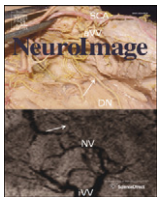
NeuroImage 61 (2012) 1300–1310

Contents lists available at SciVerse ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg



Comments and Controversies

Ten ironic rules for non-statistical reviewers

Karl Friston*

The Wellcome Trust Centre for Neuroimaging, Institute of Neurology, Queen Square, London WC1N 3BG, UK

ARTICLE INFO

Article history:
Accepted 7 April 2012
Available online 13 April 2012

Keywords:
Statistical testing
Sample-size
Effect size
Power
Classical inference

ABSTRACT

As an expert reviewer, it is sometimes necessary to ensure a paper is rejected. This can sometimes be achieved by highlighting improper statistical practice. This technical note provides guidance on how to critique the statistical analysis of neuroimaging studies to maximise the chance that the paper will be declined. We will review a series of critiques that can be applied universally to any neuroimaging paper and consider responses to potential rebuttals that reviewers might encounter from authors or editors.

Introduction

This technical note is written for reviewers who may not have sufficient statistical expertise to provide an informed critique during the peer-reviewed process, but would like to recommend rejection on the basis of inappropriate or invalid statistical analysis. This guidance follows the 10 simple rules format and hopes to provide useful tips and criticisms for reviewers who find themselves in this difficult position. These rules are presented for reviewers in an ironic way[1] that makes it easier (and hopefully more entertaining) to discuss the issues from the point of view of both the reviewer and author — and to caricature both sides of the arguments. Some key issues are presented more formally in (non-ironic) appendices.

There is a perceived need to reject peer-reviewed papers with the advent of open access publishing and the large number of journals available to authors. Clearly, there may be idiosyncratic reasons to block a paper – to ensure your precedence in the literature, personal rivalry etc. – however, we will assume that there is an imperative to reject papers for the good of the community: handling editors are often happy to receive recommendations to decline a paper. This is because they are placed under pressure to maintain a high rejection rate. This pressure is usually exerted by the editorial board (and publishers) and enforced by circulating quantitative information about their rejection rates (i.e., naming and shaming lenient editors). All journals want to maximise rejection rates, because this increases the quality of submissions, increases their impact factor and underwrites their long-term viability. A reasonably mature journal like Neuroimage would hope to see between 70% and 90% of submissions rejected. Prestige journals usually like to reject over 90% of the papers they receive. As an expert reviewer, it is your role to help editors decline papers whenever possible. In what follows, we will provide 10 simple rules to make this job easier:

Rule number one: dismiss self doubt

Occasionally, when asked to provide an expert opinion on the design or analysis of a neuroimaging study you might feel under qualified. For example, you may not have been trained in probability theory or statistics or – if you have – you may not be familiar with topological inference and related topics such as random field theory. It is important to dismiss any ambivalence about your competence to provide a definitive critique. You have been asked to provide comments as an expert reviewer and, operationally, this is now your role. By definition, what you say is the opinion of the expert reviewer and cannot be challenged — in relation to the paper under consideration, you are the ultimate authority. You should therefore write with authority, in a firm and friendly fashion.

Rule number two: avoid dispassionate statements

A common mistake when providing expert comments is to provide definitive observations that can be falsified. Try to avoid phrases like “I believe” or “it can be shown that”. These statements invite a rebuttal that could reveal your beliefs or statements to be false. It is much safer, and preferable, to use phrases like “I feel” and “I do not trust”. No one can question the veracity of your feelings and convictions. Another useful device is to make your points vicariously; for example, instead of saying “Procedure A is statistically invalid” it is much better to say that “It is commonly accepted that procedure A is statistically invalid”. Although authors may be able to show that procedure A is valid, they will find it more difficult to prove that it is commonly accepted as valid. In short, trying to pre-empt a

[1] The points made in this paper rest heavily on irony (*Irony* from the Ancient Greek ε ρωνεία *eirōneía*, meaning dissimilation or feign ignorance). The intended meaning of ironic statements is the opposite of their literal meaning.

1053-8119/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2012.04.018

prolonged exchange with authors by centring the issues on convictions held by yourself or others and try to avoid stating facts.

Rule number three: submit your comments as late as possible

It is advisable to delay submitting your reviewer comments for as long as possible — preferably after the second reminder from the editorial office. This has three advantages. First, it delays the editorial process and creates an air of frustration, which you might be able to exploit later. Second, it creates the impression that you are extremely busy (providing expert reviews for other papers) and indicates that you have given this paper due consideration, after thinking about it carefully for several months. A related policy, that enhances your reputation with editors, is to submit large numbers of papers to their journal but politely decline invitations to review other people's papers. This shows that you are focused on your science and are committed to producing high quality scientific reports, without the distraction of peer-review or other inappropriate demands on your time.

Rule number four: the under-sampled study

If you are lucky, the authors will have based their inference on less than 16 subjects. All that is now required is a statement along the following lines:

"*Reviewer: Unfortunately, this paper cannot be accepted due to the small number of subjects. The significant results reported by the authors are unsafe because the small sample size renders their design insufficiently powered. It may be appropriate to reconsider this work if the authors recruit more subjects.*"

Notice your clever use of the word "unsafe", which means you are not actually saying the results are invalid. This sort of critique is usually sufficient to discourage an editor from accepting the paper; however – in the unhappy event the authors are allowed to respond – be prepared for something like:

"*Response: We would like to thank the reviewer for his or her comments on sample size; however, his or her concerns are statistically misplaced. This is because a significant result (properly controlled for false positives), based on a small sample indicates the treatment effect is actually larger than the equivalent result with a large sample. In short, not only is our result statistically valid. It is quantitatively stronger than the same result with a larger number of subjects.*"

Unfortunately, the authors are correct (see Appendix 1). On the bright side, the authors did not resort to the usual anecdotes that beguile handling editors. Responses that one is in danger of eliciting include things like:

"*Response: We suspect the reviewer is one of those scientists who would reject our report of a talking dog because our sample size equals one!*"

Or, a slightly more considered rebuttal:

"*Response: Clearly, the reviewer has never heard of the fallacy of classical inference. Large sample sizes are not a substitute for good hypothesis testing. Indeed, the probability of rejecting the null hypothesis under trivial treatment effects increases with sample size.*"

Thankfully, you have heard of the fallacy of classical inference (see Appendix 1) and will call upon it when needed (see next rule). When faced with the above response, it is often worthwhile trying a slightly different angle of attack; for example[2]

"*Reviewer: I think the authors misunderstood my point here: The point that a significant result with a small sample size is more compelling than one with a large sample size ignores the increased influence of outliers and lack-of-robustness for small samples.*"

Unfortunately, this is not actually the case and the authors may respond with:

"*Response: The reviewer's concern now pertains to the robustness of parametric tests with small sample sizes. Happily, we can dismiss this concern because outliers decrease the type I error of parametric tests (Zimmerman, 1994). This means our significant result is even less likely to be a false positive in the presence of outliers. The intuitive reason for this is that an outlier increases sample error variance more than the sample mean; thereby reducing the t or F statistic (on average).*"

At this point, it is probably best to proceed to rule six.

Rule number five: the over-sampled study

If the number of subjects reported exceeds 32, you can now try a less common, but potentially potent argument of the following sort:

"*Reviewer: I would like to commend the authors for studying such a large number of subjects; however, I suspect they have not heard of the fallacy of classical inference. Put simply, when a study is over-powered (with too many subjects), even the smallest treatment effect will appear significant. In this case, although I am sure the population effects reported by the authors are significant; they are probably trivial in quantitative terms. It would have been much more compelling had the authors been able to show a significant effect without resorting to large sample sizes. However, this was not the case and I cannot recommend publication.*"

You could even drive your point home with:

"*Reviewer: In fact, the neurological model would only consider a finding useful if it could be reproduced three times in three patients. If I have to analyse 100 patients before finding a discernible effect, one has to ask whether this effect has any diagnostic or predictive value.*"

Most authors (and editors) will not have heard of this criticism but, after a bit of background reading, will probably try to talk their way out of it by referring to effect sizes (see Appendix 2). Happily, there are no rules that establish whether an effect size is trivial or nontrivial. This means that if you pursue this line of argument diligently, it should lead to a positive outcome.

Rule number six: untenable assumptions (nonparametric analysis)

If the number of subjects falls between 16 and 32, it is probably best to focus on the fallibility of classical inference — namely its assumptions. Happily, in neuroimaging, it is quite easy to sound convincing when critiquing along these lines: for example,

"*Reviewer: I am very uncomfortable about the numerous and untenable assumptions that lie behind the parametric tests used by the authors. It is well-known that MRI data has a non Gaussian (Rician) distribution, which violates the parametric assumptions of their statistical tests. It is imperative that the authors repeat their analysis using nonparametric tests.*"

The nice thing about this request is that it will take some time to perform nonparametric tests. Furthermore, the nonparametric tests will, by the Neyman–Pearson lemma,[3] be less sensitive than the

[2] This point was raised by a reviewer of the current paper.

[3] The Neyman–Pearson lemma states that when performing a hypothesis test, the likelihood-ratio test is the most powerful test for a given size and threshold.

Table 1
Some common effect-sizes for a one sample *t*-test: this simple model has been chosen to highlight the relationship among various forms or measures of effect size. The key things to take from this table are (i) effect sizes can be unstandardised or standardised — where standardised effect sizes require both the size of the treatment effect and its standard deviation; (ii) all standardised effect sizes can be computed from the unstandardised effect size and standard deviation — and constitute different ways of quantifying the same thing; and (iii) implicitly, all standardised effect sizes can be computed from test statistics, given the degrees of freedom or number of observations. Effect sizes can be true or estimated, where (point) estimators of unstandardised effect size and standard deviation can be based upon the data used to infer a treatment effect – giving in-sample effect sizes or based upon independent data – giving out-of-sample predictions of effect size.

Name	Form	Comment	Standardised
Model parameter	μ	The size of the treatment effect — usually reported as a model parameter or regression coefficient (here, the group mean)	No
Cohen's d	$d = \frac{\mu}{\sigma} = \frac{t}{\sqrt{n}}$	A standardised measure of effect size — the *t*-statistic divided by the square root of the number of samples	Yes
Coefficient of determination	$R^2 = \frac{\mu^2}{\mu^2+\sigma^2}$	The proportion of variance explained by the treatment effect	Yes
Correlation	$\rho = \sqrt{R^2}$	A measure of association — the square root of the coefficient of determination	Yes

original likelihood ratio tests reported by the authors — and their significant results may disappear. However, be prepared for the following rebuttal:

"*Response: We would like to thank the reviewer for his or her helpful suggestions about nonparametric testing; however, we would like to point out that it is not the distribution of the data that is assumed to be Gaussian in parametric tests, but the distribution of the random errors. These are guaranteed to be Gaussian for our data, by the Central limit theorem,*[4] *because of the smoothing applied to the data and because our summary statistics at the between subject level are linear mixtures of data at the within subject level.*"

The authors are correct here and this sort of response should be taken as a cue to pursue a different line of critique:

Rule number seven: question the validity (cross validation)

At this stage, it is probably best to question the fundaments of the statistical analysis and try to move the authors out of their comfort zone. A useful way to do this is to keep using words like validity and validation: for example,

"*Reviewer: I am very uncomfortable about the statistical inferences made in this report. The correlative nature of the findings makes it difficult to accept the mechanistic interpretations offered by the authors. Furthermore, the validity of the inference seems to rest upon many strong assumptions. It is imperative that the authors revisit their inference using cross validation and perhaps some form of multivariate pattern analysis.*"

Hopefully, this will result in the paper being declined or – at least – being delayed for a few months. However, the authors could respond with something like:

"*Response: We would like to thank the reviewer for his or her helpful comments concerning cross validation. However, the inference made using cross validation accuracy pertains to exactly the same thing as our classical inference; namely, the statistical dependence (mutual information) between our explanatory variables and neuroimaging data. In fact, it is easy to prove (with the Neyman–Pearson lemma) that classical inference is more efficient than cross validation.*"

This is frustrating, largely because the authors are correct[5] and it is probably best to proceed to rule number eight.

Rule number eight: exploit superstitious thinking

As a general point, it is useful to instil a sense of defensiveness in editorial exchanges by citing papers that have been critical of neuroimaging data analysis. A useful entree here is when authors have reported effect sizes to supplement their inferential statistics (*p* values). Effect sizes can include parameter estimates, regression slopes, correlation coefficients or proportion of variance explained (see Table 1 and Appendix 2). Happily, most authors will have reported some form of effect size, exposing themselves to the following critique:

"*Reviewer: It appears that the authors are unaware of the dangers of voodoo correlations and double dipping. For example, they report effect sizes based upon data (regions of interest) previously identified as significant in their whole brain analysis. This is not valid and represents a pernicious form of double dipping (biased sampling or the non-independence problem). I would urge the authors to read Vul et al. (2009) and Kriegeskorte et al. (2009) and present unbiased estimates of their effect size using independent data or some form of cross validation.*"

Do not be deterred by the fact that reporting effect sizes is generally considered to be good practice — the objective here is to create an atmosphere in which punitive forces could expose the darkest secrets of any author, even if they did not realise they had them. The only negative outcome will be a response along the following lines:

"*Response: We thank the reviewer for highlighting the dangers of biased sampling but this concern does not apply to our report: by definition, the effect size pertains to the data used to make an inference — and can be regarded as an in-sample prediction of the treatment effect. We appreciate that effect sizes can overestimate the true effect size; especially when the treatment effect is small or statistical thresholds are high. However, the (in-sample) effect size should not be confused with an out-of-sample prediction (an unbiased estimate of the true effect size). We were not providing an out-of-sample prediction but simply following APA guidelines by supplementing our inference ("Always present effect sizes for primary outcomes." Wilkinson and APA Task Force on Statistical Inference, 1999, p. 599).*"

In this case, the authors have invoked the American Psychological Association (APA) guidelines (Wilkinson and APA Task Force on Statistical Inference, 1999) on good practice for statistical reporting in journals. It is difficult to argue convincingly against these guidelines (which most editors are comfortable with). However, do not be too disappointed because the APA guidelines enable you to create a Catch-22 for authors who have not reported effect sizes:

"*Reviewer: The authors overwhelm the reader with pretty statistical maps and magnificent p-values but at no point do they quantify the underlying effects about which they are making an inference. For*

[4] The central limit theorem states the conditions under which the mean of a sufficiently large number of independent random variables, each with finite mean and variance, will be (approximately) normally distributed.

[5] Inferences based upon cross validation tests (e.g., accuracy or classification performance) are not likelihood ratio tests because, by definition, they are not functions of the complete data whose likelihood is assessed. Therefore, by the Neyman–Pearson lemma, they are less powerful.

example, their significant interaction would have profoundly different implications depending upon whether or not it was a crossover interaction. In short, it is essential that the authors supplement their inference with appropriate effect sizes (e.g., parameter estimates) in line with accepted practice in statistical reporting ("Always present effect sizes for primary outcomes." Wilkinson and APA Task Force on Statistical Inference, 1999, p. 599)."

When they comply, you can apply rule eight — in the fond hope they (and the editors) do not appreciate the difference between in-sample effect sizes and out-of-sample predictions.

Rule number nine: highlight missing procedures

Before turning to the last resort (rule number ten). It is worthwhile considering any deviation from usual practice. We are particularly blessed in neuroimaging by specialist procedures that can be called upon to highlight omissions. A useful critique here is:

"*Reviewer: The author's failure to perform retinotopic mapping renders the interpretation of their results unsafe and, in my opinion, untenable. Please conform to standard practice in future.*"

Note how you have cleverly intimated a failure to conform to standard practice (which most editors will assume is good practice). In most cases, this sort of critique should ensure a rejection; however, occasionally, you may receive a rebuttal along the following lines:

"*Response: We would like to thank the reviewer for his or her helpful comments: however, we like to point out that our study used olfactory stimuli, which renders the retinotopic mapping somewhat irrelevant.*"

Although you could debate this point, it is probably best to proceed to rule number ten.

Rule number ten: the last resort

If all else fails, then the following critique should secure a rejection:

"*Reviewer: Although the authors provide a compelling case for their interpretation; and the analyses appear valid if somewhat impenetrable, I cannot recommend publication. I think this study is interesting but colloquial and would be better appreciated (and assessed) in a more specialised journal.*"

Notice how gracious you have been. Mildly laudatory comments of this sort suggest that you have no personal agenda and are deeply appreciative of the author's efforts. Furthermore, it creates the impression that your expertise enables you not only to assess their analyses, but also how they will be received by other readers. This impression of benevolence and omnipotence makes your final value judgement all the more compelling and should secure the desired editorial decision.

Conclusion

We have reviewed some general and pragmatic approaches to critiquing the scientific work of others. The emphasis here has been on how to ensure a paper is rejected and enable editors to maintain an appropriately high standard, in terms of papers that are accepted for publication. Remember, as a reviewer, you are the only instrument of selective pressure that ensures scientific reports are as good as they can be. This is particularly true of prestige publications like *Science* and *Nature*, where special efforts to subvert a paper are sometimes called for.

Acknowledgments

I would like to thank the Wellcome Trust for funding this work and Tom Nichols for comments on the technical aspects of this work. I would also like to thank the reviewers and editors of NeuroImage for their thoughtful guidance and for entertaining the somewhat risky editorial decision to publish this (ironic) article.

Appendix 1. The fallacy of classical inference

These appendices revisit some of the issues in the main text in greater depth — and non-ironically. The first appendix presents an analysis of effect size in classical inference that suggests the optimal sample size for a study is between 16 and 32 subjects. Crucially, this analysis suggests significant results from small samples should be taken more seriously than the equivalent results in oversized studies. Furthermore, studies with more than 50 subjects may expose themselves to trivial effects, which can be mediated by inconsistent and low-prevalence effects. The somewhat counterintuitive implication is that overpowered studies can lose integrity and should be interpreted with caution. This loss of integrity is due to a fallacy of classical inference; which states that – with sufficient power – the null hypothesis will be rejected with probability one in the presence of a trivial effect. These points are illustrated quantitatively in terms of effect sizes and loss-functions.

Sample sizes

The question "how many subjects constitute a study?" preoccupies many fields of empirical study and yet there is a surprising small literature on the subject (Lenth, 2001). General references include Cohen (1988), Desu and Raghavarao (1990), Lipsey (1990), Shuster (1990), and Odeh and Fox (1991). See Maxwell et al. (2008) for a recent review of sample size planning in relation to parameter estimation in psychology and Friston et al. (1999) for a treatment of group studies in neuroimaging. Recently, there has been a pressure to increase sample sizes in functional neuroimaging; both in terms of editorial requirements and the incidence of large cohort studies (e.g., Lohrenz et al., 2007). What follows provides a peer-reviewed citation that allows researchers to defend themselves against the critique that their study is underpowered. This is particularly relevant for functional neuroimaging, where the cost of large sample studies can be substantial.

Sample size "must be big enough that an effect of such magnitude as to be of scientific significance will also be statistically significant. *It is just as important, however, that the study not be too big, where an effect of little scientific importance is nevertheless statistically detectable*" (Lenth, 2001; our emphasis). In what follows, we address the balance between these requirements in terms of effect sizes. In brief, we will appeal to a fallacy of classical inference to show that studies can be overpowered and that these studies are sensitive to trivial effects. The arguments presented below are elementary and general: they do not depend upon domain-specific treatment effects (e.g., activations) or levels of noise. In fact, the arguments are sufficiently simple they can be articulated with a couple of heuristics and equations. We first review the notion of standardised effect sizes and see how effect and sample size combine to affect sensitivity. This appendix concludes with a straightforward loss-function analysis that suggests there is an optimal sample size for any given study.

Sample sizes and functional neuroimaging

Between-subject analyses are now commonplace in most applications of neuroimaging. These are usually based upon the summary statistic approach (Holmes and Friston, 1998), where estimates of activations or treatment effects are harvested from within-subject (first-level) analyses and then passed to between-subject (second-

level) *t*-tests. The basic idea is to test for treatment effects that are large in relation to inter-subject variation. In classical inference, this usually proceeds by modelling each sample or subject-specific measurement as a random Gaussian deviation about a group mean effect μ. An effect is declared significant when the probability of obtaining the samples – under the null hypothesis that the mean is zero – is sufficiently small, usually $p<0.05$. However, the fallacy of this approach is that the null hypothesis is always false (because the probability of two groups being exactly the same is zero) and can always be rejected with sufficient degrees of freedom or power. To understand this more formally, one can quantify deviations from the null hypothesis — this is the effect size:

Effect size

Effect size measures the strength of a deviation from the null hypothesis. Usually, the term effect size refers to the estimate of an unknown true effect size based on the data at hand. In this appendix, we will distinguish between true and estimated effect sizes where necessary and assume that effect sizes are standardised. In other words, the (true) size of the effect is expressed relative to its (true) standard deviation:

$$d = \frac{\mu}{\sigma}. \qquad 1$$

For simplicity, we only consider effect sizes in the context of a one-sample *t*-test, noting that most classical inferences under parametric assumptions can be reduced to a *t*-test. The standardised effect size is the underlying deviation from the mean under the null hypothesis – which we will assume is zero – divided by the standard deviation over samples or subjects. Cohen's *d* can be regarded as a point estimator of (the true) effect size and is based on the sample mean and sample standard deviation. Standardised effect sizes allow us to talk in general terms without having to worry about the amplitude of treatment and random effects. Furthermore, because we are dealing with between-subject comparisons, we can ignore the distinction between fixed and mixed-effects models and the composition of random effects — these issues have been considered previously from a Bayesian and classical perspective (Friston et al., 2002, 2005, respectively).

Cohen (1988) divides effect sizes into large, medium and small. We will extend this to cover *trivial* effect sizes; where large effect sizes are about one (the mean has the same size as the standard deviation), medium effect sizes are about a half, small effect sizes are a quarter and trivial effect sizes are one eighth (see Table 2). In what sense are trivial effect sizes trivial? Consider the following example: imagine we compared the intelligence quotient (IQ) between the pupils of two schools. When comparing two groups of 800 pupils, we found mean IQs of 107.1 and 108.2, with a difference of 1.1. Given that the standard deviation of IQ is 15, this would be a trivial effect size (less than two). In short, although the differential IQ may be extremely significant, it is scientifically uninteresting — to all intents and purposes, the pupils at both schools have the same IQ. Now imagine that your research assistant had the bright idea of comparing the IQ of students who had and had not recently changed schools. On selecting 16 students who had changed schools within the past 5 years and 16 matched pupils who had not, she found an IQ difference of 11.6, where this medium effect size just reached significance. This example highlights the difference between an uninformed overpowered hypothesis test that gives very significant, but uninformative results and a more mechanistically grounded hypothesis that can only be significant with a meaningful effect size.

Another quantitative interpretation of effect size appeals to diagnosis or classification: effect sizes are small when they cannot discriminate between subjects that do and do not show an effect. This is shown in Fig. 1, where the distributions of (standardised) samples from populations with different true effect sizes are plotted, under Gaussian assumptions. If we use the optimal threshold or criterion of $d/2$ – to classify a subject as showing an effect – we can quantify

Table 2
Banding of standardised effect sizes and the associated discriminability (expressed in terms of classification accuracy) and consistency (population prevalence, under a binomial model).

Effect size	Cohen's d	Classification accuracy	Population prevalence
Large	~1	~70%	~50%
Medium	~1/2	~60%	~20%
Small	~1/4	~55%	~6%
Trivial	~1/8	~52.5%	~1%
None	0	50%	0%

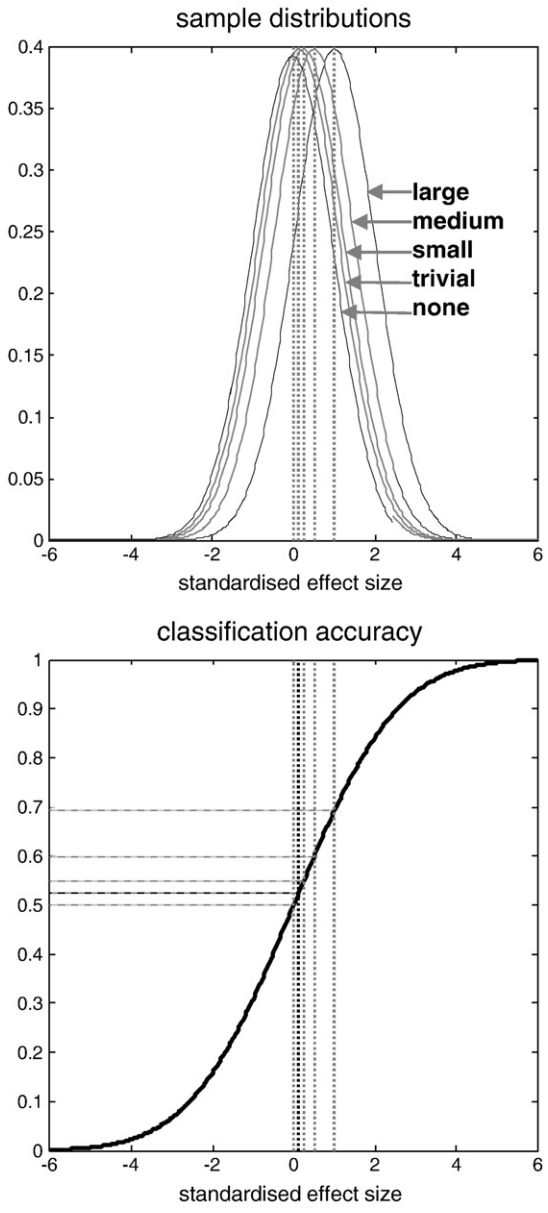


Fig. 1. Upper panel: distributions of samples from populations with different (standardised) effect sizes. These canonical effect sizes correspond to "large", "medium", "small", "trivial" and "none" (denoted by the vertical broken lines). Lower panel: classification accuracy as a function of effect size. This is the area under the distributions to the right of an optimal threshold; this threshold is half-way between the mean of each group and zero.

the accuracy of classification in terms of effect size. The accuracy is simply the area above threshold under each distribution:

$$\int_{d/2}^{\infty} \mathcal{N}(z : d, 1)dz. \qquad 2$$

The lower panel of Fig. 1 shows classification accuracy as a function of true effect size. For large effect sizes, we would correctly classify a responsive subject about 70% of the time. However, if we consider trivial effect sizes, we would be just above chance classification at about 52.5% (see Table 2). In other words, given a trivial effect size, we can (effectively) do no better than chance in deciding whether a given subject showed a treatment effect or not. Note that this way of framing the magnitude of effect sizes does not care about the source of random effects. These could be true variations in the underlying response from subject to subject or reflect noisy measurements of a consistent response. Operationally, the composition or source of these random fluctuations is irrelevant — both conspire to make trivial effect sizes too small for diagnosis or discrimination.

The quantitative nature of effect sizes can also be seen if we assume that subjects are sampled from one of two groups at random. In this binomial model, one group expresses the effect and the other does not. The random effects here depend on how subjects are selected from one group relative to another. Under this binomial model, we can express the effect size in terms of the proportion of the population that shows the effect γ. This is given by

$$d = \frac{\gamma}{\sqrt{\gamma(1-\gamma)^2 + \gamma^2(1-\gamma)}}. \qquad 3$$

Fig. 2 shows the true effect size as a function of the prevalence of subjects showing an effect. Here, an effect size of one means that only half the subjects show an effect (e.g. activation). In other words, to say that a response occurred more often than not – in a subject sampled at random – we would require a large effect size or more. A small effect size corresponds to a prevalence of about 6%. In other words, a small effect size means that a small minority of subjects exhibit an effect, under this binomial model. Crucially, only 1% of the population needs to show an effect to induce a trivial effect size. Clearly, this model of random effects violates parametric assumptions and would call for nonparametric tests. However, it serves as a useful heuristic to show that small effect sizes can arise when the population shows inconsistent, low-prevalence effects.

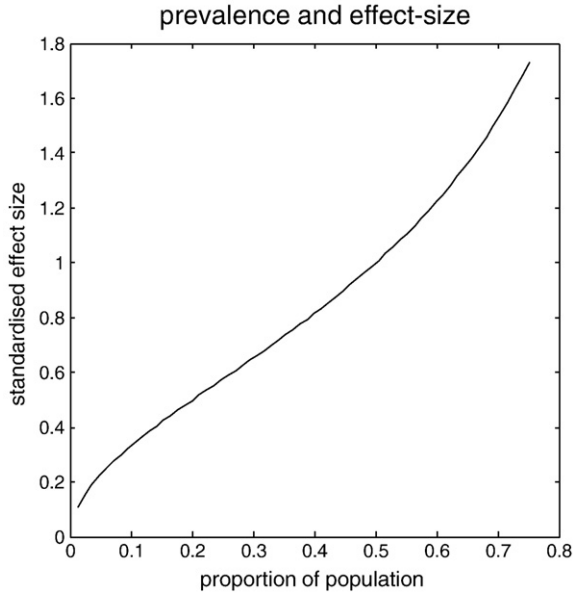


Fig. 2. Standardised effect size as a function of prevalence or the proportion of a population who express an effect. In this example, subjects are selected from one of two groups that express an effect, or not, with no measurement error (the binomial model in the main text).

The Lindley paradox and the fallacy of classical tests

So why are we worried about trivial effects? They are important because the probability that the true effect size is exactly zero is itself zero and could cause us to reject the null hypothesis inappropriately. This is a fallacy of classical inference and is not unrelated to Lindley's paradox (Lindley, 1957). Lindley's paradox describes a counterintuitive situation in which Bayesian and frequentist approaches to hypothesis testing give opposite results. It occurs when; (i) a result is significant by a frequentist test, indicating sufficient evidence to reject the null hypothesis $d = 0$ and (ii) priors render the posterior probability of $d = 0$ high, indicating strong evidence that the null hypothesis is true. In his original treatment, Lindley (1957) showed that – under a particular form of prior on the effect size – the posterior probability of the null hypothesis being true, given a significant test, approaches one as sample-size increases. This behaviour is cited when cautioning against oversized studies: "Moreover, one should be cautious that extremely large studies may be more likely to find a formally statistical significant difference for a trivial effect that is not really meaningfully different from the null" (Ioannidis, 2005). Lindley's paradox can be circumnavigated by precluding frequentist inferences on trivial effect sizes. The main point of this appendix is that we get this protection for free, if we avoid oversized studies. In what follows, we pursue this issue in terms of sensitivity of frequentist tests to trivial effects.

Sensitivity and effect sizes

Here, we examine how sensitivity under classical inference depends upon effect and sample sizes. Given a single-sample *t*-test and Gaussian assumptions about random effects, it is easy to approximate the sensitivity $(1-\beta)$ as a function of effect and sample size n. Sensitivity is simply the probability of rejecting the null hypothesis correctly for a fixed specificity, here $\alpha = 0.05$. *t*-Statistics based on samples from a population with a true effect size d have a non-central *t*-distribution $T(t:\delta,v)$ with non-centrality parameter $\delta = d\sqrt{n}$ and degrees of freedom $v = n-1$. This means sensitivity is

$$\begin{aligned} 1-\beta(d) &= \int_{u(\alpha)}^{\infty} T(t : d\sqrt{n}, n-1)dt \\ \alpha(0) &= \int_{u(\alpha)}^{\infty} T(t : 0, n-1)dt. \end{aligned} \qquad 4$$

Where $\beta(d)$ is the probability of a Type 1 error and $u(\alpha)$ is the threshold for the Students *t*-statistic that controls the false-positive rate $\alpha(0)$ or specificity under the null hypothesis that $d = 0$. This sensitivity is shown in image format in Fig. 3 (upper panel). It can be seen that sensitivity reaches 100% for high effect and sample-sizes. This iso-sensitivity line (at 60% sensitivity) shows that as sample size increases, one becomes sensitive to smaller effect sizes. This is important for two reasons.

First, it shows that if one finds a significant effect with a small sample size, it is likely to have been caused by a large effect size. This is important because it means that a significant result in a small study requires a larger effect size than the equivalent result in a large-sample study. More formally, for any given *t*-statistic, Cohen's effect size $d = t/\sqrt{n}$ decreases with the number of subjects. In other words, if your scientific report is critiqued because your significant result was based on a small number of subjects, you can point out:

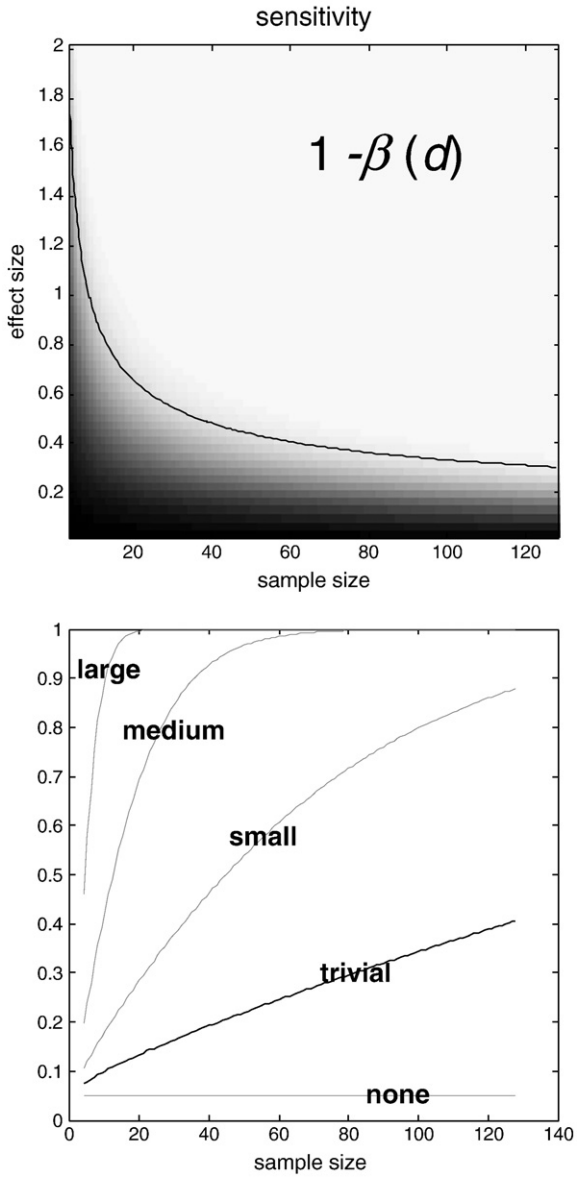


Fig. 3. Upper panel: sensitivity as a function of effect and sample-size. White corresponds to 100% sensitivity. The solid line is an iso-contour at 60% sensitivity. Lower panel: selected sensitivity curves as a function of sample-size for the four canonical effect sizes in Fig. 1.

> "*The fact that we have demonstrated a significant result in a relatively under-powered study suggests that the effect size is large. This means, quantitatively, our result is stronger than if we had used a larger sample-size.*"

The conflation of significance and power is not an infrequent mistake. Lenth discusses some common misconceptions like:

> "*Not only is it significant, but the test is really powerful!*" or "*The results are not significant … because the test is not very powerful.*" (Lenth, 2001).

The second key observation from Fig. 3 is that the sensitivity to trivial effect sizes increases with sample size (lower panel). Ultimately, with very large sample sizes, this sensitivity will reach 100%. This means that large cohort designs are sensitive to trivial effects and should be interpreted with caution. This suggests that there is some optimal compromise between under and overpowered designs:

A loss-function analysis

One way of optimising sample size is to consider what we want to make an inference about. If we are interested in non trivial effects, we could create a loss-function that placed sensitivity to large effect sizes in opposition to sensitivity to trivial effect sizes. For example, if the cost of detecting an effect was one for trivial and minus one for large effect sizes. The expected loss would be

$$\ell = \beta(1) - \beta\left(\frac{1}{8}\right). \qquad 5$$

This loss-function assumes that the gain in sensitivity to large effect sizes is offset by an increase in sensitivity to trivial effect sizes. The ensuing cost function is shown in Fig. 4 (upper panel) and has a minimum at around $n = 16$. In short, if we wanted to optimise the sensitivity to large effects but not expose ourselves to trivial effects, sixteen subjects would be the optimum number. One could argue that sensitivity to medium effects was as important as insensitivity to trivial effects (the ensuing loss-function $\ell = \beta(\frac{1}{2}) - \beta(\frac{1}{8})$ is shown in the lower panel and has a minimum at 50 subjects, which concurs

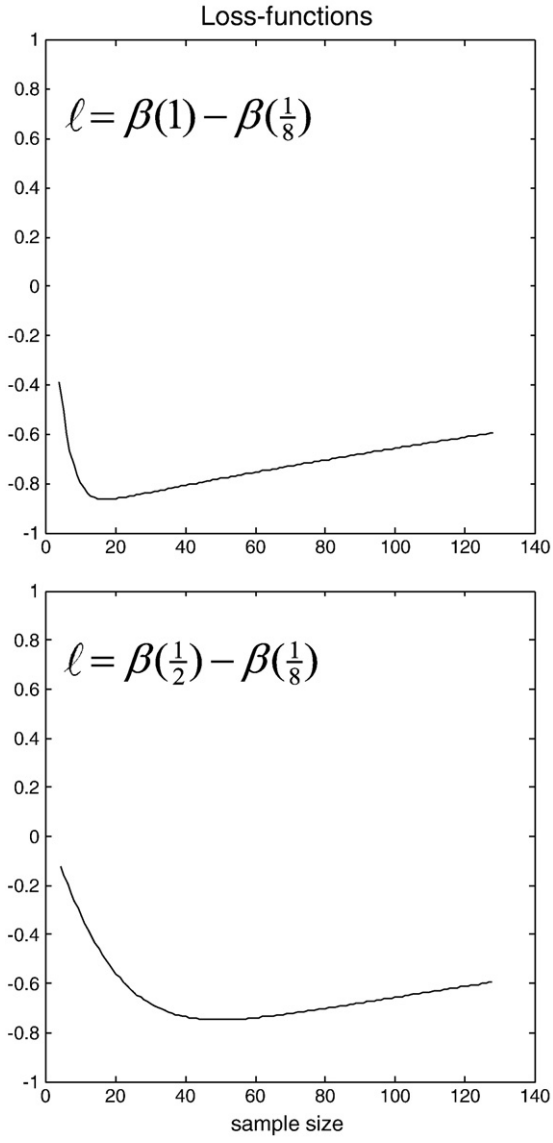


Fig. 4. Loss-functions pitting sensitivity to large (upper panel) and medium (lower panel) effects against sensitivity to trivial effects.

with the conclusions of Simon, 1987). However, recall that medium effect sizes only support a 60% classification accuracy and could be caused by a prevalence of just 20% under our binomial heuristic (see Table 2).

It is easy to estimate the standardised effect size, this is just the t-value divided by the square root of the number of subjects. This follows because the t-statistic $t = \sqrt{n}d$, where $d = t/\sqrt{n}$ is Cohen's point estimator of effect size. Fig. 5 shows $d = u(\alpha)/\sqrt{n}$ for a specificity of $\alpha = 0.05$. It can be seen that with 16 subjects one would always report effect sizes that were estimated to be about 0.4 or above (i.e. small to medium). However, with 50 subjects it is possible to report estimated effect sizes that are trivial to small. It is important to note that the true effect size could be lower because there is uncertainty about the estimate (see Appendix 2).

Protected inference

Hitherto, we have assumed that inference is based on controlling false positive rates under the null hypothesis. In the lower panel of Fig. 3, this control is reflected in the flat line (labelled 'none'), showing that the sensitivity to null effect sizes is 0.05. In other words, specificity is the same as sensitivity to null-effects. This somewhat unconventional perspective on specificity suggests something quite interesting. It means we can suppress sensitivity, not to null effects sizes but to trivial effect sizes. This is easy to do by replacing $\alpha(0)$ in Eq. (4) with

$$\alpha(d) = \int_{u(\alpha)}^{\infty} T(t : d\sqrt{n}, n-1)dt. \qquad 6$$

This fixes the sensitivity of the t-test to a constant and small level $\alpha(d)$ if the true effect size is d. Fig. 6 shows the sensitivity to different effects sizes using a specificity of $\alpha(\frac{1}{8}) = 0.05$. This application of classical inference protects against trivial effects by ensuring sensitivity to trivial effect sizes is small and does not increase with sample size. However, it means sensitivity to null effects decreases with sample size — this is the classical specificity $\alpha(0)$, shown at a different scale in the lower panel. In other words, to protect oneself against inferring an effect is present, when it is trivial, one can increase classical specificity with sample size. Under this protected inference, there is no marginal cost to increasing the number of samples because $\beta(\frac{1}{8}) = 0.05$ is fixed (see Eq. (6)) and one could sample as many subjects as possible.

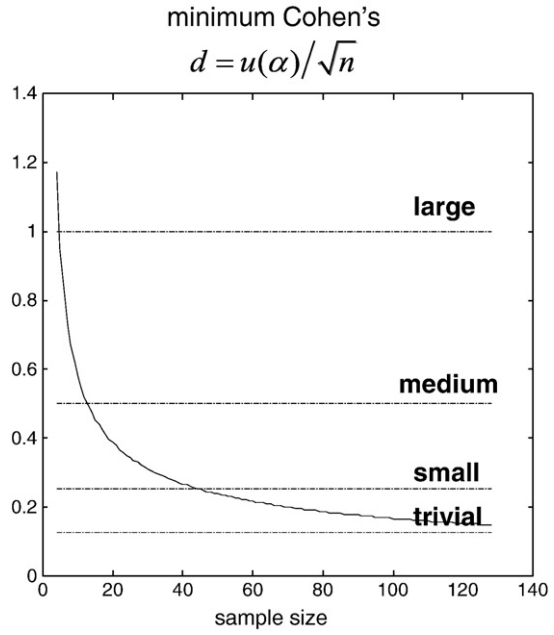


Fig. 5. Estimated effect size as a function of sample-size for t-values that control specificity at 0.05.

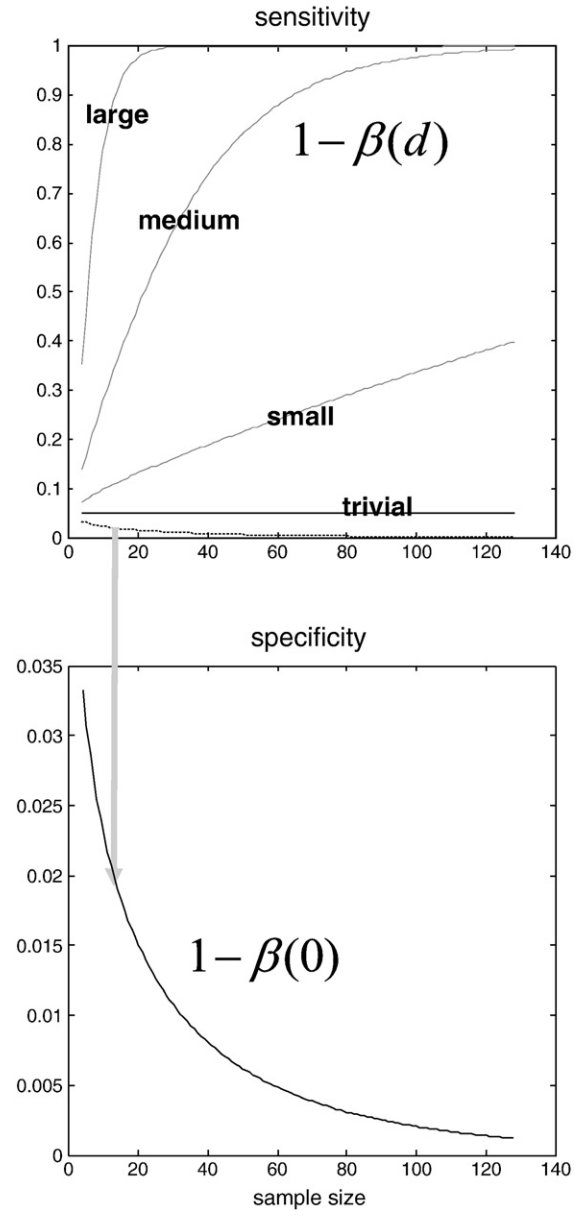


Fig. 6. Upper panel: selected sensitivity curves as a function of sample-size for the four canonical effect sizes — using protected inference that ensures the sensitivity to trivial effect sizes is 0.05. Lower panel: the equivalent sensitivity in classical terms (c.f. the sensitivity to null effects — "none" in the upper panel).

Conclusion

In conclusion, we have framed a fallacy of classical inference in terms of effect sizes and have argued that the optimum experimental design should sensitize inference to large effect sizes, while desensitizing inference to trivial effect sizes. For classical inference based on $\alpha(0)$, this leads to an optimal number of subjects of about 16. Clearly, the arguments above are heuristic and rely upon a fairly rough categorization of effect sizes. However, this treatment is general and reinforces the concept that designs can be overpowered. In short, if you cannot demonstrate a significant effect with sixteen subjects, it is probably not worth demonstrating. Having said this, the adage "you can never have enough data" is also true, provided one

takes care to protect against inference on trivial effect sizes — for example using protected inference as described above.

One could argue that small or even trivial effects are interesting, if they are expressed consistently over subjects but measured inaccurately. This is because a trivial standardised effect size could reflect a large effect size that has been dwarfed by noise. However, there is no way of knowing whether inter-subject variability is due to true variations in an effect or measurement noise. Without an independent way of partitioning the error variance, one cannot infer whether the true effect size differs from that measured. In this sense, it is probably best not to report *p*-values associated with trivial effect sizes, unless they are qualified.

Can true but trivial effect sizes can ever be interesting? It could be that a very small effect size may have important implications for understanding the mechanisms behind a treatment effect — and that one should maximise sensitivity by using large numbers of subjects. The argument against this is that reporting a significant but trivial effect size is equivalent to saying that one can be fairly confident the treatment effect exists but its contribution to the outcome measure is trivial in relation to other unknown effects — that have been modelled as random effects.

In summary, "the best-supported alternative hypothesis changes with the sample size, getting closer and closer to the null as the sample size increases. Thus, *p*-values should not be considered alone, but in conjunction with point estimates and standard errors or confidence intervals or, even better, likelihood functions" (Senn, 2001; p202).

Appendix 2. Effect sizes and predictions

In recent years, there has been some disquiet about reporting effect sizes in neuroimaging. In particular, some authors (and reviewers) have expressed concerns about reporting effect sizes in voxels or regions of interest that have been selected on the basis of their *p*-values (Kriegeskorte et al., 2009; Vul et al., 2009). It is not uncommon to ask whether the effect size should be estimated using independent data and whether the reporting of a *p*-value and effect size represents some form of biased sampling or 'double dipping'. The purpose of this appendix is to clarify the distinction between inference and estimation and distinguish between (in-sample) effect sizes and (out-of-sample) predictions. This distinction resolves tensions between reporting effect sizes and unbiased predictions of treatment effects.

Inference and (in-sample) effect sizes

We shall be concerned with two sorts of statistical procedures — inference and estimation. Classical inference means (here) the use of inferential statistics such as *p*-values to test hypotheses about treatment effects. Conversely, estimation is concerned with predicting new outcomes from data in hand — of the sort seen in machine learning and cross validation. Heuristically, inference is used to test hypotheses about the presence of a treatment effect (e.g., do subjects respond to a drug), whereas estimation is concerned with predicting a treatment effect (e.g., what effect will a drug have on a responsive patient). The goals of inference and estimation are distinct and call on different procedures.

Classical inference involves rejecting the null hypothesis based on an inferential statistic, such as a *t*-statistic and a suitably adjusted *p*-value. It is standard practice to complement the inferential statistic with an (in-sample) effect size — a practice encouraged by the APA Task Force on Statistical Inference (Wilkinson and APA Task Force on Statistical Inference, 1999). In this context, an in-sample effect size is a descriptive statistic that measures the magnitude of a treatment effect, without making any statement about whether this measure reflects a true effect. These effect sizes complement inferential statistics such as *p*-values and facilitate the quantitative interpretation of a result: see Ferguson (2009) for a fuller discussion. Crucially, the in-sample effect size is a statement about the data in hand — not about the true effect size or an estimate given independent (new) data.

As noted above, effect sizes can be standardised or unstandardised (see Table 1). Examples of standardised effect sizes include Cohen's *d*, the correlation coefficient and coefficient of determination (or proportion of variance explained). Unstandardised effect sizes include parameter estimates and regression coefficients. Standardised effect sizes express treatment effects relative to random effects and therefore have a less direct quantitative interpretation — "we usually prefer an unstandardised measure (regression coefficient or mean difference) to a standardized measure" (Wilkinson and APA Task Force on Statistical Inference, 1999, p. 599).

Clearly, large effect sizes and small *p*-values go hand-in-hand. In terms of point estimators, the in-sample effect size is proportional to the corresponding test statistic (see Table 1). This means that (on average) the in-sample effect size is an overestimate of the true effect size, because we only report in-sample effect sizes for data with large test statistics — some of which could be large by chance. This inherent bias increases for higher (more conservative) thresholds or, alternatively, small effects (Brand et al., 2008). Fig. 7 tries to illustrate this point using simulated data. It can be seen (in the upper panel) that increasing the threshold will select results that, by chance, have bigger test statistics and effect sizes. The ensuing bias can be particularly acute when performing large numbers of univariate tests (as in the analysis of neuroimaging data) because high thresholds are required to control false positive rates. In short, high thresholds necessarily inflate type II errors (false negatives) and in-sample effect sizes. This is the price paid for controlling family wise error rates when testing multiple hypotheses.

Estimation and (out-of-sample) predictions

The inherent bias above only arises because we fail to report the effect sizes when the statistical test is not significant. However, there are situations where this is unavoidable — for example in neuroimaging, where one has to select voxels or regions in which to estimate the effect size. In these situations, one can select regions using inferential statistics and estimate the true size of the effect using independent data or cross validation schemes; such as split half or leave-one-out procedures. These estimates of effect size are referred to as out-of-sample predictions and provide unbiased estimates of the true effect size. However, if some of the data are used to detect effects and the remaining data are used for out-of-sample predictions, significant effects will be missed. This is because the efficiency of the detection (inference) is comprised by only testing some of the data — by the Neyman–Pearson lemma. Fig. 7 illustrates this point and shows that one can either test for effects efficiently, accepting effect sizes are biased; or one can estimate effect sizes in an unbiased fashion, accepting that inference is inefficient. Formally, this can be regarded as an instance of the well-known bias-efficiency (variance) trade-off. Put informally, you cannot have (detect) your cake and eat (estimate) it.

Conclusion

The difference between (in-sample) effect sizes and (out-of-sample) predictions has led to some confusion about how to report statistical results properly. This confusion can be resolved by distinguishing between the use of inferential statistics to *detect* treatment effects and the use of out-of-sample predictions to *estimate* treatment effects. In short, one can either use classical inference to report significant effects in terms of *p*-values and (in-sample) effect sizes. Alternatively, one can use cross validation schemes to provide out-of-sample

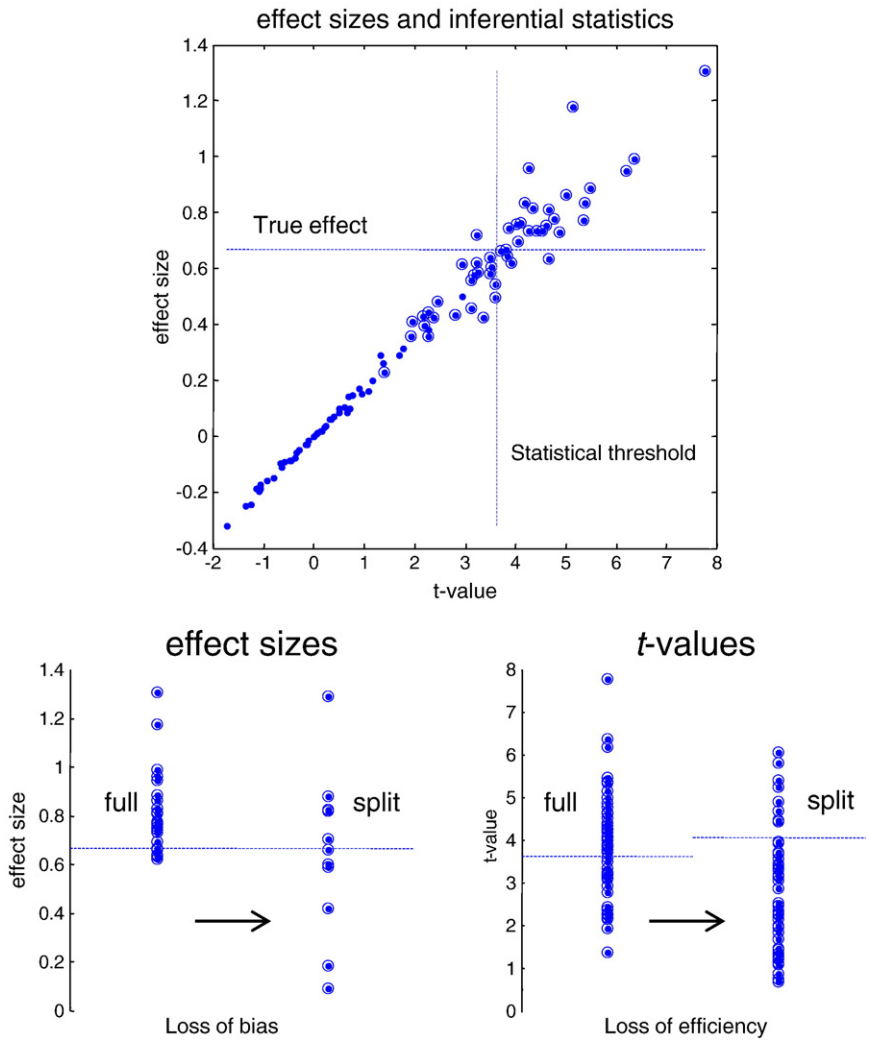


Fig. 7. An illustration of the trade-off between bias and efficiency: these results are based upon a simple treatment effect — namely, a large effect size observed in the presence of random Gaussian effects with a standard deviation of one. Data were simulated for 100 subjects (or regions), each generating 32 samples. The true effect was set at zero for half of the subjects (or regions) and two thirds for the other half. The data were then analysed using a one sample *t*-test to test the null hypothesis that the treatment effect was zero. The data were then split into two (16 sample) halves. The first (training) dataset was used to compute *t*-statistics, which were thresholded with Bonferroni correction to identify subjects (or regions) for subsequent out-of-sample estimates of the treatment effect. The upper panel shows the in-sample predictions (effect size) plotted against the inferential statistic (*t*-statistic). The circled dots correspond to subjects (or regions) that truly expressed an effect. The horizontal line corresponds to the true treatment effect and the vertical line to the Bonferroni corrected threshold on the *t*-statistic. This scatterplot shows that effect sizes and inferential statistics are highly correlated and that, necessarily, effect sizes are generally larger than the true treatment effect — when reported for tests that survive a high threshold. The lower panels compare and contrast the predictions from the classical inference (full) and a split-half (cross validation) procedure. The left panel (left dots) shows the effect size (in-sample prediction of the treatment effect) for all significant subjects or voxels. The circled dots correspond to subjects (or regions) showing true effects. The corresponding out-of-sample predictions based on parameter estimates from the second (test) dataset, selected on the basis of a significant *t*-test of the first (training) data are shown on the right. These results demonstrate that the out-of-sample predictions are an unbiased estimate of the true treatment effect — however, there are many fewer subjects (or regions) for which predictions are made, because their detection was less efficient. The right panel illustrates this in terms of inferential statistics: the dots on the left are the *t*-statistics from the full analysis of all subjects (or regions) showing a true effect. The corresponding distribution on the right shows the *t*-statistics from the same analyses of the training data in the split-half procedure. Not only is the threshold for the split-half *t*-statistic higher but also the statistics are generally lower. These results illustrate the fact that one can either have an unbiased (out-of-sample) prediction of the treatment effect, or an efficient test for discovering treatment effects but not both at the same time.

predictions of effect sizes. Crucially, one cannot do both at the same time.

The distinction between reporting in-sample and out-of-sample estimates of effect size speaks to the issue of sampling bias, known colloquially in neuroimaging as the non-independence problem or double dipping. The issue is attended by some emotive and perhaps unfortunate rhetoric — like voodoo correlations. This may reflect the specious – plausible but false – nature of the underlying argument against reporting in-sample effect sizes: it is plausible because it highlights the sampling bias inherent in reporting in-sample effect sizes. It is false because it misinterprets in-sample estimates of effect size as out-of-sample estimates. On a lighter note, the argument has produced some amusing responses (Fisher and Student, 2012 — http://www.psyencemag.org/).

References

Brand, A., Bradley, M.T., Best, L.A., Stoica, G., 2008. Accuracy of effect size estimates from published psychological research. Percept. Mot. Skills 106 (2), 645–649.
Cohen, J., 1988. Statistical Power Analysis for the Behavioral Sciences, 2nd ed. Academic Press, New York, New York.
Desu, M.M., Raghavarao, D., 1990. Sample-size Methodology. Academic Press, Boston.
Ferguson, C.J., 2009. An effect size primer: a guide for clinicians and researchers. Prof. Psychol. Res. Pract. 40 (5), 532–538.
Fisher, A.Z., Student, S.T., 2012. A triple dissociation of neural systems supporting ID, EGO, and SUPEREGO. Psyence 335, 1669.
Friston, K.J., Holmes, A.P., Worsley, K.J., 1999. How many subjects constitute a study? Neuroimage 10 (1), 1–5.
Friston, K.J., Glaser, D.E., Henson, R.N., Kiebel, S., Phillips, C., Ashburner, J., 2002. Classical and Bayesian inference in neuroimaging: applications. Neuroimage 16 (2), 484–512 Jun.
Friston, K.J., Stephan, K.E., Lund, T.E., Morcom, A., Kiebel, S., 2005. Mixed-effects and fMRI studies. Neuroimage 24 (1), 244–252.

Holmes, A., Friston, K., 1998. Generalisability, random effects and population inference. Fourth Int. Conf. Functional Mapping of the Human Brain: NeuroImage, 7, p. S754.
Ioannidis, J.P.A., 2005. Why most published research findings are false. PLoS Med. 2(8) (e124), 696–701.
Kriegeskorte, N., Simmons, W.K., Bellgowan, P.S., Baker, C.I., 2009. Circular analysis in systems neuroscience: the dangers of double dipping. Nat. Neurosci. 12 (5), 535–540.
Lenth, R.V., 2001. Some practical guidelines for effective sample-size. Determin. Am. Stat. 55 (3), 187–193.
Lindley, D.V., 1957. A statistical paradox. Biometrika 44, 187–192.
Lipsey, M.W., 1990. Design Sensitivity: Statistical Power for Experimental Research. Sage Publications, Newbury Park, CA.
Lohrenz, T., McCabe, K., Camerer, C.F., Montague, P.R., 2007. Neural signature of fictive learning signals in a sequential investment task. Proc. Natl. Acad. Sci. U.S.A. 104 (22), 9493–9498.
Maxwell, S.E., Kelley, K., Rausch, J.R., 2008. Sample-size planning for statistical power and accuracy in parameter estimation. Annu. Rev. Psychol. 59, 537–563 Review.
Odeh, R.E., Fox, M., 1991. Sample-size Choice: Charts for Experiments with Linear Models, 2nd ed. Marcel Dekker, New York.
Senn, S.J., 2001. Two cheers for P-values. J. Epidemiol. Biostat. 6, 193–204.
Shuster, J.J., 1990. CRC Handbook of Sample-size Guidelines for Clinical Trials. CRC Press, Boca Raton, FL.
Simon, R., 1987. How large should a phase II trial of a new drug be? Cancer Treat. Rep. 71 (11), 1079–1085.
Vul, E., Harris, C.R., Winkielman, P., Pashler, H., 2009. Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. Perspect. Psychol. Sci. 4 (3), 274–290.
Wilkinson, L., APA Task Force on Statistical Inference, 1999. Statistical methods in psychology journals: guidelines and explanations. Am. Psychol. 54 (8), 594–604.
Zimmerman, D.W., 1994. A Note on the Influence of Outliers on Parametric and Nonparametric Tests. J. Gen. Psychol. 121 (4), 391–402.