed tests can be justified in this case, which might be regarded as 'applied' rather than 'pure' research. And in this situation, justification for a one-sided test would be on the basis of what is of interest, as demanded by Woodman[22] and Ruxton & Neuhaeuser[23]. A classic case is the placebo-controlled drug trial:

> "Since a drug could *never* be approved when it is less effective than placebo, there is no chance for an error to occur in the approval process. Thus the probability…should *only* be evaluated when the drug effect is greater than placebo. Obviously this leads to one-tailed alternative hypotheses and the eventual computation of a one-tailed *P*-value."[24] (see also [10,11,15,25-27]).

But one-sided testing has applications in clinical trials in addition to placebo-controlled drug trials. Thus, as already noted, two one-sided tests can be used to assess bioequivalence[20]. Moreover, one-sided testing can be extended to a two-stage procedure which allows testing

for improved efficacy in addition to bioequivalence, and without the need for a Bonferroni adjustment[18]. However, while the latter procedure controls the family-wise Type I Error rate, the false discovery rate is expected to increase[28]. This problem can be alleviated by using a lower value of α for the second stage superiority test[29]. Of course there will be occasions in applied research where two-sided tests are appropriate. For example, when testing drug combinations against the individual drugs, one may be interested in the possibility that the drugs are antagonistic[27]. There will also be times when both types of test are required. Thus Goeman, Aldo and Stijnena[30] have described a three-sided test procedure in which one-sided tests are used to assess inferiority and superiority while a two-sided test is used to assess bioequivalence, again without the need for Bonferroni adjustment.

BIGGER POWER OR SMALLER *N*

A number of objections have been raised to the use of one-sided tests. Curran-Everett[12] voices what seems to be a general concern (although, like Matthews[31], he is not opposed to the use of one-sided tests in appropriate circumstances): "A one-sided *P* value makes it easier to reject the corresponding null hypothesis". In other words, for a given sample size, a one-sided test is more powerful than the two-sided alternative. That is, there is a lower probability of Type II Error. This point can be expressed in yet another way. When planning a study the maximum acceptable ('critical') Type I and Type II Error rates (α and β, respectively) must be considered. For a given choice of α and β a one-tailed comparison requires a smaller n than the two-tailed variety. Why this should be seen as arguing against the use of one-sided tests is a mystery for two reasons. Firstly, it is the Type I and Type II Error rates that have primacy, *not* the sample size. That this is so can be seen from the fact that when planning a study one *first* specifies α and β and *then* estimates the sample size required to achieve the

desired value of β (α is independent of n). Secondly, for a given n, maximizing the probability of rejecting H_0 when it is false (i.e. the power, $1-\beta$) is obviously *always* desirable. In the online supplement (section S1) I suggest that the solution to this mystery is nothing more substantial than irrelevant suspicions about the motives of investigators. It is my contention that such suspicions should be disregarded. And whatever the power of a statistical procedure, that procedure must make sense. Counting negative results as positive does not make sense. Moreover, in the context of pure research, I would argue that the prediction under test *should* be given the best possible chance of confirmation because any theory is better than no theory. If the theory is wrong it will eventually fail in competition with other theories. And in the context of placebo-controlled drug trials, a one-sided test is the more ethical choice because it reduces the number of subjects required to achieve the desired power[31].

ZERO POWER

Another widely expressed objection is that for a one-sided test, the power to detect effects in the direction opposite to that predicted or required is necessarily zero. Ruxton & Neuhaeuser[23] express it well:

> "We would expect them to have a convincing explanation for why they would treat a large observed difference in the unexpected direction no differently from a difference in the expected direction that was not strong enough to justify rejection of the null hypothesis."

The implication is that one must demonstrate that such effects are impossible, or extremely unlikely, or simply of no interest[8,9,14,22,23,32-34]. As already indicated the last of these does indeed obtain in applied research where, presumably, the direction of interest (i.e. the

required direction) will usually be clear. The problem of zero power in one direction arises mainly in pure, theory-driven research, where the predicted direction of a treatment effect is deduced from the theory. Following Goldfried[35], Lombardi & Hurlbert[14] address this issue by considering various scenarios for further testing and/or reporting of *P* values following a nonsignificant result for a one-sided test. These are: (1) report the *P* value and argue that the negative result is unimportant; (2) do a two-sided test and report its *P* value; (3) do a one-sided test in the opposite direction and report its *P* value; (4) repeat the study and then do either (2) or (3). They conclude that none of these options is acceptable and recommend that two-sided tests should almost always be used in the "collective interest". Actually it seems to me that the collective interest is best served by making one's raw data publically available *via* the internet so that others may analyze them as they see fit. The role of repeating studies is commented on in supplementary section S2.

Goldfried[35] recognized that other scenarios might be possible. I suggest there is at least one: that one-tailed tests should be treated in exactly the same way as two-tailed tests. Thus if a one-tailed test returns a *P* value of 0.99 then that is the value that should be reported and the test result declared "not significant". This is option (1) *sans* the accompanying apologetic arguments. As for the 'problem' of zero power for detecting a negative effect, this is a pseudo-problem. Citing Oakes[36], Lombardi & Hurlbert[14] note that following a two-sided test, and under the "principled assumption" that there *was* a treatment effect, one's confidence in the direction of that effect can be expressed as the quantity $(1-P/2)$ (or the quantity $(1-P/2)/(P/2)$, which they call an "odds ratio"; however a better term is "odds"[37]). This seems reasonable. Thus as the positive tail probability $P/2 \to 0$ (or, equivalently, $(1-P/2) \to 1$) we become increasingly confident that the true treatment effect was in the positive direction. But that tail probability is just the level of significance that would be reported by a one-sided test. So what Lombardi & Hurlbert[14] suggest essentially amounts to doing a one-sided test

(incidentally you can also use a closed test procedure[38]). The only difference is the presumption that $\mu \neq 0$. But, as Lombardi & Hurlbert[14] acknowledge, that assumption is *always* plausible - some would say inevitably true[36] – regardless of whether the test is one- or two-sided, or whether the test result is significant or not significant. So, following a nonsignificant one-sided test, and in the spirit of treating one- and two-sided tests in exactly the same way, we may, if only for the sake of argument, assume that there *was* a treatment effect and then ask: what is the strength of the evidence in favour of the predicted effect? As for the observed outcome in the two-sided case, the answer to that question is $(1-P)$ (or odds of $(1-P)/P$), but with P now the significance level reported by the one-side test. If $(1-P)$ is a small number (e.g. $< \alpha$) or, equivalently, the odds in favour of the opposite effect are high (e.g. $P/(1-P) > 20{:}1$), then we may reasonably conclude that *if there was an effect*, it was most likely opposite to that predicted (this follows from the assumption that $\mu \neq 0$). Anyone is then at liberty to pursue this possibility as they see fit. This approach seems in accordance with the suggestion of Koch[15] that following a one-sided test, an outcome opposite to that predicted should be treated "…in a *post hoc*, exploratory, and descriptive way rather than inferentially.". But, as discussed in the next section, perhaps we can go beyond that.

ONLY ONE TYPE OF TEST?

The essential premise of my thesis is that experimental outcomes consistent with a one-sided prediction should be regarded as positive, while those inconsistent with that prediction should be deemed negative. You may reject this premise and choose instead to regard any departure from the statistical hypothesis $\mu = 0$ as positive, regardless of the prediction under test.. This make sense to me, but the issue between us would disappear if there were no choice to be made. And in this regard Lombardi and Hurlbert[14] may have unwittingly led the way.

The fact that one can, under the assumption $\mu \neq 0$, assess the direction of a possible treatment effect following a nonsignificant one-sided test points to a fundamental difference between one- and two-sided predictions and their corresponding negations (null hypotheses). For the two-sided prediction $\mu \neq 0$ the null hypothesis is $\mu = 0$. A high *P* value cannot provide evidence for this null hypothesis because there are infinitely many alternatives $\mu \neq 0$ that are compatible with such high *P* values[1,2]. By contrast, for the one-sided prediction $\mu > 0$ the null hypothesis is $\mu \leq 0$, which can be considered as the disjunction of two statements: "$\mu = 0$ or $\mu < 0$". As in the two-sided case a high *P* value cannot provide evidence for the first statement, *but it can and does provide evidence for the second if we assume the first is false*. And we assume the first is false, i.e. we suppose $\mu \neq 0$, because we are interested in knowing what sort of effects a nonsignificant one-sided test might have missed. Of course one can regard this assumption as being just for the sake of argument, and not necessarily true. But it is tempting to take just one further step. It can be argued that, despite the use of proper controls and randomization, the statistical hypothesis $\mu = 0$ is rarely if ever *literally* true[36]. If one accepts this claim, the next logical step is obvious: *always* assume that $\mu \neq 0$ – literally, not just for the sake of argument - and express the results of all statistical tests for treatment effects (whether one- or two-sided) as the strength of the evidence for or against the predicted (one-sided) or observed (two-sided) effect. Thus for a one-sided test the strength of the evidence in favour of the predicted effect is (1–*P*), or *P* in favour of the opposite effect. For a two-sided test the strength of the evidence is (1–*P*/2) in favour of an effect in the direction indicated by the experimental outcome, and *P*/2 in favour of an effect in the opposite direction [14]. If the evidence in favour of an effect exceeds $1-\alpha$ or, equivalently, the evidence against is less than α, one may choose to accept that effect. But as already indicated the tail probability (*P*/2) of the two-sided test is equal to the tail probability (*P*) of the one-sided test.

Hence if one assumes that $\mu \neq 0$ *a priori* it makes no difference whether one does a one-sided test or two-sided test: the result is the same. Which amounts to saying that there is only one type of test. So the need to choose between one- and two-sided tests is abolished.

To summarize: the advantage of assuming *a priori* that $\mu \neq 0$ is that it avoids the often criticized statistical hypothesis $\mu = 0$ and also avoids the need to choose between one- and two-sided tests. Furthermore, the method is as powerful as a one-sided test, but also allows the detection of effects in both directions as in a two-sided test. The price to be paid for this desirable state of affairs is the abandonment of the null hypothesis. But since, I believe, many people will agree that in the real world the statistical hypothesis $\mu = 0$ is rarely (if ever) true, this hardly seems a high price to pay. From a practical point of view , α, β and P are determined in the usual way (although the P value from a two-sided test must be divided by two in order to obtain the one-sided value). There are of course some issues of a philosophical or logical nature. Specifically, there is no Type I Error because there is no null hypothesis. There is still a Type III Error, i.e. concluding there is an effect in one direction when in fact the true effect is in the opposite direction. The one-sided P value puts an upper limit on the probability of committing this error, and α is the maximum allowable value of P (although borderline cases are still recognized). There is also something akin to a Type II Error (with probability β) although it is obviously not failing to reject H_0 when it is false because there is no H_0. Rather what we are left with is an inability to determine the direction of a treatment effect with acceptable confidence should P exceed α. The method is appealing from a philosophical (i.e. Popperian) point of view because if one assumes *a priori* that $\mu \neq 0$ then that statistical hypothesis obviously doesn't need to be tested. It is then clear that what we are really testing is a theoretical prediction, not a null hypothesis.

CONCLUSIONS

From a practical standpoint, if the outcome of a statistical test is highly significant it may not matter much whether the test is one-sided or two-sided. But the correct choice becomes crucial in the case of a two-sided borderline result. Moreover, from a philosophical point of view, it may seem desirable that any method of analysis should make sense, regardless of whether the result is borderline or not. In this regard the question of whether to use a one-sided or two-sided statistical test is not about suspected - or actual - motives, or the collective interest or even statistics but simple logic: an experimental outcome cannot be both negative and positive in relation to a predicted or required outcome. That is, it cannot both refute and confirm the prediction; it cannot both satisfy and not satisfy the requirement. An appreciation of this fact leads to the following simple rule: *a one-sided prediction or requirement deserves a one-sided statistical test*. The decision whether to perform a one-sided or two-sided test should always be made on logical grounds, not statistical ones. In particular the question of statistical power should be recognized for what it is: irrelevant. Insisting that all tests for treatment effects be two-sided is not only illogical but unethical, because in a placebo-controlled drug trial it means reducing the power to detect beneficial effects for no good reason. A nonsignificant one-sided result can be assessed provisionally by assuming, for the sake of argument, that there was an effect. But if one chooses to take this assumption as *a priori* true, the problem of one-sided *versus* two-sided tests disappears; there is only one type of test. The application of the ideas developed in the present paper are illustrated in supplementary section S3, which discusses two thought experiments of Ludbrook[8] concerned with clinical trials.

ACKNOWLEDGEMENTS

The author declares that he has no competing interests.

REFERENCES

1. Hurlbert SH, Lombardi CM. Final collapse of the Neyman-Pearson decision theoretic framework and rise of the neoFisherian. *Ann. Zool. Fenn.* 2009; **46**: 311-49.
2. Walpole RE, Myers RH, Myers SL, Ye KE. *Probability and statistics for engineers and scientists*, 9th edn. Pearson, Boston, 2011.
3. Baker LF, Mudge JF. Making statistical significance more significant. *Significance* 2012; **9**:29-30.
4. Mudge JF, Baker LF, Edge CB, Houlahan JE. Setting an optimal alpha that minimizes errors in null hypothesis significance tests. *Plos One* 2012; **7**: e32734.
5. Hackshaw A, Kirkwood A. Interpreting and reporting clinical trials with results of borderline significance. BMJ 2011; 343: Epub 4 July 2011; doi:10.1136/bmj.d3340.
6. Popper KR. *The logic of scientific discovery*. Hutchinson, London, 1983.
7. Popper KR. *Conjectures and refutations: the growth of scientific knowledge*. Routledge, New York, 2002.
8. Ludbrook J. Should we use one-sided or two-sided *P* values in tests of significance? *Clin. Exp. Pharmacol. Physiol.* 2013; **40**: 357-61.
9. Ludbrook J. Second Thoughts on the sidedness of *P*. *Clin. Exp. Pharmacol. Physiol.* 2013; **40**: 589-90.
10. Peace KE. One-sided or two-sided *p* values: which most appropriately address the question of drug efficacy? *J. Biopharm. Statist.* 1991; **1**: 133-38.
11. Peace KE. Editor's reply. *J. Biopharm. Statist.* 1996; **6**: 217-18.
12. Curran-Everett D. Sides of the story. *Clin. Exp. Pharmacol. Physiol.* 2013; **40**: 593.

13. Hurlbert SH, Lombardi CM. Lopsided reasoning on lopsided tests and multiple comparisons. *Aust. N. Z. J. Stat.* 2012: **54**: 23-42.

14. Lombardi CM, Hurlbert SH. Misprescription and misuse of one-tailed tests. *Austral. Ecol*. 2009; **34**:, 447-68.

15. Koch GG. One-sided and two-sided tests and *p* values. *J. Biopharm. Statist.* 1991; **1**: 161-70.

16. Armitage, P, Berry G. *Statistical methods in medical research*. Blackwell, Oxford, 1984.

17. Cox DR, Hinkley DV. *Theoretical statistics*. Chapman and Hall, London, 1974.

18. Dunnett CW, Gent M. An alternative to the use of two-sided tests in clinical trials. *Stat. Med.* 1996; **15**: 1729-38.

19. Perneger TV. What's wrong with Bonferroni adjustments. *Br. Med. J.* 1998; **316**: 1236-1238.

20. Schuirmann DJ. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *J. Pharmacokinet. Biopharm.* 1987; **15**: 657-680.

21. Westlake WJ. Response to T. B. L. Kirkwood: Bioequivalence testing--a need to rethink. *Biometrics*. 1981; **37**: 589-594.

22. Woodman RJ. Using one-sided hypothesis tests with a clear conscience. *Clin. Exp. Pharmacol. Physiol.* 2013; **40**: 595-96.

23. Ruxton GD, Neuhaeuser M. When should we use one-tailed hypothesis testing? *Methods Ecol. Evol.* 2010; **1**: 114-17.

24. Ruberg SJ. Dose response studies. I. Some design considerations. *J. Biopharm. Statist.* 1995; **5**: 1-14.

25. Fisher LD The use of one-sided tests in drug trials: An FDA advisory committee member's perspective. *J. Biopharm. Statist.* 1991; **1**: 151-56.

26. Overall JE. A comment concerning one-sided tests of significance in new drug applications. *J. Biopharm. Statist.* 1991; **1:** 157-60.

27. Dubey SD. Some thoughts on the one-sided and two-sided tests. *J. Biopharm. Statist.* 1991; **1**: 139-50.

28. Ng TH. Simultaneous testing of noninferiority and superiority increases the false discovery rate. *J. Biopharm. Statist.* 2007; **17**: 259-64.

29. Yuan J, Tong T, Ng TH. Conditional Type I Error rate for superiority test conditioned on establishment of noninferiority in clinical trials. *Drug Inf. J.* 2011; **45**: 331-36.

30. Goeman JJ, Solari A, Stijnen T. Three-sided hypothesis testing: Simultaneous testing of superiority, equivalence and inferiority. *Stat. Med.* 2010; **29**: 2117-2125.

31. Matthews ST. One-tailed significance tests and the accounting for alpha. *Clin. Exp. Pharmacol. Physiol.* 2013; **40**: 594.

32. Drummond G. One, two, or lots of sides to a problem? 2013; *Clin. Exp. Pharmacol. Physiol.* **40**: 592.

33. Quinn GP, Keough MJ. *Experimental design and data analysis for biologists*. Cambridge University Press, Cambridge, 2002.

34. Cox DR. *Principles of statistical inference*. Cambridge University Press, Cambridge, 2006.

35. Goldfried MR. One-tailed tests and unexpected results. *Psychol. Rev.* 1959; **66**: 79-80.

36. Oakes M. *Statistical inference: A commentary for the social and behavioural sciences*. John Wiley & Sons, Chichester, 1986.

37. Bland JM, Altman DG The odds ratio. *BMJ* 2000; **320**: 1468.

38. Hauschke D, Steinijans VW. Directional decision for a two-tailed alternative. *J. Biopharm. Statist.* 1996; **6**: 211-18.

ONLINE SUPPLEMENT

On the use of one-sided statistical tests in biomedical research

Ricardo Murphy

University of Oslo, Department of Molecular Medicine,
Physiology Section, Pb. 1103, Blindern, 0317 Oslo, Norway

S1. Bigger power or smaller *n* - what's the problem?

For a given *n*, a one-sided test has a higher power for detecting an effect in the predicted direction than a two-sided test. Or, equivalently, a smaller *n* is required for a given power. But why should these attributes be regarded as objections to the use of one-sided tests? Clearly it is no answer to say "because it makes it easier to reject the corresponding null hypothesis"; that is just another way of saying "bigger power or smaller *n*". The answer seems to lie in the suspected motivation for choosing a one-sided test. That is, it is suspected that the test is chosen *only* because it has a higher power or allows a smaller *n*. Thus, according to the statistics editor of one biomedical journal:

> "…there's always the suspicion that people are just using a one-sided test to have a better chance of getting statistical significance,"

a view confirmed to me by a statistician. And from Hurlbert & Lombardi[1] we have:

> "Why risk manuscript rejection with your *P* value of 0.08 from a two-tailed test when you can quietly make a *post hoc* 'prediction', redo your test as a one-tailed one, and obtain *P* = 0.04?".

While according to Ludbrook[2]:

> "Matthews…points to the danger that clinical trialists will prefer one-sided to two-sided tests because the minimal group size for a given power will be smaller with the former".

Although in actual fact Matthews[3] suggests that if a trialist did succumb to this temptation he/she would be behaving ethically:

> "It is…more ethical to use the most powerful detection method because this reduces the number of subjects required to ascertain the effect of a given treatment. The use of one-sided tests increases statistical power. After all, the meaning of a statistically 'significant' result is really just support for further investigation. No drug ever went to market on the strength of results from a single experiment.".

Matthews makes valid points. See also Overall[4] on the ethics of one- and two-sided tests in drug trials. It is also true that the reduced *n* will save time and money in clinical trials[4,5]. But as correctly pointed out by Ruberg[5]:

> "…the cost and time savings are not the ends that justify the means, but rather a beneficial outcome of taking the appropriate scientific approach.".

Yet the issue we are presently concerned with is not ethics or cost effectiveness but *sanity*. Of course we all have our suspicions. For example, one might be forgiven for suspecting that people who raise suspicions about the integrity of others do so only in order to discredit experimental results they don't happen to like. Or to undermine the competition and so improve their chances of securing some of that limited public funding. Or because they feel their own methods of analysis are being called into question. Or because they project their own failings on to others. Or just to be perverse (perhaps they had a bad day). Do we really want science based on suspicion? I don't. I think people who express such suspicions should either present evidence in support of them or refrain from commenting. I also think that scientific discourse should be based on rational arguments and verifiable facts rather than suspicion, insinuation, amateur psychology and mind reading. But there is a deeper point here. Implicit in the concerns regarding the suspected motives of an investigator is the notion that the worth of his/her inferences should be judged *on the basis* of those suspected motives. This is pernicious nonsense. The validity of such inferences should be judged on logical grounds; the investigator's motives (whether real or imagined) have got nothing to do with it. Thus a valid argument based on a set of premises (theory, background knowledge and initial conditions[6,7]) either will or will not result in a one-sided prediction (the conclusion of the argument). If it does then that one-sided prediction should be assessed with a one-sided statistical test.

Consider the following hypothetical situation. A group submits a paper employing a one-sided test, justifying its use by arguing that a result in the negative direction is very unlikely and of no interest. A referee is not persuaded by these arguments, but points out that a logical consequence of the theory under test is a one-sided prediction, and so a one-sided test is appropriate in any case. Are we to reject the researchers' result because they did the right test for the wrong reason? Surely not. So far so good. But now suppose that, some weeks after the paper is published, a disgruntled whistle blower reveals that originally a two-sided test was performed but yielded a borderline result. So a one-sided test was done in order to get the desired significant result. Finally the researchers dreamt-up the justification for the one-sided test and submitted their manuscript. Armed with this incriminating evidence, are we now to reject the researchers' inference? Of course not! That would be completely irrational. The only valid basis for rejecting the inference resulting from the one-sided test is to find a flaw in the referee's argument, i.e. to show that the theory under test did *not* allow the direction of the treatment effect to be predicted. Thus the very notion that the suspected motives of an investigator are relevant to an assessment of the worth of his/her results is a nonsense. Of course in this example it might be argued that, given the dishonest behaviour of the researchers concerning their choice of test, their experimental data should be regarded with suspicion. Well in case you didn't get my attitude toward suspicion in science the first time

I'll repeat it for you in the context of this example. Anyone who thinks the researchers' raw data are suspect should *demonstrate* that they cannot be reproduced or keep quiet.

No one is above suspicion, least of all those who suspect others of wrong doing. But this is of no import because the credibility of scientific results is not based on the presumption that scientists are above suspicion any more than it is based on trust or faith. Rather it rests on the verifiable validity of their arguments and the intersubjective reproducibility of their experimental findings.

S2. When to repeat a study?

Suppose a theory predicts $\mu > 0$. We perform the appropriate experiment and do a one-sided *t* test which returns the nonsignificant result $P = 0.99$. We conclude there is no support for the prediction (and therefore the theory). Moreover, under the assumption that there was any effect at all, it follows immediately from the *P* value that the odds are 99:1 in favour of $\mu < 0$. What happens next depends on how creative you are. If you can't think of any explanation for a negative result then all you can do is draw attention to it, so that others may pursue it if they wish. On the other hand if you can come up with a new theory to explain it then you may wish to repeat the study to confirm the unexpected result. This is option (4) in Lombardi & Hurlbert[8] but only in the case that you can explain the negative effect. If not then there is no point in repeating the study because the results will be uninterpretable. Thus if the negative result is confirmed it will be meaningless because you don't have a theory to give it meaning. If the result is not significant then, as usual, you will not be able to conclude anything. And if the result is positive – i.e. confirming the original prediction – you will now have two contradictory results and so you still won't be able to conclude anything. It is for others to repeat the study if they can come up with an explanation for the negative effect. They at least have the possibility of obtaining a meaningful result.

I am not quite sure what to make of the contention of Lombardi & Hurlbert[8] that the main benefit of repeating a study is "a truly celestial level of 'statistical purity'". But clearly if you have a new theory that explains the unexpected result there is no point in *just* repeating the study:

> "…we require that the new theory should be *independently testable*. That is to say, apart from explaining all the *explicanda* which the new theory was designed to explain, it must have new and testable consequences (preferably consequences of a *new kind*); it must lead to the prediction of phenomena which have not so far been observed."[7].

Thus a new prediction must be deduced from your new theory and tested experimentally. If it is a good theory it will predict the direction of the expected new effect, and this will necessitate a one-sided test. Obviously if the new theory concerns something in the background knowledge (e.g. equipment malfunction) or initial conditions (e.g. tissue not in a stationary state) it makes sense to test the new prediction and/or take corrective action *before* repeating the study which produced the negative result. On the other hand, if the result does not appear to be artifactual, but it would be very expensive to repeat the study, one may choose to move on to the testing of new predictions; this is a matter of judgment. Of course if

you choose to adopt the approach advocated in the main text - i.e. assume *a priori* that $\mu \neq 0$ – then repeating the study may seem unnecessary (although independent confirmation by others is always welcome).

In the context of repeating studies one should also mention the problem of the reproducibility of experimental findings, something we hear a lot about nowadays (and rightly so). While acknowledging the legitimate concerns of Bissell[9], I am inclined to agree with Russel[10] that there should be more confirmatory studies, especially by independent research groups, because there is a need to demonstrate reproducibility, especially *intersubjective* reproducibility. I do not regard such studies as a waste of resources.

S3. Discussion of Ludbrook's thought experiments

We can apply some of the ideas developed in the present paper to two thought experiments discussed by Ludbrook[11]. In Experiment 1 a pharmacologist used a one-sided two-sample *t*-test to test the one-sided prediction (statistical hypothesis) that his newly synthesized drug ("THINNA") induces weight loss ($\mu < 0$). This prediction was confirmed at the 2.6% level of significance. Drummond[12] raised the following question:

> "Professor Ludbrook's pharmacologist used a single-tailed test, assuming that weight gain was impossible. Was that reasonable?".

It is not clear to me that he did assume this. Certainly there was no need to do so since the negated prediction – i.e. the null hypothesis $\mu \geq 0$ - clearly allows for a weight gain. In any event it makes no difference whether he did or not, or whether any such assumption was reasonable or unreasonable. If the drug really caused either no change or an increase in weight then the probability of concluding a weight loss was never going to be more than $\alpha = 0.05$, and the probability of obtaining a weight loss at least as large as that actually observed was at most $P = 0.026$. On the other hand if it had turned-out that $P >> \alpha$, then under the assumption that there was any effect at all, P would be the strength of the evidence in favour of a weight gain, and anyone would be at liberty to pursue this possibility should they think it worthwhile. What's the problem? As a matter of fact, if the pharmacologist predicted *a priori* that the drug induced weight loss on the basis of some theory of metabolism or appetite and drug action (which presumably he did, otherwise he wouldn't have known what drug to synthesize), then the preceding study on obese mice postulated by Ludbrook[11] should have been assessed with a one-sided test as well! Perhaps there are those whose telepathic powers or suspicious minds lead them to suspect that the pharmacologist only wanted to do a one-sided test in order to have a better chance of obtaining evidence favourable to his theory. But whether or not they are right about that is irrelevant. The fact is he *did* have a theory, and the one-sided prediction was deduced as a logical consequence of it. In this case the one-sided test was the appropriate one to do, as correctly surmised by Ludbrook[11], whether the pharmacologist knew it or not.

In Experiment 2 a clinical trialist used a two-sided, two-sample *t* test to assess the difference between THINNA and an established drug "SLIMMA". She obtained the borderline result $P = 0.052$ and concluded there was no significant difference between the

drugs. Was a two-sided test the right choice? According to Ludbrook[11] the trialist's "scientific hypothesis" (theory) was: "the two drugs will produce the same change in weight". What then was her statistical hypothesis? As far as I can see it was: "the two drugs will produce the same change in weight". Or in other words the scientific and statistical hypotheses are identical. And both are identical with the prediction, which happens to be a zero prediction ($\mu = 0$). Zero predictions or, more generally, *precise* predictions are outside the scope of this paper, which is concerned with the use of one-sided tests to assess the plausibility of the one-sided predictions $\mu > 0$ and $\mu < 0$. But in any event, is the prediction $\mu = 0$ really what we are interested in? I suggest not. For simplicity let us suppose that the two drugs are comparable with regard to cost and side effects, and that we are not interested in bioequivalence (see Schuirmann[13] for a discussion of the latter). So in other words this is a superiority trial; the only issue is whether the new drug is more effective than the old one. Then if μ is the difference in induced weight loss (THINNA−SLIMMA) it seems to me we are only interested in outcomes consistent with the requirement $\mu < 0$; for $\mu \geq 0$ (null hypothesis) THINNA is of no interest. The point is: *the trialist should have done a one-sided test not a two-sided test*. In case this is still not clear it may become so by recognizing that Experiment 2 is really not about testing a scientific hypothesis at all. Indeed there is no genuine scientific hypothesis but only the decision criterion $\mu < 0$. And the aim of Experiment 2 is to determine whether μ is consistent with that requirement. Had the trialist done a one-sided test she would have been rewarded with the significant result $P = 0.026$ instead of the borderline one. Of course that would not be the end of the matter, as pointed out by Matthews[3]. But here our main concern is to get this initial test right and so determine whether THINNA is worth investigating further as a possible replacement for SLIMMA.

So it turns out that both experiments, as well as the mouse trial preceding Experiment 1, should have been analyzed with one-sided tests. Indeed it seems to me that many if not most experiments deserve one-sided tests. Thus in applied work it will often be clear that effects in one direction only are of interest. Placebo-controlled drug trials provide a clear example. But even in pure, theory-driven research, as one's theories become more refined there will surely come a time – probably sooner rather than later – when one can predict the direction of a treatment effect! Indeed this should ideally be the case from the outset; a non-directional theory is not scientific in the Popperian sense because a statistically significant result is guaranteed for sufficiently large n[14]. Of course if, as suggested in the main text, one takes the statistical hypothesis $\mu = 0$ to be *a priori* false, then the whole problem of one-sided *versus* two-sided tests disappears. For example, in Experiment 2 the strength of the evidence that THINNA is better than SLIMMA is 1–0.026 = 0.974 (equivalent to odds of (1–0.026)/0.026 = 37:1). And you get this answer regardless of whether you do a one-sided test or a two-sided test. Which of course means that in effect there is only one type of test.

References

1. Hurlbert SH, Lombardi CM. Final collapse of the Neyman-Pearson decision theoretic framework and rise of the neoFisherian. *Ann. Zool. Fenn.* 2009; **46**: 311-49.
2. Ludbrook J. Second thoughts on the sidedness of *P*. *Clin. Exp. Pharmacol. Physiol.* 2013; **40**: 589-90.
3. Matthews ST. One-tailed significance tests and the accounting for alpha. *Clin. Exp. Pharmacol. Physiol.* 2013; **40**: 594.
4. Overall JE. A comment concerning one-sided tests of significance in new drug applications. *J. Biopharm. Statist.* 1991; **1:** 157-60.
5. Ruberg SJ. Dose response studies. I. Some design considerations. *J. Biopharm. Statist.* 1995; **5**: 1-14.
6. Popper KR. *The logic of scientific discovery*. Hutchinson, London, 1983.
7. Popper KR. *Conjectures and refutations: the growth of scientific knowledge*. Routledge, New York, 2002.
8. Lombardi CM, Hurlbert SH. Misprescription and misuse of one-tailed tests. *Austral. Ecol*. 2009; **34**:, 447-68.
9. Bissell M. The risks of the replication drive. *Nature*. 2015; **503**: 333-334.
10. Russel JF. If a job is worth doing, it is worth doing twice. *Nature*. 2013; **496**:7.
11. Ludbrook J. Should we use one-sided or two-sided *P* values in tests of significance? *Clin. Exp. Pharmacol. Physiol.* 2013; **40**: 357-61.
12. Drummond G. One, two, or lots of sides to a problem? 2013; *Clin. Exp. Pharmacol. Physiol.* **40**: 592.
13. Schuirmann DJ. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *J. Pharmacokinet. Biopharm.* 1987; **15**: 657-680.
14. Oakes M. *Statistical inference: A commentary for the social and behavioural sciences.* John Wiley & Sons, Chichester, 1986.