The Prevalence and Impact of Multiple Hypothesis Testing Corrections in Clinical Trials

Jack Fitzgerald

September 24, 2026

Intro

Medical regulators in major healthcare markets deem multiple hypothesis testing (MHT) corrections necessary when multiple outcomes are involved

- E.g., European Medicines Agency, UK Medicines and Healthcare Products Regulatory Agency
- In particular, the U.S. Food and Drug Administration deems family-wise error rate (FWER) corrections for all trials with multiple primary outcomes (FDA 2022, draft guidance in 2017)
- **Idea:** Either a trial needs stat. sig. results for all primary outcomes to be successful, or each primary outcome is a 'separate chance' at stat. sig. results

However, systemic evidence that academic clinical trials don't comply with these guidelines

- **Synthesis of 13 published systematic reviews on 3122 published clinical trials:** 1317 engage in MHT, only 16% apply corrections (Menyhárt & Győrffy 2025, *IJE*)

Two questions:

1. Do we observe the same non-compliance across regulated, non-academic trials?
2. If so, how much does this affect clinical trial conclusions?

This Project

I leverage aggregated data on 9653 clinical trials registered on ClinicalTrials.gov with multiple primary outcomes (MPOs) to assess the prevalence and impact of MHT corrections

- I use a systematic large language model (LLM) classification approach that determines whether *p*-values are adjusted based on *p*-value descriptions and statistical analysis plans (SAPs)

In 2025, over 56% of MPO trials, and over 35% of Phase 3+ MPO trials, didn't apply *any* MHT corrections

- These non-compliance rates were even higher in previous years

Applying FWER corrections to *p*-values of MPO trials that didn't adjust for MHT erases *all* statistically significant results in 11-15% of MPO trials, and in 9-12% of Phase 3+ MPO trials

- I highlight improvements in clinical trial reporting infrastructure to address these problems

Data

I use the Aggregate Analysis of ClinicalTrials.gov (AACT) database (CTTI 2026)

- Stores trial information, (primary) outcomes, SAP URLs, and submitted results (including *p*-values and their descriptions)

Two places where one can determine whether analytical strategies are adjusted for MHT are (1) *p*-value descriptions and (2) SAPs

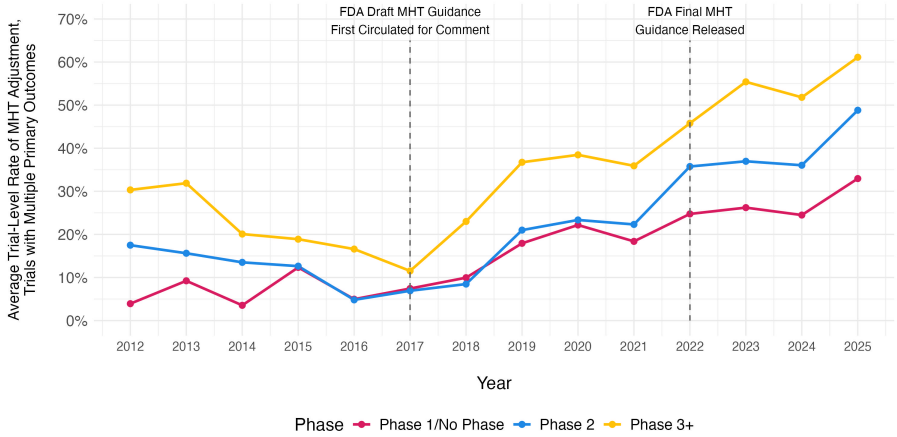
- **Challenge:** There are 9653 MPO trials in the ClinicalTrials.gov database with valid *p*-values, SAPs are typically dozens of pages long, and SAPs necessarily interact with *p*-value descriptions

LLM Classification

I use a systematic LLM classification method, passing *p*-value description text and SAP PDFs to OpenAI 6.2 Luna. To paraphrase prompts:

- ***p*-value description + SAP (27% of trials):** 'Given the SAP and this *p*-value's description, does an explicit indication make it more likely than not that this *p*-value is MHT/FWER-adjusted?'
- ***p*-value description + no SAP (30% of trials):** 'Does an explicit indication in this *p*-value description make it more likely than not that this p-value MHT/FWER-adjusted?'
- **No *p*-value description + SAP (22% of trials):** 'Does an explicit indication in the SAP make it more likely than not that a *p*-value with no description is MHT/FWER-adjusted?'
- **No *p*-value description + no SAP (21% of trials):** Automatically coded as not MHT-adjusting

Prevalence of MHT Adjustment



73% (64%) of all (Phase 3+) MPO trials do not adjust any results for MHT

What Is The Impact of FWER Adjustment?

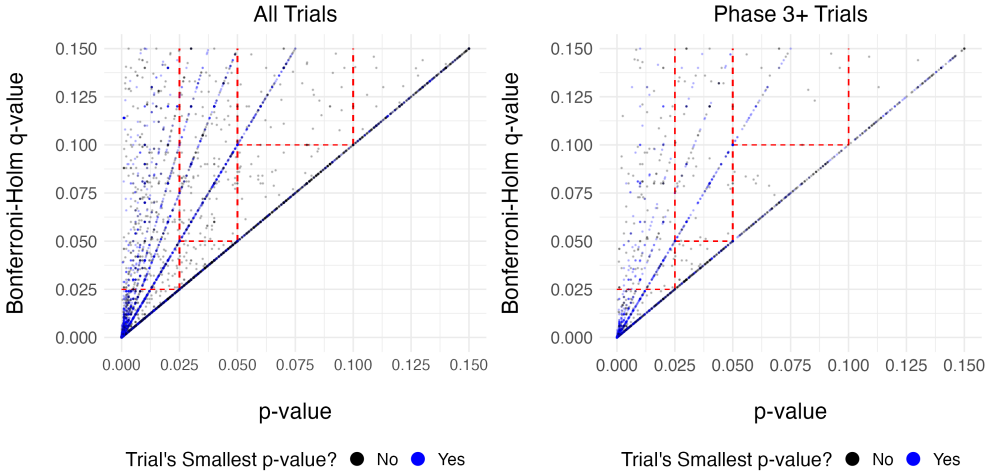
The AACT data has reported p-values for each trial arm's effect on each outcome

- For trials that don't MHT-adjust, I use Bonferroni-Holm adjustment (Holm 1979, *SJS*) to convert the family of raw, primary-outcome p-values to FWER-adjusted q-values

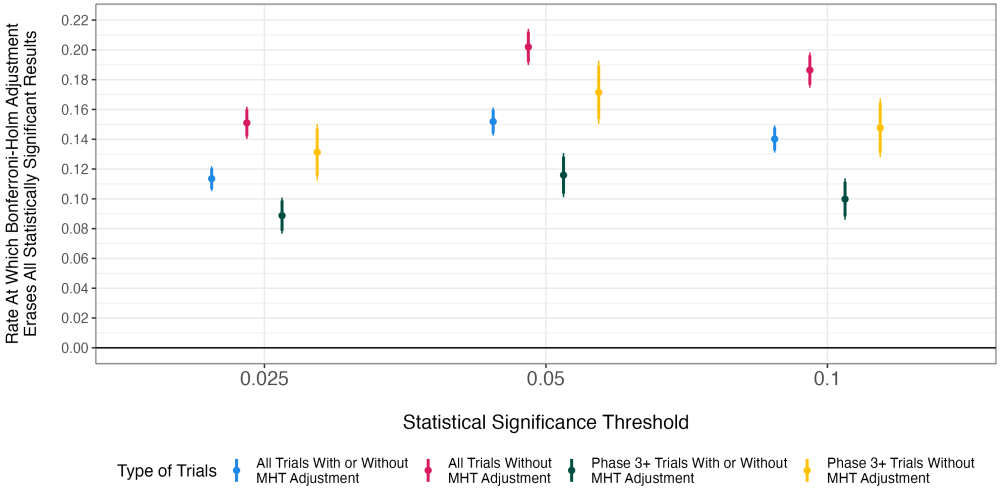
Reliable computation of Bonferroni-Holm q-values requires that I'm looking at the study's original p-values, so...

- I exclude 4095/9653 trials that threshold-report some p-values
- I do not adjust p-values for any trial that does *some* MHT adjustment for any p-value, FWER or not (minor point; 92% of trials that MHT adjust are FWER adjusting)

Visualizing Bonferroni-Holm Correction



Main Results



Implications

In Bonferroni-Holm correction, the q-value of the largest p-value is just the original p-value

- So by definition, a trial where all statistically significant results are erased by Bonferroni-Holm correction has at least one p-value above the significance threshold
- These trials are *relying on their smallest p-values* to have any statistically significant results at all
- ... but those results aren't precise enough to rule out that they arise simply due to MHT

Conclusion

This research has extremely actionable implications for medical regulators

- The FDA (2022) already has official guidelines deeming FWER adjustment necessary for all trials with multiple primary outcomes
- ... but the FDA (2022) guidelines are currently non-binding

When considering Phase 3+ trials for treatment approval, the FDA should either:

1. Require that researchers implement FWER adjustments before the trials are eligible for treatment approvals (i.e., make the guidelines binding)
2. Require that researchers report raw p-values and automatically do FWER adjustment on researchers' behalves

The FDA has considerable discretion to deem trial evidence insufficient to warrant approval (Janiaud et al. 2021, *AIM*)

- They should use this discretion to make sure that we have **precise evidence that treatments work before they're approved for public use**

Thank You For Your Attention!

These Slides

I am on the job market this year!

Website: https://jack-fitzgerald.github.io
Email: jackfitzgeraldresearch@gmail.com