Published in final edited form as:

Neuroimage. 2013 November 1; 81: 499–502. doi:10.1016/j.neuroimage.2013.02.056.

Ironing out the statistical wrinkles in “Ten Ironic Rules”

Martin A. Lindquist¹, Brian Caffo, and Ciprian Crainiceanu

Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University

Abstract

The article “Ten ironic rules for non-statistical reviewers” (Friston, 2012) shares some commonly heard frustrations about the peer-review process that all researchers can identify with. Though we found the article amusing, we have some concerns about its description of a number of statistical issues. In this commentary we address these issues, as well as the premise of the article.

1 INTRODUCTION

As statisticians working in the field of neuroimaging, it was with great interest we read the article “Ten ironic rules for non-statistical reviewers” (labeled TIR; Friston, 2012). TIR shares some commonly heard frustrations about the peer-review process that all researchers can identify with. Though we found the article amusing, we are concerned about how it presents a number of statistical concepts. These include, among other things, the discussion of power, effect size and sample size. However, we first discuss its premise.

We disagree with TIR's characterization of non-statistical reviewers. Unlike the implication of TIR, we find our non-statistical scientific colleagues to have a great deal of intuition for statistical thought, despite lacking formal training. Moreover, working in a highly collaborative environment has taught us that both experts and non-experts alike can have good and bad ideas about statistics (as well as every other field) and that the idea of sharp boundaries between domains is inaccurate and counterproductive.

We also disagree with the potential message implied by TIR for reviewers to scale back comments regarding the statistical points it raised. We stress that reviewers should not feel hesitant to raise their concerns because of the ironic critiques in TIR. In fact, many of the criticisms that TIR (seemingly) laments via sarcasm and irony are perfectly legitimate in appropriate contexts. It is often TIR's hypothetical author, not the hypothetical reviewer, who fails in appropriate statistical thinking and scientific engagement.

We approach a discussion of TIR as if the hypothetical reviewer and author were involved in a legitimate non-ironic discussion.

2 COMMENTS ON SAMPLE SIZE

Much of TIR is devoted to a defense of small sample sizes (e.g., Rules 4 and 5 and Appendix 1). A central tenet in TIR is that a significant result obtained with a small sample size is “stronger” than if it had been obtained with a larger sample size, because small sample tests cannot detect trivial or uninteresting effects.

¹Corresponding author: Martin Lindquist Johns Hopkins University Bloomberg School of Public Health Department of Biostatistics 615 N. Wolfe St. Baltimore MD 21205 mlindqui@jhsph.edu.

²This idea was originally suggested by Tal Yarkoni in a blog response to TIR (www.talyarkoni.org/blog).

According to TIR this derives from the “fallacy of classical inference”, which is based on the fact that the null hypothesis is generally false in a strict sense (i.e. the effect is never *exactly* zero). As a consequence, the null can always be rejected with sufficient sample size. We agree that researchers should be aware of these issues when performing sharp null hypothesis testing. In addition, though not explicitly mentioned in TIR, they should be aware that tests based on large sample sizes are more susceptible to biases masquerading as small effects. We also agree that if an effect is found using a small data set, then it is often well worth reporting and publishing. However, TIR goes further to suggest that having less data is better merely to avoid this quirk of sharp null hypothesis testing. We strongly disagree with this statement. After all, we believe it would be difficult to make an argument for less accurate parameter estimates and wider confidence intervals, which are other effects of using smaller sample sizes. In addition, it is more difficult to interpret statistically significant effects in small samples, as the sample size precludes from performing sensitivity analyses or checking certain assumptions. The latter include that the model was correct, there is no bias due to sampling or missing data, and there were no important unaccounted for relationships, such as confounders. Moreover, absence of replication (across data collecting sites, for example) and the possibility that there are small - but meaningful - effects (more on this below) that are missed, all plague small studies. Thus reviewers critiquing small sample sizes raise potentially legitimate concerns.

What should be made of TIRs statement that large sample sizes should not be used, as sharp null hypothesis tests may find small effects that are practically unimportant; an argument enforced by invoking the “fallacy of classical inference”? While it is undoubtedly true that as the sample size increases, smaller effects become significant, we prefer to think of this in terms of increased “statistical power”, which generally invokes more positive connotations. More data creates the possibility of detecting more subtle effects. Hypothesis tests cannot separate important, but subtle, and actually trivial effects, and in our opinion a “fallacy” only arises if one believes they can. TIR, on the other hand, considers this a serious problem and attempts to create a framework for automating the process of weeding out small effects by keeping sample sizes small. In our mind, the job of determining important effects lies in the hand of the researcher performing the test. As such, hypothesis tests should not be performed in a black box manner, as their interpretation will depend upon context and the scientific question under investigation. If one, as TIR, considers small effect sizes uninteresting, sharp null hypothesis can be abandoned for interval null hypotheses². More precisely, if $\beta_{\text{irrelevant}}$ is an effect that is considered to be practically irrelevant then the null hypothesis $H_0 : \beta = 0$ can simply be replaced with $H_0 : \beta - \beta_{\text{irrelevant}}$ and the “fallacy” disappears. In general, sharp null testing versus general alternatives reduces a great deal of information into a single binary decision that can mask the scientific evidence in favor or against the null.

So while we agree with TIR that larger sample sizes will tend to reject more subtle effects, we strongly disagree with limiting the sample size to circumvent this issue. For example, consider a researcher who seeks to estimate the average height of all adult males in the United States. Suppose the researcher samples 100 representative people from the population and uses this information to (i) estimate the average height and (ii) perform a hypothesis test to determine whether the average height is taller than some specific value, say 69 inches. Now consider that the researcher is given the option of sampling everyone in the population, i.e. performing a census. With this information the researcher can *exactly* compute the average height, thus negating the need for hypothesis testing. In this setting the statistical problem is more descriptive in nature than inferential and performing a hypothesis test is not a meaningful endeavor, as it does not provide the researcher with any additional information. While we prefer to have data on as many subjects as possible and adjust inferential procedures based on this information, TIR (appears to) suggest it is better to use a

smaller data set that allows for hypothesis testing. We have a hard time reconciling this way of thinking. Statistical tests are used to determine whether a parameter is significantly different from some pre-specified value. As the sample size increases it becomes easier to detect such differences. This isn't a fallacy in our opinion; this is simply what they are designed to do. If researchers find this problematic, we believe that they should ask different, more appropriate, questions instead of reducing the sample size. Performing hypothesis tests is not the sole goal of statistical inference. In general, in the neuroimaging context it is not clear that the approach based on separate analyses of voxels and p-values is optimal, as rejecting the hypothesis of no effect may not actually be the most interesting aspect of the analysis (see Lindquist and Gelman (2009) for further discussion).

In the remainder of this section we discuss a number of additional issues surrounding the “under-sampled study”, the “over-sampled study”, using TIRs formulation, whether Bayesian analysis circumvents the “fallacy of classical inference” and conclude with a general discussion of sample size considerations.

On small sample sizes

We believe that when the design is good and the data “perfectly” follows the specified model, then a signal found to be statistically significant in small sample sizes is very important and likely to remain significant even if more data are collected. The main problem we have with under-sampled studies is that with few subjects there is very limited room for asking questions about confounding variables (e.g., gender, age). Lessons from the epidemiology literature suggest that many large signals disappear or are strongly attenuated once large, well-designed studies were conducted and confounder corrections introduced. For these reasons, larger sample sizes are preferred. The commonly made arguments against larger sample sizes typically regard diminished statistical returns in the form of power when factoring in important non-statistical considerations, such as having to sacrifice animals, cost or treatment side effects. Hence, it is perfectly reasonable for scientific peer reviewers, who are themselves very aware of these non-statistical considerations, to inquire about the potential need for larger sample sizes.

As a potential reply to criticisms of small samples, TIR quips “We suspect the reviewer is one of those scientists who would reject our report of a talking dog because our sample size equals one!”. To be clear, this tongue in cheek example of the evidential validity of small sample hypothesis, while funny, is not an appropriate analogy. That only one talking dog is necessary to refute the hypothesis “no dogs talk” is an example of the use of simple falsifiability in deduction reasoning that bears no resemblance to the necessary complexity of required inductive evidence in an imaging study.

On large sample sizes

Equivalently, raising concerns over the consequences of very large sample sizes can also be valid. As TIR points out, large sample sizes can produce statistically significant effects of no practical importance (i.e. “trivial effects”). But what about small or subtle effects relating to important outcomes, such as death, autism and Alzheimer's disease? TIR implies that trivial effects are generally unimportant in neuroimaging studies without differentiating between trivial and subtle. In the absence of appropriate context, TIR is over generalizing; it is not difficult to envision potentially important effects that are small in neuroimaging. For example, imagine that activation in a single brain region has a subtle, but robust and reproducible, correlation with some trait of interest, or a longitudinal study shows a subtle change in connectivity over time. The information this provides may not be trivial in either its potential effect or impact, and the researcher disregards it at their own peril.

In discussing trivial effects, TIR (correctly) notes that hypothesis tests cannot separate between important, but subtle, and actually trivial effects. Alas, tests can only provide guidelines for determining effects that are significantly different from some null value. We note that much of the confusion related to this and many other of TIR's points could be alleviated via more wholistic analyses that jointly consider the outcome of hypothesis tests along with estimates, confidence intervals, exploratory plots and other summaries of the data, along with careful scientific thinking. Our suggestion is to consider the scientific goals, inferential and diagnostic claims, potential biases, issues with confounding and unobserved variables and the biological basis for the statistical model when considering the impact of a large sample size on inference.

On sample size and Lindley's Paradox

The appendix of TIR uses Lindley's Paradox to conclude that use of Bayesian analyses circumvents the so-called "fallacy of classical inference". However, we disagree with this assessment. To elaborate, Lindley's paradox (Lindley, 1957) describes a situation wherein Bayesian and frequentist approaches to hypothesis testing give rise to different results for certain types of prior distributions. The brilliant, seemingly paradoxical, example provided by Lindley has since given rise to much discussion in the statistics literature, including opposing interpretations of its implications. In Appendix 1, TIR states (non-ironically): "[Lindley's paradox] can be circumnavigated by precluding frequentist inferences on trivial effect sizes." We note that Lindley's paradox-type discrepancies between fixed error rate testing and thresholding Bayes factors can arise for a variety of reasons, and can occur for large and small sample sizes and for large and small effect sizes.

Furthermore, if one adopts the position of a frequentist employing Bayesian methodology, invoking Lindley's paradox circumvents nothing. Thresholding Bayes factors, for example, produces a hypothesis test, with its own type I error rate and power that suffers from all of the defects and benefits of hypothesis testing in general. That the Bayesian and frequentist tests yield different conclusions is a consequence of different operating characteristics (i.e. implicit alpha level).

If one adopts the perspective of a Bayesian, it is interesting to note that Lindley's example implies that as the prior is made more diffuse, the decision implied by thresholding Bayes factors increasingly favors the null, regardless of the effect size (see the appendix). Thus, thresholding Bayes factors must certainly be troubling to researchers, since, by choosing a flat enough prior, they conclude the null irrespective to the amount of evidence against the null!

This example seeks to illustrate that simply switching between statistical paradigms, yet still engaging in ostensibly the same procedure via thresholding likelihood ratios, does not adequately address the issues raised by TIR.

On sample sizes in balance

In summary, sample size discussions, both prior to conducting a study and post-hoc in peer review, should depend on a number of contextual factors and especially specifics of the hypotheses under question. A small sample size is perfectly capable of differentiating gross brain morphometry between, say, children and adults. However, thousands of participants may be necessary to detect subtle longitudinal trends associated with human brain activation patterns in disease. That is, it is information content that is important, of which number of study participants is only a proxy.

A priori sample size calculations can help greatly in this regard. However, they can easily be misleading. As an example, TIR Appendix 1 does not take into consideration multiple

testing and applies a high threshold rarely used in practice. If the threshold is lower than 5%, the required sample size increases. Unfortunately, TIR generalizes the results and implies (intentionally or not) the message that 16 participants is a benchmark sample size for neuroimaging. To be unambiguous, the specific number 16 is neither optimal, nor a reasonable rule of thumb.

3 COMMENTS ON OTHER STATISTICAL ISSUES

In Rule 6, TIR ironically criticizes reviewers questioning (distributional) model assumptions. The hypothetical author's argument regarding the Rician distribution represents a straw man factual misunderstanding on the part of the hypothetical reviewer, which should be easily corrected in the review process. However, more troubling than a slightly misinformed hypothetical reviewer is the hypothetical rebuttal, which implies that all studies employing smoothing and inter-subject linear models have Gaussian error distributions guaranteed by the Central Limit Theorem (CLT). While we mostly agree with the implied asymptotics, the tenor of the comment fails to note that: independence assumptions, degree of smoothing, number of participants and underlying distribution of the data all inform the validity of the CLT as an approximation. Moreover, the hypothetical reviewer may view the loss of information from over-smoothing as potentially more important than obtaining Gaussian error distributions.

Regardless, questioning model assumptions, distributional or otherwise, and analysis choices is an important aspect of statistical criticism. We encourage reviewers to continue to question distributional assumptions and require that authors defend them, regardless of the statistical paradigm (e.g., frequentist or Bayesian) that is used.

In Rule 7, TIR ironically criticizes reviewers for demanding cross validation. TIR invokes the Neyman-Pearson (NP) lemma as an argument. This seminal result in statistics states that when performing a significance test between two simple point hypotheses, the likelihood ratio test (which assumes a known distribution of the data under the null and alternative) is the most powerful test (see Casella and Berger, 2001, for an accessible treatment). TIR incorrectly extends this result to make statements such as “the efficiency of the detection (inference) is comprised by only testing some of the data – by the Neyman-Pearson lemma” and “it is easy to prove (with the Neyman-Pearson lemma) that classical inference is more efficient than cross validation”. Unfortunately, NP does neither, and it cannot be generalized to make these types of statements about sample size considerations and cross validation.

In addition, the statement “the non-parametric tests will, by Neyman-Pearson lemma, be less sensitive than the original likelihood ratio tests” is only true if the exact parametric assumptions of the likelihood ratio test are valid, precisely the point the hypothetical reviewer sought to make. Though a wonderful result, NP simply does not enable researchers to make the strong claims made in TIR.

In addition, the implication that significance testing and evaluating prediction errors via cross validation have equivalent inferential goals is incorrect, unless one is thinking at a level of abstraction far removed from the hypothetical reviewer's clear intention. In fact, contrasting the methods and goals of predictive (cross validated) inference versus classical is a well described topic in statistics and machine learning (Breiman, 2001).

In Rule 8, TIR does not actually address the substance of the hypothetical reviewer's comment. It only differentiates in-sample versus out-of-sample effect size estimation. The substance of the hypothetical reviewer's comment is that the effect size estimates are biased. In this case the hypothetical reviewer is asking a perfectly reasonable question and the response does not directly address the concern. That said, we are not opposed to authors

reporting in-sample effect sizes. Instead, we stress the importance of carefully declaring whether in-sample or out-of-sample effect size estimates are reported. Armed with full information, reviewers and readers can have meaningful input on potential biases and interpretation (for more information see Kriegeskorte et al., 2010, for example).

4 CONCLUSION

Peer reviewing in inter-disciplinary fields is tricky. In our own work, we routinely receive reviews asking us to expand on neuro-scientific discussions and decrease statistical discussions, while in the same review another referee asks for the exact opposite. Both concerns can be relevant, since reviewers read the manuscript from different points of view. Though a lot of emotion can bubble to the surface when initially reading reviews, we find that when revisited later, most comments were either objectively reasonable, or at least reasonable given the likely background of the referee. It is notable that all of the ironic statistical reviewer comments in TIR could arise as: legitimate criticisms, misunderstandings stemming from poor communication by the submitting author, or misstatements based on a reasonable lack of specific statistical knowledge by an otherwise expert referee.

Referees, associate editors and editors donate substantial time, almost always in good faith, to improve journals like Neuroimage. It is the authors' obligation to take reviews in equal good faith and convince the journal of a manuscript's merit. If needed, arguing with, learning from and educating the referees in the process. We hope that the message from TIR will not intimidate referees from airing out statistical defects that they believe are present in submitted manuscripts.

We recognize (as does TIR via irony) that it is difficult to be an expert on all areas of neuroimaging, the statistical aspects included. However, as statistics takes an increasingly central role in neuroimaging, it is important to promote collaboration, communication and education between statisticians, engineers, neuroscientists and others working in this complex area.

Finally we note that for readers sympathetic to the frustrations in the peer review process, thriving attempts to adapt or change it exist, such as PLoS One (<http://www.plosone.org/>) and research in public peer review (see Leek et al., 2011, for example).

Acknowledgments

We would like to thank Thomas A. Louis, Tor Wager and Charles Rohde for providing insightful comments that improved the content and presentation of our response.

APPENDIX

To illustrate Lindley's paradox, we follow closely the insightful presentation in Bernardo and Smith (Bernardo and Smith, 2000), page 394. Consider that $Y = (Y_1, \dots, Y_n)$ is a random sample from a distribution $N(\mu, \tau_0^{-1})$, where the precision (one over the variance) τ is known. Consider two models

$$M_0: \mu = \mu_0, \quad M_1: \mu \sim N(\mu_1, \tau_1^{-1}),$$

where μ_1 and τ_1 are the known mean and precision of the prior on the mean parameter.

Lindley (1957) showed that if $\bar{Y}_n = 1/n \sum_{i=1}^n Y_i$ then the Bayes factor for comparing models M_0 and M_1 is

$$B_{01}(Y, \mu_1, \tau_1, n) = \frac{P(M_0|Y)}{P(M_1|Y)} = \left(\frac{\tau_1 + n\tau_0}{\tau_1} \right)^{1/2} \frac{\exp \left[\frac{1}{2} \{ \tau_1^{-1} + (n\tau_0)^{-1} \}^{-1} (\bar{Y}_n - \mu_1)^2 \right]}{\exp \left\{ \frac{1}{2} n\tau_0 (\bar{Y}_n - \mu_0)^2 \right\}},$$

where $P(M|Y)$ denotes the posterior probability of a model M given the data Y . These calculations are made under the assumption that the prior probability of each hypothesis is equal to 0.5, though the exact same arguments hold for any choice of prior probabilities, as long as non-zero probability is assigned to model M_0 . Notice that $B_{01}(Y, \mu_1, \tau_1, n)$ approaches ∞ as τ_1 approaches 0 (with the remaining arguments held fixed), regardless of the difference between $\mu_1 - \mu_0$. There are two equally important consequences of this result:

1. For a given data set a prior that is noninformative enough (the precision parameter τ_1 is small enough) will always favor the null model M_0 .
2. For any given data set for which $\tau_0^{1/2} |\bar{Y}_n - \mu_0|$ is large enough to ensure that the null hypothesis M_0 is rejected using a frequentist approach a noninformative enough prior (small τ_1) will always favor the null hypothesis.

The apparent paradox comes from the fact that, from a frequentist perspective, Bayes factors become test statistics with an α level controlled by the dispersion of the prior, τ_1 . The dependence is counterintuitive, with more diffuse priors (smaller τ_1) corresponding to larger α levels.

References

- Bernardo, JM.; Smith, AFM. Bayesian Theory. John Wiley & Sons; Chichester, UK: 2000.
- Breiman L. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*. 2001; 16(3):199–231.
- Casella, G.; Berger, R. Statistical inference. Duxbury Press; 2001.
- Friston K. Ten ironic rules for non-statistical reviewers. *NeuroImage*. 2012
- Kriegeskorte N, Lindquist M, Nichols T, Poldrack R, Vul E. Everything you never wanted to know about circular analysis, but were afraid to ask. *Journal of Cerebral Blood Flow & Metabolism*. 2010; 30(9):1551–1557. [PubMed: 20571517]
- Leek J, Taub M, Pineda F. Cooperation between referees and authors increases peer review accuracy. *PloS one*. 2011; 6(11):e26895. [PubMed: 22096506]
- Lindley DV. A statistical paradox. *Biometrika*. 1957; 44:187–192.
- Lindquist M, Gelman A. Correlations and multiple comparisons in functional imaging: A statistical perspective (commentary on vul et al., 2009). *Perspectives on Psychological Science*. 2009; 4:310–313.