the wrong direction," from hypothesized effect to data instead of the other way around.

To illustrate this point, consider a similar problem in clinical practice: a physician reports to a patient that the result of her diagnostic test was positive. When the patient asks whether the test result could be wrong, she would be poorly served with an answer that the test is highly specific (i.e., given the absence of disease, it is very unlikely that the test result is positive). The patient's primary interest concerns the question: Given that my test result is positive, what is the probability of truly having the disease? This positive predictive value depends not only on specificity but also on sensitivity and the a priori probability of disease. A clinician weighs the evidence of a test result in view of her combined understanding of the biology of the disease, patient characteristics, and the pre-test probability of disease. That is why the interpretation is not left to the laboratory technician. Similarly, statistically significant results should be interpreted taking the prior expectation and plausibility of the null hypothesis into account. By intuition, people usually get this right. For example, consider a trial report with a statistically significant benefit obtained with a homeopathic, super-diluted remedy. People who do not believe in homeopathy are unlikely to believe the test results. Statistical testing is like interpreting a diagnostic test result by looking only at its specificity—that is, under the null hypothesis of no disease. Interpretation of statistical tests should also take into account the plausibility or likelihood of the alternative hypothesis, which depends on external or subjective knowledge. That is also why interpretation of study results should not be left to a simplistic statistical rule.

Statistical adjustment for multiple comparisons, as recommended by default in the ASN guidelines (1), results in an increased probability of false-negative results. It also undermines the interpretation of related endpoints (6). It is equivalent to a physician finding an abnormally low hemoglobin concentration in a patient but no longer judging it worthy of treatment because she also found iron deficiency. In their Figure 2, Sorkin et al. (1) show that the probability of at least 1 false-positive result occurring increases with the number of tests performed. This is true when test results are independent. Because, in practice, outcomes are typically related, the default should be to not adjust, and if adjustment is nonetheless done, it should be justified. Many other commentaries support this view, again summarized by Hurlbert et al. (3).

To assist in the interpretation of significance, the ASN guidelines (1) recommend that *P* values should be reported with a statement of the sample size, an estimate of the treatment effect, and its variability. This is 1 option, but it is very cumbersome and we do not believe that adding more statistical information would help the general reader in interpretating (non)significance. Why not demand instead that effects are reported with CIs? Contrary to what is stated in the AJCN guidelines (1), however, CIs do not give a range in which the true value of a parameter θ is expected to lie. It is not Bayesian; a 95% CI does not mean that the probability that the true value of the parameter is in the interval is 95%. Instead, as conceived by Neyman (7), a 95% CI encompasses a range of hypothesized effect sizes that have a *P* value exceeding 0.05—that is, hypothesized effect sizes within this range would be compatible with the sample estimate x_0 if the *P* value would be set at 0.05. In mathematical notation: $Pr(\theta|x_0) \neq Pr(x_0|\theta)$. Some additional pitfalls in the interpretation of CIs are outlined by Greenland et al. (4).

In conclusion, we agree that *P* values should not be banned. But, they should generally not be dichotomized, they should never be reported as (non)significant, and they should not be used unless there are good reasons for doing so. Even better is to separate results into a point estimate and its corresponding 95% CI. Because all information about statistical precision is contained in CIs, it is not necessary to additionally report *P* values.

The authors' responsibilities were as follows—HV: wrote the manuscript; EF, PvV, and AMP: edited the manuscript; and all authors: read and approved the final manuscript. The authors report no conflicts of interest.

Hans Verhoef
Edith Feskens
Pieter van 't Veer
Andrew M Prentice

From Wageningen University, Division of Human Nutrition and Health, Wageningen, The Netherlands (EF, PvV; HV, e-mail: hans.verhoef@wur.nl); and MRC Unit, The Gambia at London School of Hygiene and Tropical Medicine, Fajara, Banjul, The Gambia (AMP).

References

1. Sorkin JD, Manary M, Smeets PAM, MacFarlane AJ, Astrup A, Prigeon RL, Hogans BB, Odle J, Davis TA, Tucker KL, et al. A guide for authors and readers of the American Society for Nutrition journals on the proper use of *P* values and strategies that promote transparency and improve research reproducibility. Am J Clin Nutr 2021;114:1–6.
2. Wasserstein RL, Schirm AL, Lazar NA. Moving to a world beyond "p < 0.05." Am Stat 2019;73:1–19.
3. Hurlbert SH, Levine RA, Utts J. Coup de grâce for a tough old bull: "statistically significant" expires. Am Stat 2019;73(Suppl 1):352–7.
4. Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, Altman DG. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol 2016;31(4):337–50.
5. Bishop D. Rein in the four horsemen of irreproducibility. Nature 2019;568(7753):435.
6. Schulz KF, Grimes DA. Multiplicity in randomised trials I: endpoints and treatments. Lancet North Am Ed 2005;365(9470):1591–5.
7. Neyman J. Outline of a theory of statistical estimation based on the classical theory of probability. Phil Trans Royal Soc of London Ser A 1937;236:333–80.

doi: https://doi.org/10.1093/ajcn/nqab370.

Reply to Verhoef et al.

Dear Editor:

We thank our colleagues, Verhoef et al., for their thoughtful reply (1) to our article, "A guide for authors and readers of the American Society for Nutrition Journals on the proper use of *P* values, and transparency, to improve research reproducibility" (2).

Our colleagues state that we "directly contradict the explicit and well-considered recommendation to abandon statistical significance testing by the American Statistical Association." We do not. In the American Statistical Association (ASA) Statement on Statistical Significance and *P*-Values (3), the ASA does not state that *P* values should be banned but rather that they should be used in proper context.

JDS was supported by the Baltimore VA Medical Center GRE CC, NIA AG028747, and NIDDK P30 DK072488; CPD was supported by K24 DK104676 and P30 DK04056, BBH was supported by VA RR&D grant 5I21RX003169-02; TAD was supported by HD-085573 and HD-099080 and USDA CRIS 3092-51000-060.

> "Scientific conclusions … should not be based *only on whether a p-value passes a specific threshold* [emphasis added]. … Researchers should bring many contextual factors into play to derive scientific inferences, including the design of a study, the quality of the measurements, the external evidence for the phenomenon under study, and the validity of assumptions that underlie the data analysis. Pragmatic considerations often require binary, 'yes-no' decisions, but this does not mean that *p*-values alone can ensure that a decision is correct or incorrect."

We agree with the ASA. We state in our paper, "The *P* value alone does not fully communicate the result of an experiment," and "to allow the *P* value to be properly understood, additional information should be given. The additional information can include the sample size (and number of events for a categorical outcome), the estimated treatment effect and the precision of this estimate (e.g., SD or SE), or the estimated treatment effect and a measure its variability, such as a 95% CI."

Although we agree with our colleagues that *P* values should not be banned, we reject the assertion that we "fail to consider a *P* value as a conditional probability (i.e., the probability of findings in the sample at least as extreme as observed, given that in truth, there is no association in the sampled population)." We state that the *P* value is a conditional probability, "The *P* value gives the probability that an effect as extreme or more extreme than the observed effect (the change attributed to the intervention) would be seen if the intervention truly had no effect on the measured outcome." Our colleagues state that we "fail to address what is arguably the most important issue—namely, that most researchers interpret *P* values using flawed reasoning (i.e., it is not use, but the misuse of *P* values that is the main problem)." We agree, *P* values can be improperly interpreted! Our paper addresses the misuse of *P* values in the sections entitled "The *P* value: what it does and does not tell us" and "Common conditions that can lead to incorrect *P* values."

Our colleagues state, "To justify significance testing, ASN guidelines (1) state that medical and nutritional research often requires making a binary choice (e.g., to declare a treatment effective or not, to recommend 1 set of nutritional recommendations or another, to further investigate or move on to another question).… It is disheartening to see that, after decades of progress in thinking about these issues, this misleading and simplified approach is being promoted by ASN…. Dichotomizing *P* values implies that biology is discontinuous, which is seldom the case."

We disagree. Deciding if an experiment is a success or failure, or if a treatment is effective or not, is, by definition, a binary decision. This does not mean that a single experiment should determine one's belief about a proposed intervention, nor does it mean that the *P* value is the only metric that should be used in evaluating the result of an experiment. In our paper we say,

> "Medical and nutritional research often require making a binary choice, to declare a treatment effective or not, to recommend one set of nutritional recommendations or another, to further investigate or move on to another question. *Clinical guidelines and treatment decisions do not generally rely on a single study's outcome*, they are derived from an accumulation of effect estimates from the literature. *P values can help inform the accumulation and the subsequent binary decision*. [emphasis added]"

Our colleagues also say, "we would hope that a binary decision is not made only based on the presence or absence of an effect, but also and primarily on the magnitude of the effect." Although we agree that the magnitude of an effect should be taken into account when evaluating the results of an experiment, basing one's judgement primarily on the *magnitude* of an effect can lead to incorrect inferences. When the precision of measurement is low, the magnitude of an intervention's treatment may be misleading. The SE, CI, and *P* value can all help put the magnitude of an observed effect into perspective.

Our colleagues state,

> "Statistical adjustment for multiple comparisons, as recommended by default in the ASN guidelines (1), results in an increased probability of false-negative results. It also undermines the interpretation of related endpoints (6) …. In their Figure 2, Sorkin et al. (1) show that the probability of at least 1 false-positive result occurring increases with the number of tests performed. This is true when test results are independent. Because, in practice, outcomes are typically related, the default should be to not adjust, and if adjustment is nonetheless done, it should be justified."

We demonstrate in our Figure 2 that testing multiple hypotheses increases the type I error rate. The increase contributes to the nonreproducibility of studies. While it is true that, in our calculations for Figure 2, we assumed that tests were independent, the principle we demonstrate, that multiple testing increases the probability of false-positive results, is true regardless of whether the multiple events are independent (the probability of 1 test having a given value is independent of the outcome any other test) or dependent (the probability that 1 test has a given value is related to the value of another test). Although we could have taken into account the nonindependence of multiple tests (starting with the formula $P(A \text{ and } B) = P(A)P(B \text{ given } A)$), doing so would be of little practical value because the formula *1*) requires knowing a priori the probability that 1 test will be positive, given that the other test is positive, which is rarely known, and *2*) the probabilities typically differ for each pair of tests. Addressing the nonreproducibility of published studies, we said that "Performing multiple tests and reporting only a subset, such as those that are found to be statistically significant, should never be done, as this eliminates the ability to evaluate a significant finding in the context of its experiment-wise type-I error rate." Addressing the increase in the type I error rate caused by performing multiple tests and knowing that 1 solution may not fit all situations, we wrote, "When multiple tests are performed, if not obvious, the number of tests performed should be reported and an adjustment to the *P* value should *be considered* for multiple comparisons, including tests that are not reported. *If no adjustment for multiple comparisons is made, this should be clearly indicated and justified in the article's methods section* [emphasis added]." Justification might be proffered for no, or partial, adjustment for multiple comparisons when outcome measures are collinear—for example, change in fasting glucose concentration and hemoglobin A1c at 3 mo are reported in a study of response to therapy for diabetes. Another justification might be multiple comparisons performed exclusively among a priori–defined secondary outcome measures, as described by Armitage and Berry (4):

> "The danger of data dredging is usually reduced by the specification of one response variable, or perhaps a very small number of variables, as the primary endpoint, reflecting the main purpose of the trial. Differences in these variables between treatments is taken at their face value. Other variables are denoted as secondary endpoints. Differences in these are regarded as important but less clearly established, perhaps being subject to a multiplicity correction or providing candidates for exploration in further trials."

The decision not to adjust for multiple comparisons must be made with great care and clearly justified; multiple comparisons without appropriate adjustment and selective reporting of the results of multiple comparisons contribute to nonrepeatability of studies.

Another criticism of our paper was, "To assist in the interpretation of significance, the ASN guidelines (1) recommend that *P* values should be reported with a statement of the sample size, an estimate of the treatment effect, and its variability. This is 1 option, but it is very cumbersome, and we do not believe that adding more statistical information would help the general reader in interpretating (non)significance. Why not demand instead that effects are reported with CIs?" We disagree. We do not believe that reporting a result using a CI is less "cumbersome" than doing so with a *P* value; the burden is the same. Whether one uses *P* values or CIs, it is important to report the sample size (and when the outcome is dichotomous, the number of events). When reporting results using a *P* value, 4 values should be given: *1*) point estimate, *2*) measure of variability (typically SD or SE), *3*) *P* value, and *4*) number of events. Similarly, when reporting a result using a CI, 4 values should be given: *1*) point estimate, *2*) lower confidence limit, *3*) upper confidence limit, and *4*) number of events (and possibly a *P* value). In our paper we do not say that *P* values *must* be used, or that point estimates with confidence intervals *should not* be used. Either may be utilized but both should be used in the context of formal hypothesis generation and testing and described in sufficient detail to support reproducibility of research.

We wrote our paper to address misconceptions about the *P* value and to promote reproducibility of study findings. Our goal was to promote proper use and interpretation of *P* values and to increase authors' and readers' understanding of the strengths and weaknesses of null hypothesis testing. As stated in our paper, we agree with Verhoef et al.'s statement that "*P* values should not be banned." *P* values are a useful component of hypothesis-driven inquiry, but like any research tool, must be used with knowledge and interpreted correctly. Proper use of *P* values should be part of a larger effort to increase the reproducibility of study results.

The authors' responsibilities were as follows—JDS: wrote the manuscript; PAMS, AJM, AA, RLP, BBH, JO, TAD, KLT, CPD, and DT: edited the manuscript; and all authors read and approved the final manuscript.

Author disclosures: JDS is a member of the ASN's Statistical Review Board; MM, PAMS, AJM, and AA are Associate Editors of the *American Journal of Clinical Nutrition*; JO is Editor-in-Chief *Current Developments in Nutrition*; TAD is Editor-in-Chief of *The Journal of Nutrition*; KLT is Editor-in-Chief of *Advances in Nutrition*; CPD is the Editor-In-Chief of the *American Journal of Clinical Nutrition*; DKT is Academic Editor of the *American Journal of Clinical Nutrition,* RLP reports no conflicts of interest.

John D Sorkin
Mark Manary
Paul AM Smeets
Amanda J MacFarlane
Arne Astrup
Ronald L Prigeon
Beth B Hogans
Jack Odle
Teresa A Davis
Katherine L Tucker
Christopher P Duggan
Deirdre K Tobias

From the Baltimore VA Medical Center Geriatric Research, Education, and Clinical Center, Baltimore, MD, USA (BBH; JDS, e-mail: jsorkin@som.umaryland.edu); University of Maryland School of Medicine, Department of Medicine, Division of Gerontology, Geriatrics, and Palliative Medicine, Baltimore MD, USA (JDS); Department of Pediatrics, Washington University, St. Louis, MO, USA (MM); Division of Human Nutrition and Health, Wageningen University & Research, Wageningen, The Netherlands (PAMS); Nutrition Research Division, Health Canada, Ottawa, Ontario, Canada (AJM); Department of Biology, Carleton University, Ottawa, Ontario, Canada (AJM); Novo Nordisk Foundation, Centre for Healthy Weight, Hellerup, Denmark (AA); Independent Scholar (RLP); Department of Neurology, Johns Hopkins School of Medicine, Baltimore, MD, USA (BBH); Laboratory of Developmental Nutrition, Department of Animal Science, North Carolina State University, Raleigh, NC, USA (JO); USDA/ARS Children's Nutrition Research Center, Department of Pediatrics, Baylor College of Medicine, Houston, TX, USA (TAD); Department of Biomedical and Nutritional Research and Center for Population Health, University of Massachusetts Lowell, Lowell, MA, USA (KLT); Center for Nutrition, Division of Gastroenterology, Hepatology, and Nutrition, Boston Children's Hospital, and Department of Nutrition, Harvard TH Chan School of Public Health, Boston, MA, USA (CPD); and Division of Preventive Medicine, Brigham and Women's Hospital, and Harvard Medical School and Department of Nutrition, Harvard TH Chan School of Public Health, Boston, MA, USA (DKT).

References

1. Verhoef H, Feskens E, Van't Veer P, Prentice AM. ASN guidelines on *P* values. Am J Clin Nutr 2022;115(2):597–8.
2. Sorkin JD, Manary M, Smeets PAM, MacFarlane AJ, Astrup A, Prigeon RL, Hogans BB, Odle J, Davis TA, Tucker KL, et al. A guide for authors and readers of the American Society for Nutrition journals on the proper use of *P* values and strategies that promote transparency and improve research reproducibility. Am J Clin Nutr 2021;114(4):1280–5.
3. Wasserstein RL, Lazar NA. The ASA statement on p-values: context, process, and purpose. Am Stat 2016;70(2):129–33.
4. Armitage P, Berry G. Statistical methods in medical research. 2nd ed. Oxford (UK): Blackwell Scientific; 1987.

doi: https://doi.org/10.1093/ajcn/nqab371.