5-1-2015

# Per Family Error Rates: A Response

James F. Troendle

*Office of Biostatistics Research, National Heart, Lung and Blood Institute, Bethesda, MD*, james.troendle@nih.gov

Keshia-Lee Martin

*American University, Washington, D.C.*, keshialeemartin@gmail.com

Vance W. Berger

*National Cancer Institute, Rockville, MD*, vb78c@nih.gov

Part of the Applied Statistics Commons, Social and Behavioral Sciences Commons, and the Statistical Theory Commons

# *Invited Debate*
# Per Family Error Rates: A Response

| **James F. Troendle** | **Keshia-Lee Martin** | **Vance W. Berger** |
|---|---|---|
| Nat'l. Heart, Lung, & Blood Inst. | American University | National Cancer Institute |
| Bethesda, MD | Washington, DC | Rockville, MD |

As the authors note, the familywise error rate (FWER) is used rather often, whereas the per-family error rate (PFER) is not. Is this as it should be? It would seem that no universal answer is possible, as context determines which is more appropriate in any given application. In the general scenario of testing the benefit of an intervention, one might ideally want an error rate that aligns with the decision for benefit. In most cases the FWER does this pretty well, while allowing one to identify those endpoints for which benefit exists. The PFER does not seem to have any advantage over the FWER in this general testing scenario. Perhaps in some other scenarios the PFER might have some reasonable role.

*Keywords:* Familywise error rate, per-family error rate

## Introduction

As Berger (2004) notes, the alpha level should be selected strategically, based on the ramifications of committing a Type I error relative to a Type II error. The entire testing framework becomes more complicated when dealing with multiple hypothesis tests, and in this case various circumstances must be taken into account. Apart from choosing the proper alpha level for the specific situation, one must also define (prospectively) what constitutes a win (so to speak). Is it enough to find statistical significance on any one endpoint? Or do we instead combine the results in some way to obtain an overall finding?

The familywise Type-I error rate (FWER) is the probability of at least one Type I error in a family of hypotheses occurring, and is used rather often. The

per-family Type I error rate (PFER) is the sum of probabilities of Type I errors in the family for all hypotheses, and is almost never used in practice (Frane, 2015).

When performing multiple hypothesis tests, various circumstances must be taken into account. Apart from choosing the proper alpha level for the specific situation (preferably strategically, rather than based on the one size fits all precedent of 0.05), there is a risk that a Type I (false positive) or Type II (false negative) error may occur. The familywise Type-I error rate (FWER), the probability of at least one Type I error in a family of hypotheses occurring, is used rather often. Meanwhile, the per-family Type I error rate (PFER), the sum of probabilities of Type I errors in the family for all hypotheses, is almost completely ignored (Frane, 2015). Does the PFER deserve as much attention as the FWER receives? We do not attempt any general answer to this question, but, instead, focus on one specific application. For the commonly encountered scenario of testing the benefit of an intervention with several possible endpoints, we think there is a good reason why PFER is not used.

As the author (Frane, 2015) states, committing numerous Type I errors simultaneously is worse than committing only one, with FWER unable to differentiate between creating one Type I error and multiple Type I errors in a family of hypotheses. We suggest that the choice between controlling the FWER or the PFER should be based on the specific situation. The FWER works well for the commonly encountered scenario of testing an intervention with several possible endpoints of interest. The PFER does not appear to have any advantage over the FWER in this scenario, but perhaps in some other scenarios it might. The purpose of this response is not to determine which error rate is superior to the other, but how to establish which error rate should be controlled based on a testing situation. We first consider the scenario of testing an intervention for benefit due to any of several endpoints and then discuss the choice of alpha level.

## Tests of an intervention with multiple endpoints of interest

Consider a study designed to test whether an intervention or exposure is beneficial or detrimental to patient health, compared to some comparison condition. Suppose that benefit can be measured by using any of several endpoints. This is quite a general scenario, which applies equally to clinical trials as well as to behavioral intervention studies or in fact to many observational studies. In this case, it is easy to see that control of the FWER is sufficient to guarantee that if any endpoint is identified as significant, and if biases can be suitably removed by the study design, then either any such endpoint is truly affected by the intervention or an unlikely

event has occurred. This is also true if the PFER is controlled. However, control of the PFER is more restrictive (less powerful) than control of the FWER. Thus, there is no reason to prefer the PFER to the FWER in this general scenario.

An interesting observation about this scenario is that control of the FWER is not necessary to guarantee the type of concordance desired. One might consider testing an intersection hypothesis whose rejection corresponds with evidence of an intervention benefit. To make this clearer, suppose that there are two endpoints, and let $H_1$ ($H_2$) be the null hypothesis that the first (second) endpoint is unaffected by the intervention. If one would recommend the intervention if either endpoint is beneficial, then one really wants to claim benefit if either $H_1$ or $H_2$ are false. This argues for testing the intersection null hypothesis $H_0 = H_1 \cap H_2$. Rejection of this null hypothesis corresponds to benefit. This approach circumvents multiple comparison altogether as only a single hypothesis is tested.

The downside to this approach is that rejection of $H_0$ leaves one unable to conclude improvement on any specific endpoint. As Durkalski and Berger (2009) note, success on a composite endpoint leaves one "unable to determine which outcome is driving the claim". The other caveat to this approach is one must decide how to test $H_0$, which in general could be difficult. An adaptive testing approach could prove useful (Berger and Ivanova, 2002), but the usual solution for testing $H_0$ involves rejecting if $\min(p_1, p_2) \leq \alpha/2$, where $p_1$ ($p_2$) is the $p$-value for testing $H_1$ ($H_2$). With this solution, one is once again controlling the FWER, although in general such an approach could lead to more powerful testing procedures. This observation is a major reason why FWER is the predominantly used error rate for publications of confirmatory findings for studies that test an intervention. Bloch et al. (2001) describe one way of testing a single null hypothesis, although rejecting their null also allows one to conclude non-inferiority on all endpoints.

## Choosing an alpha level

Returning now to the strategic selection of the alpha level, we note that cancer therapy often involves both high risk and high reward. The promise of meaningful improvement is counterbalanced by the almost certain toxicity of the treatment which, in some cases, may have the potential to do more harm than good. That said, false positives and false negatives can both result in grave consequences, including illnesses left untreated, illnesses over-treated, and ultimately higher mortality rates for patients. So the calculation has to consider the relative harm likely caused by each type of error.

As one extreme example (following Berger, 2004), one may conduct a trial to determine if broccoli will prevent arthritis. If broccoli is found, rightfully or wrongfully, to prevent arthritis, then the result would simply be increased consumption of broccoli. Since broccoli is known to have other health benefits, and few (if any) drawbacks, this will still lead to substantial health benefits, regardless if it helps to treat the symptoms of arthritis. So here, a Type I error would not result in very much harm at all. Alpha can be set to a much larger level than the usual 0.05. Another example is Glucosamine and Chondroitin. Like broccoli, these substances have no known side effects and are known to be generally good for cartilage health. Despite no strong evidence of a benefit for sufferers of osteoarthritis pain, many people take Glucosamine and Chondroitin because of the low risk involved coupled with some possible benefit. Conversely, if an aggressive and highly toxic cancer treatment is found to be beneficial, then its increased use will incur additional costs and also result in toxicity, so the benefit should offset this risk, and we should be fairly certain that it does (Berger, 2004). A Type I error in this case would result in severe consequences, so alpha should be small, 0.05 or perhaps even 0.01. These are simple examples, but the concept is that alpha should be carefully considered, and not just set at the usual level of 0.05 as a matter of course (Berger & Hsieh, 2005).

## References

Berger, V. W. (2004). On the generation and ownership of alpha in medical studies. *Controlled Clinical Trials*, *25*(6), 613-619. doi:10.1016/j.cct.2004.07.006

Berger, V. W., Hsieh, G. (2005). Rethinking statistics: basing efficacy alpha levels on safety data in randomized trials. *Israeli Journal of Emergency Medicine, 5*(3), 55-60. http://isrjem.org/IJEM_Aug_AlphaLevels_Proof.pdf

Berger, V. W., Ivanova, A. (2002). Adaptive tests for ordered categorical data. *Journal of Modern Applied Statistical Methods, 1*(2), 269-280. http://digitalcommons.wayne.edu/jmasm/vol1/iss2/36/

Bloch, D. A., Lai, T. L., Tubert-Bitter, P. (2001). One-sided tests in clinical trials with multiple endpoints. *Biometrics, 57*(4), 1039-1047. doi:10.1111/j.0006-341X.2001.01039.x

Durkalski, V., Berger, V. W. (2009). Re-formulating equivalence trials as superiority trials: the case of binary outcomes. *Biometrical Journal, 51*(1), 185-192. doi:10.1002/bimj.200810499

Frane, A. V. (2015). Are Per-Family Type I Error Rate Relevant in Social and Behavioral Science? *Journal of Modern Applied Statistical Methods, 14*(1).