Sample Size for Multiple Hypothesis Testing in Biosimilar Development

Johanna Mielke\textsuperscript{a}, Byron Jones\textsuperscript{a}, Bernd Jilma \textsuperscript{b}, and Franz König \textsuperscript{b}

\textsuperscript{a}Novartis Pharma AG, Basel, Switzerland; \textsuperscript{b}Medical University of Vienna, Vienna, Austria

\textbf{ABSTRACT}
In biosimilar development, often multiple endpoints within a study, multiple doses and routes of administration or multiple studies in different populations are considered. However, a regulatory requirement is that equivalence of the biosimilar and the reference drug has to be shown for all comparisons, which would typically require a large sample size for a clinical development program. One way that the sample size can be reduced, when \( m \) null hypotheses are to be considered, is to require that only \( k < m \) null hypotheses have to be rejected to get approval. In fact, this is a realistic requirement, since despite their guidelines the European Medicines Agency (EMA) has already approved applications for biosimilars where not all primary endpoints met the equivalence criteria. In this article, we investigate the properties of the test for the success of at least \( k \) out of \( m \) endpoints and discuss several multiplicity adjustments that might be useful in practice. We illustrate the impact of multiple hypotheses testing on the sample size using three real-world examples of pharmacokinetic studies that were submitted to the EMA for the approval of biosimilars. Supplementary materials for this article are available online.

\section{1. Introduction}

A biosimilar, or follow-on biologic, is “a biological product that is highly similar to the reference product notwithstanding minor differences in clinically inactive components” (FDA 2015) and is produced as a cheaper copy of the reference product. The use of biosimilars is predicted to lead to a $250 billion saving on the costs of biologics in the United States between 2014 and 2024 (Miller 2013). Since several patents on important biologics have already expired and many more will expire in the near future (Generics and Biosimilars Initiative 2015), the development of biosimilars has recently become very important to the pharmaceutical industry.

If this development is to be done efficiently in terms of the cost and time needed to get a biosimilar to market, then keeping sample sizes to the minimum needed to achieve success is vitally important. In this article, we illustrate a methodology than can be very useful when trying to conserve sample size. Our work is motivated by the results of a systematic review of the clinical development programs for biosimilars in Europe (Mielke et al. 2016). The review showed that, even though multiple hypothesis tests were conducted in the majority of the applications for market authorization of biosimilars to the European Medicines Agency (EMA), it was apparent that often no consideration was given to the impact of this multiple testing on sample size. Indeed, this issue was shown to be present in several stages of the biosimilar development cycle, and two examples from the systematic review can be easily identified: (1) There were several studies, for example, the Pharmacokinetic/Pharmacodynamic (PK/PD)-trial undertaken for the application of Tevagratim (Lubenau et al. 2009), which is a biosimilar to Neupogen (Amgen), which compared several doses and routes of administration. In this study, two different doses (5 or 10 \( \mu \text{g/kg} \)) and subcutaneous (SC) and intravenous (IV) routes of administration were compared. (2) In some applications, multiple studies in different populations were undertaken. For example, for the application of Abasaglar (biosimilar to Lantus), the sponsor conducted one Phase III study in patients with diabetes mellitus type 1 (Blevins et al. 2015) and one Phase III study in patients with diabetes mellitus type 2 (Rosenstock et al. 2015). In addition, for PK trials, the regulatory guidelines for showing biosimilarity (CHMP 2014) published by the EMA recommends showing equivalence both for AUC and Cmax (i.e., multiple co-primary endpoints must be considered). We note that when testing for equivalence, the null hypothesis is that of “not equivalent” and the alternative hypothesis is that of “equivalent.” Defining what is meant by “equivalence” will be context dependent and how it is tested will be explained in Section 2.

In biosimilar development, it is desirable that equivalence is shown in all doses, routes of administration, patient populations, and endpoints. Because all hypotheses have to be rejected, there is no multiplicity issue and the Type I error rate is controlled. However, there is an impact on the Type II error rate that results in a higher sample size (for a specified power) if all hypotheses need to be rejected, compared to the sample size required if only a single hypothesis needs to be rejected. Whereas the regulatory agencies urge sponsors to guard against potential inflation of the Type I error rate, as this might lead to a higher risk for the consumers, the decrease in power if multiple hypotheses are tested seems to be often neglected in practice in biosimilar development. It is noted that if several studies with identical set-ups are planned and should be summarized, a meta-analysis (DerSimonian and Laird 1986) can be conducted.
which is a technique for summarizing multiple studies that does not decrease the power. However, meta-analysis is not applicable for studying different endpoints, doses, indications, or routes of administrations, which is more common in biosimilar development.

The aim of this article is to illustrate the decrease in power due to multiple testing and how its effects can be controlled in a manageable way in clinical biosimilar development. But similar considerations might be applicable for the quality assessment (CHMP 2017; Burdick, Thomas, and Cheng 2017). In the process, we extend the results of multiple testing to the situation where only a subset of the total number of null hypotheses needs to be rejected to declare success. This extension is motivated by previous successful applications to the EMA in which not all primary endpoints were successful (Mielke et al. 2016) but approval was given nevertheless: one example for this is the application for Zarzio by Sandoz (CHMP 2008), which is a biosimilar to Neupogen (Amgen): the sponsor submitted four PK/PD studies, with four different doses (1, 2.5, 5, and 10 µg/kg), and with one dose used in two routes of administration. For the lower doses, and after multiple SC doses, both Cmax and AUC did not meet the acceptances ranges. Nevertheless, the product was finally approved and equivalence was claimed for the other doses and the IV route of administration. So, in situations where a very high sample size is required to achieve a nominal power to reject all m hypotheses in a set, one possible solution is to get regulatory agreement that approval can be obtained if at least a subset of k of m hypothesis tests of lack of equivalence can be rejected.

The concept of testing for equivalence on at least k of m endpoints is particularly appealing for biosimilar development because, in contrast to the development of a new drug in which often one study is considered to be pivotal, for biosimilars a so-called “totality of the evidence” approach is used in which all studies in a program are equally important (Holzmann, Balser, and Windisch 2016). The decision is therefore not made by considering, for example, the PK studies only, but by taking into account the results of pre-clinical studies and Phase III studies as well.

It is important to note that, without any adjustment for multiplicity, and in case not all tests can be rejected, as in the example above, claiming equivalence on only k out of m tests (k < m) will inflate the Type I error rate and is therefore not an approach that will be acceptable to regulators. In this article, as an alternative to this naive approach, we describe a valid testing procedure for k out of m endpoints that controls the Type I error rate. The concept of a “k out of m” test was first introduced by Rüger (1978) and an adjustment for multiplicity was given by Hommel and Hoffmann (1988). It has also been discussed more recently for the special case of at least 2 out of 3 endpoints by Quan, Bolognese, and Yuan (2001). They proposed an adjustment of the significance level for controlling the Type I error rate for this case, but did not consider the influence on the sample size or power of the test. The special case of 2 out of 3 successes was also discussed by Ristl et al. (2016) who proposed a fallback procedure for inference on individual endpoints for situations in which not all co-primary endpoints can be rejected.

In this article, we generalize the results of Quan, Bolognese, and Yuan (2001) by conducting a simulation study to show that the increase in sample size for testing k out of m tests is dependent on the correlation between the tests. We focus specifically on the situation in which k < m since, to the best of our knowledge, no information about the operating characteristics of this approach has been published. As the Type I error rate must also be controlled, we also describe and compare multiplicity adjustment strategies for the k out of m tests situation and use, in addition to the unadjusted test, the Bonferroni adjustment (see, e.g., Wiens and Dmitrienko 2010) and the α-level adjustment for multiplicity that was first proposed by Hommel and Hoffmann (1988).

In Section 2, the statistical tests for equivalence are described. The adjustment to the significance level that is required in the case of k < m is then introduced, where the tests may be on multiple endpoints, multiple comparisons, etc. The setup and the results of the simulation study are then presented in Section 3 to illustrate the impact of multiple hypothesis testing on the required sample size. In Section 4, the increase in sample size is shown using three different real-world examples, which are taken from clinical development programs that were submitted to the EMA for getting approval as a biosimilar in Europe. In Section 5, we give some general conclusions and advice.

2. Methods

For simplification, we use the term test to refer to the multiple treatment arms, studies, or endpoints. It is assumed that m statistical hypothesis tests are to be carried out and that the equivalence margins ±Δ are identical for all tests. As described in the introduction, multiplicity issues occur in all stages of development and therefore, both parallel groups designs, which are mostly used for efficacy comparisons in patients, and cross-over designs for PK/PD studies are relevant study designs of interest. As a compromise, we focus for the simulation studies in Section 3 on the parallel groups design and for the real-world examples in Section 4 on cross-over designs. It should be noted that the choice of design has no influence on the qualitative interpretation of our results. In this section, we introduce the methodology based on the parallel groups design, the notation and test statistics for the cross-over design are given in Section 4.

Let $Y_{j,T}^{(i)}$ be the response for the ith test of subject j and for the Test treatment (T). It is assumed that $Y_{j,T}^{(i)}$ follows a normal distribution with mean $\mu_{T}^{(i)}$ and variance $\sigma_{T}^{2(i)}$. The notation and assumptions for the Reference treatment (R) are similar, with response $Y_{j,R}^{(i)}$ and parameters $\mu_{R}^{(i)}$ and $\sigma_{R}^{2(i)}$. The true difference in means between T and R for the ith test is denoted by $\delta^{(i)} = \mu_{R}^{(i)} - \mu_{T}^{(i)}$. We assume that the sample size for each group (T or R) is the same and equal to $n^{(i)}$ for the ith test. The total sample size for the ith test is denoted by $N^{(i)} = 2n^{(i)}$ for an assumed parallel groups design.

2.1. Evaluation of Equivalence for a Single Test

The objective of an equivalence trial is to show that the Test treatment is neither inferior nor superior to the Reference treatment (Ye and Yao 2012). For some prespecified equivalence margin Δ, the following hypothesis for the ith test should be tested:

$H_{0}^{(i)} : |\mu_{R}^{(i)} - \mu_{T}^{(i)}| \geq \Delta$ vs. $H_{1}^{(i)} : |\mu_{R}^{(i)} - \mu_{T}^{(i)}| < \Delta$, 

$i = 1, \ldots, m$. 

The null hypothesis can be evaluated using two one-sided tests, the TOST procedure (Schuirmann 1987), where each null hypothesis is split into two parts:

\[ H_{0(i)}^{(1)} : \mu_R^{(i)} - \mu_T^{(i)} \leq -\Delta \text{ vs. } H_{1(i)}^{(1)} : \mu_R^{(i)} - \mu_T^{(i)} > -\Delta, \]

\[ i = 1, \ldots, m; \]

\[ H_{0(i)}^{(2)} : \mu_R^{(i)} - \mu_T^{(i)} \geq \Delta \text{ vs. } H_{1(i)}^{(2)} : \mu_R^{(i)} - \mu_T^{(i)} < \Delta, \]

\[ i = 1, \ldots, m. \]

For simplicity, and in line with the work of Kong, Kohberger, and Koch (2004), we assume that the variances of the responses associated with each test are known and need not be estimated.

The corresponding test statistics are then

\[ Z_{1(i)} = \frac{\bar{y}_R^{(i)} - \bar{y}_T^{(i)}}{\sqrt{\frac{\sigma_{R(i)}^2}{n(i)} + \frac{\sigma_{T(i)}^2}{n(i)}}} + \Delta \]

and

\[ Z_{2(i)} = \frac{\bar{y}_T^{(i)} - \bar{y}_R^{(i)}}{\sqrt{\frac{\sigma_{T(i)}^2}{n(i)} + \frac{\sigma_{R(i)}^2}{n(i)}}} + \Delta, \]

where \( \bar{y}_{R(i)} \) and \( \bar{y}_{T(i)} \) being the mean values for \( T \) and \( R \), respectively. We assume

\[ \sigma_{T(i)}^2 = \sigma_{R(i)}^2 = \sigma_{(i)}^2. \]

As both test statistics follow a standard normal distribution, the null hypothesis is rejected if both \( Z_{1(i)} \) and \( Z_{2(i)} \) are larger than the \( (1 - \alpha) \) -quantile of the standard normal distribution.

### 2.2. Multiple Equivalence Tests

According to the description in the previous section, a finite number of statistical hypotheses \( H^{(1)}, \ldots, H^{(m)} \) are tested. The null hypothesis is, therefore

\[ H_0 : \text{At least } m - k + 1 \text{ null hypotheses are true,} \]

and the alternative hypothesis is

\[ H_1 : \text{Less than } m - k + 1 \text{ null hypotheses are true.} \]

Therefore, the decision for this null hypothesis is to reject if and only if

\[ \sum_{i=1}^{m} r^{(i)} \geq k, \]

where \( r^{(i)} = 1 \) if the test decision for the \( i \)-th test is to reject the null hypothesis and \( 0 \) if the null hypothesis cannot be rejected. It should be noted that the intersection-union test (Berger and Hsu 1996) and the union-intersection test (Bauer 1991) are special cases of the described test: it is the intersection-union test for \( k = m \) (all dimensions have to be successful) and the union-intersection test for \( k = 1 \) (at least one null hypothesis rejected). We will call the power to reject all null hypothesis \( (k = m) \) the conjugative power as proposed by Senn and Bretz (2007).

It is assumed that the \( m \) test statistics \( Z_{1(i)} \) and the \( m \) test statistics \( Z_{2(i)} \) both follow a multivariate normal distribution with mean values \( \mu_1 \) and \( \mu_2 \), respectively, and covariance matrix \( \Sigma = \Sigma_1 = \Sigma_2 \). The means values are given by

\[ \mu_1 = \left( \sqrt{n^{(1)}} \frac{\Delta - \delta^{(1)}}{2\sigma^{(1)}}, \ldots, \sqrt{n^{(m)}} \frac{\Delta - \delta^{(m)}}{2\sigma^{(m)}} \right), \]

\[ \mu_2 = \left( \sqrt{n^{(1)}} \frac{\Delta + \delta^{(1)}}{2\sigma^{(1)}}, \ldots, \sqrt{n^{(m)}} \frac{\Delta + \delta^{(m)}}{2\sigma^{(m)}} \right). \]

For the covariance matrix, it is assumed that the correlation between all tests is identical. Therefore, the covariance matrix is given by

\[ \Sigma = (1 - \rho) \mathbb{I} + \rho J, \]

where \( \mathbb{I} \) is the \( m \times m \) identity matrix and \( J \) is an \( m \times m \) dimensional matrix with all entries equal to 1.

### 2.3. Adjustment for Multiplicity

Here, we consider adjustments for multiplicity to control the Type I error rate in situations in which not all, but only a subset of at least \( k \) out of \( m \) hypotheses needs to be rejected. One typical example that we have already identified is the assessment of multiple endpoints.

We assume that \( H^{(1)}, \ldots, H^{(m)} \) is a family of hypotheses. The most common and easiest procedure for multiplicity adjustment that controls the familywise error rate (FWER) is the Bonferroni method (see, e.g., Wiens and Dmitrienko 2010). Each hypothesis is tested at level

\[ \alpha^* = \frac{\alpha}{m}, \quad (1) \]

where \( m \) is the number of hypotheses in the family. It is well-known that this method is very conservative (even for \( k = 1 \)) in many situations and is obviously even more conservative if \( k \geq 2 \). Using this approach will severely reduce the power of the tests and can increase the sample size to an unrealistically high value.

To overcome the problem of very conservative tests, Victor (1982) introduced, based on the work by Rüeg (1978), a concept that guarantees with a chosen certainty that not more than \( k \) hypothesis are falsely rejected. More recently, Lehmann and Romano (2005) called this idea k-FWER error rate control and defined the k-FWER as the probability to reject at least \( k \) true null hypotheses.

More formally, let \( I \subseteq \{1, \ldots, m\} \) be the set of true null hypotheses. Then, the k-FWER is given by

\[ k-\text{FWER} = P[\text{reject at least } k \text{ hypothesis } H^{(i)} \text{ with } i \in I]. \]

Lehmann and Romano (2005) proposed a simple multiplicity adjustment, which had already been proved by Hommel and Hoffmann (1988), that controls the k-FWER, that is, guarantees that

\[ k-\text{FWER} = P[\text{reject at least } k \text{ hypothesis } H^{(i)} \text{ with } i \in I] \leq \alpha, \]

with \( \alpha \) being the nominal significance level. The approach reduces the \( \alpha \)-level for the individual tests to

\[ \alpha^* = \frac{k \cdot \alpha}{m}. \quad (2) \]
We will call this approach the k-adjustment in the following. The Bonferroni adjustment is a special case of this adjustment (set k = 1). This procedure improves the approach of Quan, Bolognese, and Yuan (2001), who proposed testing at level \( \alpha/2 \) instead of level 2/3\( \alpha \) in the case where at least 2 out of 3 hypotheses need to be rejected. In Sections 3 and 4, we will give results using the Bonferroni adjustment, the k-adjustment and no adjustment to illustrate the impact of multiplicity control on the required sample size. It should be emphasized that the required sample sizes are not directly comparable because the procedures aim at controlling different error rates and therefore the Type I error rates are different. The comparisons between the different methods, however, does allow a quantification of the additional burden if a multiplicity adjustment is used. The choice of which method to use depends on the set-up of the development program.

3. The Simulation Study

The exact computation of the conjugate power for equivalence testing is difficult in situations with more than two dimensions although the distribution of the test statistic is known (Kong, Kohberger, and Koch 2004; Hsieh and Liu–Zhu 2013; Zhu 2017). This makes the exact calculation of the sample size also complicated and, therefore, the required sample size will be approximated by simulations. All simulations were performed with R version 3.2.3 (R Core Team 2015). The code for the calculation of the sample size is available as supplementary material.

3.1. Design of the Simulation Study

In this simulation study, the minimal sample size is estimated to give at least 80% power for the rejection of k out of m tests at a one-sided significance level of \( \alpha = 0.05 \) in a parallel groups design. It is assumed that the sample size, the true difference between T and R and the standard deviation of the tests is equal in each test. These values are denoted by \( n, \delta, \) and \( \sigma, \) respectively, \( \{n = n^{(1)} = \cdots = n^{(m)}\}, \delta = \delta^{(1)} = \cdots = \delta^{(m)}, \sigma = \sigma^{(1)} = \cdots = \sigma^{(m)}\). As described in Section 2.2, it is assumed that the correlation between the tests, \( \rho, \) is the same between all test statistics. It should be noted that the R code provided in the supplementary material can handle arbitrary covariance matrices and it is explained how to adjust the code in the supplementary material. Following the simulation study by Kong, Kohberger, and Koch (2004), the observations at the subject level are not generated, but the test statistics are directly simulated using \( m \) realizations of the standard normal distribution, \( e = (e^{(1)}, \ldots, e^{(m)})' \), and then transforming these to fit the specified multivariate distribution of the test statistics. More concretely, the vector of the test statistics are calculated using

\[
Z = T e + \sqrt{\frac{n}{2\sigma^2}}(\Delta - \delta) 1,
\]

where 1 is an m-dimensional vector with all entries equal to 1 and \( TT' = \Sigma \) is the Cholesky decomposition of the variance–covariance matrix of the test statistics.

The objective of the simulation study is to analyze the change in sample size dependent on several characteristics of the trial. Table 1 gives an overview of the parameter constellations used in the study. In total, 420 scenarios were simulated and the total sample size per study, \( N = 2n \), is reported. The results are based on 10,000 simulated trials per number of subjects that leads to a standard deviation of the simulated estimate of power of 0.004 for a target power of 80%.

3.2. Results

In this section, we will present the results of the simulation study. First, the results for the case in which all null hypotheses have to be rejected are given (\( k = m \)) and in the second part, the required sample size for at least k out of m hypotheses are described and the impact of the proposed multiplicity adjustments are presented.

3.2.1. Rejection of all Tests (\( k = m \))

Figure 1 shows the results for the case \( k = m \) (equivalence has to be shown for all tests, that is, it is the intersection-union test) and \( \sigma = 0.3 \). The left-hand panel is for exact equivalence (\( \delta = 0 \)) and the right-hand panel is for (slight) inequivalence (\( \delta = log(1.05) = 0.049 \)). We can see that in both panels, the higher the number of tests, the higher is also the required sample size. For example, in the case of \( \delta = log(1.05) \) and uncorrelated tests, the total sample size for one test is 76, for two tests is 102, and for five tests is 134.

We can see that in both panels high correlations reduce the sample size. For higher correlations, the required sample size for the multiple tests converges to the sample size for a single test. If the true difference between T and R increases, the required sample size also increases. Comparing both panels, we see that the shapes of the graphs look similar for both values of \( \delta \). There does not seem to be any fundamental change in behavior if the true difference between T and R changes. The results for \( \sigma = 0.15 \) are very similar (not shown).

3.2.2. Rejection of at least k out of m Tests

Figure 2 shows the required sample size for the rejection of at least \( k \leq m \) tests for a standard deviation \( \sigma = 0.3 \) and a true
difference $\delta = \log(1.05)$ (results for the other settings are comparable, and so are not shown) using the $k$-multiplicity adjustment as defined in Equation (2): if only a small proportion of tests must meet the equivalence range, the sample size is smaller than for a single endpoint. A high correlation increases the sample size in this setting. This is contrary to the results shown in Figure 1 that indicated that a higher correlation always leads to smaller required sample sizes. While this sounds contradictory at first sight, this effect can easily be explained: if we look, for example, at five highly correlated tests, the probability of one being within the equivalence margins is essentially the same as if one would only look at a single event. Therefore, there is only one chance to fulfill the requirement. If the tests are independent, the probability that one lies within the equivalence margins is higher because there a five (independent) chances that this can happen.

Figure 3 shows the impact of the different multiplicity adjustments for $m = 5$ tests. It should be noted that the sample sizes for the different adjustment methods are not directly comparable because the Type I error rate is different for the three methods. We illustrate the impact of the different multiplicity adjustments on the Type I error rate for some scenarios in the supplementary material. Nonetheless, it is interesting to analyze the change in sample size for the different adjustment methods. As expected, the sample size increases if the significance level is adjusted for multiplicity. The number of additional subjects that are required for the $k$-adjustment as defined in Equation (2) in comparison to no adjustment is moderate if a high proportion of tests ($k = 3, 4$) need to be successful, which is also the scenario that might be most relevant in practice: in the uncorrelated case, for 3 out of 5 tests, 70 subjects are required with adjustment, whereas 58 are necessary without adjustment. For 4 out of 5
tests, 90 instead of 80 subjects would be required. The difference between the Bonferroni adjustment is much bigger, for example, for 4 out of 5 tests, 130 subjects need to be enrolled.

Figure 3 also shows the behavior of the adjustments for correlations close to \( \rho = 1 \): For the Bonferroni adjustment and the unadjusted test, the functions converge to each other for correlations close to \( \rho = 1 \) independently of \( k \). For the Bonferroni adjustment, they converge to the sample size that is required for a single test with significance level \( \alpha/5 \), whereas the unadjusted tests reach the value of a single test with significance level \( \alpha \).

The results for the \( k \)-adjustment method are different: for correlations close to \( \rho = 1 \), the sample sizes are still very different for different values of \( k \): The reason for that is the adjustment in the \( \alpha \)-level: for example, in the case of a correlation of 1, the \( \alpha \)-level for \( k = 1 \) and \( m = 5 \) would be adjusted to \( \alpha^* = \alpha/5 \). The dashed line uses \( \alpha^* = \alpha \) as the significance level. Therefore, if five tests are considered, the sample size at a correlation of \( \rho = 1 \) is expected to be highest for 1 out of 5 endpoints and lowest for 5 out of 5 endpoints, which is confirmed in the simulation study. A similar observation was made by Senn and Bretz (2007) for \( k = 1 \) and multiple values of \( m \); they reported that the power function for \( k = 1 \) and \( m = 2 \) and \( m = 10 \), for example, crosses if the correlation increases, which is equivalent to the crossing of the sample size function. They concluded that in the case of \( k = 1 \), the power can be improved by adding more tests in the case of a correlation of \( \rho \leq 0.7 \). As shown here, the effect for fixed \( m \) and different values of \( k \) is less strong, nonetheless in the case of \( m = 5 \) tests and correlation larger than 0.8, the penalization of the \( \alpha \)-level adjustment leads to a higher required sample size for \( k = 1 \) than for \( k = 3 \) if the \( k \)-adjustment method is used.

4. Examples

We consider three examples of PK trials undertaken for getting approval as a biosimilar in Europe. We will use the same statistical framework as defined in Section 2. But as PK trials are considered, the study design is, in contrast to the previous sections, not a parallel groups design, but a cross-over design: every subject takes the Test and the Reference product and therefore acts as his or her own control (Jones and Kenward 2015). The analysis is based on the individual subject differences between T and R for subject \( j \) and for test \( i (i = 1, \ldots, m) \), which are defined as \( d_{j}^{(i)} = Y_{j,k}^{(i)} - Y_{j,T}^{(i)} \). The corresponding test statistics for the cross-over design are

\[
Z_1^{(i)} = \frac{\hat{d}^{(i)} + \Delta}{\sqrt{\frac{\hat{\sigma}^2}{N}}}
\]

and

\[
Z_2^{(i)} = \frac{-\hat{d}^{(i)} + \Delta}{\sqrt{\frac{\hat{\sigma}^2}{N}}},
\]

where \( \hat{d}^{(i)} \) is the mean value and \( \hat{\sigma}^2 \) is the variance of \( d_j^{(i)} \) \((j = 1, \ldots, N)\). As previously, the total sample size is denoted by \( N \) and the variance is assumed to be known.

So, it is only the formula for the variance of the estimated treatment difference that has changed compared to our exposition in terms of the parallel groups design, there are no other fundamental differences in the application of the previously described methodology if a cross-over design is used instead of a parallel groups design.

Most of the information we use is taken from the European public assessment reports (EPARs) that are published online (EMA 2017). In the EPARs, often neither the point estimates nor the standard deviations are given, only the confidence intervals (CIs) are stated. The point estimates \( \hat{\theta} \) and the standard deviations \( \hat{\sigma} \) that are required for the sample size estimation can be calculated using the stated CIs limits (CI\(_L\), CI\(_U\)) and the total sample size \( N \). In a \( 2 \times 2 \) cross-over design, it can be assumed that the \((1 - 2\alpha)\)-confidence limits were calculated using the formulas

\[
\text{CI}_L = \hat{\theta} - \frac{\hat{\sigma}}{\sqrt{N}} t_{\frac{1-\alpha}{2},N-2} \quad \text{and} \quad \text{CI}_U = \hat{\theta} + \frac{\hat{\sigma}}{\sqrt{N}} t_{\frac{1-\alpha}{2},N-2},
\]
where $t_{1−α,N−2}$ is the $1−α$ quantile of the $t$-distribution on $N−2$ degrees of freedom. For simplicity, we will also use this formula as an approximation for higher-order cross-over designs (i.e., designs with more than two periods or more than two treatments). Then, the point estimates $\hat{p}$ can be calculated by

$$\hat{p} = \frac{\text{Cl}_1 + \text{Cl}_u}{2}.$$  

The estimated standard deviation $\hat{s}$ is given by

$$\hat{s} = \sqrt{\frac{N}{t_{1−α,N−2}} (\text{Cl}_u − \hat{p}).}$$

The required total sample size $N$ for 80% power for the rejection of $k$ out of $m$ hypotheses are calculated and compared to the reported sample size in the study. It should be noted that the comparison between the sample size estimation based on observed parameters and the reported sample size is for illustration purposes only and not a method that can be applied in practice.

In the following examples, the required total sample sizes for scenarios with $k < m$ will be shown after applying the three adjustments for multiplicity that were described in Section 2.3. We will use a short-hand notation $a(b, c)$ where $a$ is the total sample size using the $k$-adjustment method as defined in Equation (2), $b$ is the total sample size required with the Bonferroni adjustment (see Equation (1)), and $c$ is the result without any adjustment. As discussed previously, no adjustment is done if $k = m$ (i.e., equivalence has to be achieved for all tests).

### 4.1. Multiple Co-Primary Endpoints and Treatment Regimens (Abasaglar)

The PK/PD-study (code 14L-MC-ABEM) that was undertaken for the approval of Abasaglar (Eli Lilly Regional Operations GmbH) is an example of a study with multiple co-primary endpoints (AUC, Cmax) and multiple treatment regimens (two different doses): 24 healthy volunteers (23 completers) received a single SC injection of 0.3 U/kg and 0.6 U/kg. The study design consisted of four sequences and four periods but no details regarding the order of treatments in the sequences were given. Equivalence should be shown for both AUC and Cmax and the two doses. Therefore, four null hypothesis need to be rejected. No details about the sample size calculation are publicly available.

Table 2 shows the outcome of the study (CHMP 2009). The CI for AUC for the higher dose does not lie completely within the acceptance range of 80–125%. The sponsor was therefore not able to show equivalence for all endpoints and claimed in the EPAR that this was due to the small sample size. It is acknowledged that this study was considered to be a supportive study only and that is why the study was most likely not powered to show equivalence for all four comparisons. Nevertheless, we would like to use this study to illustrate the impact of multiple hypotheses testing.

Considering the observed variabilities and point estimators, we find that a standard deviation of 0.5 and a ratio between the test and reference product of log(1.05) (on the log-scale) are reasonable assumptions for a sample size calculation. Since all four endpoints are measured within the same subject, the tests can be assumed to be correlated. However, the degree of correlation might not be known a priori in practice. As shown in Section 3.2.1, a conservative approach is, if all endpoints need to be rejected, to assume a correlation of 0. Then, the power to reject all four hypotheses is only 2%. For 80% power, 88 subjects would have been necessary that is nearly four times the actual sample size. If a constant correlation of $\rho = 0.5$ is assumed, this number reduces to 80 subjects. If $\rho = 0.9$ can be justified then 66 subjects would have been sufficient.

Typically, a priori, no information on the correlation structure is available and therefore a conservative approach, that assumes a correlation of 0, could be used. However, this approach gives a sample size that is very high for a PK study. Therefore, as an alternative, we consider powering the study for $k = 3$ out of 4 endpoints (with $\rho = 0$). Then, only 56 (52, 78) subjects are needed for 80% power, which is much closer to the sample size used by the sponsor and can be considered a realistic sample size in practice. We note that, as already shown in simulations in Section 3.2.2, the use of the $k$-adjustment method does not lead to an extreme increase in sample size if a high percentage of tests need to be successful.

### 4.2. Multiple PK Studies (Zarzio)

For the application of Zarzio (Sandoz), four PK studies were submitted (CHMP 2008). All studies were $2 \times 2$ cross-over trials. Table 3 shows an overview of the studies. Study EP06-103 contained two different doses. Therefore, two comparisons between T and R were studied. In total, five null hypotheses should be rejected to claim equivalence for all doses and both routes of administration. AUC and Cmax were considered as the primary PK endpoints. In this section, we will focus on Cmax only. To our knowledge, no details about the sample size calculation are publicly available.

In the EPAR (CHMP 2008) it is stated, that for the “lower doses and after multiple s.c. dose of 5 μg/kg, AUC and Cmax failed to meet the bioequivalence criteria.” The sponsor “claimed that the observed differences were due to differences in the levels of the test and reference product.” The results are given on the original scale in percent based on the completers of the study. The estimated standard deviation $\hat{s}$ was calculated using the confidence interval width and is given on the log-scale. CI represents the confidence interval limits.

**Table 3.** Overview of studies undertaken for the approval of Zarzio (Sandoz).

| Study code | Dose μg/kg | Route of administration | N  |
|------------|------------|-------------------------|----|
| EP06-101   | 10         | SC                     | 40 |
| EP06-102   | 5          | IV                     | 26 |
| EP06-103   | 2.5        | SC                     | 28 |
| EP06-105   | 5          | SC                     | 28 |
| EP06-105   | 1          | SC                     | 24 |

The given sample sizes $N$ are the planned total sample sizes per study, no comment is made in the EPAR (CHMP 2008) on the number of completed assessments.
of purity of the two products, leading to a systematic bias toward an apparently increased bioavailability for the reference product. Therefore, Cmax was adjusted to ELISA-detectable doses. Nonetheless, one out of the five comparisons failed to meet the equivalence margins (see Table 4).

The planned sample sizes seem to lead to low power for the observed variability and differences even if the studies are considered separately. For example, for the lower dose in study EP06-103, the achieved power for this study is only 41% if the observed standard deviation and difference is used. It seems likely that the observed differences and standard deviations were much larger than considered during the planning of the study.

The point estimators and the estimated variabilities vary extremely between the studies. If we assume that these were the true values and known before the study was planned then clearly an equal allocation of subjects would not have been optimal if the total sample size should be minimized because some studies would be extremely overpowered (see, for the case of two tests, the discussion in Varga, Tsang, and Singer (2017)). So presuming that the unequal variances are, in fact, the true state of nature, we allow, in this example, for unequal sample sizes. It is assumed that the test statistics are uncorrelated. The allocation that minimizes the total sample size, while achieving 80% conjunctive power for the PK study program, leads to a total of 300 subjects, with 92 subjects in study EP06-101, 6 in study EP06-102, 102 in study EP06-103 for the low dose, and 32 for the high dose and 68 subjects in study EP06-105 (see Table 5). It has to be noted that our determination of unequal allocation is for theoretical purposes only, although, if unequal variances were anticipated, it is likely that an unequal allocation would be used in practice. However, in this particular study, there seemed to be no reason a priori to expect unequal variances or treatment differences:

for example, study EP06-105 and the lower dose in study EP06-103 both used low doses and the same route of administration. Nonetheless, the post-hoc calculated sample size for the lower dose in study EP06-103 is much higher, which is unlikely to be predictable beforehand. If an equal allocation was intended (and zero correlation between test statistics is assumed), 96 subjects per study would have been required. This leads to 480 subjects in total. The increase is much larger than shown in the simulation study in Section 3.2.1. The reason is the heterogeneity between the observed differences and variabilities and the sample size is driven by the parameters of the lower dose of study EP06-103. Table 5 also gives the required sample sizes if it is assumed that the true difference is \( \delta = \log(1.05) \), but the observed standard deviations are used. It is shown that in this case, equal allocation leads to approximately the same sample size that was used in the study. The range of the sample sizes in the case of unequal allocation is also reduced.

A total sample size of 480 subjects is unrealistic in a PK setting. If the studies were powered for 4 out of 5 tests, with equal allocation to the studies and no correlation between the test statistics, only 210 (330, 200) subjects would be required and for 3 out of 5 tests, even 120 (170, 90) subjects would be sufficient. The decrease is stronger than shown in the simulation study in Section 3.2.2 and that can again be explained by the high heterogeneity between the studies.

The PK assessment for Zarzio actually consisted of several consecutive trials and so the assessment could have been stopped for futility at the end of each stage. Indeed, Bauer (1989) proposed a testing procedure that allows stopping early for success or futility if only success for \( k \) out of \( m \) studies is required. In the case of planned consecutive trials, this might be a useful strategy. For sequential trials, it does not seem to be common to adjust the sample size such that the probability for the complete or, as seen here, for a specific part (PK studies) of the clinical development program is controlled. In this example, we demonstrated the increase in sample size if the sponsor would like to have a specific certainty that not only a single study, but the complete PK study program is successful.

### 4.3. Multiple Comparison to a Control (Grastofil)

The study GCSF-SUIN-055BOI-3FA that was used for the application of Grastofil (Apotex Europe B.V.) is an example of a PK study in which multiple groups were compared with a control (Jilma et al. 2014): the test product was compared both to the

---

**Table 4. Overview of the results of the PK-studies (Cmax) conducted for the approval of Zarzio after adjustment to the ELISA-detectable doses as given in the EPAR (CHMP 2008).**

| Study code | \( \bar{\delta} \) (CI) | \( \hat{\delta} \) |
|------------|----------------|----------------|
| EP06-101   | 88.84 (82.49, 95.67) | 0.2782 |
| EP06-102   | 98.82 (95.76, 101.98) | 0.0937 |
| EP06-103-I | 86.61 (78.63, 95.40) | 0.2999 |
| EP06-103-h | 95.91 (88.73, 103.67) | 0.2414 |
| EP06-105   | 88.50 (81.55, 96.03) | 0.2333 |

The estimated standard deviation \( \hat{\delta} \) was calculated as defined in Section 4 by using the confidence intervals that were reported in the EPAR and is given on the log-scale. CI represents the confidence interval limits. Point estimators and confidence intervals are given in percent on the original scale. EP06-103-I: part of the study with low dose (2.5 \( \mu \)g/kg), EP06-103-h: part of the study with high dose (5 \( \mu \)g/kg).

---

**Table 5. Reported and post-hoc estimated total sample sizes for the PK studies for Zarzio (Sandoz).**

| Study code | \( \delta \): Observed differences | \( \delta \): \( \log(1.05) \) |
|------------|-------------------------------|-------------------------------|
|             | R (power) | E (power) | U (power) | E (power) | U (power) |
| EP06-101    | 40 (76.84) | 96 (97.92) | 92 (97.61) | 26 (93.65) | 26 (93.65) |
| EP06-102    | 26 (100)  | 96 (100)  | 6 (99.98)  | 26 (100)  | 6 (99.89)  |
| EP06-103-I  | 28 (41.46) | 96 (82.74) | 102 (84.75) | 26 (90.57) | 26 (92.58) |
| EP06-103-h  | 28 (98.88) | 96 (100)  | 32 (99.60) | 26 (97.98) | 24 (97.16) |
| EP06-105    | 24 (91.57) | 96 (99.39) | 68 (97.22) | 26 (98.39) | 20 (95.99) |
| total       | 146 (28.84) | 480 (80.52) | 300 (80.04) | 130 (81.93) | 104 (80.04) |

EP06-103-I: part of the study with low dose (2.5 \( \mu \)g/kg), EP06-103-h: part of the study with high dose (5 \( \mu \)g/kg). R (power): sample sizes in the EPAR and post-hoc power for the single study, E (power): post-hoc required sample sizes using equal allocation and post-hoc power for individual study, U (power): post-hoc required sample sizes with unequal allocation that leads to minimal required total sample size and post-hoc power. \( \delta \) gives the assumed true differences between T and R. Power is given in percent. No correlation between the test statistics is assumed in the calculation of the total power.
product approved by the Food and Drug Administration (FDA) and by the EMA. It is desirable that equivalence is shown both to the FDA and EMA approved product and for AUC and Cmax. Therefore, four null hypotheses must be rejected. In contrast to the previous examples, the correlation structure between the tests is induced by design and partially known. It is noted that the set-up in this example is similar to the one discussed by Zheng, Wang, and Zhao (2012), however, we also consider that all tests have to be successful. Zheng, Wang, and Zhao (2012) studied comparisons of several formulations. Thus, only claiming equivalence on at least one formulation was required in their example.

A three-period cross-over design with six sequences was used. The primary endpoints were AUC and Cmax. In total, 48 subjects were randomized (8 subjects per sequence). No details about the sample size calculation were given in the EPAR (CHMP 2013) or in the related publication (Jilma et al. 2014). The results of the study are shown in Table 6. All CIs lie fully within the bioequivalence margins.

For calculating the power, if all four tests are taken into account, the partially known correlation matrix can be used:

\[
\Sigma_1 := \begin{pmatrix}
1 & \rho_1 & \rho_2 \\
\rho_1 & 1 & \rho_3 \\
\rho_2 & \rho_3 & 1
\end{pmatrix}.
\]

The entries \(\rho_1, \ldots, \rho_4\) describe the correlation between AUC and Cmax and are compound specific. Using this matrix as the correlation matrix and the estimates obtained in the study, the power to claim equivalence in all four tests is 69% for \(N = 48\) subjects in the study. To increase the power to 80%, 60 subjects would have been necessary, which is an increase of 25% in the reported number of subjects. Interestingly, this is independent of the chosen correlations \(\rho_1, \ldots, \rho_4\). This is due to the much higher variability of Cmax in comparison to AUC (see Table 6). Without any prior knowledge about the structure of the covariance matrix, the most conservative choice is to use the identity matrix as the correlation matrix. Then, 64 subjects would have been necessary, showing that knowledge about the covariance matrix should be incorporated into the sample size estimation to avoid an unnecessary increase in sample size.

Using the full level \(\alpha\) is typical for situations, where separate analysis plans are used for each of the two regions concerned, that is, one for EMA and one for the FDA (Seldrup 2011). The regulatory requirement for a specific region is fulfilled if equivalence is established against the reference product in that region (regardless of what the outcome of the comparison to the reference product of the other region would be). Still, if the company wants to have high chances to get approval in both regions, the power to be successful in both comparisons should be considered. As discussed by Phillips et al. (2013), in such a situation where the primary analysis differs between regions, no multiplicity adjustments with regards to the different analysis for the different regulators/regions might be needed.

For simplification, we now focus on Cmax and assume that this was the only primary endpoint. Then, the correlation matrix is completely known and given by

\[
\Sigma_2 := \begin{pmatrix}
1 & 1/2 & 1/2 & 1 \\
1/2 & 1 & 1 & 1/2 \\
1/2 & 1 & 1 & 1/2 \\
1 & 1/2 & 1/2 & 1
\end{pmatrix}.
\]

For Cmax only and with this correlation matrix, again \(N = 60\) subjects would have been required confirming the previous statement that the sample size for jointly testing AUC and Cmax is predominantly driven by the higher variability of Cmax. Theoretically, if a company was happy with the approval in any of the two regions, it could determine the sample size using the power to claim equivalence on one out of two tests for Cmax, but not adjusting the significance level for the reasons discussed above and assuming the known correlation structure. Then the sample size would be decreased to 32 subjects. With the Bonferroni or the \(k\)-adjustment, 42 subjects would have been necessary.

If the rejection of at least one of the two tests for Cmax is sufficient, also the Dunnett test could be considered. This test is an adjustment for multiple testing that is specifically designed for many-one comparisons and uses the known correlation structure and is therefore less conservative than the adjustment described in Section 2.3 (Dunnett 1955). Using this test, 40 subjects would have been necessary. The difference between the adjusted sample size calculation and the Dunnett test is rather small showing that, in this case, the used adjustments are not very conservative.

5. Conclusion

Correctly determining the required sample size is an important step in all clinical trials. In trials that are conducted for showing biosimilarity of a new product to an already approved biological medicine, often multiple endpoints, multiple doses or routes of administration, or multiple studies in different patient populations are considered. It is desirable that equivalence can be claimed for all tests. However, if more than one statistical hypothesis test is carried out, the study or drug development program needs to be powered for multiple testing to avoid the study failing because of lack of power.

By simulation and for a fixed power of 80% to reject all hypotheses, we have shown that the increase in sample size due to testing multiple hypotheses can be extreme, especially if no correlation between the test statistics can be assumed and many tests are carried out. EMA has already approved applications with PK and PD studies where not all primary endpoints met the equivalence criterion. Therefore, one possible solution, if the required sample size is not feasible, is to power the study so that at least \(k\) out of \(m\) tests have to meet the equivalence criterion. In this article, we used the adjustment for multiplicity that was first proposed by Hommel and Hoffmann (1988) and compared this approach to the Bonferroni adjustment and to the
unadjusted test. The properties of this approach were assessed in a simulation study. Interestingly, in contrast to the setting in which all tests have to be successful, a high correlation does not necessarily decrease the required sample size in all cases: if only a low proportion of tests needs to be successful (e.g., 2 out of 5), having a high correlation might lead to a higher number of subjects, whereas for a high proportion of successful tests (e.g., 4 out of 5), it decreases the number of subjects. This makes the sample size calculation more complex because the worst case scenario is not identical for all scenarios and might therefore not be the simple case of uncorrelated test statistics as one might expect. Extensive simulation studies using different assumptions on the correlation structure are therefore necessary to correctly determine the sample size for the \( k \) out of \( m \) scenarios. Although the different multiplicity adjustments are not directly comparable because the Type I error rate is not the same for the methods, we quantified the change in sample size for the different adjustment methods to illustrate the price of a multiplicity adjustment. Interestingly, the adjustment that was proposed by Hommel and Hoffmann (1988) leads only to a small increase in sample size, especially for scenarios in which a high proportion of endpoints has to be successful (e.g., 3 out of 5, 4 out of 5).

In practice, in the case in which not all hypotheses could be rejected, the sponsors claimed equivalence on a subset of the endpoints. It should be emphasized that without any adjustment this approach leads to an inflation of the Type I error rate whereas the proposed tests control \( k \text{-FWER} \) for the \( k \) out of \( m \) test if the adjustment by Hommel and Hoffmann (1988) or the Bonferroni adjustment is used. For correlated test statistics efficiency could be further improved by incorporating the correlation structure also into the multiplicity adjustment (e.g., Bretz et al. 2011; Xie 2012; Bullen and Obuchowski 2017).

The handling of multiple hypotheses testing in practice and the impact on power and sample size was illustrated using three PK studies that were submitted to EMA for approval as a biosimilar (Abasaglar, Zarzio, Grastofil). The main resources were EPARs. The examples showed that if multiple hypothesis testing is taken into account, the required sample size might be unrealistically high: for example, for a study for the approval of Abasaglar (Eli Lilly), 88 subjects instead of 24 subjects would be necessary if no correlation between the endpoints can be assumed. For the application for Zarzio in which five separate PK studies in different doses and routes of administration were performed, even 480 instead of the originally planned 146 subjects would be necessary. In the third example, the application of Grastofil (Apotex Europe B.V.), the test product was compared to the FDA approved product and to the EMA approved product using two endpoints which would require 60 instead of 48 subjects for 80% power to claim equivalence to the two reference products even if the partially known correlation structure between the endpoints is used. However, if it would be considered acceptable to fail on one out of these studies, the sample size reduces to 56 subjects for Abasaglar, to 210 for Zarzio and to 38 for Grastofil if the \( k \text{-FWER} \) is controlled with the adjustment proposed by Hommel and Hoffmann (1988), and these are reasonable sample sizes in practice.

All the examples showed that the increase in sample size due to multiple testing is often not considered in practice. The reported sample size is much lower than required for the multiple comparisons if a reasonable value of power should be achieved. This can lead to severely underpowered studies and to the failure of development programs. However, it is also acknowledged that in practice, this issue does not seem to be reported as a major problem. Nonetheless, we recommend that multiple testing be accounted for when calculating sample sizes for biosimilar studies. If the sample size is too high, because success is required on all tests, then a practical compromise is to control success only on a subset of tests. We have given examples where this approach can achieve reasonable power with practical sample sizes.

It should be emphasized that the concept of testing for at least \( k \) out of \( m \) endpoints fits well with the concept of “totality of the evidence” that is especially advocated by the FDA (FDA 2015). In their guideline, it is explained that if differences are found between the biosimilar and the reference product, these differences might be acceptable if the total application is convincing. Therefore, it is conceivable that a sponsor, for example, plans to show equivalence on only \( k \) out of \( m \) PK endpoints, but adds some additional analytical studies to compensate for that.

Supplementary Material

Type I error rate: Illustration of the impact of the three adjustment methods on the Type I error rate for selected scenarios. (pdf)

R code: All major functions that were used for the calculations in this article are provided. Examples show how a sample size calculation can be performed. (pdf)

Acknowledgments

The authors gratefully acknowledge the funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 633567 and from the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract number 999754557. The opinions expressed and arguments employed herein do not necessarily reflect the official views of the Swiss Government.

Funding

H2020 Marie Skłodowska-Curie Actions [633567]; Swiss State Secretariat for Education, Research and Innovation (SERI) [999754557].

ORCID

Bernd Jilma http://orcid.org/0000-0001-5652-7977
Franz König http://orcid.org/0000-0002-6893-3304

References

Bauer, P. (1989), “Sequential Tests of Hypotheses in Consecutive Trials,” Biometrical Journal, 31, 663–676. [46]
Bauer, P. (1991), “Multiple Testing in Clinical Trials,” Statistics in Medicine, 10, 871–890. [41]
Berger, R. L., and Hsu, J. C. (1996), “Bioequivalence Trials, Intersection-Union Tests and Equivalence Confidence Sets,” Statistical Science, 11, 283–319. [41]
Blevins, T., Dahl, D., Rosenstock, J., Ilag, L., Huster, W., Zielonka, J., Pollom, R., and Prince, M. (2015), “Efficacy and Safety of LY2963016 Insulin Glargine Compared With Insulin Glargine (Lantus®) in Patients With

Illustration of the impact of the three adjust-
ment methods on the Type I error rate for selected scenarios. (pdf)

R code: All major functions that were used for the calculations in this article are provided. Examples show how a sample size calculation can be performed. (pdf)
Type 1 Diabetes in a Randomized Controlled Trial: The ELEMENT 1 Study," *Diabetes, Obesity and Metabolism*, 17, 726–733. [39]

Bretz, F., Posch, M., Klimm, E., Klinglmueller, F., Maurer, W., and Kolmeyer, K. (2011), "Graphical Approaches for Multiple Comparison Procedures Using Weighted Bonferroni, Simes, or Parametric Tests," *Biometrical Journal*, 53, 894–913. [48]

Bulewicz, J. A., and Obuchowski, N. A. (2017), "Correcting for Multiple Testing During Diagnostic Accuracy Studies," *Statistics in Biopharmaceutical Research*, 9, 243–248. [48]

Burдick, R. K., Thomas, N., and Cheng, A. (2017), "Statistical Considerations in Demonstrating CMC Analytical Similarity for a Biosimilar Product," *Statistics in Biopharmaceutical Research*. [40]

CHMP (2008), "Zarzio: EPAR - Public Assessment Report," available at [http://www.ema.europa.eu/docs/en_GB/document_library/EPAR_-_Public_assessment_report/human/000917/WC500046528.pdf] [40,45]

— (2009), "Abasria: EPAR - Public Assessment Report," available at [http://www.ema.europa.eu/docs/en_GB/document_library/EPAR_-_Public_assessment_report/human/002835/WC500175383.pdf] [45]

— (2013), "Grastofil: EPAR - Public Assessment Report," available at [http://www.ema.europa.eu/docs/en_GB/document_library/EPAR_-_Public_assessment_report/human/002150/WC500154066.pdf] [47]

— (2014), "Guideline on Similar Biological Medicinal Products Containing Biotechnology-Derived Proteins as Active Substance: Non- Clinical and Clinical Issues," available at [http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2015/01/WC500180219.pdf] [39]

— (2017), "Draft Reflection Paper on Statistical Methodology for the Comparative Assessment of Quality Attributes in Drug Development," available at [http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2017/03/WC500224995.pdf] [40]

DerSimonian, R., and Laird, N. (1986), "Meta-Analysis in Clinical Trials," *Controlled Clinical Trials*, 7, 177–188. [39]

Dunnett, C. W. (1955), "A Multiple Comparison Procedure for Comparing Several Treatments with a Control," *Journal of the American Statistical Association*, 50, 1096–1121. [47]

EMA (2017), "EPAR database," available at [http://www.ema.europa.eu/docs/en_GB/document_library/EPAR_search.jsp?mid=WCO01ac058001t124] [44]

FDA (2015), "Scientific Considerations in Demonstrating Biosimilarity to a Reference Product," available at [http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM291128.pdf] [39,48]

Generics and Biosimilars Initiative (2015), "Biological Patents Expiries," available at [http://www.gabionline.net/Biosimilars/General/Biologicals-patent-expiries] [39]

Holzmann, J., Balser, S., and Windisch, J. (2016), "Totality of the Evidence at Work: The First U.S. Biosimilar," *Expert Opinion on Biological Therapy*, 16, 137–142. [40]

Hommel, G., and Hoffmann, T. (1988), "Controlled Uncertainty," in *Multiple Hypotheses Testing/ Multiple Hypotheses Testing*, New York: Springer, pp. 154–161. [40,41,47,48]

Hsieh, C.-Y., and Liu, J.-P. (2013), "An Approximate Approach to Sampling Size Determination for the Equivalence Hypothesis," *Journal of Biopharmaceutical Statistics*, 23, 526–538. [42]

Jilma, B., Jagiello-Gruszfeld, A., Tomczak, P., Gadgil, H., Orlik, G., Desai, K., Catalano, T., and Dowd, J. (2014), "Demonstration of Clinical Comparability of the Biosimilar Filgrastim to Neupogen, in Terms of Safety and Efficacy, in Healthy Volunteers and Patients Receiving Myelosuppressive Chemotherapy," *European Oncology & Haematology*, 10, 107–115. [46,47]

Jones, B., and Kenward, M. G. (2015), *Design and Analysis of Cross-Over Trials* (3rd ed.), Boca Raton, FL: Chapman and Hall/CRC Press. [44]

Kong, L., Kohberger, R. C., and Koch, G. G. (2004), "Type I Error and Power in Noninferiority/Equivalence Trials With Correlated Multiple Endpoints: An Example From Vaccine Development," *Journal of Biopharmaceutical Statistics*, 14, 893–907. [41,42]