Contents lists available at ScienceDirect

Preventive Medicine

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

Review Article

To curb research misreporting, replace significance and confidence by compatibility

A Preventive Medicine Golden Jubilee article

Sander Greenland [a], Mohammad Ali Mansournia [b,*], Michael Joffe [c]

[a] Department of Epidemiology, Department of Statistics, University of California, Los Angeles, USA
[b] Department of Epidemiology and Biostatistics, School of Public Health, Tehran University of Medical Sciences, Tehran, Iran
[c] Department of Epidemiology & Biostatistics, Imperial College London, United Kingdom

ARTICLE INFO

Keywords:
Compatibility
Confidence interval
Credible interval
Hypothesis test
P-value
S-value
Significance test
Statistical inference

ABSTRACT

It is well known that the statistical analyses in health-science and medical journals are frequently misleading or even wrong. Despite many decades of reform efforts by hundreds of scientists and statisticians, attempts to fix the problem by avoiding obvious error and encouraging good practice have not altered this basic situation. Statistical teaching and reporting remain mired in damaging yet editorially enforced jargon of "significance", "confidence", and imbalanced focus on null (no-effect or "nil") hypotheses, leading to flawed attempts to simplify descriptions of results in ordinary terms.

A positive development amidst all this has been the introduction of interval estimates alongside or in place of significance tests and *P*-values, but intervals have been beset by similar misinterpretations. Attempts to remedy this situation by calling for replacement of traditional statistics with competitors (such as pure-likelihood or Bayesian methods) have had little impact. Thus, rather than ban or replace *P*-values or confidence intervals, we propose to replace traditional jargon with more accurate and modest ordinary-language labels that describe these statistics as measures of compatibility between data and hypotheses or models, which have long been in use in the statistical modeling literature. Such descriptions emphasize the full range of possibilities compatible with observations. Additionally, a simple transform of the *P*-value called the surprisal or S-value provides a sense of how much or how little information the data supply against those possibilities. We illustrate these reforms using some examples from a highly charged topic: trials of ivermectin treatment for Covid-19.

1. Introduction

The persistence of statistical misinterpretations is astonishing, especially those which one might have thought to have been laid to rest generations ago. Foremost is the misreporting of the relation of "statistical significance" and "confidence intervals" to reality. In the early 1900s warnings against concluding "no effect observed" because the observations were "not statistically significant" could be found alongside warnings against the converse fallacy of equating "statistical significance" to an important effect (Pearson, 1906; Boring, 1919; Tyler, 1931). Such cautions have been repeated hundreds of times since in commentaries, editorials and guidelines for progressive journals (Rothman, 1978; Rothman, 1986; Altman and Bland, 1996; Schmidt and Hunter, 1997; Sterne and Davey Smith, 2001; Gigerenzer, 2004; Greenland, 2011; Higgs, 2013; Gigerenzer and Marewski, 2015; Goodman, 2016; Greenland et al., 2016; Wasserstein and Lazar, 2016; Amrhein et al., 2017; Greenland, 2017; Wasserstein et al., 2019). Yet prominent medical journals continue to publish articles describing results with *P*-values above 0.05 as showing "no effect", and describing 95% interval estimates including the null value (of 0 for differences, 1 for ratios) as if they indicate there is no effect, when in fact the results leave considerable uncertainty about the actual effect size.

If, by analogy with medical practice, researchers adopt the motto "first do no harm to knowledge", then it is our duty to discourage authors and journalists from making unjustifiable claims about what studies show even if those claims align with what we think is true, and to

* Corresponding author at: Department of Epidemiology and Biostatistics, School of Public Health, Tehran University of Medical Sciences, PO Box: 14155-6446, Tehran, Iran.
E-mail address: mansournia_m@sina.tums.ac.ir (M.A. Mansournia).

https://doi.org/10.1016/j.ypmed.2022.107127
Received 22 June 2022; Accepted 27 June 2022

vigorously protest and oppose evidence distortions enforced by journals. We have thus been campaigning to revise statistical terminology to better reflect the realities of what standard procedures do and don't tell us. In doing so, we are not trying to ban any statistic or add any new methods. Rather, we seek to end harmful practices like using *P*-values or interval estimates as arbiters of what in ordinary terms is or isn't "significant", what we can be confident about, or what is or is not present or important. Our motivation is that these statistics only display one very narrow dimension of compatibility between analysis data and the assumptions of our analysis procedures (Greenland et al., 2016; Amrhein et al., 2019; Greenland, 2019b; Rafi and Greenland, 2020), making them far from sufficient for decisions about whether to pursue a possible effect or advise a treatment.

2. An example: Trials of ivermectin

In the pandemic year 2020, the question of whether inclusion of ivermectin in protocols for the prevention or treatment of Covid-19 became a highly charged topic, leading to many randomized trials of various protocols. Unfortunately, several of these studies were of doubtful quality, in some instances even appearing to have fraudulent data (Hill et al., 2022); for such cases, debate about their statistical presentation is pointless. We will thus consider a widely cited trial of low-dose ivermectin on symptoms of mild Covid-19 (López-Medina et al., 2021) for which such concerns were not raised, so that we may focus on proper interpretation of its statistical results. It reported:

> The median time to resolution of symptoms was 10 days (IQR, 9–13) in the ivermectin group compared with 12 days (IQR, 9–13) in the placebo group (hazard [rate] ratio for resolution of symptoms, 1.07 [95% CI, 0.87 to 1.32]; $P = 0.53$ by log-rank test). By day 21, 82% in the ivermectin group and 79% in the placebo group had resolved symptoms.

The outcomes were thus only slightly better in the ivermectin group: $(12-10)/12 = 17\%$ reduction in median resolution time, 7% higher rate of resolution, and $(82–79)/82 = 4\%$ more resolution at day 21. The abstract concluded "a 5-day course of ivermectin, compared with placebo, did not significantly improve the time to resolution of symptoms." While this sentence is ambiguous about the meaning of "significantly" and might be understood to mean only that the *P*-value exceeded 0.05, we think most readers would still take the quote to mean the actual effect was clinically insignificant or of no practical importance.

For such readers the quote could be entirely misleading because the actual effect in the trial is subject to considerable statistical uncertainty. The results are simply too imprecise to provide assurances about absence, presence, or even the direction of an effect; in particular, while the statistics fail to show that ivermectin has an effect, they also fail to show that the ivermectin protocol used in the trial did not improve time to resolution to a clinically important extent if a 30% increase in the resolution rate would qualify as "important". Instead, they refute the notion that the ivermectin protocol had the sort of dramatic effects reported by some earlier trials (Hill et al., 2022).

The authors and journal may have emphasized failure to show an effect and neglected to mention failure to show no effect in order to oppose the misguided promotion of ivermectin as a miracle drug. Their choice may then be seen as one of whether reporting should be guided completeness and accuracy vs. more behavioral-political goals. As if to illustrate this point, a popular medical newsletter covered the trial with a headline "Ivermectin Disappoints in Mild COVID-19— Colombian trial flop" and asserted that the ivermectin patients "did no better than a placebo group" (Walker, 2021). Such claims misrepresent the statistical results, for they take no account of the "random" in randomized trial and how that leaves uncertainty in the results (albeit a more measurable uncertainty than in observational studies). For example, the 95% "confidence" interval (CI) for the ratio of resolution rates for ivermectin vs. placebo extends from 0.87 to 1.32, which means that every rate ratio inside the interval has $p > 0.05$. Thus, using the conventional 5% significance cutoff adopted in the paper, we could say that the results are most compatible with or provide little evidence against anything from a 13% lower to a 32% higher symptom-resolution rate from the ivermectin treatment protocol used in the trial.

To translate the statistics into qualitative clinical terms, we'd have to agree on what percent improvement would be considered "clinically significant". If that were 25% and we accepted the traditional 0.05 criterion for deciding how to report this trial, then the results were indecisive: The *P*-value for a 25% improvement is 0.14, very compatible with the data, while even 30% higher and a 10% *lower* recovery rate for ivermectin also have $p > 0.05$ and thus are reasonably compatible with the data by the 0.05 convention. But then, would attention be maximized by a more accurate headline such as "Trial Fails to Provide Definitive Answers for Ivermectin in Mild Covid-19"? We suspect not, and believe that attracting attention via spinning of ambiguous results into definite "findings" is one culprit behind traditional statistical misinterpretations that pivot on hard thresholds for "significance". Nonetheless, we get $p = 0.0015$ for a 50% higher recovery rate; hence we could also say the trial is not very compatible with even that large of an effect for the treatment protocol. Therefore, a headline of "Trial Finds Evidence Against Large Benefits of Ivermectin in Mild Covid-19" would be defensible if 50% were considered large but 30% were not.

While the problem of ignoring statistical uncertainty has been lamented for many decades, the pandemic has accelerated the harm it brings to medical research and practice, not only through misreporting of results but also by aggravating distrust of science. One way it does so is by creating the appearance of conflict where there is none, fueling conspiratorial accounts in which positive reports of ivermectin are automatically rejected, ignored, or dismissed while negative reports sail into high-profile journals and are trumpeted in medical news. Yet, among other trials we do find "significant" reports such as (Ahmed et al., 2021), which states:

> Viral clearance in the 5-day ivermectin group was significantly earlier compared to the placebo group on days 7 and 14 (hazard ratio (HR) 4.1, 95% CI 1.1–14.7 ($p = 0.03$) and HR 2.7, 95% CI 1.2–6.0 ($p = 0.02$)).

The measured outcome in this study is a lab endpoint (viral clearance) instead of symptom resolution, and that or other differences may explain why this study seemed "positive" and the other seemed "negative". But with regards to clinical significance, any claim of conflict would be an illusion produced by the 0.05 cutpoint: This clearance study is far less precise than the Lopez-Medina symptom study, and thus compatible with effects ranging from doubtful to huge clinical significance, overlapping the symptom study within the crucial range of borderline effects.

To summarize our message: There are exceptional studies designed to provide definitive results, and if they succeed in gathering valid and highly precise data they may achieve that goal. But the vast majority of studies do not supply enough reliable information to provide the sharp conclusions suggested by accompanying headlines, even if their integrity is unquestioned. In the spirit of honest reporting, such studies should be treated as contributing evidence to the pool available for research synthesis rather than as supplying definitive results. We thus see a need for improving statistical presentations that encourage more realistic, modest goals, and we will now describe how to do that with the language of compatibility.

3. Simple approaches to reform

Statistical reforms need not be as extreme as (say) replacing frequentist with Bayesian methods. But they do need to recognize how seriously all but the most sophisticated readers take verbal descriptions and labels. The first step is to understand how labelling *P*-values as "significance levels" and interval estimates as "confidence intervals"

Box 1
How to calculate *P*-values for alternative hypotheses.

Suppose one wants a P-value for an alternative hypothesis to the null, say the hypothesis that the log hazard-rate ratio ln(HR) for the treated (or equivalently, the treatment coefficient in a proportional-hazards model) equals some alternative β. If b and SE are estimates of ln(HR) and the standard error of b, then an approximate 2-sided P-value for the hypothesis that the true hazard ratio is exp(β) can be found by looking up the *Z*-statistic $z = |b - \beta|/SE$ in a table of two-tail normal percentiles. Note usual P-value for the null hypothesis that ln(HR) = 0 can be found from the same formula by using β = 0, which makes $z = |b|/SE$. If all that is available is the estimate exp(b) of the hazard ratio HR and its 95% compatibility ("confidence") limits HR_L and HR_U, SE can be estimated from $SE = \ln(HR_U/HR_L)/3.92$, where 3.92 is 2•1.96 and 1.96 is the 95% two-tail percentile for a normal distribution. The same methods apply when using an odds ratio OR or a risk ratio RR and its limits in place of a hazard ratio HR.

Now suppose d and SE are estimates of a difference D in risks and the standard error of d. An approximate 2-sided P-value for the hypothesis that the true difference D is an alternative δ can be found by looking up the Z-statistic $z = |d - \delta|/SE$ in a table of two-tail normal percentiles. The standard error for d can be estimated from 95% limits D_L and D_U as $SE = (D_U - D_L)/3.92$. The same approach can be used for differences in approximately normal means, subject to substituting t-distribution percentiles for normal percentiles with appropriate degrees of freedom.

Box 2
Gauging compatibility with *P*-values and evidence with S-values (surprisals).

Suppose we have a P-value p for testing a hypothesis H (whether null or alternative) derived under a set of auxiliary assumptions A such as proper randomization and concealment of treatment assignment. An example is $p = 0.14$ for testing H: "25% improvement from ivermectin". The modern Fisherian *P*-value was conceived in a falsification framework in which "testing" meant "attempting to refute H" and so is often described as a measure of evidence against the tested hypothesis H. But the P-value is quantitatively reversed for this description: If all the auxiliary assumptions were correct, $p = 1$ would represent no evidence against H and p = 0 would represent evidence completely contradicting H (complete refutation of H). Hence we refer to p as gauging the compatibility of the data with H, given A, so that p = 1 represents complete compatibility and p = 0 represents complete incompatibility. If we have any doubts about the auxiliary assumptions, we can and should shift this interpretation to say p gauges the compatibility of the data with the combination of the target hypothesis H *and* the background assumptions A (Greenland, 2019b).

With this shift, a defect of *P*-values that remains is their scaling: P-values of 0.999 and 0.95 are as far apart as P-values of 0.001 and 0.05, but the former pair exhibits a trivial compatibility difference while the compatibility difference for the latter pair is huge. One measure of the evidence in p that addresses both this scaling problem and falsificationist goals is to ask what p would correspond to in a simple mechanical experiment (Greenland, 2019b; Rafi and Greenland, 2020; Cole et al., 2021). For example, if we made independent tosses of a coin to test H: "the tosses aren't loaded for heads", 3 heads in a row would give p = 1/8 = 0.125, whereas 4, 5, 6, 7, and 8 heads in a row would give p = 1/16, 1/32, 1/64, 1/128, 1/256 which round to roughly $p = 0.06, 0.03, 0.02, 0.01, 004$. In general, the number of heads in a row needed to get p is the binary S-value $s = \log_2(1/p) = -\log_2(p)$, also called the *Shannon information* or *surprisal* provided by p against H and A; its units (number of tosses) are called bits, a contraction of "binary digits" (Greenland, 2019b; Rafi and Greenland, 2020).

In the Lopez-Medina trial (López-Medina et al., 2021), we get p = 0.14 for H: "25% higher resolution rate" and $s = -\log_2(0.14) = 2.8$, meaning that P-value provides <3 bits (coin-tosses worth) of information against a 25% improvement from ivermectin; but we get $p = 0.0015$ for H: "50% higher resolution rate" and $s = -\log_2(0.0015) = 9.4$, >9 bits of information against 50% improvement from ivermectin (about the same amount of information that 9 heads in a row would provide against "the tosses aren't loaded for heads"). In contrast, the Ahmed et al. trial (Ahmed et al., 2021) reported $p = 0.03$ and 0.02 for no difference in viral clearance at days 7 and 14, which correspond roughly to s = 5 and 6 bits of information against there being no effect of ivermectin on clearance.

have fueled confusion and misinterpretations of the sort described elsewhere (Higgs, 2013; Greenland et al., 2016). Thus, a simple first step in reform is to replace the confusing and misleading jargon of "significance" and "confidence" with descriptions based on the logically weaker notion of compatibility of hypotheses or models with data (Amrhein et al., 2019; Greenland, 2019a; Greenland, 2019b), a term long used in theoretical statistics (Box, 1980; Bayarri and Berger, 2000; Robins et al., 2000). This leads to presenting *P*-values as measures of compatibility without reference to cutpoints - especially when the cutpoint is almost always the 0.05 default, rather than justified from error-cost considerations. It also leads to presenting interval estimates as compatibility intervals: Rather than instilling confidence in the results, the intervals should remind the reader that even the best observations are highly compatible with a broad range of possibilities.

Intervals still depend on the arbitrary cutpoint for inclusion (again, almost always the $p > 0.05$ default, which produces the 95% "confidence" claim). To reduce misimpressions from this dichotomization, we further advise that *P*-values be shown for alternatives as well as for null hypotheses, especially alternatives that would be considered minimally significant *in a clinical sense* (Poole, 1987; Amrhein et al., 2019; Greenland, 2019b; Rafi and Greenland, 2020; Cole et al., 2021; Amrhein and Greenland, 2022). As shown in Box 1, this is easily done using published confidence limits or common software outputs. A Stata command and an R function for finding *P*-values for alternative regression coefficients are provided in Appendix A.

Box 2 supplies more precise descriptions of how we can use P-values to gauge compatibility with different possibilities, along with the binary surprisal or *S-value* (negative base-2 log P-value) $s = -\log_2(p)$ to gauge evidence against those possibilities (Greenland, 2019b; Rafi and Greenland, 2020; Cole et al., 2021; Amrhein and Greenland, 2022). Table 1 shows the P-values and S-values from the Lopez-Medina trial (López-Medina et al., 2021) for several possible values for the hazard ratio.

As can be seen, a 50% higher resolution rate (HR = 1.50) with ivermectin has p = 0.0015, corresponding to roughly s = 9 bits of information against such a strong beneficial effect (see Box 1) (Greenland,

Table 1
P-values and binary S-values (surprisals) from the Lopez-Medina trial of ivermectin and Covid-19 resolution rate (López-Medina et al., 2021) for several hazard-rate ratio (HR) values.* Point estimate of 1.07 is where $p = 1$ and $s = 0$; 95% limits of 0.87, 1.32 are where $p = 0.05$ and $s = 4.32$.

Hazard ratio:	0.80	0.90	1.00	1.10	1.25	1.50
P-value p	0.0062	0.10	0.52	0.79	0.14	0.0015
S-value s	7.3	3.3	0.93	0.33	2.8	9.4

* Based on Wald (Z-score) statistic. S-value is $-\log_2(p)$.

2019b; Cole et al., 2021; Amrhein and Greenland, 2022). Thus the Lopez-Medina data do appear to refute the notion that ivermectin as administered in the trial provides large benefit.

4. Conclusions

The introduction of statistical methods (especially for experimental design) has contributed enormously to scientific progress, as well as to quality control and other industrial and engineering applications. Nonetheless, many researchers have come to recognize that statistical "significance" and "confidence" have done extensive damage to research reporting in exchange for claimed benefits about "error rates" – rates that in many settings apply only in idealized theory, with poor connection to actual errors in research (Greenland, 2017; Amrhein et al., 2019; McShane et al., 2019; Hirschauer, 2022) – and have consequently endorsed calls to abandon these labelings (Higgs, 2013; Amrhein et al., 2019; Greenland, 2019a; Greenland, 2019b; McShane et al., 2019; Wasserstein et al., 2019; Rafi and Greenland, 2020; Cole et al., 2021; Amrhein and Greenland, 2022).

We expect that these calls will continue to be opposed by traditionalists who believe words don't matter and that established usage deserves special respect simply for being established. We disagree, as we see much of the enduring problem as stemming from statistical tradition, and thus see a break with tradition as long overdue. Tradition is no scientific justification and is a poor substitute for criteria like logic and evidence. And the evidence is telling: For some 70 years research reporting has been subject to an uncontrolled experiment in enforced statistical jargon and conventions, leading to striking distortions in descriptions of study results.

It has long been argued (Yates, 1951; Rothman, 1978; Rothman, 1986; Altman and Bland, 1996; Sterne and Davey Smith, 2001; Greenland et al., 2016; Amrhein et al., 2017; Greenland, 2017; Amrhein et al., 2019; Calin-Jageman and Cumming, 2019; Wasserstein et al., 2019; Rafi and Greenland, 2020) that appropriate reporting should emphasize uncertainties in estimates, rather than falsely dichotomize "findings". Such reporting often produces the reaction among researchers that the results sound vague and unhelpful. Unfortunately, it is a fact of health and medical sciences that single studies of real patients can only reach tentative conclusions. Besides sample-size limitations (which limit statistical precision), trial interpretation must face deviations from ideals such as treatment nonadherence, withdrawals from study, and failure to include all patients who in practice would receive the treatment. We aim for reforms that will lead to clearer recognition of uncertainty and tentativeness, and that help explain why claims of "contradictory findings" are often mere products of identifying "findings" with dichotomous declarations.

Yet, as seen in ongoing editorial policies,[1] there is resistance and even vigorous counter-reformation which continues to enforce "significance" cutpoints and declares studies "negative" if they fail to meet them. Of course, hardened defenses of harmful practices and resistance to change should be familiar to all students of medical history, ranging from 19th century denials of germ theory to 20th century resistance to a role for microbial agents in certain cancers. It is time for the research community to recognize that statistical practice is suffering from the same sort of resistance to reform, and actively fight for accurate plain-language reporting of results by challenging referees and editors who continue to force destructive conventions on authors. There are now numerous citations that can be supplied as part of that challenge (Rothman, 1986; Altman and Bland, 1996; Sterne and Davey Smith, 2001; Gigerenzer, 2004; Greenland, 2011; Gigerenzer and Marewski, 2015; Goodman, 2016; Greenland et al., 2016; Wasserstein and Lazar, 2016; Amrhein et al., 2017; Greenland, 2017; Amrhein et al., 2019; Calin-Jageman and Cumming, 2019; Greenland, 2019b; McShane et al., 2019; Wasserstein et al., 2019; Rafi and Greenland, 2020; Cole et al., 2021; Amrhein and Greenland, 2022).

[1] E.g., from *JAMA Instructions for Authors*JAMA Network, 2022, Instructions for Authors. under "Key Points – Findings": "Report basic numbers only but state if results are statistically significant or not significant".

Contributors

All authors were involved in the concept and design of the study. SG wrote the first draft of the manuscript, and all authors contributed to subsequent revisions of the manuscript and approved the final version. SG is the guarantor.

Funding

None declared.

Declaration of Competing Interest

None declared.

Acknowledgements

The authors are most grateful for helpful review comments from Peter Bacchetti, John Carlin, Martin Wolkewitz, Valentin Amrhein, and an anonymous referee.

Appendix A. Stata command and R function for finding P-values for general hypotheses about regression coefficients

Suppose we are interested in finding an approximate P-value for the hypothesis $\beta = b$ in the logistic regression model of $\text{logit}(\pi) = \alpha + \beta x$ where $\pi = E(Y|X = x)$. The Stata and R codes are as follows:

Stata:

```
logit Y X.
test _b[X] = b
```

Where b is replaced by the particular value corresponding to the association of interest; for example b would be replaced by 0.693, the natural logarithm of 2, if one wanted the P-value for an odds ratio of 2 from a 1-unit increment in X.

R:

```
model<−lrm(Y ~ X, data = data name).
summary(multcomp::glht(model, "X = b"))
```

Where the glht function is from multcomp package.

The Stata and R codes above are applicable to a variety of regression models beyond the logistic, including other generalized linear models as well as mixed-effects models and survival models, simply by changing the model designator from "logit" in Stata or"lrm" in R. For example, if using a Cox proportional-hazards model for the time-to-event T, failure indicator D, and the same explanatory variable X, the first Stata line becomes.

```
stset T, fail(D).
stcox X, nohr.
```

and the first R line becomes.

```
model<−coxph(Surv(T, D) ~ X, data = data name).
```

References

Ahmed, S., Karim, M.M., Ross, A.G., Hossain, M.S., Clemens, J.D., Sumiya, M.K., Phru, C. S., Rahman, M., Zaman, K., et al., 2021. A five-day course of ivermectin for the treatment of COVID-19 may reduce the duration of illness. Int. J. Infect. Dis. 103, 214–216.

Altman, D.G., Bland, J.M., 1996. Absence of evidence is not evidence of absence. Aust. Vet. J. 74, 311.

Amrhein, V., Greenland, S., 2022. Discuss practical importance of results based on interval estimates and entire p-value functions, not point estimate and null p-values. J. Inf. Technol. (in press).

Amrhein, V., Korner-Nievergelt, F., Roth, T., 2017. The earth is flat (p> 0.05): significance thresholds and the crisis of unreplicable research. PeerJ 5 (e3544).

Amrhein, V., Greenland, S., McShane, B., 2019. Retire statistical significance. Nature 567, 305–307.

Bayarri, M., Berger, J.O., 2000. P values for composite null models. J. Am. Stat. Assoc. 95, 1127–1142.

Boring, E.G., 1919. Mathematical vs. scientific significance. Psychol. Bull. (16), 335.

Box, G.E., 1980. Sampling and Bayes' inference in scientific modelling and robustness (with discussion). J. Royal Stat. Soc. 143, 383–404.

Calin-Jageman, R.J., Cumming, G., 2019. The new statistics for better science: ask how much, how uncertain, and what else is known. Am. Stat. 73, 271–280.

Cole, S.R., Edwards, J.K., Greenland, S., 2021. Surprise! Am. J. Epidemiol. 190, 191–193.

Gigerenzer, G., 2004. Mindless statistics. J. Socio-Econ. 33, 587–606.

Gigerenzer, G., Marewski, J.N., 2015. Surrogate science: the idol of a universal method for scientific inference. J. Manag. 41, 421–440.

Goodman, S.N., 2016. Aligning statistical and scientific reasoning. Science 352, 1180–1181.

Greenland, S., 2011. Null misinterpretation in statistical testing and its impact on health risk assessment. Prev. Med. 53, 225–228.

Greenland, S., 2017. Invited commentary: the need for cognitive science in methodology. Am. J. Epidemiol. 186, 639–645.

Greenland, S., 2019a. Are confidence intervals better termed "uncertainty intervals"? No: call them compatibility intervals. BMJ 366:15381.

Greenland, S., 2019b. Valid p-values behave exactly as they should: some misleading criticisms of p-values and their resolution with s-values. Am. Stat. 73, 106–114.

Greenland, S., Senn, S.J., Rothman, K.J., Carlin, J.B., Poole, C., Goodman, S.N., Altman, D.G., 2016. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Am. Stat. 73 (suppl 1), 106–114.

Higgs, M.D., 2013. Do we really need the s-word? Am. Sci. 101, 6–9.

Hill, A., Mirchandani, M., Pilkington, V., 2022. Ivermectin for COVID-19: Addressing Potential Bias and Medical Fraud. Open Forum Infectious Diseases 9 (2) ofab645.

Hirschauer, N., 2022. Unanswered questions in the p-value debate. Significance 19, 42–44.

JAMA Network, 2022. Instructions for Authors.

López-Medina, E., López, P., Hurtado, I.C., Dávalos, D.M., Ramirez, O., Martínez, E., Díazgranados, J.A., Oñate, J.M., Chavarriaga, H., et al., 2021. Effect of ivermectin on time to resolution of symptoms among adults with mild COVID-19: a randomized clinical trial. JAMA 325, 1426–1435.

McShane, B.B., Gal, D., Gelman, A., Robert, C., Tackett, J.L., 2019. Abandon statistical significance. Am. Stat. 73, 235–245.

Pearson, K., 1906. Note on the significant or non-significant character of a sub-sample drawn from a sample. Biometrika 5, 181–183.

Poole, C., 1987. Confidence intervals exclude nothing. Am. J. Public Health 77, 492–493.

Rafi, Z., Greenland, S., 2020. Semantic and cognitive tools to aid statistical science: replace confidence and significance by compatibility and surprise. BMC Med. Res. Methodol. 20, 244.

Robins, J.M., van der Vaart, A., Ventura, V., 2000. Asymptotic distribution of p values in composite null models. J. Am. Stat. Assoc. 95, 1143–1156.

Rothman, K.J., 1978. A show of confidence. N. Engl. J. Med. 299, 1362–1363.

Rothman, K.J., 1986. Significance questing. Ann. Intern. Med. 105, 445–447.

Schmidt, F.L., Hunter, J.E., 1997. Eight common but false objections to the discontinuation of significance testing in analysis of research data. In: Harlow, L.L., Mulaik, S.A., Steiger, J.H. (Eds.), What if there Were no Significance Tests? Erlbaum, Mahwah, NJ, pp. 37–63.

Sterne, J.A., Davey Smith, G., 2001. Sifting the evidence-what's wrong with significance tests? Bmj 322, 226–231.

Tyler, R.W., 1931. What is statistical significance? Educ. Res. Bull. 115–142.

Walker, M., 2021. Ivermectin disappoints in mild COVID-19— Colombian trial flop. MedPage Today Mar. 4.

Wasserstein, R.L., Lazar, N.A., 2016. The ASA statement on p-values: context, process, and purpose. Am. Stat. 70, 129–133.

Wasserstein, R.L., Schirm, A.L., Lazar, N.A., 2019. Moving to a world beyond "p< 0.05". Am. Stat. 73, 1–19.

Yates, F., 1951. The influence of statistical methods for research workers on the development of the science of statistics. J. Am. Stat. Assoc. 46 (19-34), 32–34.