Why are two mistakes not worse than one? A proposal for controlling the expected number of false claims

Thomas Jaki[1,*], Alice Parry[1]
[1]Department of Mathematics and Statistics, Lancaster University, U.K.
* To whom correspondence should be addressed: E-mail: jaki.thomas@gmail.com

Abstract

Multiplicity is common in clinical studies and the current standard is to use the familywise error rate to ensure that the errors are kept at a prespecified level. In this paper we will show that, in certain situations, familywise error rate control does not account for all errors made. To counteract this problem we propose the use of the expected number of false claims (EFC). We will show that a (weighted) Bonferroni approach can be used to control the EFC, discuss how a study that uses the EFC can be powered for co-primary, exchangeable and hierarchical endpoints and show how the weight for the weighted Bonferroni test can be determined in this manner.

Keywords expected number of false claims (EFC); familywise error rate; hierarchical endpoints; multiplicity.

1 Introduction

Multiplicity arises frequently in clinical trials, for example, when testing sequentially, considering multiple treatment arms or multiple endpoints. In the context of confirmatory clinical trials, the guidelines on multiplicity from the European Medicines Agency [1], clearly advocate controlling the familywise error rate (FWER) in the strong sense [2]. Let the number of hypotheses of interest be m and m_0 be the (unknown) number of true null hypotheses. Table 1 defines the standard notation for a multiple hypotheses testing problem [e.g. 2].

Table 1: Standard notation in multiple hypotheses testing.

Hypotheses	Rejected	Not Rejected	Total
True	V	U	m_0
False	S	T	m - m_0
Total	W	R	m

The FWER is then given by $P(V > 0)$. In this article, we will argue that controlling the FWER, though essential in many cases, can be insufficient protection against error inflation and propose that the expected number of rejections, $E(V)$, is more appropriate in some settings. To illustrate the point, consider a diabetes study that investigates if a treatment has an effect on the HbA1c level and/or quality of life. The corresponding null hypotheses can be written as

$$H_H : \quad \theta_H \leq 0$$
$$H_Q : \quad \theta_Q \leq 0$$

where θ_i is the effect (for example difference in change from baseline) for endpoint i (H corresponding to HbA1c and Q to quality of life). Table 2 shows the proportion of times 0, 1 or 2 mistakes are observed under the set of hypothesis discussed above assuming that our endpoints are independent and using a one-sided level of 0.05 for each endpoint.

Table 2: Proportion of times different number of incorrect rejections occur.

Number of rejections	0	1	2
Proportion	0.9025	0.095	0.0025

The probability of incorrectly rejecting both hypotheses is small in this case, the consequences of making both mistakes can, however, be drastic. Consider, for example, a trial that investigates a treatment for two different indications. Making two mistakes in this context means that the treatment could become available to (and taken by) two different patient populations and hence potentially expose a much larger number of patients to an ineffective, possibly even harmful, treatment.

From these results, one can see, that the FWER is $1 - 0.9025 = 0.0975$ since no attempt was made to control the FWER at a specific level. An alternative metric of interest is the expected number of false rejections, $E(V)$. Denoting $\{I_i = 1\}$ as the event that hypothesis H_i is wrongly rejected and $\{I_i = 0\}$ that either H_i is not rejected or that H_i is not true then, $E(V) = E(I_1 + ... + I_m) = E(I_1) + ... + E(I_m) = P(I_1 = 1) + ... + P(I_m = 1)$. Consequently the expected number of rejections is at $2 * 0.05 = 0.10$ slightly larger than the FWER. By rewriting we find that find that

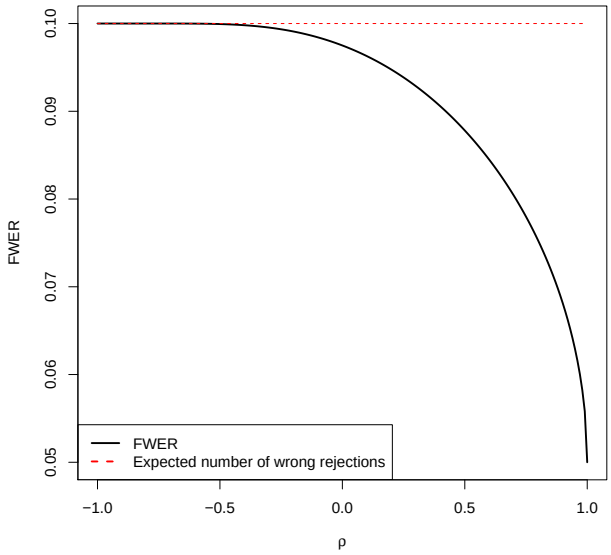
$$\begin{aligned} E(V) &= P\,(\text{incorrectly reject } H_H \text{ and retain } H_Q) \\ &\quad + P\,(\text{retain } H_H \text{ and incorrectly reject } H_Q) \\ &\quad + 2 * P\,(\text{incorrectly reject } H_H \text{ and } H_Q) \\ &= \text{FWER} + P\,(\text{incorrectly reject } H_H \text{ and } H_Q) \end{aligned}$$

and hence it becomes apparent, that designing our hypothetical study to control the FWER treats incorrectly concluding an effect on HbA1c equal to incorrectly concluding an effect on quality of life equal to incorrectly concluding an effect on HbA1c *and* quality of life. This immediately begs the question: Why is making two mistakes not worse than making one?

At this point one could argue, that one can live with the very small probability of making two mistakes and hence not consider the problem any further. Looking at the FWER and the expected number of wrong rejections for the set of hypotheses above under the normal model for varying correlation in Figure 1, however, clearly shows that the expected number of wrong rejections becomes substantial as correlation increases. In particular for correlations close to one, the expected number of wrong rejections is almost twice the FWER.

After motivating the potential shortcoming of using the FWER, we will formalize our proposal, the expected number of false claims (EFC), in the next section. We then continue to show a simple way to control the EFC and discuss powering studies based on it (Section 3). We illustrate the methods for different structures of hypotheses and finish with a brief discussion.

Figure 1: Comparing the FWER and expected number of wrong rejections for different correlation between endpoints. Calculations are based on the joint multivariate normal distribution obtained using the R package `mvtnorm` [3].



2 The expected number of false claims

In the motivating example, we have assumed that each wrong rejection has unwanted consequences. In many multiple testing situations, however, making an additional mistake is of no further consequence and hence the distinction between one or more mistakes irrelevant. For example in a dose finding setting where it is desired to determine the minimum effective dose (MED), incorrectly rejecting the hypothesis that a particular dose is ineffective is of no further consequence if a dose below has already incorrectly been declared effective. In some sense, the crucial mistake − a wrong dose being determined as the MED − happens as soon as you make one wrong rejection. Additional wrong rejections have no further impact on this wrong decision. This observation was also utilized in [4] to control the FWER when estimating the MED. For our purposes, it is therefore essential to firstly introduce the notion of a claim, a single hypothesis or set of hypotheses, whose rejection will result in a consequential decision. As the name suggests, we are thinking here of rejections that are necessary to add a label claim to a product, although the applications of this notion goes beyond this specific application (see Section 4).

Let l_i be the number of hypotheses that need to be rejected to make claim i and define the event "making claim i" as $\{C_i\} = \{\text{reject } l_i \text{ or more } H \in \mathcal{K}_i\}$ for some set of relevant null hypotheses, $\mathcal{K}_i$. Most commonly rejection of all relevant hypotheses would be required to make a claim and hence we are focusing on these cases in the remainder of the manuscript. Note also that in the situation of each claim being based on a single hypothesis and assuming that the null hypothesis is true implies $\{C_i\} = \{I_i = 1\}$.

For the previous example we define $\mathcal{K}_1 = \{H_H\}$ and $\mathcal{K}_2 = \{H_Q\}$ so that $\{C_1\} = \{\text{reject } H_H\}$ and $\{C_2\} = \{\text{reject } H_Q\}$. A different, possibly more realistic, example could investigate the same two endpoints, but only be interested in the quality of life endpoint if an effect on HbA1c has been established. In that case we would have $\mathcal{K}_1 = \{H_H\}$ and $\mathcal{K}_2 = \{H_H, H_Q\}$ so that

$\{C_1\} = \{\text{reject } H_H\}$ and $\{C_2\} = \{\text{reject } H_H \text{ and } H_Q\}$.

With this definitions in mind and supposing that M claims are possible, we can now define the expected number of false claims (EFC) as

$$EFC = \sum_{i=1}^{M} \max_{\boldsymbol{\theta}_{0m} \in \boldsymbol{\Theta}_{0m}} P(\{C_m\}|\boldsymbol{\theta}_{0m})$$

where $\boldsymbol{\Theta}_{0m}$ denotes all possible parameter configurations relating to the hypotheses in $\mathcal{K}_m$ that are consistent with the respective null hypotheses for the mth claim.

Going back to the first example where both individual hypothesis themselves result in a claim, we can easily find the EFC as 0.1. In the second example where claim 2 can only be made if both hypothesis are rejected, the EFC under the assumption of independence is smaller at $0.05 + 0.0025 = 0.0525$ due to the fact that the second claim is much harder to achieve.

2.1 Related error rates

Now that we have introduced our proposal it is worth pointing out some relationships between other proposals in the literature. The first point to make in this respect is that no method exists that explicitly considers claims. The per family error rate (PFER) discussed in [5] defined there as $\frac{\text{number of erroneous rejections}}{\text{number of families}}$ or more formally described as $E(V)$ is a special case of our proposal. The fundamental difference between the PFER and the EFC is that for the former any wrongly rejected hypothesis is counted while the EFC only consider cases where at least l_i hypotheses in $\mathcal{K}_i$ are rejected. To clarify this difference further, consider the second example given where $\mathcal{K}_2 = \{H_H, H_Q\}$. In this setting incorrect rejection of H_H *and* H_Q is necessary for it to contribute to the EFC while either one of them would be counted in the PFER.

If we focus on the situation where only a single hypothesis is required for making a claim (i.e. $\{C_i\} = \{I_i = 1\}$ and consequently $EFC = E(V)$), then we have already shown in section 1 that for two hypothesis the EFC is related to the FWER in the following manner $EFC = E(V) = FWER + P\,(\text{incorrectly reject } H_1 \text{ and } H_2)$ with similar results easily obtainable for more hypotheses. Another related error rate in this case is the false discovery rate (FDR) [6] defined as $E(\frac{V}{R})$. From this definition it is apparent that the FDR is bound between 0 and 1 which could be viewed as an advantage while computationally it is slightly more complex as the case of $R = 0$ needs to be considered.

2.2 Controlling the EFC

In [7] it is noted that "...[in a clinical study] the claim-wise error rate is probably the most important attribute to control..." and one way to achieve this is by controlling the EFC at a certain level, say η. This is, in fact, quite easily achieved by simply splitting the overall level, η, equally between the individual probabilities, that is applying a Bonferroni adjustment to each claim probability. More specifically, it is easy to see that ensuring

$$\max_{\boldsymbol{\theta}_{0m} \in \boldsymbol{\Theta}_{0m}} P(\{C_m\}|\boldsymbol{\theta}_{0m}) \leq \frac{\eta}{M} \qquad m = 1, \ldots, M$$

will guarantee the overall level η.

Although this is a very simple approach, it may not be very practical as rarely all claims will be equally important. A more realistic way to control the EFC therefore uses a weighted Bonferroni adjustment [8] that allows more weight to be assigned to more important claims. More specifically, ensuring that

$$\max_{\boldsymbol{\theta}_{0m} \in \boldsymbol{\Theta}_{0m}} P(\{C_m\}|\boldsymbol{\theta}_{0m}) \leq w_m \eta \qquad m = 1, \ldots, M$$

with $w_m \geq 0$ such that $\sum_{m=1}^{M} w_m = 1$ will clearly also control the EFC.

2.3 Examples

The general concept of the EFC as well as methods to control it are fairly straightforward. In many practical situations, the nature of the individual claims do, however, require particular care when controlling the EFC. In this section we will provide 3 illustrative examples to show how EFC control can be achieved. For all illustrations we assume normally distributed endpoints and use the `mvtnorm` package [3] for the computations. We will use $\eta = 0.05$ and one-sided hypotheses for superiority. Results for two-sided hypotheses follow the same patterns except that they are symmetric around a correlation of zero.

2.3.1 Co-primary endpoints

The first case we want to discuss, though only for completeness, is the situation where multiple primary variables are required to describe a clinical benefit. The CPMP guidance for Alzheimer's disease [9], for example, stipulates that a treatment must show an effect on a cognitive endpoint and a functional endpoint. Consequently, even though there are two hypotheses to be tested, only a single claim is investigated. In order to control the EFC it is therefore sufficient to ensure that the probability of making this claim (i.e. rejecting both hypothesis) is controlled at η. Current practice in this situation is to require each hypothesis to be rejected at level η so that the EFC is clearly below the desired level in this case.

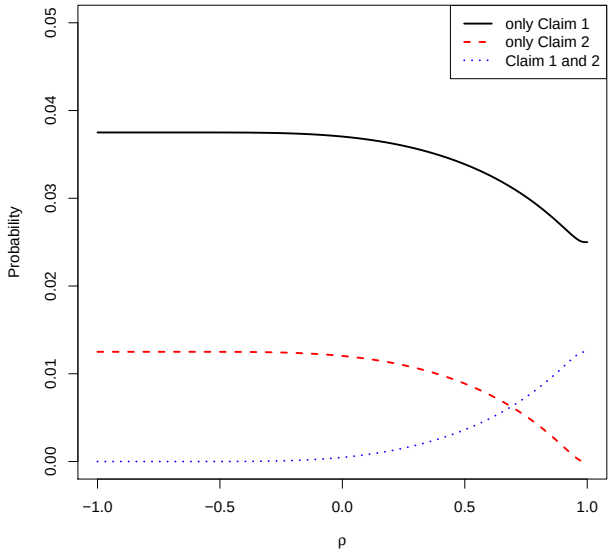
2.3.2 Exchangeable claims

In some situations multiple independent claims are possible for one treatment. We term this case exchangeable claims since we are envisaging the case were claims are not dependent on each other and hence making either claim would be considered a success. For this illustration we return to the motivating example which considered showing an effect on HbA1c or quality of life (or both) a success. Despite considering both claims a success in this setting let us assume that making a claim on HbA1c is more important. Consequently we can construct a testing strategy that controls the EFC at level η by testing H_H at level $w_1\eta$ and H_Q at $(1-w_1)\eta$.

Figure 2 shows the probability of making only claim 1, only claim 2 and both for $w_1 = 0.75$ as the correlation changes. Notable is that the probability of making both claims is negligible for negative correlations, but becomes substantial for strong positive correlations. The EFC which is simply the sum of the probabilities of making exactly one claim plus twice the probability of making both claims is exactly 0.05 as desired.

Although the situation described where either a reduction in HbA1c levels or quality of life are of primary interest (i.e. indifference about which of the two is improved) is probably not very frequently encountered, there are many related settings where exchangeable claims occur. For example in the context of regulatory approval, conditional approval of the treatment (Claim

Figure 2: Probability of making claims when the claims are exchangeable across different correlations. A weight of $w_1 = 0.75$ is used.



1) and full approval (Claim 2) would fall into this framework. Licensing a treatment for different indications or decisions about (disjoint) subgroups can also be framed as exchangeable claims.

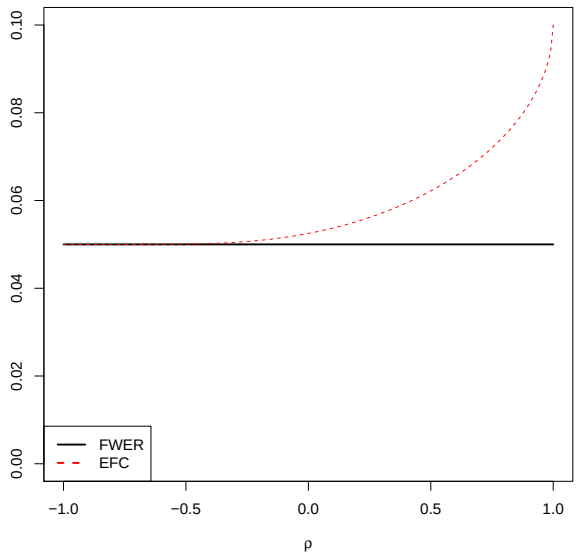
2.4 Hierarchical endpoints

The final case we want to discuss concerns the frequently used hierarchical testing strategies. In the context of malaria clinical trials, for example, cure is the most important (and primary) endpoint. Most treatments for malaria are, however, effective so that other measures to distinguish treatments are frequently of interest. Consider, for example, prevention of new infections as a secondary endpoint. In such a case one would usually employ a hierarchical testing strategy that only tests the secondary endpoint if the primary has been rejected. Note that it is possible to distinguish different infections in malaria [10] so that it is not necessary to actually cure a patient to establish if a new infection has occured. Putting this situation into the context of claims and using subscripts C for the cure endpoint and P for the prevention endpoint, we have $\{C_1\} = \{\text{reject } H_C\}$ and $\{C_2\} = \{\text{reject } H_C \text{ and } H_P\}$.

Consider first the expected number of false claims when the following fixed sequence test procedure [11] is used: The prevention hypothesis is only tested at full level if the primary hypothesis has been rejected at full level. In Figure 3 it is easy to see the false sense of security using FWER control can give. In this example, the FWER is controlled at the desired level of 0.05, the EFC, however, is up to twice as large. In the context of our example, this means that in addition to allowing an error rate of 0.05 for the primary cure hypothesis we also allow an up to 0.05 chance of concluding a preventative effect when there is none.

To control the EFC at level η we can, however, once more use the weighted Bonferroni test. Since claiming cure is clearly more important than claiming a preventative effect, we will use a weight of $w_1 = 0.8$ here. This means that we can test $\{C_1\}$ at level $w_1\eta$ which implies that we can test H_C at the same level as well as it is the only hypothesis relevant for this claim.

Figure 3: FWER and EFC for 2 hierarchical endpoints using a fixed sequence test for FWER control.



To achieve EFC control we then also need to test $\{C_2\}$ at level $(1-w_1)\eta$ which more precisely means that we need to ensure that under the null $P(\text{reject } H_C \text{ and } H_P) \leq (1-w_1)\eta$. It is clear, that there are different ways to achieve this. One approach that will ensure that this probability is controlled for any correlation ρ is to recognize, that it is maximized for $\rho = 1$. Consequently testing H_P at level $(1-w_1)\eta$ will ensure that $P(\text{reject } H_C \text{ and } H_P) \leq (1-w_1)\eta$ holds. Figure 4 shows the realized EFC and FWER for $w_1 = 0.8$ when using this method. As previously the EFC and FWER are essentially identical for negative correlations while the difference is increasing as the correlation increases. The EFC is below the desired nominal level of 0.05 and only exhausts the full level for perfect positive correlations due to the use of the worst case configuration and the dependence between claims. Alternative approaches that ensure $P(\text{reject } H_C \text{ and } H_P) \leq (1-w_1)\eta$ that incorporate the correlation could be used instead to ensure exhaustion of the error level.

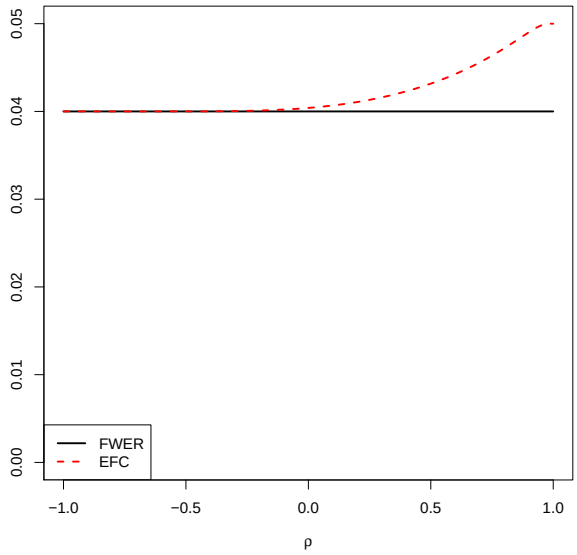
3 Power

Having established our proposal, we now consider powering studies that are designed to control the expected number of false claims and show how setting power constraints can be used to determine the weight of the weighted Bonferroni test. As before we will differentiate between the three different types of endpoints/claims as power has different implications for these different settings. Throughout, however, we will use $1-\beta_i$ to describe the power we wish to have to make claim i and denote the vector of parameters in $\mathcal{K}_m$ with a particular effect of interest as $\boldsymbol{\theta}_{1m}$.

3.1 Co-primary endpoints

Co-primary endpoints, as discussed above, occur in the situation where multiple primary variables are required to describe a clinical benefit. Consequently only a single claim − which is established based on several endpoints − is of interest. A natural way to power such a study

Figure 4: EFC with $w_1 = 0.8$ and $w_2 = 0.2$ using a Bonferroni adjustment for EFC control.



is to ensure that the probability of rejecting all hypotheses necessary for making the claim of interest is sufficiently large for worthwhile effects. In particular one would power such a study to ensure that $P(\text{reject } l_i \text{ or more } H \in \mathcal{K}_1 | \boldsymbol{\theta}_{11}) = P(\{C_1\}|\boldsymbol{\theta}_{11}) \geq 1 - \beta_1$.

3.2 Exchangeable claims

Looking at the situation where each claim separately can be viewed as a success, the initial thought is to ensure adequate power for making at least one claim. Previously, however, we have argued, that even for exchangeable claims there does exist a difference in how important it is to make a claim. In the context of provisional versus full approval, for example, it is clearly superior to obtain full approval. To account for this situation we have previously allowed the claims to carry different weights − how these weights are arrived at, however, was left open. Our proposal now is to power the study to implicitly determine the weights given to each claim. In particular, for a situation of M exchangeable claims, we propose to determine the sample size of the study, n, and the weights $w_1, \ldots, w_{M-1}$ simultaneously through solving the system

$$\begin{aligned} P(\{C_1\}|\boldsymbol{\theta}_{11}) &\geq 1 - \beta_1 \\ P(\{C_2\}|\boldsymbol{\theta}_{12}) &\geq 1 - \beta_2 \\ \vdots &\quad \vdots \\ P(\{C_M\}|\boldsymbol{\theta}_{1M}) &\geq 1 - \beta_M. \end{aligned} \tag{1}$$

To give an example, consider a situation with two exchangeable claims each comprised of a single hypothesis (e.g. the previous example of HbA1c and quality of life). Suppose further that the standardized effect of interest for the first and second hypothesis is 0.5 and 0.4, respectively. Table 3 shows the required sample size for the two different ways to power the study using an EFC of 0.05 and assuming a correlation of 0.5 between endpoints.

Unsurprisingly, the sample size required when looking to make at least one claim is (substantially) lower, then when requiring a certain power for each claim. More weight is given to

Table 3: Sample size required per arm (n) and optimal weight (w_1) for two exchangeable claims and standardized effects of $(0.5, 0.4)$ for EFC of 0.05 and a correlation of 0.5 between endpoints. The first line powers the study to have power of 90% to make at least one claim. The lower part of the table uses one power constraint for each claim for a variety of power constraints.

	$1-\beta_1$	$1-\beta_2$	w_1	n
At least one claim	0.9	NA	0.82	68
Separate power	0.9	0.9	0.14	113
	0.9	0.8	0.35	92
	0.9	0.7	0.55	82
	0.8	0.9	0.05	109
	0.8	0.8	0.17	84
	0.8	0.7	0.33	71

the first claim in this situation due to the larger desired effect on the first claim. When using separate powers, the weight associated with each claim does adjust with the required power and also the anticipated effect (not shown) as expected. It is notable that the weight on the first claim when requiring equal power for each claim is below 0.5 (in contrast to being above 0.5 for making either claim) to counteract the smaller effect in the second endpoint by giving the second claim more weight.

3.3 Hierarchical endpoints

A natural way to power a study using a hierarchical structure is to associate a certain power with each claim in the structure as before. This setting yields the same system of equations as given in (1). The fundamental difference between them is how the different claims, $\{C_i\}$ are defined and hence which hypotheses need to be rejected to make each claim. Note that a special case of this proposal is to use $\beta_i = 1,\ i = 2, \ldots, M$ in which case the study is powered only for the first claim − a solution often employed when using FWER control.

To illustrate powering for hierarchical claims, we will use a similar setting to the one described in the previous section. Consider a situation with two hierarchically ordered claims, each comprised of a single hypothesis (e.g. the primary claim is on a reduction HbA1c level while a secondary claim is on quality of life). Suppose further that the standardized effect of interest for the first endpoint is 0.5 and 0.4 for the second. Table 4 shows the required sample size using an EFC of 0.05 and assuming a correlation of 0.5 between endpoints for a variety of power constraints.

The sample size required when only powering for the first claim is identical to a standard 2-sample z-test. In this situation it is, however, notable that not all weight is given to the first claim. This is due to requiring the sample size to be an integer. In fact, for $n = 138$ any weight between 0.98 and 1 will satisfy the power requirement. Similarly multiple choices for w_1 are often also available for other situations considered. We have simply used the smallest value of w_1 satisfying the power constraint in all of our evaluations. Notice also that for the hierarchical structure, it is not always possible to satisfy the power constraint exactly. Depending on the effect size and required powers, one of the powers may be larger than the desired value due to the correlation between the claims. To see this, consider a case where two claims in a hierarchical procedure are required to have 90% power. In order to achieve 90% power for the second claim the power for the first claim must be larger than 90% as there is a chance of making claim 1

Table 4: Sample size required per arm (n) and optimal weight (w_1) for two hierarchical claims with standardized effects of $(0.5, 0.4)$ for EFC of 0.05 and a correlation of 0.5 between endpoints. The realized powers are given in the columns $1 - \hat{\beta}_i$.

$1-\beta_1$	$1-\beta_2$	w_1	n	$1-\hat{\beta}_1$	$1-\hat{\beta}_2$
0.9	0	0.98	69	0.900	0.226
0.9	0.9	0.21	130	0.957	0.900
0.9	0.8	0.24	101	0.902	0.800
0.9	0.7	0.48	85	0.900	0.706
0.8	0	0.98	50	0.801	0.133
0.8	0.9	0.21	130	0.963	0.900
0.8	0.8	0.24	101	0.902	0.800
0.8	0.7	0.24	83	0.832	0.700

but not claim 2 as long as the two are not perfectly correlated. The realized powers for each claim are provided in the additional columns labeled $1 - \hat{\beta}_i$ in the table. Compared to the results for exchangeable endpoints, the sample size is increased for the hierarchical setting as soon as we do require some power for the second claim as expected. For the considered setting, the realized power for the first claim tends to be larger than the desired minimum which is, in part, due to the second claim requiring making claim 1.

4 Discussion

In this paper we have introduced the expected number of false claims (EFC), which is designed to ensure that all relevant mistakes are properly accounted for. We have also shown that a weighted Bonferroni adjustment can be used to control the EFC at the desired level and illustrated how powering studies based on the EFC can be used to determine the weights of the weighted Bonferroni adjustment. Although we have focused throughout this work on cases were the EFC is different from the familywise error rate, both concepts are equivalent when only one claim is sought or when claims are mutually exclusive. At the same time the aim of this work is not to claim that the EFC is superior to the FWER, but rather show that in some situations, such as hierarchically structured questions, it might be more appropriate. A generalized FWER (gFWER) has also been suggested [12, 13]. This gFWER is designed to account for the willingness to tolerate more than one false rejections due to the high volume of hypotheses to be tested, for example, in genomic trials as long as the number of is controlled, i.e. pre-defined. The EFC approaches the issue of more errors from the other side. Rather than allowing more mistakes, the EFC focuses on properly accounting for all errors made. As an immediate consequence the conventional levels of significants used (i.e. 0.05) may be to stringent for situations with many claims.

Throughout most of the paper we have focused on the case where two endpoints are of interest. The concept of the EFC is, however, applicable in many more settings. For example, in the context of regulatory approval, obtaining conditional approval versus full approval of a treatment naturally falls into the framework discussed. Similarly, we believe that the concept of EFC is quite natural for the development of a treatment for several indications or multiple populations.

5 Acknowledgements

We would like to thank the annonymous reviewers for their helpful suggestions that improved the clarity of the manuscript notably. This report is independent research arising from Prof. Jaki's Career Development Fellowship (NIHR-CDF-2010-03-32) and Senior Research Fellowship (NIHR-SRF-2015-08-001) supported by the National Institute for Health Research. The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health.

References

[1] Committee for Proprietary Medicinal Products. *Points to Consider on Multiplicity Issues in Clinical Trials*. The European Agency for the Evaluation of Medicinal Products, 2002. last visited 23/09/2013.

[2] F Bretz, T Hothorn, and P Westfall. *Multiple Comparisons Using R*. CRC Press, 2011.

[3] Alan Genz, Frank Bretz, Tetsuhisa Miwa, Xuefei Mi, Friedrich Leisch, Fabian Scheipl, and Torsten Hothorn. *mvtnorm: Multivariate Normal and t Distributions*, 2013. R package version 0.9-9995.

[4] Jason C Hsu and Roger L Berger. Stepwise confidence intervals without multiplicity adjustment for doseresponse and toxicity studies. *Journal of the American Statistical Association*, 94(446):468–482, 1999.

[5] JW Tukey. The collected works of john w. tukey, vol. viii. multiple comparisons: 1948–1983. *Chapman & Hall, New York*, 1953.

[6] Y Benjamini and Y Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

[7] A Phillips, C Fletcher, G Atkinson, E Channon, A Douiri, Thomas Jaki, J Maca, D Morgan, J Roger, and P Terrill. Multiplicity: Discussion points from the Statisticians in the Pharmaceutical Industry multiplicity expert group. *Pharmaceutical Statistics*, 2013. Published online; doi: 10.1002/pst.1584.

[8] Y Benjamini and Y Hochberg. Multiple hypotheses testing and weights. *Scandinavian Journal of Statistics*, 24:407–418, 1997.

[9] CHMP/EWP. *Guideline on medicinal products for the treatment of Alzheimer's disease and other dementias*. Committee for medicinal products for human use. CPMP/EWP/553/95 Rev. 1, London, UK, 2008. accessed 2 Oct 2013.

[10] T Jaki, A Parry, K Winter, and I Hastings. Analysing malaria drug trials on a per-individual or per-clone basis: a comparison of methods. *Statistics in medicine*, 32(17):3020–3038, 2013.

[11] BL Wiens. A fixed sequence Bonferroni procedure for testing multiple endpoints. *Pharmaceutical Statistics*, 2(3):211–215, 2003.

[12] G Hommel and T Hoffmann. Controlled uncertainty. In P Bauer, G Hommel, and E Sonnemann, editors, *Multiple Hypothesenprüfung/Multiple Hypotheses Testing*, pages 154–161. Springer, 1988.

[13] E Lehmann and J Romano. Generalizations of the familywise error rate. *Annals of Statistics*, 33(3):1138–1154, 2005.