Comment on "Multiple Endpoints in Clinical Trials: Guidance for Industry"[*]

Paul Niehaus UC San Diego
Davide Viviano Stanford GSB
Kaspar Wüthrich UC San Diego

December 6, 2022

The guidance document "Multiple Endpoints in Clinical Trials: Guidance for Industry" (FDA (2022), docket ID FDA-2016-D-4460—henceforth, the Guidance) proposes guidelines for researchers to follow when reporting to the FDA the results of clinical trials that involve testing more than one hypothesis. This is a welcome effort to provide much-needed clarity in an area of statistical practice that has become quite confused (and confusing!), and we welcome the opportunity to contribute comments on the proposal.

One particular strength of the note is the way it so often connects statistical practice back to the decision rules the FDA must follow in order to consistently implement Federal legislation. For example, it often discusses the mapping from hypothesis rejections to regulatory decisions in explaining the rationale for particular methods. This grounding of statistics in decision-making is very helpful, as it lets us examine in a systematic way what testing procedures (i.e., mappings from data to hypothesis rejections) are likely to lead to good regulatory decisions—understood here to be decisions that control the probability of approving drugs that do *not*, in fact, have the requisite benefits, while otherwise maximizing the chance of approving those that do. This is precisely the approach we have taken in our own recent work on multiple hypothesis testing (Viviano et al., 2022).

In keeping with this idea, our comments below focus on identifying statistical procedures that control the probability of mistaken approval decisions at a desirable level while also minimizing the probability of mistaken rejections. We discuss two broad areas in which we believe additional clarity and/or consistency would be beneficial, or where it may be in the public interest to consider alternative procedures without sacrificing control over the rate of mistaken approvals.

[*]Email: `pniehaus@ucsd.edu`, `dviviano@stanford.edu`, `kwuthrich@ucsd.edu`

1 Multiple endpoints

A common scenario in which multiple testing considerations arise is that in which there are multiple endpoints, and we desire to make a regulatory decision based on whether there are or are not effects on one or more of these. The Guidance draws an important distinction between *primary* and *secondary* endpoints, and then draws a further helpful distinction between two scenarios with respect to the primary endpoints:

(1) "When Demonstration of Treatment Effects on All of Two or More Distinct Endpoints Is Necessary to Establish Clinical Benefit (Co-Primary Endpoints)"

(2) "When Demonstration of a Treatment Effect on at Least One of Several Primary Endpoints Is Sufficient"

We comment on each scenario in turn.[1]

1.1 When effects on all primary endpoints must be established

In scenario (1), the Guidance recommends testing individual endpoints at level α and not making any multiplicity adjustments. The reasoning for this recommendation is that the drug "will not be considered effective without demonstration of a treatment effect on all of these disease features" so that "there is no multiplicity problem when the study is designed to demonstrate efficacy on all of the separate endpoints." (p. 9) The Guidance then discusses whether to increase α to compensate for the loss of power incurred by testing more than one hypothesis and concludes:

> Increasing α for each co-primary endpoint is not acceptable because doing so may undermine the ability to interpret a treatment effect on each disease aspect considered critical to show that the drug is effective in support of approval. (p. 9)

This is certainly true as stated: if the goal is to obtain assurance that there are effects on *each* endpoint, then the chance of mistakenly rejecting each null hypothesis should be controlled at level α. But if an effect must be found on *all* co-primary endpoints in order to warrant approval, then this can imply control of the probability of mistakenly approving the drug at levels well *below* α.

[1] This distinction is related to the distinction between conjunction testing or intersection-union testing (reject the joint null if all tests are significant) and disjunction testing or union-intersection testing (reject the joint null hypothesis if at least one test is significant) in the literature on multiple testing (see, e.g., Rubin, 2021, for a discussion).

To illustrate, suppose first the case of two completely independent one-sided tests. Then the probability of rejecting both at the 0.025 level, and thus mistakenly declaring success, is $0.025 \times 0.025 = 0.000625$. If instead one uses $\alpha = \sqrt{0.025} \approx 0.158$ for both tests, the resulting Type I error is 0.025% as desired, but the power is much higher. Of course, this particular α-adjustment would not be appropriate across all scenarios, as the size of the adjustment that controls the overall probability of a mistaken approval will depend on the degree of dependence between the tests. But context-specific exact α-adjustments are fairly straightforward to implement if desired.[2]

1.2 When effects on at least one primary endpoint must be established

In scenario (2), the Guidance explains there is a multiplicity problem and recommends using testing procedures that control the family-wise Type I error rate (FWER). The Appendix then describes several statistical methods for doing so, with a thoughtful discussion of their advantages and disadvantages relative to each other. These all have in common the implied assumption of *separate testing*. In other words, they presume that the researcher will first test each null hypothesis separately (using procedures that ensure that the FWER is controlled at level α), and then reject the aggregate hypothesis of no effect on *any* endpoint if any one of these separate tests rejects. In this sense, separate tests are *indirect*.

It is currently not clear whether the recommendation of this indirect, two-step procedure is intended to preclude *direct tests* of the joint null hypothesis that all effects are zero. The latter can often be done using a simple F-test, for example. Such a test necessarily controls the probability of mistakenly rejecting the null of no effects. And direct tests can be more powerful in certain circumstances—though not all—than indirect ones. In other words, allowing for the use of direct tests has the potential to reduce the probability of mistaken rejections while still controlling the probability of mistaken approvals at the same level as the methods discussed in the Appendix.[3]

Set against this potential benefit, one complication of using a direct approach is that

[2] This is the case if, for example, one is willing to use estimators that can be cast in a regression setting. In this case, the estimated covariance matrix between the estimators provides the necessary information.

[3] The specific choice of method for testing the joint null should be guided by prior knowledge of the (positive) effects researchers expect to observe (the "alternative hypothesis" in statistical jargon). For instance, if the effects are expected to be large on single endpoints but not all, researchers may want to use the largest t-score among all tests as the test statistic (and adjust α appropriately). On the other hand, small positive effects on *all* endpoints would justify F-tests. This follows by standard properties of statistical tests (Lehmann and Romano, 2005).

rejecting the null need not imply a finding of *beneficial*, as opposed to harmful, effects. In addition to rejecting the null of an F-test, for example, one would also want to check that the direction of the effects was as desired.[4] It would be useful to understand the FDA's perspective on this issue—whether there are approaches built around direct tests that are suitable, or whether (and if so, why) the recommendation is to eschew these entirely.

2 Different types of multiplicity

In addition to the distinctions discussed above involving multiple endpoints, the Guidance also references additional forms of multiplicity. It would be useful to sharpen the distinctions between these and clarify what procedures are appropriate for each.

The language on this point in the current Guidance is somewhat varied. The title states that it is about multiple *endpoints*, and the initial discussion (e.g., Section II.A) maintains this focus. But later, the Guidance refers to "multiple endpoints *and analyses* [emphasis added]" (p. 6), and says that descriptions with respect to additional attributes including, for example, "multiple subject subgroups based on demographic or other characteristics" (p. 5) similarly must meet the requirements that "appropriate adjustments for multiple endpoints and analyses can be selected, prespecified, and applied" (p. 6), while noting that what exactly is appropriate may be context-specific.

Distinguishing between these types of multiplicity would help improve policy-making in the public interest. Consider subgroup analysis: for example, deciding whether or not to approve a drug for use by women when estimated impacts on both women and men are observed. It is not *necessary* to demonstrate effects for both women and men in order to justify approving the drug for use by women. And it is not *sufficient* to demonstrate effects on at least one of women or men to do so: if effects are demonstrated only for men, this is not sufficient to justify use by women. Neither of the core scenarios contemplated by the guidance note (i.e., those in III.C.1. or III.C.2) are thus applicable to this class of decision problem.

As this example suggests, the first step to developing appropriate protocols is to precisely articulate the relationship between hypotheses test(s) and regulatory decision(s). The

[4] A related issue is that direct approaches do not tell researchers *which* hypotheses were or were not rejected, but when the regulatory decision depends on whether or not an effect has been detected on at least one endpoint, this does not matter for that decision. It may, of course, be of independent interest for informing future work, in which case it might make sense to first specify a direct test of the joint null on which the approval decision would be based, and then separate tests of individual component hypotheses which would be rated as "exploratory."

Guidance already does a laudable job of this in some places, e.g., in distinguishing between different ways that multiple endpoints might influence a single decision (III.C.1. v.s. III.C.2). It would be helpful to extend this reasoning to also examine the cases of multiple sub-populations and interventions (e.g., dosages), which will typically imply multiple decisions.[5] As noted above, this is the spirit of the exercise we conduct in Viviano et al. (2022), where we provide a theoretical framework in which the ultimate decision-maker (analogous here to the public-interest regulator) optimally chooses to require one set of procedures for dealing with the case of multiple interventions or subpopulations, and another set for the case of multiple endpoints. Rubin (2021) draws related distinctions. The point here is not the detailed conclusions of either analysis, but rather the broad point that both analyses support the same common-sense intuition: it is typically in the public interest to handle different types of multiplicity differently.

References

FDA, "Multiple Endpoints in Clinical Trials: Guidance for Industry," Technical Report, U.S. Department of Health and Human Services, Food and Drug Administration October 2022.

Lehmann, Erich L. and Joseph P. Romano, *Testing statistical hypotheses*, Springer, 2005.

Rubin, Mark, "When to adjust alpha during multiple testing: A consideration of disjunction, conjunction, and individual testing," *Synthese*, 2021, pp. 1–32.

Viviano, Davide, Kaspar Wüthrich, and Paul Niehaus, "(When) should you adjust inferences for multiple hypothesis testing?," *arXiv preprint arXiv:2104.13367*, 2022.

[5] Another way to see the need to handle multiple decisions differently is via reductio ad absurdum. Consider *all* the statistical tests that inform *all* the decisions made by the FDA. The probability of at least one Type I error across all these decisions is approximately 100%. But this does not mean that the FDA should control the FWER across all studies; if it did, it would almost never approve a new treatment. Clearly, some framework is needed that explains when and how different decisions are or are not interrelated in such a way that multiple testing adjustment is in the public interest.