# Review of: "When to Adjust Alpha During Multiple Testing"

Jelle Goeman[1]

1 Leiden University Medical Center (LUMC)

Potential competing interests: The author(s) declared that no potential competing interests exist.

A persistent misunderstanding of multiple testing methods haunts the epidemiology literature. I believe it originates from the most common explanation of the need for multiple testing methods in practitioners' handbooks, which often reads something like this: "suppose that none of the colors of jelly bean is truly causal for acne, then the probability that one of the colors is mistakenly found to be associated is 64%". A reader could mistake this sentence for the definition of the studywise error rate, leading the reader to believe that multiple testing methods are exclusively concerned with the joint studywise null hypothesis, the hypotheses that none of the colors of jelly bean causes acne. Rubin seems to be one of these, and I know from his papers that Greenland, who also reviewed this paper,[1] is victim to the same confusion. However, the sentence from the handbook was always meant as an example of a studywise error, not as a definition.

The studywise (a.k.a. strong familywise) error rate is in fact the probability that the study leads to an erroneous conclusion, not just under the joint studywise null hypothesis but under any combination of true and false hypotheses.[2][3][4][5][6] Let me explain this using the jelly bean example. Suppose that, unknown to the researcher, purple and teal jelly beans truly cause acne while all the other 18 colors do not. If the researcher performs hypothesis tests for all 20 colors, some colors may result in a significant result, while others do not. For example, the tests may reject the null hypotheses for teal and mauve. The resulting paper, claiming to have found an association between both teal and mauve jelly beans and acne will have incurred one false negative (type II error) by failing to reject the null hypothesis for purple, and one false positive (type I error) by mistakenly rejecting the null hypothesis for mauve. In many contexts, false positives are seen as worse outcomes than false negatives.[7] This is arguably also the case here, since the news about "bad" mauve jelly beans will attract much more attention than the lack of a finding for purple, which is likely hidden away somewhere deep in the methods and results sections. We say that a study makes a studywise error if the analysis of the data resulted in at least one false positive result, i.e. if the resulting paper presents one or more false scientific claims. The studywise error rate is the probability of the occurrence of such a false claim. If the studywise error rate is always correctly controlled at 5%, the guarantee is that at least 95% of studies do not result in any false positive scientific claims, regardless of the number of hypotheses that were true or false in these studies. This is a very useful guarantee.

From the preceding paragraph it is clear that the studywise error rate is relevant whenever researchers want to avoid false positive results, and not just when they are interested in the joint studywise null hypothesis. Then, why does the joint studywise null hypothesis, the situation that all null hypotheses for all jelly beans are true, pop up so frequently in layman's explanations of the multiple testing problem? This is simply because the studywise error rate is often largest when all null hypotheses are true. In the jelly bean example we find that unadjusted testing results in a false positive claim with 64% probability if none of the jelly bean colors truly cause acne, but with 60% probability if 2 of the 20 do truly cause acne. The joint studywise null hypothesis is the worst case scenario, but it is not the only situation in which studywise errors may arise.

There is also a logical inconsistency in Rubin's recommendation that a multiple testing correction would be needed only to make statements about the joint studywise null hypothesis, which are often vague and uninformative, but not for much more informative claims about specific null hypotheses. Surely, more precise claims require more precise evidence? In Rubin's logic we may claim that teal and mauve jelly beans cause acne if we find significant results for these colors without correction for multiple testing; however, we are forbidden to conclude from this that at least one color of jelly bean causes acne, since we did not correct for multiple testing. This is a very confusing logic to do inference with. In contrast, valid multiple testing methods respect logic: they allow all inferences to be followed up by their logical implications.[8]

Like all statistical methods, multiple testing methods should never be applied mindlessly. There are situations in which the studywise error rate is appropriate, and there are situations in which other error rates are more meaningful.[9] The jelly bean example, however, seems to me like a clear case for the studywise error rate. We see from the publication in the final panel that the significant result is singled out and presented out of context. If a false positive result, therefore, it would get excessive spotlight. Only if all significant and non-significant results would have been presented side by side, with equal emphasis, unadjusted testing could have been argued for. Instead, since the research started from a non-significant overall test, the set-up as presented in the cartoon even looks like an unplanned subgroup analysis, a classic example of *p*-hacking.

If the studywise error rate is not the most meaningful error rate, there is good reason not to use it. A very bad reason not to use the studywise error rate, however, is to dismiss it on the basis of an avoidable misunderstanding of its definition. This is what Rubin does. It is sad to see an old and debunked[10][11] myth revived once more in this paper.

## References

1. ^*Sander Greenland. (2022). Review of: "When to Adjust Alpha During Multiple Testing".doi:10.32388/d4zmiz.*
2. ^*Xinping Cui, Thorsten Dickhaus, Ying Ding, Jason C. Hsu. (2021).An Overview of Multiple Comparisons. doi:10.1201/9780429030888-1.*
3. ^*(1987). Multiple Comparison Procedures. doi:10.1002/9780470316672.*
4. ^*Thorsten Dickhaus. (2014). The Problem of Simultaneous Inference. doi:10.1007/978-3-642-45182-9_1.*
5. ^*Frank Bretz, Torsten Hothorn, Peter Westfall. (2016).Multiple Comparisons Using R. doi:10.1201/9781420010909.*
6. ^*Sandrine Dudoit, Mark J. van der Laan. (2008).Multiple Testing Procedures with Applications to Genomics. doi:10.1007/978-0-387-49317-6.*
7. ^*Jelle J. Goeman. (2016). Randomness and the Games of Science. doi:10.1007/978-3-319-26300-7_5.*

8. ^Jelle J. Goeman, Aldo Solari. (2010). The sequential rejection principle of familywise error control. Ann. Statist., vol. 38 (6). doi:10.1214/10-aos829.
9. ^Yoav Benjamini, Yosef Hochberg. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society: Series B (Methodological), vol. 57 (1), 289-300. doi:10.1111/j.2517-6161.1995.tb02031.x.
10. ^Jelle J. Goeman, Aldo Solari. (2014). Multiple hypothesis testing in genomics. Statist. Med., vol. 33 (11), 1946-1978. doi:10.1002/sim.6082.
11. ^R. Bender, S. Lange. (1999). Multiple test procedures other than Bonferroni's deserve wider use. BMJ, vol. 318 (7183), 600-600. doi:10.1136/bmj.318.7183.600a.