Philosophy of Science (2024), **00**, 1–11
doi:10.1017/psa.2024.13

PSA PHILOSOPHY OF SCIENCE ASSOCIATION
PHILOSOPHY OF SCIENCE

SYMPOSIA PAPER

Bamboozled by Bonferroni

Conor Mayo-Wilson

Department of Philosophy, University of Washington, Seattle, WA, USA
Email: conormw@gmail.com

(Received 21 April 2023; revised 09 February 2024; accepted 12 April 2024)

Abstract

When many statistical hypotheses are evaluated simultaneously, statisticians often recommend adjusting (or "correcting") standard hypothesis tests. In this article, I (1) distinguish two senses of adjustment, (2) investigate the prudential and epistemic goals that adjustment might achieve, and (3) identify conditions under which a researcher should *not* adjust for multiplicity in the two senses I identify. I tentatively conclude that the goals of scientists and the public may be misaligned with the decision criteria used to evaluate multiple-testing regimes.

Imagine a pharmaceutical company spends years developing a new cancer treatment. Because of the expense of drug development, the company collects extensive data during human trials. In particular, researchers collect data about hundreds of health outcomes other than cancer. When the data are analyzed, researchers find that treatment is associated with a reduction in breast cancer. Here's an instance of a more general question:

> **Question:** Should the pharmaceutical researchers alter their methods for analyzing the cancer data because the treatment for efficacy was assessed in *many* other ways?

According to many statisticians and scientists, the answer is yes. Let *multiplicity* refer to the act of evaluating many statistical hypotheses simultaneously. When multiplicity occurs, many statisticians and scientists recommend "correcting"[1] *p*-values so as to reduce the number of false-positive results.[2] Although Bayesian

[1] Henceforth, I say "adjust" rather than "correct" so as to avoid suggesting that adjustment is good or obligatory.

[2] See Lehmann and Romano (2008, chap. 9). The most common classical techniques are Bonferroni's method and the procedure of Benjamini and Hochberg (1995).

statisticians reject the use of *p*-values, many likewise argue that one's statistical methods should be adjusted for multiplicity.[3] This raises a very general question:

> **Central Question:** Under what conditions, if any, should statistical methods be adjusted for multiplicity? In what way should they be adjusted? And why?

The central question is important because as our computational power grows, so does our ability to evaluate thousands of policy-relevant statistical hypotheses in a matter of minutes.

Although statisticians have investigated the reliability of many adjustment procedures, few have clarified the central question. What exactly is adjustment? Can "adjustment" be defined without reference to particular statistical methods? If "adjusting" means "changing reported *p*-values," then devout Bayesian statisticians never adjust for multiplicity, as they avoid calculating *p*-values![4] So is there a sense of "adjustment" that renders classical and Bayesian approaches comparable?

The normative dimensions of the central question have also yet to be clarified. In what sense "should" one adjust for multiplicity? Is adjustment *rationally* required to achieve certain goals? If so, which goals? Is adjustment *epistemically* required to respect one's evidence? Is it *scientifically* required by norms of scientific inquiry? Is it *ethically* obligatory? If adjustment is not obligatory, is it permissible or good in any sense?[5]

Finally, answers to those normative questions depend on *who* or *what* is adjusting. Researchers can adjust reported *p*-values. But so can journal editors. Grant-giving agencies—like the National Institutes of Health (NIH)—can also adjust for multiplicity in various ways. Which, if any, of these decision-making bodies should adjust?

The main contribution of this article is to (1) distinguish two senses of adjustment, (2) investigate the prudential and epistemic goals that adjustment might achieve, and (3) formulate more precise versions of the central question. I also prove a new theorem characterizing when adjustment is *impermissible*. I tentatively conclude that there is a mismatch between the goals of scientists (both individually and collectively) and the guarantees of existing adjustment procedures. This article, thus, is a call for further research: We must either prove existing adjustment methods achieve goals of actual scientific interest or develop alternative procedures.

1 Basic model

To distinguish types of adjustment, I introduce a model. Suppose N hypotheses are under investigation. Assume that any subset of the N hypotheses might be true. Let $\Theta = \{0,1\}^N$ be the set of all binary strings/vectors of length N. A vector $\theta \in \Theta$,

[3] See Berry and Hochberg (1999) and Scott and Berger (2006) for discussions of Bayesian approaches to multiplicity.

[4] See Rubin (2021) for an attempt to answer the central question when "adjustment" is interpreted narrowly about significance levels.

[5] Philosophers have focused on evidential questions. See Kotzen (2013) and Mayo (2018). In contrast, most statisticians employ a quasi-decision-theoretic framework, which seems most suited for questions of rationality.

https://doi.org/10.1017/psa.2024.13 Published online by Cambridge University Press

therefore, specifies which of the N hypotheses are true and which are false. Let $H_k = \{\theta \in \Theta : \theta_k = 0\}$ be the set of vectors that say the kth hypothesis is true.

Suppose that for each hypothesis H_k, there is some experiment X_k that *could* be conducted (or observation that *could* be made); researchers believe X_k could be informative about whether H_k holds. Formally, X_k is a random variable, and for each $\theta \in \Theta$, let $\mathbb{P}_\theta(X_1, \ldots, X_N)$ denote the probability measure that specifies the chances of various experimental outcomes.

For simplicity, assume that for all $\theta \in \Theta$, the N experiments are *mutually independent* with respect to $\mathbb{P}_\theta$. In symbols, let $\vec{X} = \langle X_{i_1}, X_{i_2}, \ldots, X_{i_k} \rangle$ be a random vector, representing some subset of the N experiments. Then for all sequences $\vec{x} = \left(x_{i_1}, \ldots x_{i_k}\right)$ representing the outcome of those $k \leq N$ experiments,

$$\mathbb{P}_\theta\left(\vec{X} = \vec{x}\right) = \prod_{j \leq k} \mathbb{P}_\theta\left(X_{i_j} = x_{i_j}\right) \tag{1}$$

Further, suppose that the truth or falsity of the H_k entirely determines the probabilities of the possible outcomes of the kth experiment; that is, for all $k \leq N$ and all $r \in \{0, 1\}$, there is a probability distribution $\mathbb{P}_{k,r}$ such that $\mathbb{P}_\theta(X_k = x_k) = \mathbb{P}_{k,\theta_k}(X_k = x_k)$. Together with the assumption of mutual independence, this entails that

$$\mathbb{P}_\theta\left(\vec{X} = \vec{x}\right) = \prod_{j \leq k} \mathbb{P}_{i_j,\theta_{i_j}}\left(X_{i_j} = x_{i_j}\right) \text{for all } \theta \in \Theta. \tag{2}$$

To assess whether a decision-maker should adjust for multiplicity, compare two types of situations. In the first, the decision-maker learns the outcome of a proper subset of the N tests.

For simplicity, suppose that the researcher learns only the value of X_1. In the second, she learns the values of *all* N variables. Say that the decision-maker should adjust for multiplicity if her (1) beliefs or (2) decisions about H_1 should differ in those two situations. Let's clarify those two senses of "adjustment."

2 Belief

For the Bayesian, beliefs are modeled by posterior probabilities, and so a Bayesian adjusts for multiplicity if there is a value x_1 of X_1 such that

$$P(H_1|X_1 = x_1) \neq P(H_k|X_1 = x_1, \ldots X_N = x_N) \tag{3}$$

for all values $x_2, \ldots x_N$ of $X_2, \ldots X_N$ for which $P(X_1 = x_1, \ldots X_N = x_N) > 0$. One could distinguish a weaker sense of adjustment, whereby equation (3) holds for *some* values of $X_2, \ldots X_N$. For critics of Bayesianism, one can replace the probability functions in equation (3) with another object representing belief.[6]

Should one ever adjust for multiplicity, in the strong sense just identified? Yes. Consider a Bayesian researcher who regards the hypotheses as dependent, in that learning about one hypothesis provides evidence about another. For example, suppose our hypothetical pharmaceutical researchers consider two hypotheses:

[6] For instance, one might represent belief using orderings (Mayo-Wilson and Saraf 2022), ranking functions (Spohn 2012), Dempster–Shafer functions (Dempster 1968), and so forth.

https://doi.org/10.1017/psa.2024.13 Published online by Cambridge University Press

(1) The treatment is not effective in 33-year-old women, and (2) the treatment is not effective in 34-year-old women. A researcher might reasonably believe that the first hypothesis is true if and only if the second is. If so, acquiring data about 33-year-old women would provide evidence about the efficacy of the treatment for 34-year-old women. Here's a toy model to illustrate such adjustment.

Example 1: Suppose each X_k is a binary random variable that represents a test to retain or reject H_k. Assume there are $\alpha, \beta \in (0, 1)$ such that for all $\theta \in \Theta$,

$$\mathbb{P}_\theta(X_k = 1) = \begin{cases} \alpha & \text{if } \theta_k = 0 \\ 1 - \beta & \text{if } \theta_k = 1 \end{cases}.$$

That is, each test X_k has a Type I error of α and Type II error of β.

To model a researcher who believes the hypotheses to be dependent, suppose that the researcher assigns positive probability to precisely two vectors in Θ, namely, $\mathbf{0} = \langle 0, \ldots, 0 \rangle$, which says each H_k is true, and $\mathbf{1} = \langle 1, \ldots, 1 \rangle$, which says each H_k is false. If $\pi = P(\mathbf{0}) = 1 - P(1)$ represents the researcher's prior degree of belief that all hypotheses are true, then her posterior probability in H_1 if she learns only that the first test is negative equals the following:

$$P(H_1 | X_1 = 0) = \frac{\pi \cdot (1 - \alpha)}{\pi \cdot (1 - \alpha) + (1 - \pi) \cdot \beta}. \tag{4}$$

In contrast, if she learns *two* tests are negative, her posterior is as follows:

$$P(H_1 | X_1 = 0, X_2 = 0) = \frac{\pi \cdot (1 - \alpha)^2}{\pi \cdot (1 - \alpha)^2 + (1 - \pi) \cdot \beta^2} \tag{5}$$

Finally, if she learns the second test is positive, her posterior will be as follows:

$$P(H_1 | X_1 = 0, X_2 = 1) = \frac{\pi \cdot (1 - \alpha) \cdot \alpha}{\pi \cdot (1 - \alpha) \cdot \alpha + (1 - \pi) \cdot \beta \cdot (1 - \beta)} \tag{6}$$

If $0 < \pi < 1$, then equation (4) equals both equation (5) and equation (6) if and only if $\alpha = (1 - \beta)$. If $\alpha \neq (1 - \beta)$, therefore, the Bayesian researcher adjusts for multiplicity in the strong sense defined in equation (3).[7]

□

Example 1 illustrates the commonsense idea that when one believes two hypotheses stand or fall together, evidence for/against one hypothesis is evidence for/against the other. Thus, a Bayesian researcher will adjust for multiplicity. Similarly, if the researcher believes that evidence for one hypothesis is evidence *against* another, she will adjust for multiplicity, as can be shown by analogous calculations.

[7] Technically, our definition of adjustment compares the case in which the researcher learns X_1 to the case in which she learns the value of *all* N variables. The previous equations show the researcher adjusting when there are precisely $N = 2$ hypotheses under investigation, but similar calculations show adjustment is necessary when $N > 2$ as well.

https://doi.org/10.1017/psa.2024.13 Published online by Cambridge University Press

In short, if a researcher believes several hypotheses are dependent, she will typically adjust her beliefs for multiplicity. Conversely, if the researcher regards the hypotheses as mutually *independent*, then she will *not* adjust for multiplicity; in that case, it is easy to check that $P(H_1|X_1) = P(H_1|X_1, \ldots X_N)$—again, assuming equation (1) holds.[8]

On the one hand, these results about the relationship between adjustment and dependence in the toy Bayesian model are not surprising. They illustrate the intuition that a researcher who wants to know what to believe about the effects of cigar smoking (i) will typically adjust her belief if she acquires data about the effects of cigarette smoking but (ii) will *not* adjust her beliefs if she acquires data about implicit bias.

On the other hand, the results begin to answer the central question. In particular, they answer the objection that there is no principled way to determine when to adjust (Perneger 1998). This objection is typically leveled against classical methods—like Bonferroni's or Benjamini–Hochberg's—that recommend adjusting significance thresholds downward as the number of hypotheses increases. Yet the objection applies equally to a simple objective Bayesian method that I discuss later; the method adjusts for multiplicity by uniformly decreasing the prior probabilities assigned to hypotheses as the number of hypotheses grows.

According to critics, the justification of such methods implies that one should adjust/"correct" for any chosen set of hypotheses. But that's absurd because one would be required to adjust for *every statistical hypothesis that has ever been formulated*. This motivates thinking that the answer to the central question is, "One should never adjust for multiplicity, and intuitions to the contrary are misleading."

The toy results show how simple Bayesian thinking can partially answer the objection. Prior evidence or background theory may tell us that certain hypotheses are dependent, and in such cases, belief adjustment will almost certainly be necessary. Further research should investigate whether the most common classical adjustment methods (see subsection 3.2) can ever be interpreted as reflecting belief adjustment.

One might object that the aforementioned definition of adjusting "belief" is too simple to model some common statistical practices. The problem is that the *same* probability measure P appears on both sides of equation (3). So the definition is inapplicable for assessing whether "objective" Bayesian methods require adjustment.

Recall that objective Bayesians maintain that the prior probability that one assigns to hypothesis H may vary with the hypothesis space in which H is embedded. For example, consider an attempt to identify which genes are associated with which heritable diseases. For each gene and disease under investigation, researchers may investigate a hypothesis $H_{g,d}$ of the form "Gene g is associated with the disease d." In an objective Bayesian analysis, each hypothesis $H_{g,d}$ will typically receive lower prior probability if there are $20,000$ genes under investigation than it would receive if there were $10,000$ genes under consideration.

[8] See Berry and Hochberg (1999, 218) for the calculation and discussion of the conditions under which mutual independence of the hypotheses is reasonable. The hypotheses are mutually independent with respect to P if $P(\cap_{i\in I}\theta_i = r_i) = \Pi_{i\in I}P(\theta_i = r_i)$ for all $I \subseteq \{1, \ldots, N\}$ and all binary vectors $(r_i)_{i\in I}$.

https://doi.org/10.1017/psa.2024.13 Published online by Cambridge University Press

I will not compare the merits of objective versus subjective Bayesian analysis.[9] But simple objective Bayesian adjustment methods deserve further scrutiny. Imagine our hypothetical pharmaceutical researcher wonders about the effect of El Niño on the stock market. The mere contemplation of a new hypothesis should not automatically cause the researcher to become less confident in the efficacy of the new cancer treatment.

Yet considering additional—logically independent—hypotheses can affect an objective Bayesian's prior probabilities if those probabilities are chosen in a mechanical fashion as a function of the number of hypotheses.

Objective Bayesians might respond that a prior distribution need not represent anyone's *beliefs*.[10] Rather, a prior should be treated as part of a *decision rule*. I agree, and I consider decision-making in the next section. For now, note that it is similarly implausible that a pharmaceutical researcher should adjust her decisions about the efficacy of the cancer treatment after contemplating El Niño. Saying that the researcher's prior need not represent her beliefs does not explain why adjustment is not necessary.

3 Decision

Scientists are rarely satisfied with an answer to the question, "What should I believe?" They also want to know, "What should I do?" For instance, an experimentalist might want to know which experiment she should conduct next.

Imagine that for each hypothesis H_k, there is some set of acts A_k that the researcher might take. For instance, a researcher might announce that the hypothesis H_k has been rejected or that it's been retained. She might collect more data about H_k or cease an experiment. And so on.

I call elements of A_k *component acts*, and I define a *strategy* to be a set S of component acts such that for all k, either $S \cap A_k$ is a singleton or empty. That is, at most, one act can be taken with respect to a hypothesis. A *decision rule* d maps subsets of (values of) the observable variables $X_1, \ldots X_N$ to strategies. I require that $d\left(X_{k_1} = x_{k_1}, \ldots, X_{k_m} = x_{k_m}\right)$ contains precisely one element from each of the sets A_{k_m}. That requirement says that a decision rule specifies actions only with respect to hypotheses for which the researcher has collected data, and that if researcher observes X_k, then she must take some action in A_k.

I say that a decision rule d adjusts for multiplicity if there is some x_1 such that

$$d(x_1) \notin d(x_1, \ldots x_N) \tag{7}$$

for all values $x_2, \ldots x_N$ of $X_2, \ldots X_N$.

Do any plausible decision rules require adjusting? Again, yes. For a Bayesian, reporting one's posterior probabilities is a decision. So belief adjustment is a special case of decision adjustment. A better question is, "Can there be decision adjustment without belief adjustment, and what goals, if any, does decision adjustment achieve?"

Before discussing the standard approach for evaluating testing procedures (in terms of the family-wise error rate or false-discovery rate), I begin with the most naive, decision-theoretic approach for answering these questions. The naive approach

[9] See Goldstein (2006) and Berger (2006) for opposing views.

[10] See Gelman and Shalizi (2013) for alternative interpretations of prior probabilities used in Bayesian analyses.

https://doi.org/10.1017/psa.2024.13 Published online by Cambridge University Press

is worth sketching because (1) it is, I think, the correct approach when it can be employed,[11] and (2) it helps one identify the oddness of the goals that are presumed in standard discussions of adjustment.

3.1 A naive approach

Suppose a researcher assigns a utility $u(S,\theta)$ to each strategy S and vector $\theta \in \Theta$ specifying which of the N hypotheses are true. If we fix a vector $\theta \in \Theta$, then the researcher's expected utility (with respect to $\mathbb{P}_\theta$) can be defined straightforwardly, whether she decides to observe one variable or all N variables:[12]

$$\mathbb{E}^1_\theta[d] = \sum_{x_1 \in \mathcal{X}_1} \mathbb{P}_\theta(X_1 = x_1) \cdot u(d(x_1), \theta)$$

$$\mathbb{E}^N_\theta[d] = \sum_{\vec{x} \in \mathcal{X}} \mathbb{P}_\theta\big(\vec{X} = \vec{x}\big) \cdot u(d(\vec{x}), \theta).$$

Here, $\mathcal{X}_1$ is the range of X_1, and $\mathcal{X}$ is the range of the random vector $\vec{X} = (X_1, \ldots, X_N)$. One can now apply standard decision-theoretic terms to identify different senses in which a decision rule is good or bad.

For instance, a researcher might desire a *maximin* decision rule, that is, a rule d such that $\min_{\theta\in\theta}\mathbb{E}^j_\theta[d] \geq \min_{\theta\in\theta}\mathbb{E}^j_\theta[e]$ for all decision rules e, where $j = 1$ or $j = N$. Alternatively, she might be a Bayesian; that is, she might always select a (subjective) expected-utility-maximizing strategy with respect to her posterior. Recall that the subjective expected utility of a strategy S with respect to a measure P is given by the following:

$$\mathbb{E}_P[S] := \sum_{\theta \in \Theta} P(\theta) \cdot u(S, \theta). \tag{8}$$

Thus, there is a Bayesian who will adjust for multiplicity if there is a probability measure P, utility function u, and experimental outcomes $\vec{x} = (x_1, \ldots, x_N) \in \mathcal{X}$ such that three conditions hold:

1. $P\big(\vec{X} = \vec{x}\big) > 0$;
2. a_1 maximizes $\mathbb{E}_{P(\cdot|X_1=x_1)}[a]$ over all $a \in A_1$; and
3. $a_1 \notin S$ for some S that maximizes $\mathbb{E}_{P(\cdot|\vec{X}=\vec{x})}[T]$, where T ranges over strategies containing a component act in every A_k.

We can now make the central question more precise in a second way. For which utility functions do standard nonprobabilistic decision rules like maximin adjust for multiplicity in the sense of equation (7)? Similarly, for which priors and utility functions does an expected-utility maximizer adjust for multiplicity?

For simplicity, assume that a decision-maker's utilities are *separable* across component acts in the following sense.[13] Assume that, for each hypothesis H_k, there is

[11] See Muller et al. (2006) for a defense of this decision-theoretic approach.

[12] For simplicity, I assume all of the sets in this article are finite, including Θ, the ranges of the random variables $X_1, \ldots X_n$, and the range of all decision rules. Under appropriate measure-theoretic assumptions, the sums in the article can be replaced with integrals if one is interested in extending these ideas to continuous spaces.

[13] See Cohen and Sackrowitz (2005) for a similar assumption.

https://doi.org/10.1017/psa.2024.13 Published online by Cambridge University Press

a "component" utility function $u_k : A_k \times \{0, 1\} \to \mathbb{R}$ that specifies the utilities $u(a, 0)$ and $u(a, 1)$ of taking action $a \in A_k$ when H_k is true and false, respectively. Further, suppose that the utility of a strategy $u(S, \theta)$ in state θ is the sum of the utilities of component acts, that is:

$$u(S, \theta) = \sum_{k \leq N} \sum_{a \in S \cap A_k} u_k(a, \theta_k). \tag{9}$$

Utilities are separable when (a) the decision-maker can take component acts in parallel, and (b) payoffs for taking different component acts do not interact. Such assumptions are most plausible when two conditions are met. First, acts are cheap, or the decision-maker has plentiful resources (and so pursuing multiple projects in parallel is not prohibitively costly). Second, the hypotheses concern unrelated phenomena (so that the important theoretical consequences of a conjunction of hypotheses is the union of the theoretical consequences of the conjuncts). If the decision-maker is a grant-making institution like the National Science Foundation (NSF) or NIH, then utilities associated with projects in different scientific fields are plausibly separable. The size of the institution makes funding projects in parallel possible, and it is rare to find results in two disparate scientific fields that, when taken together, yield important insights that neither result yields by itself.

The next theorem suggests that when utilities are separable, adjustment is never obligatory, and it is sometimes impermissible.[14]

Theorem 1. Suppose utilities are separable in the sense of equation (9). Then there are maximin rules that do not adjust for multiplicity. If in addition, the hypotheses of Θ are mutually independent with respect to P, then one can maximize (subjective) expected utility with respect to P without adjusting. It follows that if the maximin rule is unique, then no decision rule that adjusts is maximin. Similar remarks apply to expected-utility maximization.

One might object that individual scientists will rarely have separable utilities for the reasons identified earlier. Component acts are often costly: Pursuing one project typically comes at the expense of pursuing another. And even if the component acts are cheap (e.g., making an announcement), it is rare that scientists investigate hypotheses that are so unrelated that if the conjunction were true, no further important insights would follow. Scientists are highly specialized, and thus they typically study hypotheses that are related.

However, I have not identified the necessary conditions for separability; utility functions might be (approximately) separable for other reasons. More importantly, theorem 1 yields sufficient conditions for nonadjustment, not necessary ones. So a suspicion that theorem 1 is rarely applicable does not justify decision adjustment for individual researchers. The theorem shifts the burden to providing a positive argument for adjustment.

The reader might speculate that given the extensive research on multiplicity, statisticians have (i) identified utility functions that plausibly represent the interests

[14] See the online supplemental materials for a proof.

https://doi.org/10.1017/psa.2024.13 Published online by Cambridge University Press

of scientists and (ii) shown that common adjustment procedures are uniquely maximin, or expected-utility maximizing with respect to those utility functions. Unfortunately, that's not the case. Some classical procedures for multiple testing are, in fact, *inadmissible* (i.e., weakly dominated) for plausible utility/loss functions.[15] Thus, the criteria used to justify standard classical testing procedures are more complex than they might initially seem; I turn to those criteria now.

3.2 Family-wise error rates and false-discovery rates

Classical approaches to multiple testing typically aim to control either the *family-wise error rate* (FWER), which is the probability that a series of tests yields at least one false positive, or the *false-discovery rate* (FDR), which is the expected *proportion* of rejected null hypotheses that are true.

Statisticians routinely say that the FWER is rarely of interest. I agree. The FWER is almost always maximized when all null hypotheses are false. But in many applications, researchers know that at least one null hypothesis is false. Consider again genome-wide association studies that investigate the associations between thousands of genes and multiple heritable diseases. If at least one disease is known to be heritable and genes are the mechanism for inheritance, then there must be at least one gene that is associated with at least one disease!

Thus, some researchers now insist that multiple-testing regimes should control the FDR. If the FDR is identical to one's loss function, are existing regimes maximin? Do they ever minimize subjective expected loss? The answer to both questions is clearly no. One minimizes the FDR (or FWER) by retaining all null hypotheses. Thus, as is standard in classical hypothesis testing, existing multiple-testing procedures typically (i) fix a threshold for FDR and (ii) attempt to maximize power (i.e., the probability of a false negative) subject to the constraint that the FDR is below the threshold. Assuming utility is identified with (some kind of) power, statisticians have identified testing regimes that are maximin among the set of procedures that maintain FDR and/or FWER below a threshold.[16]

I will not rehearse standard objections to maximin reasoning,[17] nor to the bizarre two-step procedure in which one first culls testing procedures using FDR and then applies maximin. Instead, I emphasize that the decision criteria just described (1) treat all *null* hypotheses equally, (2) treat null hypotheses *differently* from alternatives, and (3) ignore effect sizes. However, there are virtually no circumstances in which such equal treatment and dismissal of effect size reflects either scientific or public interest.

Consider a recent influential genome-wide study in which researchers tested roughly 14,000 genes for associations with seven common diseases, which included bipolar disorder and Crohn's disease (Wellcome Trust Case Control Consortium 2007). Although the authors of the study reported adjusted *p*-values, they also laudably applied many statistical techniques, incorporated background genetic knowledge, and avoided

[15] Again, see Cohen and Sackrowitz (2005).

[16] Just as there are multiple notions of "Type I error" when many hypotheses are tested (e.g., FWER or FDR), so there are multiple notions of "power" that might be invoked, such as the probability of *at least one* false negative, the "average" power, and more. For a discussion and proof of the optimality of certain classical procedures, see see Rosset et al. (2022).

[17] See Savage (1954, chaps. 9 and 10), for example.

https://doi.org/10.1017/psa.2024.13 Published online by Cambridge University Press

making policy recommendations based solely on adjusted p-values. Why did they not simply apply a testing procedure with good power subject to the control of FDR?

All seven diseases they considered are serious, but the incidence of each varies widely, as do the cost and efficacy of available treatments. From a public health perspective, therefore, it would be inappropriate to treat every hypothesis of the form "Gene g is associated with disease d" equally and to ignore the strength of such associations.

One might object that the severity of the diseases does not affect the *evidence* for the various hypotheses. Does adjustment somehow reflect one's evidence?

Answering that question is beyond the scope of this article; I lack the space to explore the relationship among evidence, belief, and decision.[18] But I am skeptical of both (a) the importance of the question and (b) an answer that involves classical procedures that control FDR or FWER.

Concerning (a), philosophers and scientists alike should be wary of directives to ignore the suffering caused by diseases and instead coldly evaluate only the evidence for empirical hypotheses. I admit that a subjective expected-utility analysis of genome-wide studies seems daunting. I have no idea how to define a prior over a roughly 100,000-dimensional (i.e., approximately $7 \cdot 14,000$) parameter space that incorporates expert knowledge. Nor do I have any idea how to define a utility function that balances considerations of the severity and incidence of different diseases. But I stress that mechanical use of multiple-testing procedures amounts to a refusal to engage with questions of ethical importance, not an answer.

Concerning (b), like many classical procedures, decision criteria that first cull tests by FWER or FDR treat null hypotheses differently from the alternatives. But if evidential strength is divorced from pragmatic and ethical considerations, it is hard to see how the asymmetric treatment of null and alternative hypotheses could reflect anything evidential: What could distinguish a hypothesis H from its negation $\neg H$, evidentially speaking?

4 Conclusions

The goals of scientists and of the public may be misaligned with the decision criteria used to evaluate multiple-testing regimes. Thus, I urge two broad projects for future research.

First, in scientific contexts in which large numbers of statistical hypotheses can be tested, scientists and philosophers must study the interests of the affected parties. The differential funding provided for medical research—in comparison to academic philosophy, for instance—is typically justified by its social importance. Scientists should make good on that promise to advance collective interests.[19]

Second, statisticians must prove that existing testing procedures advance the interests of affected parties, or they must develop alternative procedures altogether. Otherwise, we all stand to be bamboozled by Bonferroni.

Supplementary material. For supplementary material accompanying this paper visit https://doi.org/10.1017/psa.2024.13

[18] Royall (1997) clearly distinguishes questions about belief, decision, and evidence.
[19] See also Longino (1990), Kitcher (2003), and Douglas (2009).

https://doi.org/10.1017/psa.2024.13 Published online by Cambridge University Press

References

Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society: Series B (Methodological)* 57 (1):289–300. doi:https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

Berger, James. 2006. "The Case for Objective Bayesian Analysis." *Bayesian Analysis* 1 (3):385–402. doi:https://doi.org/10.1214/06-BA115.

Berry, Donald A., and Yosef Hochberg. 1999. "Bayesian Perspectives on Multiple Comparisons." *Journal of Statistical Planning and Inference* 82 (1–2):215–27. doi:https://doi.org/10.1016/S0378-3758(99)00044-0.

Cohen, Arthur, and Harold B. Sackrowitz. 2005. "Characterization of Bayes Procedures for Multiple Endpoint Problems and Inadmissibility of the Step-Up Procedure." *The Annals of Statistics* 33 (1):145–58. doi:https://doi.org/10.1214/009053604000000986.

Dempster, Arthur P. 1968. "Upper and Lower Probabilities Generated by a Random Closed Interval." *Annals of Mathematical Statistics* 39 (3):957–66.

Douglas, Heather. 2009. *Science, Policy, and the Value-Free Ideal*. 1st ed. Pittsburgh: University of Pittsburgh Press.

Gelman, Andrew, and Cosma Rohilla Shalizi. 2013. "Philosophy and the Practice of Bayesian Statistics." *British Journal of Mathematical and Statistical Psychology* 66 (1):8–38. doi:https://doi.org/10.1111/j.2044-8317.2011.02037.x.

Goldstein, Michael. 2006. "Subjective Bayesian Analysis: Principles and Practice." *Bayesian Analysis* 1 (3):403–20. doi:https://doi.org/10.1214/06-BA116.

Kitcher, Philip. 2003. *Science, Truth, and Democracy*. Oxford: Oxford University Press.

Kotzen, Matthew. 2013. "Multiple Studies and Evidential Defeat." *Nous* 47 (1):154–80. doi:https://doi.org/10.1111/j.1468-0068.2010.00824.x.

Lehmann, Erich L., and Joseph P. Romano. 2008. *Testing Statistical Hypotheses*. 3rd ed. New York: Springer Science & Business Media.

Longino, Helen E. 1990. *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton, NJ: Princeton University Press.

Mayo, Deborah G. 2018. *Statistical Inference as Severe Testing*. Cambridge: Cambridge University Press.

Mayo-Wilson, Conor, and Aditya Saraf. 2022. "Robust Bayesianism and Likelihoodism." *arXiv preprint*. https://doi.org/10.48550/arXiv.2009.03879.

Muller, Peter, Giovanni Parmigiani, and Kenneth Rice. 2006. "FDR and Bayesian Multiple Comparisons Rules." https://biostats.bepress.com/jhubiostat/paper115/.

Perneger, Thomas V. 1998. "What's Wrong with Bonferroni Adjustments." *BMJ: British Medical Journal* 316 (7139):1236–38. doi:https://doi.org/10.1136/bmj.316.7139.1236.

Rosset, Saharon, Ruth Heller, Amichai Painsky, and Ehud Aharoni. 2022. "Optimal and Maximin Procedures for Multiple Testing Problems." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 84 (4):1105–28. doi:https://doi.org/10.1111/rssb.12507.

Royall, Richard. 1997. *Statistical Evidence: A Likelihood Paradigm. Vol. 71 in Monographs on Statistics and Applied Probability*. Boca Raton, FL: Chapman & Hall/CRC.

Rubin, Mark. 2021. "When to Adjust Alpha during Multiple Testing: A Consideration of Disjunction, Conjunction, and Individual Testing." *Synthese* 199 (3–4):10969–11000. doi:https://doi.org/10.1007/s11229-021-03276-4.

Savage, Leonard J. 1954. *The Foundation of Statistics*. Mineola, NY: Dover.

Scott, James G., and James O. Berger. 2006. "An Exploration of Aspects of Bayesian Multiple Testing." *Journal of Statistical Planning and Inference* 136 (7):2144–62. doi:https://doi.org/10.1016/j.jspi.2005.08.031.

Spohn, Wolfgang. 2012. *The Laws of Belief: Ranking Theory and Its Philosophical Applications*. Oxford: Oxford University Press.

Wellcome Trust Case Control Consortium. 2007. "Genome-Wide Association Study of 14,000 Cases of Seven Common Diseases and 3,000 Shared Controls." *Nature* 447:661–78. doi:https://doi.org/10.1038/nature05911.

Cite this article: Mayo-Wilson, Conor. 2024. "Bamboozled by Bonferroni." *Philosophy of Science*. https://doi.org/10.1017/psa.2024.13