A model of multiple hypothesis testing[*]

Davide Viviano — Harvard University
Kaspar Wüthrich — University of Michigan
Paul Niehaus — UC San Diego

March 17, 2026

Abstract

Multiple hypothesis testing practices vary widely, without consensus on which are appropriate when. This paper provides an economic foundation for these practices designed to capture leading examples, such as regulatory approval on the basis of clinical trials. MHT adjustments are appropriate in our framework to the extent that research costs are invariant to the number of hypotheses. Control of average size, as for example via a Bonferroni correction, emerges in the limit case where all costs are fixed; in the opposite limit, where costs vary in proportion to the hypothesis count, no correction is needed. We illustrate implications by calculating explicit critical values using data on actual costs in the drug approval process and in program evaluation research; these suggest that some MHT adjustment is warranted in these applications, but not as much as implied by standard practice.

Keywords: Bonferroni, family-wise error rate, multiple subgroups, multiple treatments, research costs

JEL Codes: C12

[*] Earlier versions of this paper were circulated under the title: "(When) should you adjust inferences for multiple hypothesis testing?" (with and without the parentheses). First version on arXiv: 27 April 2021. Viviano and Wüthrich contributed equally to this work. Email: `dviviano@fas.harvard.edu`, `kasparwu@umich.edu`, `pniehaus@ucsd.edu`

1 Introduction

Hypothesis testing plays a prominent role in evidence-based decision-making. Typically, researchers report results from more than one test, and there has recently been increasing interest in and debate over whether the testing procedures they employ should reflect this in some way—that is, whether to apply some form of multiple hypothesis testing (MHT) adjustment. As a concrete example, consider pharmaceutical companies reporting the results of clinical trials to regulators when seeking approval to market new drugs: the U.S. regulator (the Food and Drug Administration, FDA) recently released guidelines calling for MHT adjustments on the grounds that omitting them could "increase the chance of false conclusions regarding the effects of the drug" (Food and Drug Administration, 2022). Analogous concerns arise in other settings, including experimental program evaluation in economics. As a result, a number of procedures for MHT adjustment have been proposed, and their statistical properties are well-understood (see, e.g., Romano et al., 2010, for an overview).

What is less clear is whether and when these procedures are economically desirable. That is, under what conditions does MHT adjustment lead to better decision-making from the point of view of the actor designing the process? The answer is far from obvious. It is certainly true, for example, that without MHT adjustments, the chance of making at least one type I error increases with the number of tests. But this is analogous to the truism that the more decisions one makes, the more likely one is to make at least one mistake. It is indisputable, but sheds no light on the pertinent questions, which are whether and how the rule for making individual decisions should change with the total number being made.

This paper provides a framework for analyzing such questions. We focus, in particular, on whether and when MHT adjustments arise as a solution to incentive misalignment between a researcher and a mechanism designer. Our interest in this case reflects two primary considerations. The first is substantive: incentives are clearly an issue in real-world cases of interest (e.g., in the drug approval process, which we will use as a running example). The instinctive concern many seem to have is that without MHT adjustments, the researcher would have an undue incentive to test many hypotheses in the hopes of getting lucky. We would like to formalize and scrutinize that intuition. And the second is pragmatic: to have a theory of MHT adjustments, we must have a theory that rationalizes hypothesis testing at standard levels in the first place, which (as we discuss below) is hard to do convincingly in a non-strategic setting (e.g., Tetenov, 2012, 2016).

Specifically, we study a model in which a benevolent social planner chooses norms with respect to MHT adjustments, taking into account the way this shapes researchers' incentives. The model embeds two core ideas. First, social welfare is affected by the summary recommendations (in particular, hypothesis tests) contained in research studies, and the planner also cares about the more generic benefits to society and to the researcher of conducting research per se.[1] Second, while this makes the research a public good, the costs of producing it are borne privately by the researcher. She decides whether or not to incur these costs and conduct a (pre-specified) experiment based on the private returns to doing so. The planner must, therefore, balance the goals of (i) limiting the possibility of harm due to mistaken conclusions and (ii) motivating the production of research. We represent these preferences with a utility function that includes both ambiguity-averse and expected-utility components (in the spirit of, for example, Gilboa and Schmeidler, 1989; Banerjee et al., 2020), which (we show) turn out to have intuitive connections with the statistical concepts of size control and power. We focus on cases where multiplicity takes the form of testing multiple *treatments* or estimating effects within multiple *sub-populations*;[2] multiple *outcomes* are an economically distinct case covered in an earlier version of the paper (Viviano et al., 2025).

We start by characterizing optimal hypothesis testing protocols. We show that the class of optimal protocols is the class of maximin optimal and unbiased protocols, where maximin optimality is closely connected to size control while unbiasedness requires the power of the protocols to exceed their size. We then prove that separate t-tests, which are ubiquitous in applied work, are maximin optimal and unbiased, and we provide an explicit characterization of the optimal critical value in terms of the researcher's costs.[3]

We next characterize the role of multiplicity, drawing two broad conclusions. First, it

[1] We describe the case where hypothesis *rejections* affect welfare; under a straightforward reinterpretation the framework can also accommodate situations in which "precise null" results do so.

[2] These forms of multiplicity are common in practice. For example, the majority of the clinical trials reviewed in Pocock et al. (2002, Table 1) tested for effects in more than one subgroup. In economics, 27 of 124 field experiments published in "top-5" journals between 2007 and 2017 feature factorial designs with more than one treatment (Muralidharan et al., 2025).

[3] We focus on one-sided t-tests in the main text and consider two-sided tests in Appendix B.1.

is generically optimal to adjust testing thresholds (i.e., critical values) for the number of hypotheses. A loose intuition is as follows. The worst states of the world are those in which the status quo of no treatment is best; in these states, a research study has only a downside, and it is desirable to keep the benefits from experimentation low enough that the researcher chooses not to experiment. If the hypothesis testing protocol were invariant to the number of hypotheses being tested, then for sufficiently many hypotheses, this condition would be violated: the researcher's expected payoff from false positives alone would be high enough to warrant experimentation. Some adjustment for hypothesis count may thus be needed. This logic aligns fairly well with the lay intuition that researchers should not be allowed to test many hypotheses and then "get credit" for false discoveries. Interestingly, the same logic immediately implies that critical values should adjust for other factors that influence cost such as the sample size, though these have not attracted the same degree of attention.

Second (and as this suggests), economic fundamentals—in particular, the research cost function—determine exactly how much adjustment is required. When hypotheses are equally important, for example, the optimal critical values for the separate t-tests are given by

$$t(J, \Sigma) = \Phi^{-1}\left(1 - \frac{C(J, \Sigma)}{b|J|}\right), \tag{1}$$

where J is the set of hypotheses tested (with $|J|$ denoting its cardinality), Σ captures features of the experimental design such as the sample size, $C(J, \Sigma)$ the cost of the experiment, and b the benefit to the researcher of rejecting a null. When research costs are fixed, so that C is invariant to J, this implies a Bonferroni correction.[4] When costs scale in proportion to the number of hypotheses, on the other hand, *no* MHT adjustment is required. Intuitively, the researcher has no undue incentive to test many hypotheses in this scenario because doing so is costly. This cost-based perspective also helps to clarify confusion about the boundaries of MHT adjustment and whether researchers should adjust for multiple testing *across* different studies. It suggests that MHT adjustments may be appropriate when there are cost complementarities across studies but not otherwise.[5]

To illustrate the quantitative implications of the model, we apply it to our running

[4] Including, for example, in subgroup analysis contexts where experimentation costs are sunk.

[5] The broader principle is that optimal MHT adjustments depend on how exactly hypotheses interact. Our base model emphasizes interactions in the research cost; Appendix B.2 considers interactions of other kinds, through non-linearities in the researcher's payoff and through interactions in the planner's objective.

example, regulatory approval by the FDA. Applying the formulae implied by the model to published data on the cost structure of clinical trials, we calculate adjusted critical values that are neither as liberal as unadjusted testing, nor as conservative as those implied by some of the procedures in current use. If the appropriate level in the single-hypothesis case is 5%, for example, then the optimal level according to our formulae is 3.2% with two tests, 2.6% with three tests, and tends to 1.4% as $|J| \to \infty$. By comparison, the level implied by Sidak's correction (Šidák, 1968) for controlling the Family-Wise Error Rate (FWER) under independence (and, up to rounding, also Bonferroni), is 2.5% for two tests, 1.7% for three tests, and tends to zero as $|J| \to \infty$.[6] These results suggest both that some adjustments are warranted but also that standard practices may be overly-conservative. Moreover, because costs also scale with the sample size, optimal adjustments must be less conservative for larger samples in order to induce researchers to incur the correspondingly larger costs.

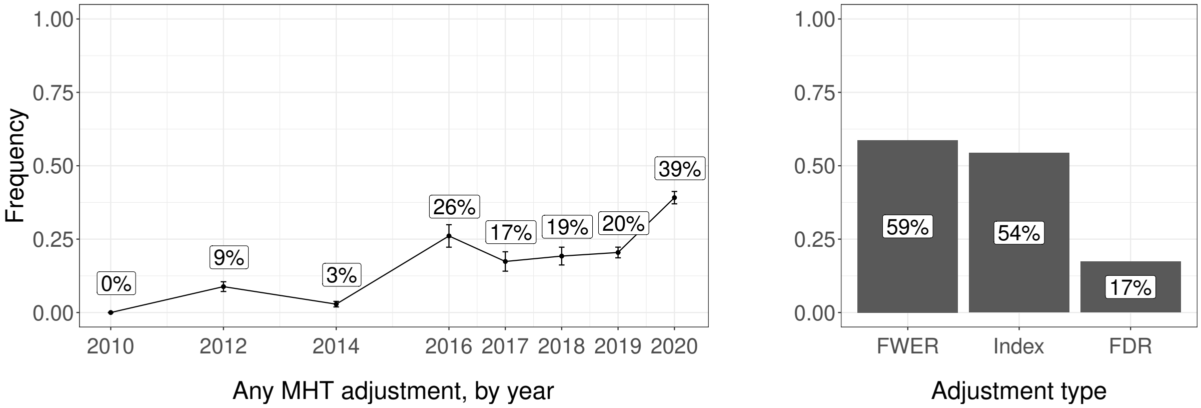
It is also natural to wonder about applicability to economic research. The share of experimental papers published in "top 5" journals that used some form of MHT adjustment grew rapidly, from 0% in 2010 to 39% in 2020, so that there is now wide variability in whether (and how) these papers adjust (see Figure 1) and little consensus on what the norms should be. In Appendix A, we develop an additional empirical application to experimental program evaluation, using a unique dataset on the costs of projects submitted to the Abdul Latif Jameel Poverty Action Lab (J-PAL) from 2009 to 2021 which we assembled for this purpose. In this application, we find that the estimated adjustments implied by our model are less conservative than FWER control using Sidak's correction, but only slightly so. This is because the relationship between costs and number of treatment arms, while significant, is relatively weak in this setting, with a cross-sectional elasticity of approximately 15%.

Our paper draws inspiration from other work using economic models to select statistical procedures, in which researchers' preferences and incentives drive the analysis. This includes work on scientific communication (e.g., Andrews and Shapiro, 2021; Frankel and Kasy, 2022), several aspects of which have been studied in more recent papers including Bates et al. (2022, 2023) and Kasy and Spiess (2023).[7] None of these papers analyze multiple hypothesis testing,

[6] The Sidak correction is a natural benchmark because it is exact for controlling the FWER with independent tests, and FWER control is common in practice (see Figure 1).

[7] See also Chassang et al. (2012); Banerjee et al. (2017); Spiess (2018); Henry and Ottaviani (2019);

Figure 1: Multiple hypothesis testing adjustment in “top-5” experimental papers



Notes: The left-hand panel reports the share of experimental papers (both field and lab experiments) that conduct at least one MHT adjustment, including both indexing and control of compound error rates, by year of publication. (Note that almost all experimental studies have more than one hypothesis). The right-hand panel reports the frequency of each adjustment type, pooling across years. Adjustment types are not mutually exclusive. FWER: Family-Wise Error Rate control. Index: Methods based on creating an index using multiple outcomes. FDR: False Discovery Rate control. Authors’ calculations based on a review of publications in the *American Economic Review* (excluding Papers and Proceedings), *Econometrica*, the *Journal of Political Economy*, the *Quarterly Journal of Economics*, and the *Review of Economic Studies*.

however. The most closely related paper is the insightful work by Tetenov (2016), who shows that t-tests are maximin optimal and uniformly most powerful in the single-hypothesis case. Our extension to a multiple-hypothesis setting requires us to deal with two major challenges. First, the notions of maximin optimality and the corresponding theoretical results are more complex because the effects of different treatments may have opposite signs. Second, within the (large) class of maximin optimal protocols, none uniformly dominates all others, requiring us to develop new notions of optimality suitable to the context.

Our paper also relates to an extensive literature at the intersection between decision theory and hypothesis testing, dating back to Wald (1950) and Robbins (1951). Previous work has motivated notions of compound error control in single-agent non-strategic environments; see in particular Kline et al. (2022, 2024) for recent examples in economics based on a Bayesian interpretation of the false discovery rate (FDR), as well as Storey (2003), Lehmann and Romano (2005), Efron (2008b), and Hirano and Porter (2020) for further examples.[8]

Banerjee et al. (2020); Williams (2021); McCloskey and Michaillat (2022); Yoder (2022).

[8] The literature on statistical treatment choice has similarly focused for the most part on non-strategic planner problems. See Manski (2004) and Tetenov (2012) as well as Hirano and Porter (2009); Kitagawa and Tetenov (2018); Athey and Wager (2021) for recent contributions.

We complement this literature (as well as the statistical literature discussed below) by developing a model that explicitly incorporates the incentives and constraints of the researchers. Relative to the decision-theoretic approach, this has two main advantages. First, it lets us characterize *when* MHT adjustments are appropriate—and also when they are *not*—as a function of measurable features of the research process. Second, it allows us to justify and discriminate between different notions of compound error (e.g., average error rate or the FWER) in the same framework based on these same economic fundamentals.

Finally, we aim to provide guidance for navigating the extensive statistical literature on MHT. This literature provides procedures for controlling particular notions of compound error,[9] but few statistical optimality results (e.g., Spjotvoll, 1972; Lehmann et al., 2005; Romano et al., 2011), and none in which MHT procedures address an incentive problem. We maximize a different (social planner's) objective, subject to incentive compatibility constraints. We also draw on List et al. (2019)'s helpful distinction between different types of multiplicity, and show how these lead to different optimal testing procedures.

2 Model

We study MHT in a game between a social planner who chooses statistical procedures and a representative experimental researcher with private incentives. In our running example, we can think of the planner as a regulator (e.g., the FDA) who defines testing protocols for studies submitted in support of applications for the approval of new drugs, and the researcher as a pharmaceutical company running a pre-specified clinical trial of a new drug hoping to obtain such approval. Multiple testing issues arise whenever research informs multiple decisions. We focus on settings with multiple *treatments* (e.g., multiple drugs) or different *subpopulations* (e.g., multiple demographic groups for which a drug may be approved); for brevity we will refer throughout to treatments, taking this to refer to multiplicity of both types.

To say something coherent about MHT, a framework must be able to rationalize conventional (single) hypothesis testing in the first place. This is known to be a challenging problem, requiring non-trivial restrictions on the research process (see, e.g., Section 1 in Tetenov, 2016,

[9] See Efron (2008a) and Romano et al. (2010) for overviews.

for a discussion)—in particular, strong asymmetries to match the inherently asymmetric nature of null hypothesis testing. For example, Tetenov (2012) shows that justifying testing at conventional levels in a single agent model with minimax regret requires extreme degrees of asymmetry: statistical tests at the 5% (1%) level correspond to decision-makers placing 102 (970) times more weight on type I than type II regret. Here the asymmetry necessary for rationalizing hypothesis testing will arise naturally from the planner's desire to prevent the implementation of harmful treatments.

2.1 A game between a researcher and a social planner

We consider a two-stage game between the planner and the researcher. In the first stage, the planner prescribes and commits to a hypothesis testing protocol, restricting how the researcher can report findings. In the second stage, given this protocol, the researcher decides whether or not to run one of several possible experiments by comparing the private benefits of experimentation to the private costs. Importantly, these private benefits may differ from the planner's objective. Unless noted otherwise, we will assume that the researcher's preferences are common knowledge and that she is not allowed to mis-characterize them.

Hypothesis testing protocols take as input the data from the experiment and output multiple binary findings indicating whether the treatments were found to be effective. These findings, in turn, affect social welfare: we will interpret a finding as equivalent to the planner's decision to implement the corresponding treatment. Since the planner selects the hypothesis testing protocol, this is equivalent to the planner pre-committing to a decision rule and the researcher truthfully reporting the decisions it implies, given the observed data.

2.1.1 The researcher's problem

The researcher takes the hypothesis testing protocol as given and decides, before observing data, whether and how to experiment. We first describe the experiment and hypothesis testing protocol and then introduce the researcher's optimization problem.

Experiment. Let $\mathcal{J}$ denote the finite set of all combinations of non-exclusive treatments that can be tested in an experiment (the "power set"), with $\emptyset \in \mathcal{J}$ denoting no experimentation. The parameter vector $\theta \in \Theta$, where Θ is a compact parameter space, captures the effects corresponding to all possible combinations of treatments in $\mathcal{J}$.

An experiment consists of a set of treatments $J \in \mathcal{J}$ and a design $\Sigma \in \mathcal{S}(J)$, which are chosen by the researcher. Here $\mathcal{S}(J)$ is the set of all possible designs given J. If the researcher experiments, she draws a vector of statistics X from a distribution $F_{\theta,J,\Sigma}$, indexed by J, θ, and Σ. The design Σ summarizes all the relevant features of the distribution of X the researcher can choose ex-ante, such as the sample size of the experiment. The researcher pre-specifies and reports J and Σ before running the experiment, so that they become common knowledge. J and Σ may depend on the researcher's prior knowledge and private incentives, but not on the realized statistics X. We thus abstract from issues of p-hacking and selective reporting. This case is relevant for considering decision-making at the FDA, for example, which requires pre-registration.[10]

Remark 2.1 (Mutually exclusive treatments)**.** We focus on settings where the treatments may not be mutually exclusive. This is relevant in the regulatory approval process context, for example, when there are multiple subgroups and the pharmaceutical companies receive separate approvals for each group, or when there are different drugs that can be provided to the same set of individuals. That said, our framework also accommodates settings with mutually exclusive treatments, such as competing drugs for treating the same condition. First, if the researcher can report multiple findings and each treatment will be implemented with an (exogenous) probability and these probabilities sum to one, the results in Sections 3 and 4 apply due to the linearity of the planner's objective defined below. Second, if the researcher is only allowed to report one finding, the resulting model is isomorphic to the one discussed at the end of Appendix B.2. □

Hypothesis testing protocols. As described above, the researcher first chooses and pre-specifies an experiment (J, Σ). She then runs the experiment, as a result of which the vector of statistics X is realized and becomes publicly available. The results from the experiment

[10] Specifically, the summary of the Final Rule for Clinical Trials Registration and Results Information Submission (42 CFR §11) on ClinicalTrials.gov states that "[r]egistration is required for studies that meet the definition of an 'applicable clinical trial' (ACT) and either were initiated after September 27, 2007, or initiated on or before that date and were still ongoing as of December 26, 2007" (NIH National Library of Medicine, nd). Registration must specify, among other things, the intervention(s), primary outcomes, and intended enrollment and study design (Code of Federal Regulations, 2024, 42 CFR §11.28).

are reported in the form of a vector of non-exclusive *findings* or *recommendations*,

$$r(X; J, \Sigma) = \big(r_1(X; J, \Sigma), \ldots, r_{|J|}(X; J, \Sigma)\big)^\top \in \{0,1\}^{|J|}, \tag{2}$$

where $r_j(X; J, \Sigma) = 1$ if and only if the treatment corresponding to J_j is found to be effective, with J_j denoting the j^{th} entry of J. If $J = \emptyset$, no findings are reported, $r(X, \emptyset, \Sigma) = 0$. If there are no findings ($r(X, J, \Sigma) = 0$), the status quo prevails. We will refer to r as a *hypothesis testing protocol.*

To simplify notation when describing the researcher's payoff and welfare below, it is useful to introduce the selector function $\delta(r(X; J, \Sigma); J) \in \{0,1\}^{2^{|J|}-1}$. Each entry of $\delta(r(X; J, \Sigma); J) = (\delta_1(r(X; J, \Sigma); J), \ldots, \delta_{2^{|J|}-1}(r(X; J, \Sigma); J))$, corresponds to one of the $2^{|J|} - 1$ possible combinations of the J treatments. Specifically, for $k = 1, \ldots, 2^{|J|} - 1$, $\delta_k(r(X; J, \Sigma); J) = 1$ if the treatment combination k is found to be effective and $\delta_k(r(X; J, \Sigma); J) = 0$ otherwise. We let $\delta(r; \emptyset) = 0$ for all r. By definition of δ, we have that

$$\sum_{k=1}^{2^{|J|}-1} \delta_k(r(X; J, \Sigma); J) \in \{0,1\}.$$

Here $\delta(\cdot; J)$ only takes into account combinations of the treatments in the set J, ignoring combinations not studied in the experiment. Example 2.1 provides an illustration of δ.

Example 2.1. The researcher uses linear regression to estimate the effects of treatments $J \in \mathcal{J}$ on an outcome Y based on an experiment with n units. Let $D_{i,j} = 1$ if unit i received treatment j and $D_{i,j} = 0$ otherwise. Suppose that $J = \{1,2\}$ and that

$$Y_i = \mu + \theta_1 D_{i,1} + \theta_2 D_{i,2} + \varepsilon_i, \ \varepsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \eta^2), \tag{3}$$

where μ, θ_1, and θ_2 are unknown parameters and η^2 is known. Let $X \sim \mathcal{N}(\theta, \Sigma)$, where X is the OLS estimator of (θ_1, θ_2) and where, specializing notation, in this example the design Σ denotes the covariance matrix of X. An example of a hypothesis testing protocol is separate (one-sided) t-testing, $r(X; \{1,2\}, \Sigma) = (1\{X_1/\sqrt{\Sigma_{1,1}} \geq t\}, 1\{X_2/\sqrt{\Sigma_{2,2}} \geq t\})^\top$.

In this example, the power set $\mathcal{J}$ is $\mathcal{J} = \{\emptyset, \{1\}, \{2\}, \{1,2\}\}$, and J is an element of $\mathcal{J}$. The selector function δ is defined as follows (suppressing the dependence of $r(X; J, \Sigma)$ on

Σ): $\delta(r(X;\{1\});\{1\}) = r(X;\{1\})$, $\delta(r(X;\{2\});\{2\}) = r(X;\{2\})$,

$$\begin{aligned}\delta(r(X;\{1,2\});\{1,2\}) &= \Big(\delta_1(r(X;\{1,2\});\{1,2\}), \delta_2(r(X;\{1,2\});\{1,2\}), \delta_3(r(X;\{1,2\});\{1,2\})\Big)\\ &= \Big(r_1(X;\{1,2\})(1-r_2(X;\{1,2\}), r_2(X;\{1,2\})(1-r_1(X;\{1,2\}), r_1(X;\{1,2\})r_2(X;\{1,2\})\Big).\end{aligned}$$

That is, the first (second) entry of $\delta(r(X;\{1,2\});\{1,2\})$ is equal to one if treatment 1 (treatment 2) is found to be effective but not treatment 2 (treatment 1), and the last entry is equal to one if both treatments are found to be effective. □

The researcher's objective. For each (J,Σ), the researcher takes as given the corresponding hypothesis testing protocol $r(\cdot;J,\Sigma)$, which is chosen by the planner in the first stage. For simplicity, we assume that the researcher knows θ, but our main results continue to hold when the researcher is imperfectly informed and has a prior about θ (see Section 5.1).

We consider settings where the researcher's and the planner's objective are misaligned. We model misalignment using researcher's utility of the form $p(J)^\top\delta(r(X;J,\Sigma)) - c_\theta(J,\Sigma)$ where $p_k(J) \geq 0$, $k = 1,\ldots,2^{|J|}-1$, is the benefit from getting approval for the kth combination of treatments, conditional on the set of treatments J being tested, and $c_\theta(J,\Sigma)$ captures the costs of research, which can be a function of (J,Σ,θ).[11] Taking expectations over X yields the following class of expected researcher payoff functions,

$$\beta_r(\theta,J,\Sigma) = \underbrace{\sum_{k=1}^{2^{|J|}-1} p_k(J)\int \delta_k(r(x;J,\Sigma),J)dF_{\theta,J,\Sigma}(x)}_{\text{expected benefits}} - \underbrace{c_\theta(J,\Sigma)}_{\text{costs}}, \quad c_\theta(J,\Sigma)\geq 0 \quad \forall(\theta,J,\Sigma). \tag{4}$$

Thus, $\beta_r(\theta,J,\Sigma)$ captures the net benefits of experimenting. We let $c_\theta(J,\Sigma) \geq 0$ for all (J,Σ,θ) and $c_\theta(\emptyset,\Sigma) = 0$ for all (θ,Σ), so that the researcher's net benefits are normalized to zero if no experiment is conducted. Misalignment arises because the researcher's payoff is different from the planner's objective. In the following, whenever we consider settings where the costs do not depend on θ, we will write $c_\theta(J,\Sigma) \equiv C(J,\Sigma)$ for a function $C(J,\Sigma)$ that does not depend on θ.

The researcher chooses which treatments to study and how to design the experiment so

[11] For instance, we could write $c_\theta(J,\Sigma) = C(J,\Sigma) - b(J,\theta)$ for some function $b(\theta,J)$ of (J,θ) that is the part of benefits that depend on θ ($b(\theta,J)$ can be an implicit function of p) and the costs $C(J,\Sigma)$ as a function of the design. Practically speaking however this requires being able to measure $b(J,\theta)$ in addition to the costs.

as to maximize her net benefits. Formally, the researcher's problem is

$$(J^*_{r,\theta}, \Sigma^*_{r,\theta}) \in \arg \max_{J \in \mathcal{J}, \Sigma \in \mathcal{S}(J)} \beta_r(\theta, J, \Sigma), \tag{5}$$

where $J^*_{r,\theta} = \emptyset$ corresponds to no experimentation. To state the theoretical results, we also define the experiment the researcher would choose when forced to run an experiment,

$$\left(J^+_{r,\theta}, \Sigma^+_{r,\theta}\right) = \arg \max_{J \in \mathcal{J} \setminus \emptyset, \Sigma \in \mathcal{S}(J)} \beta_r(\theta, J, \Sigma). \tag{6}$$

We impose a standard tie-breaking rule: whenever the researcher is indifferent regarding whether to experiment (i.e., when $\beta_r(\theta, J^*_{r,\theta}, \Sigma^*_{r,\theta}) = 0$), she experiments if the planner's utility (defined below) is weakly positive. Let $e^*_r(\theta)$ indicate whether the researcher experiments, $e^*_r(\theta) = 1\left\{J^*_{r,\theta} \neq \emptyset\right\}$.

Example 2.1 continued. Suppose that $\mathcal{J} = \{\emptyset, \{1\}, \{2\}, \{1,2\}\}$. Let $p(\{1\}) \in \mathbb{R}_+$, $p(\{2\}) \in \mathbb{R}_+$, and $p(\{1,2\}) \in \mathbb{R}^3_+$ denote the researcher's (vector of) benefits from getting approval, conditional on the set of treatments J being tested. In this example, the researcher's payoff for $J = \{1,2\}$ equals

$$\begin{aligned}\beta_r(\theta, \{1,2\}, \Sigma) =& P(r_1(X; \{1,2\}, \Sigma) = 1, r_2(X) = 0|\theta) p_1(\{1,2\}) \\ &+ P(r_2(X; \{1,2\}, \Sigma) = 1, r_1(X; \{1,2\}, \Sigma) = 0|\theta) p_2(\{1,2\}) \\ &+ P(r_1(X; \{1,2\}, \Sigma) = 1, r_2(X; \{1,2\}, \Sigma) = 1|\theta) p_3(\{1,2\}) - C(\{1,2\}, \Sigma).\end{aligned}$$

□

2.1.2 The planner's problem

The social planner chooses a hypothesis testing protocol $r \in \mathcal{R}$ to maximize her utility, where $\mathcal{R}$ is the class of all (pointwise measurable) protocols. That is, $\mathcal{R}$ contains all protocols typically found in practice, including (and not limited to) standard t-tests. The planner's utility will depend on the welfare effects of implementing the recommended treatments as well as a measure of the more generic benefits to society and to the researcher of conducting research per se.

Welfare. Welfare depends on whether the researcher experiments and on her findings if she does. To define welfare, for $k = 1, \ldots, 2^{|J|} - 1$, let $u_k(\theta; J)$ denote the effect on welfare that would result from implementing the combination of treatments k.

Given $r(X; J, \Sigma)$, the overall welfare is $u(\theta; J)^\top \delta(r(X; J, \Sigma); J)$. We normalize $u(\theta; \emptyset) = 0$

for all $\theta \in \Theta$. That is, welfare is equal to zero under the status quo when no experimentation occurs. For a given experiment (J, Σ), the expected welfare is

$$v_r(\theta, J, \Sigma) = \int \delta(r(x; J, \Sigma); J)^\top u(\theta; J) dF_{\theta,J,\Sigma}(x). \tag{7}$$

If the researcher does not experiment, $v_r(\theta, \emptyset, \Sigma) = 0$ for all $\theta \in \Theta$.

Example 2.1 continued. Suppose that $\mathcal{J} = \{\emptyset, \{1\}, \{2\}, \{1, 2\}\}$ and denote by θ_1 and θ_2, the welfare from implementing treatment 1 and 2, respectively. In this case, $u(\theta; \emptyset) = 0$, $u(\theta; \{1\}) = \theta_1$, and $u(\theta; \{2\}) = \theta_2$. If there are no interaction effects, so that the welfare effect from implementing both treatments is equal to the sum of effects from implementing each of them separately, then

$$u(\theta; \{1, 2\}) = (\theta_1, \theta_2, \theta_1 + \theta_2). \tag{8}$$

□

The planner's objective. We consider a planner who wishes to increase welfare while also limiting the possibility of harm due to mistaken conclusions, and to encourage research. Specifically, the planner chooses r to maximize

$$U(r; \lambda, \pi) = \min_{\theta \in \Theta} v_r(\theta, J^*_{r,\theta}, \Sigma^*_{r,\theta}) + \lambda \int e^*_r(\theta)\pi(\theta)d\theta, \tag{9}$$

where $\lambda \geq 0$ and $\pi(\theta) \geq 0$ for all $\theta \in \Theta$. The first component, which depends on which treatments are actually implemented, captures the desire to raise welfare while limiting harm using a standard ambiguity-averse (maximin) formulation. The second component depends on whether or not the researcher experiments. (If π is a probability density, that second component is equal to the probability of experimentation since $\int e^*_r(\theta)\pi(\theta)d\theta = \int 1\{J^*_{r,\theta} \neq \emptyset\}\pi(\theta)d\theta$). It can be interpreted as capturing the benefits of scientific research per se, and as internalizing some aspects of the researcher's utility. In Appendix B.3, we formalize the latter interpretation, showing that protocols that are optimal under U remain approximately optimal if the second component is replaced by $\int \beta^*_r(\theta)\pi(\theta)d\theta$, the expected researcher utility (if π is a density). The parameter λ allows us to trade-off each of these components. We show below that under suitable assumptions on π, the two components of U have intuitive connections to the statistical concepts of size control and power. Moreover, U admits optimal protocols that do not depend on λ and π. This is important because the relative importance of the two components may be difficult to determine and choosing high-dimensional weights π is difficult and often somewhat arbitrary. Working with U thus provides a cogent rationale for testing protocols that control size, have non-trivial power, and do not depend on λ and π.

In the regulatory approval example, the structure of the planner objective U is motivated by regulators such as the FDA being tasked by legislators with several distinct objectives (e.g., U.S. Food and Drug Administration, ndb). Each component of Equation (9) relates to a distinct objective. The first captures the desire to avoid implementing harmful treatments, as for example under the "do no harm" principle (since, we show, our framework naturally rationalizes one-sided hypothesis testing). The second captures the broader value of scientific research, which need not be directly related to the immediate regulatory decision being made.[12] The relative importance of these two objectives is generally not specified, however, which motivates focusing on protocols that are optimal for all $\lambda \geq 0$.

The weighting of ambiguity-averse and expected-utility components in the planner objective U echoes a long tradition in economic theory (e.g., Gilboa and Schmeidler, 1989; Banerjee et al., 2020). The planner objective U differs from (but, as we discuss below, approximates) the objectives in Gilboa and Schmeidler (1989) and Banerjee et al. (2020), which, in our notation, correspond to

$$U'(r; \lambda, w) = \min_{\theta \in \Theta} v_r(\theta, J^*_{r,\theta}, \Sigma^*_{r,\theta}) + \lambda \int v^*_r(\theta, J^*_{r,\theta}, \Sigma^*_{r,\theta}) w(\theta) d\theta, \tag{10}$$

for weights $w(\theta)$. The second component of U' captures the welfare from implementing the treatments, whereas the second component of U captures a preference for experimentation. The objective U' has a decision-theoretic interpretation (Gilboa and Schmeidler, 1989) and is related to Huber's ε-contamination model (see Banerjee et al., 2020). However, Appendix B.3 shows that exact solutions under U' do not necessarily guarantee size control (as the solution depends on λ and w). By contrast, working with the planner objective U allows us to justify notions of size control and power and to obtain optimal protocols that do not depend on w and λ, while retaining an approximate decision-theoretic justification.

[12] As the international guidelines for clinical trials state, for example, "the rationale and design of confirmatory trials nearly always rests on earlier clinical work carried out in a series of exploratory studies" (Lewis, 1999). More broadly, the results of one study may lead to new conceptual insights or scientific hypotheses which are valuable independent of any immediate clinical application.

3 Optimal hypothesis testing protocols

In this section, we characterize optimal hypothesis testing protocols without imposing additional functional form restrictions on the researcher's payoff or the planner's utility.

3.1 Null space, alternative space, and notions of optimality

Null space and alternative space. For a given set of treatments J, define the (global) *null space*, the set of parameters for which the welfare effect of implementing any combination of treatments is negative, as

$$\Theta_0(J) = \left\{\theta : u_k(\theta; J) < 0 \text{ for all } k \in \{1, \ldots, 2^{|J|} - 1\}\right\} \subseteq \Theta. \tag{11}$$

Similarly, define the null space given the treatments chosen by the researcher ex-ante after excluding the option not to experiment, $J^+_{r,\theta}$, as

$$\Theta_0^*(r) = \left\{\theta : u_k(\theta, J^+_{r,\theta}) < 0 \text{ for all } k \in \{1, \ldots, 2^{|J^+_{r,\theta}|}\}\right\}.$$

Moreover, define the (global) *alternative space*, the set of parameters for which welfare is always positive, as

$$\Theta_1(J) = \left\{\theta : u_k(\theta; J) \geq 0 \text{ for all } k \in \{1, \ldots, 2^{|J|} - 1\}\right\} \subseteq \Theta. \tag{12}$$

A graphical illustration of $\Theta_0(J)$ and $\Theta_1(J)$ is provided in Figure 2. We impose the following assumption.

Assumption 3.1 (Non-emptiness of null and alternative space)**.** Let $\bigcap_{J \in \mathcal{J} \setminus \emptyset} \Theta_0(J) \neq \emptyset$ and $\bigcap_{J \in \mathcal{J} \setminus \emptyset} \Theta_1(J) \neq \emptyset$ (and therefore $\Theta_0(J), \Theta_1(J) \neq \emptyset$ for all $J \neq \emptyset$).

Assumption 3.1 states that for all combinations of treatments J, welfare is strictly negative for some values of θ and weakly positive for some other values of θ.[13]

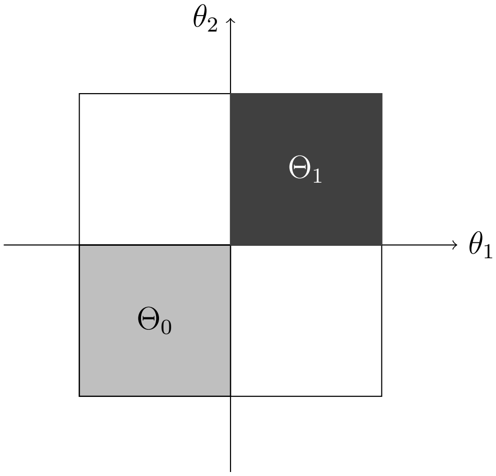
Finally, denote by $\bar{\Theta}_1$ the set of parameters for which welfare is weakly positive for each choice of treatments J, $\bar{\Theta}_1 = \bigcap_{J \in \mathcal{J} \setminus \emptyset} \Theta_1(J)$.

Notions of optimality. The main notion of optimality we consider is uniform global optimality. We say that a protocol r^* is *uniformly globally optimal* if

$$r^* \in \arg\max_{r \in \mathcal{R}} \left\{\min_{\theta \in \Theta} v_r(\theta, J^*_{r,\theta}, \Sigma^*_{r,\theta}) + \lambda \int_\Theta e^*_r(\theta')\pi(\theta')d\theta'\right\}, \quad \forall \lambda \geq 0, \pi \in \Pi, \tag{13}$$

[13] This precludes interventions that everyone believes are sure to do good or those sure to cause harm, which are not of interest and would be precluded by, for example, rules regarding research ethics.

Figure 2: Graphical illustration



Notes: Graphical illustration of the null space $\Theta_0(J) = \{\theta \in \Theta : \theta_1 < 0 \text{ and } \theta_2 < 0\}$ and the alternative space $\Theta_1(J) = \{\theta \in \Theta : \theta_1 \geq 0 \text{ and } \theta_2 \geq 0\}$ for $J = \{1,2\}$, with additive welfare as in (8).

for a given set Π. Uniformly globally optimal protocols do not depend on λ and π, which is important in practice, as argued above. In Appendix B.3, we discuss the relationship between uniform global optimality and alternative notions of optimality.

We also introduce two additional definitions that will be helpful for characterizing uniformly globally optimal protocols: maximin optimality and unbiasedness. We say that r^* is *maximin optimal* if it maximizes the planner's objective (9) for $\lambda = 0$, that is,

$$r^* \in \arg\max_{r \in \mathcal{R}} \min_{\theta \in \Theta} v_r(\theta, J^*_{r,\theta}, \Sigma^*_{r,\theta}).$$

We say that r^* is *unbiased* if

$$\beta^*_{r^*}(\theta) \geq 0 \text{ for all } \theta \in \bar{\Theta}_1, \text{ where } \beta^*_{r^*}(\theta) = \max_{J \in \mathcal{J} \setminus \emptyset, \Sigma \in \mathcal{S}(J)} \beta_{r^*}(\theta, J, \Sigma). \tag{14}$$

Here $\beta^*_r(\theta)$ is the largest net benefit the researcher can achieve when conducting an experiment. We use the term "unbiased" because, as we show below, this definition has a close connection to the definition of unbiased tests in the hypothesis testing literature, where a test is called unbiased if its power exceeds its size (Lehmann and Romano, 2005).

3.2 Characterizing optimal protocols

Here, we characterize uniformly globally optimal and maximin optimal protocols. We impose the following assumption on the planner's weights π.

Assumption 3.2 (Weights over $\bar{\Theta}_1$)**.** Suppose that $\pi \in \Pi$, where Π is the class of functions π satisfying $\pi(\theta) \geq 0$ for all $\theta \in \Theta$ and $\int_{\Theta} \pi(\theta) d\theta = \int_{\bar{\Theta}_1} \pi(\theta) d\theta > 0$.

Assumption 3.2 restricts the support of the planner's weights to the alternative space $\bar{\Theta}_1$. In other words, the planner desires to promote experimentation if she expects that treatments will generate a positive welfare effect and are therefore worth exploring. She derives no benefit (but also no harm), on the other hand, from exploration of parts of the parameter space in which some treatment effects are negative.

Using classical hypothesis testing terminology, Assumption 3.2 allows for arbitrary alternative hypotheses over the positive orthant, including those in Section 4 of Romano et al. (2011) and Chapter 9.2 of Lehmann and Romano (2005). Economically speaking, one can think of this assumption as ensuring that the components of the planner's utility function (9) cleanly separate the two motives we wish to capture: avoiding harm and pursuing benefit. Doing so has the benefit that the components of utility will then map directly into the classical statistical concepts of size and power. As we discuss in more detail in Remark 3.1, Assumption 3.2 is necessary to justify testing protocols that control size, and we thus see Assumption 3.2 as natural when size control is a desideratum.

The following proposition characterizes uniformly globally optimal protocols.

Proposition 3.1 (Necessary and sufficient conditions for uniform global optimality)**.** *Suppose that Assumptions 3.1 and 3.2 hold and that a maximin optimal and unbiased protocol exists. Then a protocol is uniformly globally optimal if and only if it is maximin optimal and unbiased.*

Proof. See Appendix D.1. □

Proposition 3.1 shows that uniformly globally optimal protocols are maximin optimal and unbiased. It assumes that a maximin and unbiased protocol exists. While the existence of such protocols is not guaranteed in general, we will show in Section 4 that such protocols exist in leading cases. In the remainder of this section, we study maximin optimality and unbiasedness in more detail and show that these properties are related to notions of size control and power of testing protocols.

The next proposition provides necessary and sufficient conditions for maximin optimality. It generalizes Proposition 1 in Tetenov (2016) (discussed in Appendix C) to a setting in which $|J| > 1$ and where researchers can choose the experimental design and hypotheses to test.

Proposition 3.2 (Necessary and sufficient conditions for maximin optimality)**.** *Let Assumptions 3.1 hold. A protocol r^* is maximin optimal if and only if*

$$\begin{aligned} \beta_{r^*}(\theta, J^*_{r^*,\theta}, \Sigma^*_{r^*,\theta}) &\leq 0 \qquad && \forall \theta \in \Theta_0^*(r^*) \\ v_{r^*}(\theta, J^*_{r^*,\theta}, \Sigma^*_{r^*,\theta}) &\geq 0 && \forall \theta \in \Theta \setminus \Theta_0^*(r^*). \end{aligned} \tag{15}$$

Proof. See Appendix D.2. □

Proposition 3.2 shows that maximin optimality is equivalent to two conditions. First, as in the case with a single hypothesis (i.e., $|J| = 1$), maximin protocols deter experimentation over the null space $\Theta_0^*(r^*)$, where each configuration of treatments reduces welfare. This captures a notion of size control, as we discuss below. Second, welfare must be non-negative for $\theta \in \Theta \setminus \Theta_0^*(r^*)$. This condition requires that if some treatments reduce welfare there must be others that compensate for them; it always holds in the single-hypothesis case but is non-trivial in the MHT case.

Building on Proposition 3.1 and the sufficient conditions in Proposition 3.2, the following corollary provides sufficient conditions for uniform global optimality stated in terms of the global null and alternative spaces.

Corollary 3.1 (Sufficient condition for uniform global optimality)**.** *Let Assumption 3.1 hold. A protocol r^* is maximin optimal if, for all $J \in \mathcal{J} \setminus \emptyset$ and $\Sigma \in \mathcal{S}(J)$,*

$$\beta_{r^*}(\theta, J, \Sigma) \leq 0 \quad \forall \theta \in \Theta_0(J) \tag{16}$$

$$v_{r^*}(\theta, J, \Sigma) \times 1\{\beta_{r^*}(\theta, J, \Sigma) > 0\} \geq 0 \quad \forall \theta \in \Theta \setminus \Theta_0(J). \tag{17}$$

If, in addition, Assumption 3.2 holds and r^* satisfies, for all $J \in \mathcal{J} \setminus \emptyset, \Sigma \in \mathcal{S}(J)$,

$$\beta_{r^*}(\theta, J, \Sigma) \geq 0 \qquad \forall \theta \in \Theta_1(J), \tag{18}$$

then it is unbiased and therefore uniformly globally optimal.

Proof. See Appendix D.3. □

Conditions (16) and (17) in Corollary 3.1 mimic those in Proposition 3.2, but are required to hold for all potential choices of J and Σ. These conditions may be easier to check than those in Proposition 3.2 because they are stated in terms of $\Theta_0(J)$ rather than $\Theta_0^*(r^*)$, which is itself a function of the set of treatments pre-specified by the researcher in response to r^*. Note that in Condition (17), we need weakly positive welfare only when the researcher finds it beneficial to experiment, since otherwise the researcher will not experiment and hence welfare will be zero. Condition (18) captures a stronger notion of unbiasedness that implies unbiasedness as defined in Equation (14).

The theoretical results in this section establish close connections between uniform global optimality on the one hand, and size control and the power of testing protocols on the other. Specifically, maximin optimality captures a notion of size control, whereas unbiasedness captures a notion of power.

Example 3.1 (Average size control and unbiasedness with linear researcher benefits)**.** Suppose that $J = \{1, 2\}$ and that welfare is additive as in Equation (8). Then the null space for $J = \{1, 2\}$ is $\Theta_0(\{1, 2\}) = \{\theta \in \Theta : \theta_1 < 0 \text{ and } \theta_2 < 0\}$ and the alternative space is $\Theta_1(\{1, 2\}) = \{\theta \in \Theta : \theta_1 \geq 0 \text{ and } \theta_2 \geq 0\}$. Figure 2 provides a graphical illustration. Suppose further that the researcher's payoff is

$$\beta_r(\theta; J, \Sigma) = b \sum_{j \in J} P(r_j(X; J, \Sigma) = 1|\theta) - C(J, \Sigma)$$

for some constant $b > 0$. This payoff is a special case of the payoff in Equation (4), where $p(\{1\}) = p(\{2\}) = b, p(\{1, 2\}) = (b, b, 2b)$. Condition (16) in Corollary 3.1 implies that

$$P(r_1^*(X; J, \Sigma) = 1|\theta_1, \theta_2) + P(r_2^*(X; J, \Sigma) = 1|\theta_1, \theta_2) \leq C(J, \Sigma)/b, \quad \forall \theta_1 < 0, \theta_2 < 0. \quad (19)$$

Equation (19) is a size control requirement (i.e., a restriction on the probability of reporting false discoveries). Note that the right-hand side $C(J, \Sigma)/b$ is not a function of r. Here, r^* satisfies Condition (18) in Corollary 3.1 if and only if

$$P(r_1^*(X; J, \Sigma) = 1|\theta_1, \theta_2) + P(r_2^*(X; J, \Sigma) = 1|\theta_1, \theta_2) \geq C(\{1, 2\}, \Sigma)/b, \ \forall \theta_1 \geq 0, \theta_2 \geq 0. \ (20)$$

Equation (20) can be interpreted as a power criterion: it requires the power of the protocol r^* to (weakly) exceed the cost-to-benefit ratio of the experiment whenever $\theta \geq 0$. Taken together, Conditions (19) and (20) imply (under suitable continuity assumptions) that $P(r_1^*(X;J,\Sigma)=1|\theta_1,\theta_2)+P(r_2^*(X;J,\Sigma)=1|\theta_1,\theta_2)=C(J,\Sigma)/b$ for $\theta_1=\theta_2=0$. In Section 4, we show that separate t-tests satisfy these optimality criteria. □

Example 3.2 (Weak FWER control with nonlinear research benefits)**.** Consider the setup in Example 3.1, but suppose instead that the researcher's payoff is

$$\beta_r(\theta;J,\Sigma)=bP\left(\sum_{j\in J}r_j(X;J,\Sigma)\geq 1|\theta\right)-C(J,\Sigma)$$

for some constant $b>0$, which is a special case of the payoff in Equation (4) with $p_j(J)=b$ for all j. Such a payoff might arise, for instance, if the firm conducting a trial needs at least one success in order to be solvent (or, when applied to academic publishing, if a researcher needs at least one significant result in order to publish). Condition (16) in Corollary 3.1 implies that

$$P(\max_j r_j^*(X;J,\Sigma)=1|\theta_1,\theta_2)\leq C(J,\Sigma)/b,\quad \forall\theta_1<0,\theta_2<0. \tag{21}$$

That is, it implies weak control of the FWER (which coincides with the positive FDR for $\theta\in\Theta_0(J)$). See Appendix B.2 for additional details. □

As discussed in Section 2, rationalizing hypothesis testing (let alone multiple testing) is difficult in practice. The results and examples in this section show that one can write down a coherent economic objective function that rationalizes the standard statistical practice of choosing protocols that both control size and have non-trivial power. Specifically, optimal protocols must guarantee size control (encoded in the maximin optimality requirement) and also guarantee sufficient power against alternatives (encoded in the unbiasedness). Expressing these requirements in the form of an optimization problem has the benefit that it will allow us to then link the notions of size control and compound error rates directly to economic fundamentals, depending on the researcher's private costs and benefits and on how these scale with the number of hypotheses.

Remark 3.1 (Assumption 3.2 and size control)**.** Restrictions on Π such as those in Assumption 3.2 are necessary to justify hypothesis testing protocols that control size (are maximin optimal). Suppose instead that Π contains all possible positive weights integrating to one (i.e., all prior distributions) on $\Theta\setminus\Theta_0(J)$ (the same reasoning applies if we simply consider all priors on $\Theta_0(J)$). Then there are no r^* satisfying Equation (13) (for all $\lambda\geq 0,\pi\in\Pi$).

To see this, fix a prior $\pi = 1\{\theta = (-1, 0, \ldots, 0)\}$ on $\Theta \setminus \Theta_0(J)$. Then for sufficiently large λ, assuming the researcher's payoff is strictly increasing in the number of findings, the protocol $r(X; J, \Sigma) = (1, \ldots, 1)$ dominates any maximin protocol, since when λ is large enough the planner would prefer the researcher to experiment even when the resulting welfare effects are negative. In this example, the cost of approving harmful treatments is outweighed by the benefits of incentivizing experimentation for sufficiently large λ. Therefore, there exist combinations of π and λ such that the planner chooses protocols that do not control size and instead incentivizes experimentation regardless of the value of θ. Restricting Π to the strictly positive orthant guarantees that optimal protocols control size (are maximin optimal) for all λ and motivates sufficiently powerful protocols only when θ is expected to have positive effects. □

Remark 3.2 (Deterrence of experimentation under the null)**.** Proposition 3.2 implies that experiments testing welfare-reducing treatments never occur in equilibrium, which might seem unrealistic. This result is a consequence of the simplifying assumption that the researcher has perfect information about θ. Section 5.1 extends our analysis to settings where the researchers has a prior about θ, characterizing testing protocols that are maximin optimal with respect not just to point mass priors (equivalent to the problem in the main model) but with respect to arbitrary priors. In this scenario optimal protocols deter experimentation by a researcher who is *certain* that some treatments are welfare-reducing, but may not deter testing by one who believes that treatment is very likely to be welfare-increasing but possibly harmful. □

3.3 Robustness to researcher constraints in the design choice

So far, we have assumed that after observing the protocol r, the researcher can choose any experiment (J, Σ) with $J \in \mathcal{J}$ and $\Sigma \in S(J)$. In some applications, however, researchers may face constraints that restrict the menu of treatments and designs they can implement. Because the planner may not know ex ante which experiments are feasible, we consider a refined notion of global optimality, referred to as *design-robust global optimality*, that builds in robustness to design constraints.

Definition 3.1 (Design-robust global optimality)**.** Consider a constrained environment in which the researcher can choose $(J, \Sigma) \in \mathcal{D}$ from a constrained set $\mathcal{D} \subseteq \{(J, \Sigma) : J \in \mathcal{J}, \Sigma \in \mathcal{S}(J)\}$. For a given protocol r and parameter θ, let $\left(J^*_{r,\theta}(\mathcal{D}), \Sigma^*_{r,\theta}(\mathcal{D})\right) \in \arg\max_{(J',\Sigma')\in\mathcal{D}} \beta_r(\theta, J', \Sigma')$. We say that a protocol is *design-robust globally optimal* if

$$r^* \in \arg\max_{r\in\mathcal{R}} \left\{ \min_{\theta\in\Theta} v_r(\theta, J^*_{r,\theta}(\mathcal{D}), \Sigma^*_{r,\theta}(\mathcal{D})) + \lambda \int_{\Theta} 1\{J^*_{r,\theta}(\mathcal{D}) \neq \emptyset\}\, \pi(\theta')\, d\theta' \right\},$$

for all $\lambda \geq 0$, $\pi \in \Pi$, and $\mathcal{D} \subseteq \{(J, \Sigma) : J \in \mathcal{J}, \Sigma \in \mathcal{S}(J)\}$.

Definition 3.1 requires protocols to be optimal for every potential set of experiments (J, Σ) available to the researcher. Without the refinement of global optimality in Definition 3.1, the planner can afford to select protocols that are not unbiased (and hence may lead to low power) for designs that she expects the researcher not to choose. With it, she must consider the possibility that the researcher may be forced to choose any design, and thus must ensure that protocols are always unbiased.

Definition 3.1 is appealing in settings where the planner has limited ex-ante knowledge of the researcher's feasibility constraints, and thus wishes to mandate a protocol that performs well uniformly across all designs.

The following proposition provides a characterization of design-robust globally optimal protocols.

Proposition 3.3. *Let Assumptions 3.1 and 3.2 hold. Then r^* is design-robust globally optimal if and only if Equations (16), (17), and (18) hold for all $J \in \mathcal{J} \setminus \emptyset$ and $\Sigma \in \mathcal{S}(J)$, assuming such a protocol exists.*

Proof. See Appendix D.4. □

Proposition 3.3 shows that the sufficient conditions for global optimality in Corollary 3.1 become necessary once we strengthen the notion of global optimality to design-robust global optimality.

3.4 Robustness to uncertainty about the researcher's payoff

So far, we have assumed that the planner knows the researcher's payoff. The following proposition shows that maximin optimality for protocols satisfying Equations (16) and (17) is preserved if the planner knows only an upper bound on the researcher's payoff. Uniform global optimality is preserved under additional restrictions on Π.

Proposition 3.4 (Robustness to unknown researcher payoffs)**.** *Let Assumptions 3.1 and 3.2 hold. Then*

(i) Any protocol r^* satisfying Equations (16) and (17) under payoff function $\beta_{r^*}(\theta, J, \Sigma)$ is also maximin optimal for any $\beta'_{r^*}(\theta, J, \Sigma)$ such that $\beta'_{r^*}(\theta, J, \Sigma) \leq \beta_{r^*}(\theta, J, \Sigma)$ for all $\theta \in \Theta, J \in \mathcal{J}, \Sigma \in \mathcal{S}(J)$.

(ii) Any protocol r^* satisfying the conditions in Corollary 3.1 under payoff function $\beta_{r^*}(\theta, J, \Sigma)$ is uniformly globally optimal for any $\beta'_{r^*}(\theta, J, \Sigma)$ such that $\beta'_{r^*}(\theta, J, \Sigma) \leq \beta_{r^*}(\theta, J, \Sigma)$ for all $\theta \in \Theta, J \in \mathcal{J}, \Sigma \in \mathcal{S}(J)$ and any positive function $\pi \in \tilde{\Pi}$, where $\tilde{\Pi} = \{\pi : \int_{\tilde{\Theta}_1(r^*)} \pi(\theta) d\theta = \int_\Theta \pi(\theta) d\theta\}$, $\tilde{\Theta}_1(r^*) = \{\theta : \beta'_{r^*}(\theta, J^*_{r^*,\theta}, \Sigma^*_{r^*,\theta}) \geq 0\}$.

Proof of Proposition 3.4. See Appendix D.5. □

Proposition 3.4(i) demonstrates an important robustness property of our maximin optimality results in settings where the researcher's payoff function is unknown. Proposition 3.4(ii) states that protocols that are uniformly globally optimal with respect to an upper bound $\beta_r(\theta, J, \Sigma)$ are also uniformly globally optimal for weights $\pi \in \tilde{\Pi}$. That is, uniform global optimality is preserved when considering a (weakly) smaller class of alternatives. For example, for the optimal separate t-tests in Section 4, the set of weights $\tilde{\Pi}$ is a subset of the set of weights on strictly positive treatment effects. Note that $\tilde{\Theta}_1(r^*)$ and $\tilde{\Pi}$ do not need to be known for the planner to implement the optimal protocol.

Proposition 3.4 is particularly useful when applied to settings in which the planner's uncertainty about the researcher's payoff hinges on the researcher's costs.

Corollary 3.2 (Unknown cost function)**.** *Let the conditions in Proposition 3.4 hold, and suppose that the researcher's costs do not depend on θ, so that $c_\theta(J, \Sigma) = C(J, \Sigma)$. Consider $\beta_r(\theta, J, \Sigma) - \beta'_r(\theta, J, \Sigma) = C'(J, \Sigma) - C(J, \Sigma)$, for some $C'(J, \Sigma) \geq C(J, \Sigma)$. Then any maximin and unbiased protocol r^* under β_r is also maximin under β'_r and uniformly globally optimal for any $\pi \in \tilde{\Pi}$, with $\tilde{\Pi}$ defined in Proposition 3.4.*

Proof. The proof follows directly from Proposition 3.4. □

Corollary 3.2 states that in settings with uncertainty over the true cost function $C'(J, \Sigma)$, the planner may use sensible lower bounds $C(J, \Sigma) \leq C'(J, \Sigma)$. This result is important in empirical applications such as the ones we consider in Section 6.

4 Optimal protocols under linearity and normality

Which (if any) specific hypothesis testing protocols are uniformly globally optimal? The answer depends on the functional form of the researcher's payoff, the functional form of welfare, and the distribution of X. Here, we show that separate t-tests are optimal in settings with a linear researcher payoff function (Assumption 4.1), a linear welfare function (Assumption 4.2), and a normally distributed vector of statistics X (Assumption 4.3).

4.1 Assumptions

Linearity assumptions. Let ω denote a vector of weights and define $\bar{\omega}(J) = \sum_{j=1}^{|J|} \omega_{J_j}$ for $J \in \mathcal{J}$. These weights will let us capture factors that affect the importance of the different treatments symmetrically from the point of view of both the researcher and the planner; they also nest the case in which all treatments are equally important ($\omega_{J_j} = 1 \ \forall j$). We consider the following assumption on the researcher's payoff.

Assumption 4.1 (Linear payoff)**.** The researcher's payoff is (recall that the entry $r_j(X; J, \Sigma)$ corresponds to treatment J_j)

$$\beta_r(\theta, J, \Sigma) = b \int \sum_{j=1}^{|J|} \omega_{J_j} r_j(x; J, \Sigma) dF_{\theta, J, \Sigma}(x) - C(J, \Sigma), \tag{22}$$

where $\omega_j \in (0, \infty)$ for all j, $b > 0$, $b\bar{\omega}(J) > C(J, \Sigma) > 0$ for all (J, Σ), and $C(J, \Sigma)$ does not depend on θ.

The payoff function (22) in Assumption 4.1 is a special case of the general payoff function (4), and the condition $b\bar{\omega}(J) > C(J, \Sigma)$ guarantees that the experiment (J, Σ) is a relevant option; otherwise the researcher would never conduct this experiment, regardless of r. Note that Assumption 4.1 rules out interactions between treatments in the researcher's utility, which we discuss in Appendix B.2. In addition, we assume that the researcher costs do not depend on θ and write them as $C(J, \Sigma)$.

We consider the following linearity assumption on welfare.

Assumption 4.2 (Linear welfare)**.** For all $r \in \mathcal{R}$,

$$\delta(r(x; J, \Sigma); J)^\top u(\theta; J) = \sum_{j=1}^{|J|} \omega_{J_j} r_j(x; J, \Sigma) u(\theta; J_j),$$

where $u(\theta; J_j)$ is the welfare from implementing treatment J_j.

Assumptions 4.1 and 4.2 capture a setting in which the researcher's and planner's objectives differ: the expected researcher's payoff depends on the (weighted) expected number of findings, whereas the welfare component of the planner's objective depends on the welfare effects that such findings generate. Before stating results, we provide an interpretation of these assumptions in the context of our leading example, the drug approval process.

With *multiple subgroups*, we interpret $r_j(X; J, \Sigma)$ as indicating whether the drug was found to be effective for subgroup J_j, which is of size ω_{J_j}. The component $b\omega_{J_j}$ denotes the expected profits from selling the drug to subgroup J_j, where b denotes the average per-sale profit. Assumption 4.1 then states that researchers care about the sum of the expected profits they can earn by selling the drug to each of the subpopulations for which its use is approved. Our specification of welfare in Assumption 4.2, meanwhile, corresponds to the utilitarian welfare from approving the drug, where $u(\theta; J_j)$ denotes the per unit treatment effect on members of subgroup J_j. The economic import of the assumption is that there are no spillovers between different subgroups.

With *multiple treatments* the interpretation is similar, but here each treatment denotes a different drug in the same market. ω_{J_j} denotes the expected number of individuals that would purchase and use drug J_j if approved, and $u(\theta; J_j)$ is the effect of drug J_j on those individuals. As before, b denotes the average per-sale profit. The economic import of Assumption 4.2 is that the sets of people who would use the different drugs are disjoint (or that the drugs do not exhibit interaction effects). Symmetry between the planner's and researcher's weights here implies that per-user profit and per-user consumer surplus are proportional across drugs. Without this property it is no longer necessarily the case that the separate t-tests defined in Equation (23) below guarantee uniformly non-negative welfare (and therefore are maximin optimal); we do not characterize optimal testing protocols in that scenario, but note that this could be a fruitful direction for future work.

Normality assumption. We focus on the leading case where X is normally distributed, which is motivated by standard normal approximations (e.g., Berry-Esseen bounds). Let θ_J denote the subvector of θ corresponding to treatments J.

Assumption 4.3 (Normality and homogeneous variances)**.** For each experiment (J, Σ), let $X \sim \mathcal{N}(\theta_J, \Sigma)$ and suppose that $u(\theta, j) = \theta_j$, with $\theta_j \in [-M, M]$ for a positive constant $M > 0$, so that X_j is an unbiased normally distributed signal of the welfare effect of treatment j. The class of designs $\mathcal{S}(J)$ is such that $\Sigma_{i,i} = \Sigma_{j,j} \in (\underline{\gamma}, \bar{\gamma})$ for all $i, j \in \{1, \ldots, |J|\}$ and some constants $0 < \underline{\gamma} < \bar{\gamma} < \infty$.

Assumption 4.3 imposes that the vector of statistics X is normally distributed, centered around the vector of welfare effects θ_J and with finite sample covariance matrix Σ.[14] The researcher can choose the covariance matrix, but we restrict the class of designs she can choose from to designs in which the variances (the diagonal entries of Σ) are positive and equal to each other. This implies that the researcher can choose the overall sample size of the experiment, for example, but is constrained in allocating sample across experimental arms. See Section 5.2 for an extension to designs with heterogeneous variances.

4.2 Optimality of t-tests

The following proposition shows that separate t-tests are maximin optimal under Assumptions 4.1, 4.2, and 4.3 and uniformly and design-robust globally optimal if, in addition, Assumption 3.2 holds.

Proposition 4.1 (Optimality of separate t-tests)**.** *Let Assumptions 4.1, 4.2, and 4.3 hold. Consider the testing protocol* $r^t(X; J, \Sigma) = \left(r^t_1(X; J, \Sigma), \ldots, r^t_{|J|}(X; J, \Sigma)\right)^\top$*, with*

$$r^t_j(X; J, \Sigma) = 1\left\{\frac{X_j}{\sqrt{\Sigma_{j,j}}} \geq t(J, \Sigma)\right\}, \forall j \in J, \tag{23}$$

for a threshold $t(J, \Sigma) \in \mathbb{R}$*. Then* r^t *is maximin optimal if and only if* $t(J, \Sigma) \geq \Phi^{-1}\left(1 - \frac{C(J,\Sigma)}{b\bar{\omega}(J)}\right)$ *for all* (J, Σ)*. If, in addition, Assumption 3.2 holds, then* r^t *is maximin optimal and unbiased (and thus uniformly globally optimal) if and only if it also satisfies* $t(J, \Sigma) = \Phi^{-1}\left(1 - \frac{C(J,\Sigma)}{b\bar{\omega}(J)}\right)$ *for at least some* (J, Σ) *with* $J \neq \emptyset$*. Finally,* r^t *is also design-robust globally optimal if and only if* $t(J, \Sigma) = \Phi^{-1}\left(1 - \frac{C(J,\Sigma)}{b\bar{\omega}(J)}\right)$ *for all feasible* $J \neq \emptyset$ *and* Σ*.*

[14] Because Σ is the finite sample covariance matrix, it is proportional to the inverse of the square root of the sample size in the experiment.

Proof. See Appendix D.6. □

Proposition 4.1 shows that separate one-sided t-tests with critical values $t \geq \Phi^{-1}\left(1 - \frac{C(J,\Sigma)}{b\bar{\omega}(J)}\right)$ are maximin optimal, where we write t in lieu of $t(J, \Sigma)$ to simplify notation. The key technical step in the proof is to show that under this protocol welfare is non-negative even when parameters have different signs (Equation (17)). Proposition 4.1 further shows that standard separate one-sided t-tests are maximin optimal and unbiased and thus uniformly globally optimal by Proposition 3.1 and also design-robust globally optimal.[15] While t-tests with all thresholds larger than $\Phi^{-1}(1 - \frac{C(J,\Sigma)}{b\bar{\omega}(J)})$ are maximin optimal, such tests are not uniformly globally optimal. Only t-tests with threshold $t = \Phi^{-1}(1 - \frac{C(J,\Sigma)}{b\bar{\omega}(J)})$ for at least some (J, Σ) are uniformly globally optimal, and only t-tests with threshold $t = \Phi^{-1}(1 - \frac{C(J,\Sigma)}{b\bar{\omega}(J)})$ for all experiments (J, Σ) are design-robust globally optimal. This demonstrates that uniform global optimality is a refinement of maximin optimality, restricting attention to protocols with sufficient power for at least some experiment, and that design-robust global optimality is in turn a refinement of uniform global optimality, restricting attention to protocols with sufficient power for all experiments.

Proposition 4.1 also shows that whether and to what extent the level of these separate tests should depend on the number of hypotheses being tested depends on the structure of the research production function $C(J, \Sigma)$ and on $\bar{\omega}(J)$, and in particular on how they vary with J. For example, suppose that $\omega_j = 1$ for all j (all treatments are equally important) so that $\bar{\omega}(J) = |J|$. If $C(J, \Sigma) = \alpha$ for some constant α then a Bonferroni correction is optimal. This corresponds to a stylized case in which the costs of experimentation are fixed regardless of the number of treatments tested ($|J|$) or the precision of the estimates (Σ). If, on the other hand, $C(J, \Sigma) = \alpha|J|$ then the optimal level of the test is α, irrespective of $|J|$. This might correspond, for example, to a case in which there are no fixed costs and testing each additional treatment requires the same increment to the sample size. The former case arguably captures the lay intuition that if the researcher can test many hypotheses in

[15] It may seem surprising that the result in Proposition 4.1 holds for any λ. Intuitively, once we focus on Π defined in Assumption 3.2, we can always find a maximin protocol that guarantees experimentation for each value of θ in the positive orthant. As a result, no matter how much weight the planner puts on her subjective utility, at the optimum, any optimal protocol will maximize separately (and therefore jointly) maximin welfare and subjective utility from experimentation.

the hopes of securing some private benefit, then the planner should require a hypothesis testing protocol that discourages this. In the latter case it is still true that the researcher obtains a higher expected reward from taking on projects that test more hypotheses, ceteris paribus, but the appropriate correction for this is already "built in" to the costs of conducting research, so that no further correction is required.

While our original motivation for obtaining the result in Proposition 4.1 was to study the consequences of multiplicity ($|J|$), it follows immediately that, because the costs $C(J, \Sigma)$ may depend on the design Σ in addition to J, the optimal critical values may as well. For example, if the researcher can choose the number of treatments to test and also the sample size $\bar{n}$ to use per treatment, then her cost structure might take the form $C(J, \Sigma) = c_f + c_{|J|}|J| + c_{\bar{n}}|J|\bar{n}$, where c_f is a fixed cost (e.g., the costs of staff scientists), $c_{|J|}$ a cost that varies with the number of treatments tested (e.g., the cost of training clinical staff on various treatment protocols), and $c_{\bar{n}}$ a cost per experimental subject (e.g., recruitment costs). In this case (normalizing $b = 1$ and again assuming for simplicity that $\bar{\omega}(J) = |J|$) we obtain optimal thresholds $t = \Phi^{-1}(1 - c_f/|J| - c_{|J|} - c_{\bar{n}}\bar{n})$ which are decreasing in the (per-treatment) sample size $\bar{n}$ as well as in the cost per treatment. Intuitively, to the extent that large-sample experiments are more costly to the researcher to run, the planner need worry less about discouraging the researcher from running such experiments when doing so would not be socially optimal. This point follows from exactly the same economic logic that rationalizes adjusting testing thresholds with respect to J, but has not (to our knowledge) come up in past discussions of MHT adjustment. In that sense, it illustrates the value of working out the economic logic underlying MHT adjustment carefully, to make sure we have fully grasped the consequences of any implicit assumptions.

The practical value of Proposition 4.1 lies in the fact that it connects optimal testing protocols to measurable properties of the cost function $C(J, \Sigma)$. We illustrate this in Section 6.1 where we develop the application to clinical trials, using publicly available data on moments of their cost structure to derive specific testing thresholds. Appendix A provides a second illustration, applying the framework to experimental program evaluation research in economics using unique data from the Jameel Poverty Action Lab (J-PAL). Readers primarily interested in implications for practice may wish to skip ahead to these exercises.

Remark 4.1 (One-sided vs. two-sided tests)**.** Under the assumptions in this section, separate one-sided t-tests are uniformly globally optimal. This is because the status quo remains in place if the researcher does not report any findings. This may be one reason that one-sided tests have been seen as appropriate in the drug approval context.[16] When there is uncertainty about the planner's action when no findings are reported, on the other hand, our model can justify two-sided hypothesis testing. We report this result in Appendix B.1. □

Remark 4.2 (Average size control and FWER control)**.** When the researcher's payoff is linear (Assumption 4.1) and $\omega_j = 1$ for all j, Proposition 4.1 implies that we can find uniformly globally optimal protocols that impose average size control, $b/|J| \sum_{j \in J} P(r_j^*(X; J, \Sigma) = 1 | \theta = 0) \leq C(J, \Sigma)/|J|$. Many popular MHT corrections do not directly target average size control and, thus, will generally not be optimal in our model. For example, if $C(J, \Sigma)$ is constant then classical Bonferroni correction is optimal in our model, since it satisfies average size control, but common refinements of Bonferroni such as Holm (1979)'s method are not. This result is driven by the linearity of the researcher's payoff function; Bonferroni corrections may not be optimal with nonlinear payoff functions (but are maximin optimal for all payoff functions dominated by a linear payoff function; see Proposition 3.4). We discuss the impact of non-linearities for the design of optimal protocols in Appendix B.2. □

5 Main extensions

Here we present two extensions of our main results: a variant of our model where the researcher knows θ only imperfectly (Section 5.1) and a relaxation of the variance homogeneity requirement in Assumption 4.3 (Section 5.2). Appendix B presents additional extensions.

5.1 Imperfectly informed researchers

So far, we have assumed that the researcher is perfectly informed and knows θ. Here, we show that our main results continue to hold in settings where the researcher has imperfect

[16] For example, former FDA advisor Lloyd Fisher writes that "For drugs that may be tested against placebos, with two positive trials required (as in the United States), it is argued that from both a regulatory and pharmaceutical industry perspective, one-sided tests at the 0.05 significance level are appropriate. In situations where only one trial against a placebo may be done (for example, survival trials), one-sided tests at the 0.025 level are appropriate in many cases."(Fisher, 1991)

information in the form of a prior about θ.[17] Denote this prior by $\pi' \in \Pi'$, where Π' is the class of all distributions over Θ.[18] The prior π' captures knowledge about θ that is available to the researcher but not to the planner.

We assume that the vector of statistics X is drawn from a normal distribution conditional on θ, where θ itself is drawn from the prior π' with $\int_\Theta \pi'(\theta) d\theta = 1$:

$$X \mid \theta \sim \mathcal{N}(\theta, \Sigma), \quad \theta \sim \pi', \quad \pi' \in \Pi',$$

where Σ is positive definite and assumed to be known after being chosen by the researcher.

The researcher acts as a Bayesian decision-maker and chooses

$$\left(J^*_{r,\pi'}, \Sigma^*_{r,\pi'}\right) \in \arg \max_{J \in \mathcal{J}, \Sigma \in \mathcal{S}(J)} \int \beta_r(\theta, J, \Sigma) d\pi'(\theta).$$

The researcher's prior is correctly specified, and welfare is given by

$$\bar{v}_r(\pi', J, \Sigma) = \int v_r(\theta, J, \Sigma) d\pi'(\theta).$$

Under imperfect information, we define maximin protocols with respect to the prior π'.

Definition 5.1 (Π'-maximin optimal)**.** We say that r^* is Π'-maximin optimal if and only if

$$r^* \in \arg\max_{r \in \mathcal{R}} \inf_{\pi' \in \Pi'} \bar{v}_r(\pi', J^*_{r,\pi'}, \Sigma^*_{r,\pi'}).$$

Definition 5.1 generalizes the notion of maximin optimality in Section 3, which is stated in terms of the parameter θ. When Π' contains only point mass distributions, the two notions of maximin optimality are equivalent.

The next proposition shows that one-sided t-tests with appropriately chosen critical values are maximin optimal under imperfect information.

Proposition 5.1 (Maximin optimality)**.** *Let Assumptions 4.1, 4.2, and 4.3 hold. Then the protocol*

$$r^t_j(X; J, \Sigma) = 1\left\{ \frac{X_j}{\sqrt{\Sigma_{j,j}}} \geq \Phi^{-1}\left(1 - \frac{C(J, \Sigma)}{b\bar{\omega}(J)}\right) \right\}, \quad \forall j \in J,$$

is Π'-maximin optimal.

[17] In the single-hypothesis testing case, Tetenov (2016) gives results under imperfect information. However, these results rely on the Neyman-Pearson lemma, which is not applicable to multiple tests.

[18] The assumption that Π' is unrestricted is made for simplicity. For our theoretical results, we only need that the class of priors Π' contains at least one element that is supported on the null space $\bigcap_{J \in \mathcal{J} \setminus \emptyset} \Theta_0(J)$.

Proof. See Appendix D.7. □

Proposition 5.1 shows that the maximin optimality of separate t-tests continues to hold under imperfect information. Intuitively, maximin optimality is preserved here because the worst-case prior is a point mass prior over θ, so that the same reasoning as under perfect information applies. This follows from classical results on linear programs (see Appendix D.7). Uniform global optimality of r^t follows as a corollary of Proposition 4.1.

5.2 Heterogeneous variances

Assumption 4.3 restricts the class of designs the researcher can choose from to designs where all X_j have the same variance, $\Sigma_{i,i} = \Sigma_{j,j}$. This assumption may seem strong since one might expect the researcher to choose a design with unequal variances, especially if she believes that the outcomes are more variable under some treatments than under others. Here, we consider settings where the researcher can choose designs with heterogeneous variances. Specifically, we study a version of our model in which the researcher can choose the sample size for each treatment arm, and thus the variances of the test statistics, but has only imperfect knowledge of the underlying heterogeneous outcome variances in the treatment and control groups. See Remark 5.1 for a discussion of settings with heterogeneous variances where the researcher has full information and can thus choose the sample size as a function of such variances.

We assume that for a given (J, Σ), $X \sim \mathcal{N}(\theta_J, \Sigma)$, where for each j, we interpret θ_j as a ratio between the treatment effect τ_j and $\sigma_j := \sqrt{\sigma_{1j}^2/2 + \sigma_{0j}^2/2}$, where σ_{1j}^2 and σ_{0j}^2 are the variances of the treated and the control outcomes in the experiment, respectively. Define the vector of sample sizes in an experiment with treatments J as $n(J) = \{n_{1j}, n_{0j}\}_{j \in J}$. Here, n_{1j} and n_{0j} are the number of treated and control units in arm j (with n_{0j} potentially constant across j if all arms share the same control group). In this model, $\Sigma_{j,j} = \frac{\sigma_{1j}^2}{\sigma_j^2 n_{1j}} + \frac{\sigma_{0j}^2}{\sigma_j^2 n_{0j}}$.

Suppose that the researcher only knows θ, the effect measured in standard deviations τ_j/σ_j for each j, but not $(\tau_j, \sigma_{1j}^2, \sigma_{0j}^2)$ separately. To encode such uncertainty, we assume that the researcher has a prior $(\tau_j, \sigma_{1j}^2, \sigma_{0j}^2) \sim \mathcal{P}_{\theta,j}$, which depends on θ. Here $u(\theta, \{j\}) = \mathbb{E}[\tau_j|\theta]$ is the expected treatment effect τ_j given θ under the prior $\mathcal{P}_{\theta,j}$. We assume that (with a slight abuse of notation) $C(J, \Sigma) = C(J, n(J))$ (where $c_\theta(J, \Sigma) = C(J, \Sigma)$ is constant in θ), so that the costs are known to the researcher. That is, the costs can depend on the sample sizes in the experiment, $n(J)$, but not on the unknown to the researcher variances $\{\sigma_{1j}^2, \sigma_{0j}^2\}_{j\in J}$.

The following assumption summarizes the model with unknown heterogeneous variances.

Assumption 5.1 (Model with unknown heterogeneous variances)**.** Suppose that for all j, $\theta_j = \tau_j/\sigma_j$ for some $\sigma_j > 0$, with $\theta_j \in [-M, M]$, for a positive constant $M > 0$. Let $\sigma_j^2 = \sigma_{1j}^2/2 + \sigma_{0j}^2/2$, where σ_{0j}^2 and σ_{1j}^2 are bounded away from zero almost surely. Let $(\tau_j, \sigma_{1j}^2, \sigma_{0j}^2) \sim \mathcal{P}_{\theta,j}$ and $u(\theta, \{j\}) = \mathbb{E}[\tau_j|\theta]$. For an experiment with treatments J, let $X|(\tau,\sigma) \sim \mathcal{N}(\theta_J, \Sigma)$, with $\Sigma_{j,j} = \frac{\sigma_{1j}^2}{\sigma_j^2 n_{1j}} + \frac{\sigma_{0j}^2}{\sigma_j^2 n_{0j}}$ for all $j \in J$. The class of designs $\mathcal{S}(J)$ is such that $n_{1j}, n_{0j} \leq n$ for a finite constant $n < \infty$ and the researcher can choose $\{n_{1j}, n_{0j}\}_{j\in J}$. Finally, assume that $C(J, \Sigma) = C(J, n(J))$.

We consider a setting in which the researcher has limited knowledge, captured by the assumption that $(\sigma_{1j}, \sigma_{0j})$ have a common expectation across j, conditional on θ.

Assumption 5.2 (Common expectation)**.** For some $\bar{\sigma} > 0$, $\mathbb{E}[\sigma_j|\theta] = \bar{\sigma}$ for all j.

Under Assumption 5.2, the researcher expects the standard deviations to be the same before running the experiment. Importantly, however, Assumption 5.2 allows the realized variances to be heterogeneous.

Under Assumption 5.1, $\{n_{1j}, n_{0j}\}_{j\in J}$ are chosen by the researcher before running the experiment and observed by the planner. As a result, the planner can de-facto mandate any choice of sample sizes by only rewarding designs that maximize her utility.

We say that the design Σ has a *sample-equalizing allocation* if $n_{1j} = n_{0j} = \bar{m}$ for all $j \in \{1, \dots, |J|\}$ and for some constant $\bar{m} \in (0, n]$ (that can be chosen by the researcher). The next proposition shows that designs with sample-equalizing allocations are optimal.

Proposition 5.2 (Optimality of separate t-tests)**.** *Let Assumptions 4.1, 4.2, 5.1, and 5.2 hold. Then, the testing protocol $r^{\mathrm{N}}(X)$, where*

$$r_j^{\mathrm{N}}(X; J, \Sigma) = \begin{cases} 1\left\{\frac{X_j}{\sqrt{\Sigma_{j,j}}} \geq t^*(J, n(J))\right\}, \quad \forall j \in J & \textit{if } \Sigma \textit{ has a sample-equalizing allocation} \\ 0 & \textit{otherwise} \end{cases} \tag{24}$$

is maximin and unbiased if $t^*(J, n(J)) = \Phi^{-1}\left(1 - \frac{C(J,n(J))}{b\bar{\omega}(J)}\right)$.

Proof. See Appendix D.8. □

Proposition 5.2 shows that separate one-sided t-tests with critical values $t = \Phi^{-1}\left(1 - \frac{C(J,n(J))}{b\bar{\omega}(J)}\right)$ based on sample-equalizing allocations are maximin and unbiased. The planner "forces" the researcher to choose designs with sample-equalizing allocations by not rewarding any results from experiments with other designs. Intuitively, since the researcher expects σ_j to be the same for all j, the optimal protocol equalizes the expected standard errors by requiring a sample-equalizing allocation. Proposition 5.2 provides guidance both on which testing protocol to implement and which design to incentivize.

Remark 5.1 (Researcher knows and can choose Σ)**.** If the researcher knows the outcomes' variances and can choose Σ strategically (i.e., by choosing sample sizes as a function of the outcome variances), separate one-sided t-tests with critical values $t = \Phi^{-1}\left(1 - \frac{C(J,\Sigma)}{b\bar{\omega}(J)}\right)$ based on variance-equalizing allocations are uniformly globally optimal (by Proposition 4.1). A variance-equalizing allocation is an allocation such that $\sigma_{0j}^2/n_{0j} + \sigma_{1j}^2/n_{1j} = \sigma^2$ for some constant $\sigma^2 > 0$. This result follows because if the researcher knows and can choose Σ, then forcing her to impose common variances preserves maximin and uniformly global optimality by arguments in the proof of Proposition 4.1. □

6 Empirical illustration and broader applicability

This section discusses the scope for applying and implementing the framework's implications. Section 6.1 considers our running example, the regulatory approval process, while Section 6.2 comments on potential applications to program evaluation in economics.

6.1 Empirical illustration

Proposition 4.1 showed that the policymaker's preferred MHT adjustments hinge on how research costs $C(J, \Sigma)$ vary with the set of chosen treatments J and the experimental design Σ. In particular, denote the level of the separate t-tests in Proposition 4.1 as

$$\alpha(J, \Sigma) := \frac{C(J, \Sigma)}{b\bar{\omega}(J)}.$$

The planner will therefore want to obtain information about $C(J, \Sigma)$, b, and $\bar{\omega}(J)$ to compute this level. In the FDA approval context with multiple subgroups, we can think of b as the expected profit per customer, and $\bar{\omega}(J)$ as the total number of customers that would buy the drug if it were approved for every subgroup. To build intuition, it will be helpful to impose the simplifying assumption $\bar{\omega}(J) = |J|$, which holds for example if each subgroup j receives equal weight $\omega_j = 1$.

Adjustment factor in general form. Let $\bar{C}$ denote the cost of a benchmark experiment with a single treatment.[19] Without loss of generality we can write

$$\alpha(J, \Sigma) = \bar{\alpha} \times \frac{C(J, \Sigma)}{|J|\bar{C}}, \tag{25}$$

where $\bar{\alpha} = \bar{C}/b$ denotes the size of the hypothesis test in the benchmark experiment. This formulation shows that the appropriate size for tests in a study with $|J|$ hypotheses can be calculated as the product of two quantities. The first is the size of the optimal test in the benchmark, single-hypothesis case. The second is the MHT *correction factor* $[C(J, \Sigma)/\bar{C} \times 1/|J|]$, which captures how the cost per test varies as the number of hypotheses tested grows (keeping in mind that this may affect the design Σ as well as J). Unless all costs are fixed ($C(J, \Sigma) = \bar{C}$) this correction factor will differ from the standard Bonferroni correction factor $1/|J|$. Notice also that if costs are strictly proportional to the number of hypotheses ($C(J, \Sigma) = \bar{C} \times |J|$) then standard inference without adjustment for MHT is optimal.

Choice of $\bar{\alpha}$. There are competing benchmarks one might consider for $\bar{\alpha}$, the test size for a benchmark study with a single treatment. FDA guidelines currently recommend a size of 2.5% for one-sided single hypothesis tests (Food and Drug Administration, 2022), but Tetenov (2016), using data on the costs and expected profits from Phase III trials, proposes a value of 15%. Given this, and the dispersion in the cost of trials for different drugs (Grabowski et al., 2002), we provide results for a range of values between 2.5% and 15%.

Modeling costs. In principle the regulator could evaluate the MHT adjustment term in (25) separately for different categories (e.g., therapeutic classes) or even using data on each study individually. They might require pharmaceutical companies to declare the fixed and variable costs of a study (information about which is often contained in contracts with the hospital or contract research organization organizing a trial) when pre-registering it. Here we wish to illustrate the potential consequences of doing so using existing, published estimates

[19] This benchmark experiment could, for instance, be an experiment with the minimum sample size for a Phase III trial according to FDA guidance (see U.S. Food and Drug Administration (nda)).

of moments of the cost structure of clinical trials. This requires that we model the cost function. We consider a simple formulation with both fixed and variable costs:

$$C(J, \Sigma) = c_f + c_v \sum_{j \in J} n_j, \tag{26}$$

where c_f is a fixed cost invariant to $|J|$ and c_v is a variable cost. This is a special case of the specification in Section 4.2, where we assume here for simplicity that variable costs vary in proportion to the number of subjects n_j. Additional costs that vary with $|J|$, independent of n_j, could be accommodated with the appropriate data, and would imply adjustments less conservative than those reported below.

It will be convenient to rewrite this expression (without loss of generality) as

$$C(J, \Sigma) = c_f + c_v |J| \bar{m} \frac{\bar{n}}{\bar{m}}, \quad \bar{n} = \frac{1}{|J|} \sum_{j \in J} n_j,$$

where $\bar{n}$ is the average sample size across subgroups in the trial in question and $\bar{m}$ is the average overall sample size of the single arm benchmark experiment (with size $\bar{\alpha}$). With an abuse of notation, we can then write the optimal level as a function $\alpha\left(\bar{\alpha}, |J|, \frac{\bar{n}}{\bar{m}}\right)$ of the size for the benchmark single-hypothesis experiment $\bar{\alpha}$, the number of treatment arms $|J|$, and the ratio of the average subgroup sample size to the benchmark experiment size $\bar{n}/\bar{m}$.

Cost calibration. Sertkaya et al. (2016), using data on the costs of 31,000 pharmaceutical clinical trials conducted in the United States between 2004 and 2012, estimate that the average fixed costs of a Phase 3 trial were 46% of the average total cost, with the rest varying either directly with the number of subjects enrolled or with the number of sites at which they were enrolled.[20] As an approximation, we set $\bar{m}\bar{J}$ equal to the average historical overall sample size across clinical trials. It follows that $c_f/(c_f + c_v \bar{m}\bar{J}) = 0.46$. Using $\bar{J} = 3$ based on the tabulations in Pocock et al. (2002) yields an MHT correction factor of $(\frac{\bar{n}}{\bar{m}} + 2.56/|J|)/3.56$.[21]

[20] According to Sertkaya et al. (2016, Table 2), average variable costs (i.e., the per-patient and per-site costs) were USD 10,826,880, and average total costs were USD 19,890,000, so that the fraction of fixed costs is $(19,890,000 - 10,826,880)/19,890,000 \approx 0.46$.

[21] We use the median estimate multiplied by the probability of reporting more than one subgroup. The critical values are not particularly sensitive to $\bar{J}$; if for example we fix $\bar{\alpha}(1) = 0.025$ and double $\bar{J}$ from 3 to 6 this decreases $\alpha(2)$ from 0.016 to 0.015, $\alpha(3)$ from 0.013 to 0.011, and $\alpha(\infty)$ from 0.007 to 0.004.

Inserting this into (25), we thus arrive at

$$\alpha\left(\bar{\alpha}, |J|, \frac{\bar{n}}{\bar{m}}\right) = \bar{\alpha} \times \left[\underbrace{\frac{1+2.56/|J|}{3.56}}_{\text{Multiplicity adjustment}} + \underbrace{\frac{1}{3.56} \times \left(\frac{\bar{n}}{\bar{m}} - 1\right)}_{\text{Sample size per arm}} \right]. \tag{27}$$

The correction factor here has two terms. The first is a "pure" correction for multiple hypothesis testing, accounting for the influence of the number of treatment arms $|J|$. The second corrects for the effects of sample size on study cost: studies with sample sizes larger than $\bar{m}$ ($\bar{n} > \bar{m}$) are more expensive to run and thus require less strict testing thresholds.

Table 1 illustrates the implications quantitatively. It tabulates the test level implied by (27) for a range of values of $|J|$ (rows), $\bar{\alpha}$ (columns), and $\bar{n}/\bar{m}$ (column groups). For example, for studies with a benchmark sample size ($\bar{n} = \bar{m}$) and assuming $\bar{\alpha} = 0.15$ as in Tetenov (2016), those with $|J| = 2$ would use size 0.096, those with $|J| = 3$ would use 0.078, and so on, asymptoting to 0.042 at $|J| = \infty$. If instead we set $\bar{\alpha} = 0.025$, consistent with FDA guidance, then studies with $|J| = 2$ would use 0.016, those with $|J| = 3$ would use 0.013, and so on, asymptoting to 0.007 at $|J| = \infty$. These thresholds are more conservative than unadjusted ones, but less conservative than those implied by Bonferroni corrections ($\bar{\alpha}/|J|$).

For further comparison, the last two columns of Table 1 report the adjustment corresponding to FWER control based on the Sidak correction (Šidák, 1968). The Sidak correction, which sets the level of each test to $1 - (1 - \bar{\alpha})^{1/|J|}$, is a useful benchmark because it is exact with independent tests (as for the case of subgroup analyses). It implies more conservative inferences than our tabulated values. For instance, with $\bar{\alpha} = 0.025$ and $|J| = 9$, the optimal level is 0.009 while that under the Sidak correction is 0.003.

The table also illustrates how the adjustment factor depends on the (relative) average per-treatment sample size $\bar{n}/\bar{m}$. The fifth, sixth and seventh columns vary this while holding $\bar{\alpha}$ fixed at 0.025. When $\bar{n}$ is smaller than $\bar{m}$ we require *more* stringent size control, while for larger $\bar{n}$ we require less stringent size control. For example, with $|J| = 2$, studies with half the benchmark sample size would use a 0.012 threshold, while studies with double the benchmark sample size would use 0.023.

The results in Table 1 should be read as illustrative of what might happen if the FDA were to require cost disclosure and base testing thresholds on the disclosed trial-specific costs,

Table 1: Critical values as functions of hypothesis count and sample size

	$\bar{n}/\bar{m} =$	100%				50%	150%	200%	$\alpha^{0.025}_{\text{Šidák}}$	$\alpha^{0.05}_{\text{Šidák}}$
$\lvert J \rvert$	$\bar{\alpha} =$	0.025	0.05	0.10	0.15	0.025			0.025	0.050
1		0.025	0.050	0.100	0.150	0.021	0.029	0.032	0.025	0.050
2		0.016	0.032	0.064	0.096	0.012	0.019	0.023	0.013	0.025
3		0.013	0.026	0.052	0.078	0.009	0.017	0.020	0.008	0.017
4		0.012	0.023	0.046	0.069	0.008	0.016	0.019	0.006	0.013
5		0.011	0.021	0.042	0.064	0.007	0.015	0.018	0.005	0.010
6		0.010	0.020	0.040	0.060	0.006	0.014	0.017	0.004	0.009
7		0.010	0.019	0.038	0.058	0.006	0.014	0.017	0.004	0.007
8		0.009	0.019	0.037	0.056	0.006	0.013	0.016	0.003	0.006
9		0.009	0.018	0.036	0.054	0.006	0.013	0.016	0.003	0.006
∞		0.007	0.014	0.028	0.042	0.004	0.013	0.014	0.000	0.000

Notes: This table tabulates optimal critical values obtained from (27) for different numbers of hypotheses ($|J|$) and (relative) sample sizes ($\bar{n}/\bar{m}$), all given an assumed critical value for the benchmark case of a single-hypothesis experiment ($\bar{\alpha}$). $\alpha_{\text{Šidák}}$ is the optimal level implied by the Sidak correction (Šidák, 1968), $1-(1-\bar{\alpha})^{1/|J|}$. The Sidak correction is exact for controlling the FWER with independent tests.

which could be obtained, for example, from privately-owned contract data (Sertkaya et al., 2016). If instead it were to base testing thresholds directly on the number of treatments $|J|$ and samples sizes $\bar{n}$, using the values indicated in the table, this would create incentives for gaming—e.g., reporting results obtained from a single experiment (in the sense that c_f was incurred only once) as if they came from multiple distinct experiments (implying that c_f was incurred more than once). Detecting and deterring such gaming might be easier in some cases than in others. Tests of the same drug in different populations, for example, could be matched based on the chemical formula of the compound in question; tests of different compounds on the same population might be harder to match.

6.2 MHT within economics

Hypothesis testing norms are also a salient issue within economics, given the publication trends noted in Figure 1. Our framework's assumptions arguably correspond most closely to economics papers that report the results of experimental program evaluations, as in that case there are clear policy decisions that the research is explicitly designed to inform (i.e., whether or not to implement or scale the programs being evaluated). Indeed, researchers often conduct studies like these in collaboration with implementation partners, such as governments or NGOs, precisely in order to evaluate the impact of treatments the partners are considering. The paper's findings may thus affect social welfare because, in addition to potentially being published in an academic journal, they can influence those decisions. Pre-specification of the analysis to be conducted in a pre-analysis plan (corresponding to our assumption that researchers pre-specify their tests) is now common in this genre of work (Miguel, 2021). And it is also common for such "policy experiments" to test more than one treatment as part of the same study. Muralidharan et al. (2025) document at least 27 such experiments published in top-5 journals alone between 2007 and 2017.[22]

With these points in mind, we also conducted a second quantitative application to program evaluation experiments in development economics, using unique data on their costs, sample sizes, and treatment arm counts which we obtained from the universe of funding proposals submitted to the Abdul Latif Jameel Poverty Action Lab (J-PAL) from 2009 to 2021. In the interests of brevity, we describe the data and analysis in depth in Appendix A and briefly restate the main findings here. We estimate that research costs are significantly and substantially less than proportional to the number of treatments tested, with elasticities ranging from 0.13 to 0.22.[23] But they are also not invariant to scale: projects with more arms cost significantly more ($p < 0.05$). As a result the appropriate testing thresholds vary with the number of treatment arms. They are similar to but slightly less conservative than those that result from a Bonferroni correction and those implied by Sidak's correction (which is exact for controlling the FWER for independent tests). Finally, the testing thresholds also vary moderately with the sample size, with larger samples implying (ceteris paribus) less conservative procedures.

This analysis focuses on a specific type of multiplicity, namely multiplicity of treatments. Economists also often deal with multiple outcomes. These do not necessitate multiple tests; indeed, researchers often aggregate the outcomes into summary statistics instead. An earlier version of this paper (Viviano et al., 2025) studied this problem within our framework, showing that optimal rules r^* test for effects on an index formed using statistical weights (similar to Anderson, 2008) when the outcomes are noisy proxies for some common underlying measure, but using economic weights (as for example in Bhatt et al., 2024) when they capture

[22] Their list includes only studies with interaction arms, so provides a lower bound on the total number of multi-armed evaluations.

[23] We obtain these estimates from descriptive regressions; they need not be causal to characterize the cost function $C(J, \Sigma)$ in our model, provided that function is invariant to r.

distinct components of the planner's utility. The fact that multiple outcomes justify different techniques than do multiple treatments (or subgroups) is noteworthy in the context of historical narratives about MHT practices: the multiple-treatment case—genetic association testing in particular—has often been cited to motivate new MHT procedures (see Dudoit et al., 2003; Efron, 2008a, for reviews), while the multiple-outcomes case seems to have been bundled with it subsequently, and less intentionally.

One could also move away from the frequentist paradigm (which we have presumed) entirely, towards a Bayesian alternative. Proposals to control the FDR are interesting in this regard. Several papers have pointed out a Bayesian rationale for doing so: controlling the (positive) FDR can be interpreted as rejecting hypotheses with a sufficiently low posterior probability (e.g., Storey, 2003; Gu and Koenker, 2020; Kline et al., 2022). In fact, these arguments apply even in the case of a *single* hypothesis. The essential idea is to balance the costs of false positives and false negatives, rather than prioritize size control at any (power) cost. We thus interpret these arguments less as support for a particular solution to the MHT problem per se, and more as a reminder of the merits of Bayesian approaches generally.

Acknowledgments

We are grateful to the Editor and anonymous referees for constructive comments that helped improve the paper. We are also grateful to Nageeb Ali, Isaiah Andrews, Tim Armstrong, Oriana Bandiera, Sylvain Chassang, Kevin Chen, Tim Christensen, Graham Elliott, Stefan Faridani, Will Fithian, Paul Glewwe, Peter Hull, Guido Imbens, Lawrence Katz, Toru Kitagawa, Pat Kline, Michael Kremer, Michal Kolesar, Ivana Komunjer, Damian Kozbur, Lihua Lei, Adam McCloskey, Konrad Menzel, Francesca Molinari, Jose Montiel-Olea, Ulrich Mueller, Mikkel Plagborg-Møller, David Ritzwoller, Joe Romano, Adam Rosen, Jonathan Roth, Andres Santos, Azeem Shaikh, Jesse Shapiro, Joel Sobel, Sandip Sukhtankar, Yixiao Sun, Elie Tamer, Aleksey Tetenov, Winnie van Dijk, Tom Vogl, Quang Vuong, Michael Wolf, and seminar participants for valuable comments, and to staff at J-PAL, Sarah Kopper and Sabhya Gupta in particular, for their help accessing data. Aakash Bhalothia and Muhammad Karim provided excellent research assistance. Viviano is also affiliated with Y-RISE, and Wüthrich is also affiliated with CESifo. All errors are our own.

Funding

Niehaus and Wüthrich gratefully acknowledge funding from the UC San Diego Academic Senate. Viviano gratefully acknowledges support from the Griffin Fund at Harvard University and from NSF Grant SES 2447088.

Non-financial support was provided by J-PAL in the form of the data used in Appendix A, which we used under the provisions of a Data Use Agreement (DUA) between J-PAL and UC San Diego. Key provisions are that we are allowed to publish data sufficient to meet the requirements of scholarly journals, subject to various provisions, that J-PAL may curate these data using its established data publication processes to remove any Confidential Data, and that J-PAL may review drafts of the paper prior to publication to ensure that no Confidential Data are disclosed.

Conflict of Interest

Niehaus is co-chair of the Science for Progress Initiative at J-PAL, which provided the data used in Appendix A. The role is uncompensated. He is not an officer, director or board member of J-PAL.

Data availability statement

The data and code underlying this research are available on Zenodo at `https://doi.org/10.5281/zenodo.18825708`.

References

Anderson, M. L. (2008). Multiple inference and gender differences in the effects of early intervention: A reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects. *Journal of the American Statistical Association 103*(484), 1481–1495.

Andrews, I. and J. M. Shapiro (2021). A model of scientific communication. *Econometrica 89*(5), 2117–2142.

Athey, S. and S. Wager (2021). Policy learning with observational data. *Econometrica 89*(1), 133–161.

Banerjee, A., S. Chassang, and E. Snowberg (2017). Chapter 4 – Decision theoretic approaches to experiment design and external validity. In A. V. Banerjee and E. Duflo

(Eds.), *Handbook of Field Experiments*, Volume 1 of *Handbook of Economic Field Experiments*, pp. 141–174. North-Holland.

Banerjee, A. V., S. Chassang, S. Montero, and E. Snowberg (2020). A theory of experimenters: Robustness, randomization, and balance. *American Economic Review 110*(4), 1206–30.

Bates, S., M. I. Jordan, M. Sklar, and J. A. Soloff (2022). Principal-agent hypothesis testing. *arXiv:2205.06812*.

Bates, S., M. I. Jordan, M. Sklar, and J. A. Soloff (2023). Incentive-theoretic bayesian inference for collaborative science. *arXiv:2307.03748*.

Bhatt, M. P., S. B. Heller, M. Kapustin, M. Bertrand, and C. Blattman (2024). Predicting and preventing gun violence: An experimental evaluation of READI chicago. *The Quarterly Journal of Economics 139*(1), 1–56.

Chassang, S., G. Padro I Miquel, and E. Snowberg (2012). Selective trials: A principal-agent approach to randomized controlled experiments. *American Economic Review 102*(4), 1279–1309.

Code of Federal Regulations (2024). What constitutes clinical trial registration information? https://www.ecfr.gov/current/title-42/part-11/section-11.28. 42 CFR §11.28.

Dudoit, S., J. P. Shaffer, and J. C. Boldrick (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science 18*(1), 71–103.

Efron, B. (2008a). Microarrays, empirical bayes and the two-groups model. *Statistical Science 23*(1), 1–22.

Efron, B. (2008b). Simultaneous inference: when should hypothesis testing problems be combined? *Annals of Applied Statistics 2*(1), 197–223.

Fisher, L. D. (1991). The use of one-sided tests in drug trials: an fda advisory committee member's perspective. *Journal of Biopharmaceutical Statistics 1*(1), 151–156.

Food and Drug Administration (2022, October). Multiple endpoints in clinical trials guidance for industry. Document ID: FDA-2016-D-4460-0024, URL: https://www.regulations.gov/document/FDA-2016-D-4460-0024.

Frankel, A. and M. Kasy (2022). Which findings should be published? *American Economic Journal: Microeconomics 14*(1), 1–38.

Gilboa, I. and D. Schmeidler (1989). Maxmin expected utility with non-unique prior. *Journal of mathematical economics 18*(2), 141–153.

Grabowski, H., J. Vernon, and J. A. DiMasi (2002). Returns on research and development for 1990s new drug introductions. *Pharmacoeconomics 20*, 11–29.

Gu, J. and R. Koenker (2020). Invidious comparisons: Ranking and selection as compound decisions. arXiv:2012.12550.

Henry, E. and M. Ottaviani (2019). Research and the approval process: the organization of persuasion. *American Economic Review 109*(3), 911–55.

Hirano, K. and J. R. Porter (2009). Asymptotics for statistical treatment rules. *Econometrica 77*(5), 1683–1701.

Hirano, K. and J. R. Porter (2020). Chapter 4 - Asymptotic analysis of statistical decision rules in econometrics. In S. N. Durlauf, L. P. Hansen, J. J. Heckman, and R. L. Matzkin (Eds.), *Handbook of Econometrics, Volume 7A*, Volume 7 of *Handbook of Econometrics*, pp. 283–354. Elsevier.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics 6*, 65–70.

Kasy, M. and J. Spiess (2023). Optimal pre-analysis plans: Statistical decisions subject to implementability.

Kitagawa, T. and A. Tetenov (2018). Who should be treated? Empirical welfare maximization methods for treatment choice. *Econometrica 86*(2), 591–616.

Kline, P., E. K. Rose, and C. R. Walters (2022). Systemic discrimination among large us employers. *The Quarterly Journal of Economics 137*(4), 1963–2036.

Kline, P. M., E. K. Rose, and C. R. Walters (2024, April). A discrimination report card. Working Paper 32313, National Bureau of Economic Research.

Lehmann, E. L. and J. P. Romano (2005). *Testing statistical hypotheses.* Springer Science & Business Media.

Lehmann, E. L., J. P. Romano, and J. P. Shaffer (2005). On optimality of stepdown and stepup multiple test procedures. *The Annals of Statistics 33*(3), 1084 – 1108.

Lewis, J. A. (1999). Statistical principles for clinical trials (ich e9): an introductory note on an international guideline. *Statistics in medicine 18*(15), 1903–1942.

List, J. A., A. M. Shaikh, and Y. Xu (2019). Multiple hypothesis testing in experimental economics. *Experimental Economics 22*(4), 773–793.

Manski, C. (2004). Statistical treatment rules for heterogeneous populations. *Econometrica 72*(4), 1221–1246.

McCloskey, A. and P. Michaillat (2022). Incentive-compatible critical values. Working Paper 29702, National Bureau of Economic Research.

Miguel, E. (2021). Evidence on research transparency in economics. *Journal of Economic Perspectives 35*(3), 193–214.

Muralidharan, K., M. Romero, and K. Wüthrich (2025, 05). Factorial designs, model selection, and (incorrect) inference in randomized experiments. *The Review of Economics and Statistics 107*(3), 589–604.

NIH National Library of Medicine (n.d.). FDAAA 801 and the Final Rule. Retrieved December 17, 2024, from https://clinicaltrials.gov/policy/fdaaa-801-final-rule#trials-registered.

Pocock, S. J., S. E. Assmann, L. E. Enos, and L. E. Kasten (2002). Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practiceand problems. *Statistics in medicine 21*(19), 2917–2930.

Robbins, H. (1951). Asymptotically subminimax solutions of compound statistical decision problems. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*. The Regents of the University of California.

Romano, J. P., A. Shaikh, and M. Wolf (2011). Consonance and the closure method in multiple testing. *The International Journal of Biostatistics 7*(1).

Romano, J. P., A. M. Shaikh, and M. Wolf (2010). Hypothesis testing in econometrics. *Annual Review of Economics 2*(1), 75–104.

Sertkaya, A., H.-H. Wong, A. Jessup, and T. Beleche (2016). Key cost drivers of pharmaceutical clinical trials in the United States. *Clinical Trials 13(2)*, 117–126.

Šidák, Z. (1968). On multivariate normal probabilities of rectangles: their dependence on correlations. *The Annals of Mathematical Statistics*, 1425–1434.

Spiess, J. (2018). Optimal estimation when researcher and social preferences are misaligned. Working Paper.

Spjotvoll, E. (1972). On the optimality of some multiple comparison procedures. *The Annals of Mathematical Statistics 43*(2), 398–411.

Storey, J. D. (2003). The positive false discovery rate: a Bayesian interpretation and the q-value. *The Annals of Statistics 31*(6), 2013–2035.

Tetenov, A. (2012). Statistical treatment choice based on asymmetric minimax regret criteria. *Journal of Econometrics 166*(1), 157–165.

Tetenov, A. (2016). An economic theory of statistical testing. Working Paper.

U.S. Food and Drug Administration (n.d.a). Step 3: Clinical research. Retrieved January 30, 2024, from https://www.fda.gov/patients/drug-development-process/step-3-clinical-research.

U.S. Food and Drug Administration (n.d.b). What we do. Retrieved December 2, 2024, from https://www.fda.gov/about-fda/what-we-do#responsibilities.

Viviano, D., K. Wuthrich, and P. Niehaus (2025). A model of multiple hypothesis testing. arXiv preprint 2104.13367v8.

Wald, A. (1950). *Statistical decision functions*. Wiley.

Williams, C. (2021). Preregistration and incentives. SSRN 3796813.

Yoder, N. (2022). Designing incentives for heterogeneous researchers. *Journal of Political Economy 130*(8), 2018–2054.