In her recent editorial, Deborah Mayo argues that journal editors "should avoid taking sides" regarding "heated disagreements about statistical significance tests." Particularly, they should not impose bans on statistical methods, such as Wasserstein et al.'s (2019) proposed ban on citing statistical significance and using p value thresholds. Were journal editors to adopt such proposals, Mayo argues, they would be acting under a conflict of interest (COI) of a special kind: an "intellectual" conflict of interest. Mayo's invocation of COIs invites consideration of how disputes over statistical method arise and are treated, and of COIs as a problem statistical methods must address. Here we consider these questions from the fields of experimental high energy physics (HEP) and oncology.

A potential objection to Mayo's claim is that journal editors are entrusted with decision-making power precisely to adopt and enforce standards on publication. Journal editors are responsible for making informed, reasoned judgments about standards that distinguish credible research conclusions. This requires making personal judgments about methodological standards. To have an intellectual interest in a policy is simply to think it is a good idea. Journal editors should act on good ideas when they have them.

This objection neglects a crucial feature of the dispute over significance testing and the "statistics wars." These disputes involve, in Mayo's words, "philosophical presuppositions." These concern fundamental aspects of scientific inquiry like "what is the purpose of a statistical test?" and "do the beliefs of investigators matter to the results of inquiry?" Philosophical debates are often thought to be never-ending or unresolvable in principle, and thus tend to have a bad reputation among non-philosophers. Perhaps some are, but even in those cases (and we don't think this is one), progress in clarifying what is at stake and eliminating non-viable positions is possible. So long as competing methodological approaches rest upon differing philosophical

presuppositions, to preclude the use of an approach as a matter of editorial policy would foreclose on the possibility of engaging that philosophical dispute at the level of scientific practice. The consequences of that foreclosure for a scientific discipline would be impoverishing.

Moreover, our consideration of statistical issues in HEP and oncology indicate that *encouragement* of philosophical reflection on statistical methods is both viable and important for addressing COIs that arise in the conduct of research. Although the field of conservation biology differs from both oncology and HEP in numerous significant respects, some points drawn from our examination of these fields are relevant: (1) statistical methods and standards must evolve to respond to problems that arise in practice, particularly with respect to potential biases and COIs; (2) methods and standards adopted collectively by a scientific community are also susceptible to criticism and revision by that community; and (3) the specific policy of banning significance and thresholds would discourage methods that require explicit treatment of error probabilities and thus (properly used) provide useful checks on COIs.

In HEP, there is a working principle governing the publication of discovery (or "observation") claims requiring *p* values corresponding to at least a 5σ deviation from the null prediction. The objection under consideration here would hold that the editors of *Physical Review* ought to be free to impose the 5σ requirement on discovery claims, even though there are some members of the HEP community who are critical of that policy, or who reject even the use of significance calculations.

However, the 5σ requirement for discovery evolved in the context of the existing statistical practices of HEP physicists, rather than being imposed by editorial fiat (Franklin, 2013). As such, it is an outcome of debates about standards of evidence within a default frequentist practice, in response to the problem posed by the interest that physicists have in

discovery (Staley, 2017a). That interest can come into conflict with an interest in avoiding erroneous discovery claims. HEP physicists realized that this COI, in combination with the problems of multiple trials and imperfectly understood sources of systematic error, required a higher standard for discovery. The adoption of the 5σ standard as a solution was predicated on the established usefulness of significance as a means of establishing statistical discrepancy between the data and the null prediction. The standard is the subject of regular criticism; calls for the replacement or revision of it are common and may become the locus of a new consensus (e.g., Lyons, 2013). Both the success of the standard and its susceptibility to revision are enabled by not having been imposed upon the HEP community by journal editors.

Moreover, the 5σ standard does not preclude alternative statistical methodologies. Frequentist analyses remain dominant in HEP, although the field has long incorporated Bayesian devices in a piecemeal manner. Justifications for such eclecticism tend to be local and pragmatic (Staley, 2017b). Statistical methodology is a lively topic of debate in the HEP community, facilitated by the PHYSTAT group, which holds regular workshops and conferences with contributions from physicists and statisticians, including frequentists and Bayesians (e.g., Prosper & Lyons, 2011). The ideas of both frequentist and Bayesian advocates find their way into practice and into print.

In the field of oncology, some are flirting with Bayesian statistics to move on from statistical significance testing and the use of *p* values. In fact, what many consider the world's preeminent cancer center, MD Anderson, has a strong Bayesian group that completed 2 early phase Bayesian studies in radiation oncology that have been published in the most prestigious cancer journal —*The Journal of Clinical Oncology* (Liao et al., 2018; Lin et al., 2020). This brings about the hotly contested issue of subjective priors and in medicine, one thinks about

Spiegelhalter's classic 1994 paper mentioning reference, clinical, skeptical, or enthusiastic priors, also using an example from radiation oncology (Spiegelhalter et al., 1994) to make his case. What if in practice, there is ample evidence that the subject matter experts have major COIs and biases so that their priors cannot be trusted? A debate raging in oncology is whether non-invasive radiation is as good as surgery for early stage lung cancer. Postoperative morbidity from surgery can range from 19-50% and 90-day mortality from 0–5% (Chang et al., 2021), making radiotherapy highly attractive as there are numerous reports hinting at equal efficacy with far less morbidity. Unfortunately, 4 major clinical trials were unable to accrue patients for this important question. Why could they not enroll patients? If a patient is referred to radiation oncology and treated with radiation, the surgeon loses out on the revenue, and vice versa. Dr. David Jones, a surgeon, notes there was no "equipoise among enrolling investigators and medical specialties… Although the reasons are multiple… I believe the primary reason is financial" (Jones, 2015). Dr. Hanbo Chen, a radiation oncologist, notes in his meta-analysis of multiple publications looking at surgery vs radiation that overall survival was associated with the specialty of the first author who published the article (Chen et al., 2018). Perhaps the pen is mightier than the scalpel!

Many of the specialists are highly biased with exposed COIs. Are we to trust their priors? Will the errors really be contained within the posteriors from these priors? In order to overcome this, you say that you want to use an uninformative or weakly informative prior as a statistical method to judge incoming data? Then what's the point of having prior knowledge, in this case the surgeons' and radiation oncologists' priors who are the subject matter experts, if you are not willing to use them? Indeed, as Mayo notes "It may be retorted that implausible inferences will indirectly be blocked by appropriate prior degrees of belief (informative priors), but this misses the crucial point. The key function of statistical tests is to constrain the human tendency to

selectively favor views they believe" (Mayo, 2021). If this statement holds for appropriate prior degrees of belief, how much more is it relevant when we can show that those involved have inappropriate prior degrees belief?

It is not the intent of his commentary to tell conservation biologists how to apply these lessons to their field, nor to deny the importance of editorial policies. As a large, interdisciplinary, and heterogeneous field, conservation biology will not easily achieve consensus of the sort achieved in HEP (approximate thought it may be). Agreed upon standards might arise only piecemeal, in sub-disciplines within the field. Such challenges do not lend themselves to solution by editorial dictate, however, and especially not one that encourages reliance on Bayesian priors with the potential to give rise to the problems that emerge in the oncological examples here considered.

Editors and the decisions they make are of course crucial to the enforcement of the methodological standards that do emerge from practice in a given field and are also essential for weeding out questionable research practices, regardless of statistical methodology. These roles, however, must not be conflated with the prohibition of a statistical methodology that, when properly employed, has significant potential for solving the kinds of COI problems we have discussed. Far better to promote engagement in the philosophical/scientific debates needed to understand what a given statistical method is and is not capable of, as the editors of this journal have done by publishing Deborah Mayo's editorial.

Chang, J. Y., Mehran, R. J., Feng, L., Verma, V., Liao, Z., Welsh, J. W., Lin, S. H., O'Reilly, M. S., Jeter, M. D., Balter, P. A., McRae, S. E., Berry, D., Heymach, J. V., Roth, J. A., & STARS Lung Cancer Trials Group. (2021). Stereotactic ablative radiotherapy for operable stage I non-small-cell lung cancer (revised STARS): Long-term results of a single-arm, prospective trial with prespecified comparison to surgery. *The Lancet Oncology*, 22(10), 1448–1457.

Chen, H., Laba, J. M., Boldt, R. G., Goodman, C. D., Palma, D. A., Senan, S., & Louie, A. V. (2018). Stereotactic Ablative Radiation Therapy Versus Surgery in Early Lung Cancer: A Meta-analysis of Propensity Score Studies. *International Journal of Radiation Oncology, Biology, Physics*, 101(1), 186–194.

Franklin, A. (2013). *Shifting standards: Experiments in particle physics in the twentieth century*. Pittsburgh: University of Pittsburgh Press.

Jones D.R. (2015) Do we know bad science when we see it? *The Journal of Thoracic and Cardiovascular Surgery*, 150(3):472-473.

Liao, Z., Lee, J. J., Komaki, R., Gomez, D. R., O'Reilly, M. S., Fossella, F. V., Blumenschein, G. R., Heymach, J. V., Vaporciyan, A. A., Swisher, S. G., Allen, P. K., Choi, N. C., DeLaney, T. F., Hahn, S. M., Cox, J. D., Lu, C. S., & Mohan, R. (2018). Bayesian Adaptive Randomization Trial of Passive Scattering Proton Therapy and Intensity-Modulated Photon Radiotherapy for Locally Advanced Non-Small-Cell Lung Cancer. *Journal of Clinical Oncology*, 36(18), 1813–1822.

Lin, S. H., Hobbs, B. P., Verma, V., Tidwell, R. S., Smith, G. L., Lei, X., Corsini, E. M., Mok, I., Wei, X., Yao, L., Wang, X., Komaki, R. U., Chang, J. Y., Chun, S. G., Jeter, M. D., Swisher,

S. G., Ajani, J. A., Blum-Murphy, M., Vaporciyan, A. A., … Mohan, R. (2020). Randomized Phase IIB Trial of Proton Beam Therapy Versus Intensity-Modulated Radiation Therapy for Locally Advanced Esophageal Cancer. *Journal of Clinical Oncology*, 38(14), 1569–1579.

Lyons, L. (2013). Discovering the significance of 5σ. arXiv:1310.1284.

Mayo, D. G. (2021). The statistics wars and intellectual conflicts of interest. *Conservation Biology*: in press.

Prosper, H. B., & Lyons, L. (Eds.). (2011). *Proceedings, PHYSTAT 2011 Workshop on Statistical Issues Related to Discovery Claims in Search Experiments and Unfolding*, CERN, Geneva, Switzerland 17-20 January 2011. CERN.

Spiegelhalter, D. J., Freedman, L. S., & Parmar, M. K. B. (1994). Bayesian Approaches to Randomized Trials. *Journal of the Royal Statistical Society*. Series A (Statistics in Society), 157(3), 357–416.

Staley, K. (2017a). Decisions, decisions: Inductive risk and the Higgs boson. In: Elliott, K. & Richards, T. (Eds.), *Exploring Inductive Risk.* New York: Oxford University Press. pp. 37–55.

Staley, K. (2017b). Pragmatic warrant for frequentist statistical practice: The case of high energy physics. *Synthese*, 194, 355–376.

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a World Beyond "p < 0.05." *The American Statistician*, 73(sup1), 1–19.