

UC Santa Cruz

UC Santa Cruz Previously Published Works

Title

Equivalence testing for multiple groups

Permalink

<https://escholarship.org/uc/item/1sr9h2v3>

Journal

Stat, 13(1)

ISSN

2049-1573

Authors

Pourmohamad, Tony

Lee, Herbert KH

Publication Date

2024

DOI

10.1002/sta4.645

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at <https://creativecommons.org/licenses/by/4.0/>

Peer reviewed

ORIGINAL ARTICLE

Equivalence Testing for Multiple Groups

Tony Pourmohamad¹ | Herbert K. H. Lee²¹Department of Data and Statistical Sciences, Genentech, Inc., California, United States²Department of Statistics, University of California, Santa Cruz, California, United States

Correspondence

Tony Pourmohamad, 1 DNA Way, South San Francisco, CA, 94080.

Email: pourmoht@gene.com

Summary

Testing for equivalence, rather than testing for a difference, is an important component of some scientific studies. While the focus of the existing literature is on comparing two groups for equivalence, real world applications arise regularly that require testing across more than two groups. This paper reviews the existing approaches for testing across multiple groups and proposes a novel framework for multigroup equivalence testing under a Bayesian paradigm. This approach allows for a more scientifically meaningful definition of the equivalence margin and a more powerful test than the few existing alternatives. This approach also allows a new definition of equivalence based on future differences.

KEYWORDS:

Bayesian, equivalence margin, TOST, Wellek-Welch test

1 | INTRODUCTION

Increasingly, the goal of many scientific studies is to demonstrate comparability between multiple groups. For example, in a preclinical toxicology study it may be of interest to show that there are no differences in toxicities associated with varying dosage levels of a pharmaceutical drug (Parasuraman 2011). Another example is in manufacturing where a manufacturer may wish to show that multiple devices (or multiple factories) produce products of an equivalent quality (Richter & Richter 2002). Determining comparability (i.e., equivalency) between groups is of interest, yet most statistical hypothesis testing procedures are not designed to handle this scenario. Rather, most statistical hypothesis testing procedures are designed with an emphasis on establishing differences between groups as the typical hypothesis testing strategy seeks to reject a null hypothesis of equivalency. Unfortunately, this hypothesis testing paradigm often leads scientists and researchers to erroneously conclude that non-rejection of the null hypothesis is the same as establishing equivalency amongst groups.

Instead, equivalence testing should be the primary method for testing hypotheses of equivalence. In contrast to traditional hypothesis testing procedures, equivalence tests reverse the role of the null and alternative hypotheses such that the null hypothesis assumes that there is a meaningful effect between groups (i.e., groups are not equivalent) versus the alternative hypothesis that there is no meaningful effect between groups (i.e., groups are equivalent). Here, a prespecified lower and upper equivalence margin are used to define what a meaningful effect is. Any effect that is outside of the equivalence margins would be deemed meaningful, while any effect within the lower and upper equivalence margin would not be considered meaningful thus establishing equivalency of the groups.

In perhaps the simplest case, Schuirmann (1987) developed the two one-sided tests (TOST) procedure for establishing equivalence amongst two groups via an intersection-union test (Berger 1982; Berger & Hsu 1996). A highly popular procedure amongst researchers and scientists, the TOST uses a set of composite null hypotheses to establish equivalency amongst two groups, e.g., if we can reject the following two one-sided null hypotheses: $H_{01} : \Delta \leq -\theta_L$ and $H_{02} : \Delta \geq \theta_U$ (where $-\theta_L$ and θ_U are the lower and upper equivalence margins, respectively), then we can conclude that $-\theta_L < \Delta < \theta_U$ and thus equivalency has been established. Now, equivalence testing has come a long way since the inception of the TOST procedure (see Wellek (2010) and the references therein for a comprehensive list), yet, most of the methodological advancements have been with regard to equivalence tests for two groups, or for equivalence testing where there are multiple outcomes of interest still amongst two groups (for example, Ghosh & Gönen 2008; Pallmann & Jaki 2016; Quan, Bolognese, & Yuan 2001). With the former in mind, this article contributes significantly to the development of a general framework for multigroup equivalence testing. Under a Bayesian paradigm, we propose a novel

framework for multigroup equivalence testing that will address the limitations of existing methods, and will be shown to provide a more flexible and interpretable framework for determining comparability between multiple groups. Moreover, our multigroup equivalence testing strategy will allow us to create a new option for defining equivalence, in terms of future observations, which is practical to implement because of the Bayesian framework.

Currently, there are few available resources in the literature dedicated to equivalence testing for multiple groups (for example, see Chapter 7 of Wellek 2010). As detailed in Wellek (2010) and Koh and Cribbie (2013), the Wellek-Welch test is one of the only methods currently in the literature specifically developed for multigroup equivalence testing. However, as we will show in this paper, the Wellek-Welch test suffers from a lack of interpretability of its associated equivalence margin, an issue that will not be inherent to our novel multigroup equivalence testing framework. We demonstrate how the lack of interpretability can decrease the power of the Wellek-Welch test. And although there are not more methodologies directly aimed at solving multigroup equivalence testing problems, there are still other approaches that can be used to address the equivalence testing issue. For example, it is often common practice to consider using a hypothesis testing strategy for demonstrating the equivalence of two groups, say for example the TOST procedure, multiple times in order to test pairwise comparisons of interest, and then to use some sort of correction procedure for multiple testing (Lauzon & Caffo 2009; Miller 1981). Issues associated with multiple testing procedures are well known (for example, inflation of type I error rates or loss of statistical power), yet, under our proposed Bayesian multigroup testing strategy this will not be an issue as only a single hypothesis test will be performed. In fact, it will be shown that our Bayesian multigroup equivalence testing strategy can obtain higher statistical power as compared to correcting for multiple hypothesis tests using a traditional TOST procedure (Section 3).

Lastly, there have been several recent procedures developed for equivalence testing under a Bayesian paradigm (Hoyda, Counsell, & Cribbie 2019; Kruschke 2018; van Ravenzwaaij, Monden, Tendeiro, & Ioannidis 2019), yet all of these approaches have focused on the scenario of testing two groups. These Bayesian approaches can be roughly broken into two categories: 1) ones that rely on traditional equivalence testing concepts like the TOST or the equivalence margin (Hoyda et al. 2019; Kruschke 2018), and 2) those that use Bayes factors (Hoyda et al. 2019; Morey & Rouder 2011; van Ravenzwaaij et al. 2019). In the case of the latter category, one could envision using Bayes factors for ANOVA models (Rouder, Morey, Speckman, & Province 2012) (or similarly some sort of constrained statistical models) for testing equivalence of multiple groups, however, this is not our preferred extension to the multigroup testing case. Rather, we view Bayesian methods that rely on traditional equivalence testing concepts to be the superior choice for moving to the multigroup case for several reasons. First, we believe that equivalence should be judged based on the equivalence margin. The equivalence margin is an easy to interpret value, often on the same scale as the data, and is rooted by construction in the scientific problem. The equivalence margin is specified by the scientists or researchers and defines what a practically meaningful difference is between groups, which we feel is more appropriate than an arbitrary cut-off for a Bayes factor (or a p-value). Secondly, Bayes factors tend not to be the natural scale at which researchers and scientists think about making decisions, and furthermore, setting an appropriate decision value based on the Bayes factors can be a tough challenge. And lastly, in the case of equivalence testing for two groups, it has been shown that non-Bayes factor approaches possess favorable qualities over Bayes factor approaches (Liao, Midya, & Berg 2019), for example, Bayes factors are heavily dependent on prior specification while non-Bayes factor approaches tend to be far more robust. It is with these reasons in mind that we will construct our Bayesian multigroup equivalence testing procedure by using and expanding on traditional equivalence testing concepts.

The remainder of this article is organized as follows. In Section 2, we describe the existing Wellek-Welch test and introduce our proposed new approach. Section 3 provides simulated and real examples to demonstrate differences between approaches. Lastly, Section 4 finishes with some discussion about future directions of research and concluding remarks.

2 | EQUIVALENCE TESTING

Let us assume that there are k independent groups with corresponding continuous measurements $x_{i1}, \dots, x_{in_i} \sim N(\mu_i, \sigma_i^2)$ where n_i is the sample size of the i th group, $i \in \{1, \dots, k\}$. Of interest is establishing equivalency for c pairwise comparisons of interest amongst the k groups. For example, there may be $k = 4$ insulin devices being tested where one device is already available on the market and serves as a "gold standard" to test the other three devices against for equivalency (i.e., $c = 3$ comparisons), or perhaps all $c = k!/(2(k-2)!)$ pairwise comparisons may be of interest. In general, we can represent this multigroup equivalency testing question with the following null and alternative hypotheses:

$$H_0 : \Delta \notin (-\theta_L, \theta_U) \quad \text{versus} \quad H_A : \Delta \in (-\theta_L, \theta_U), \quad (1)$$

where $\theta_L, \theta_U \in \mathbb{R}^c$ are vectors containing the lower and upper equivalence margins, respectively, for each of the c pairwise comparisons of interest. Here, Δ is a parameter vector of interest that defines the level at which equivalency should be judged. For example, Δ could correspond to the average difference between groups, or the difference (or ratio) of standard deviations between groups, or any other metric of interest. When H_0 is rejected, we can declare equivalency amongst all the pairwise comparisons of interest, and similarly, when H_0 is accepted we can declare that at least one of the pairwise comparisons are not equivalent. Our strategy for rejecting or accepting H_0 will be discussed in Section 2.2 once our Bayesian statistical model has been described.

Irrespective of statistical strategy, the determination of the equivalence margin is the most critical step in equivalence testing. It is a domain-specific quantity that is used to mathematically formalize the hypothesis of comparability (e.g., that two groups are equivalent if the absolute difference in population means is less than the equivalence margin). As such, the equivalence margin is a prespecified quantity (chosen before the collection of data) and should be determined by subject matter experts (Lakens, Scheel, & Isager 2018). The value of a comparability study depends on how well the equivalence margin can be justified in terms of previous knowledge (historical data or subject matter expertise), practical considerations (cost of, say, data collection), and knowledge of the scale of measurement (inherent method variability).

2.1 | Wellek-Welch Test

As one of the only equivalence testing strategies in the literature designed specifically for multigroup equivalence testing, the Wellek-Welch test (Koh & Cribbie 2013) is an extension of the original Wellek test (Wellek 2010) to the case of heteroscedasticity, and is an omnibus test for determining overall equivalence amongst all of the k groups simultaneously. The statistical test is carried out by first calculating the following quantity:

$$F' = \left(\frac{\sum_{i=1}^k w_k (\bar{x}_k - \bar{x}_\cdot)^2}{k-1} \right) / \left[1 + \frac{2(k-2)}{k^2-1} \sum_{i=1}^k \left(\frac{1}{n_k-1} \right) \left(1 - \frac{w_k}{\sum_{i=1}^k w_k} \right)^2 \right], \quad (2)$$

where $w_k = n_k/s_k^2$, $\bar{x}_\cdot = \sum_{i=1}^k w_k \bar{x}_k / \sum_{i=1}^k w_k$, and $\bar{x}_k$ and s_k^2 are the sample mean and sample variance of the k^{th} group, respectively. From (2), the following test statistic can be defined as

$$\hat{\phi}^{2'} = \left(\frac{k-1}{\bar{n}} \right) F', \quad (3)$$

where $\bar{n}$ is the average sample size across all groups. F' is approximately distributed as a non-central F distribution with $k-1$ degrees of freedom for the numerator, and df' degrees of freedom for the denominator where

$$df' = (k^2-1) / \left[3 \sum_{i=1}^k \left(\frac{1}{n_k-1} \right) \left(1 - \frac{w_k}{\sum_{i=1}^k w_k} \right)^2 \right]. \quad (4)$$

For the Wellek-Welch test, testing for multigroup equivalency can be carried out using the following null and alternative hypotheses:

$$H_0 : \phi^{2'} \geq \varepsilon^2 \quad \text{versus} \quad H_A : \phi^{2'} < \varepsilon^2. \quad (5)$$

Here, $\varepsilon > 0$ is considered to be the equivalence margin associated with the statistical test, and (much like a traditional F statistic) ε^2 can be thought of as a ratio of between-group to within-group variability. Equivalence of all of the k groups is established when the null hypothesis H_0 is rejected, and the null hypothesis is rejected at its corresponding α -level if $\hat{\phi}^{2'} < \phi'_{\text{crit}}$, where

$$\phi'_{\text{crit}} = \left(\frac{k-1}{\bar{n}} \right) F'_{k-1, df'; \alpha}(\bar{n}\varepsilon^2). \quad (6)$$

The Wellek-Welch test is an attractive solution to the multigroup equivalence testing problem given the ease of its straightforward calculations, and the fact that it does not require multiple hypothesis testing. However, there are several concerns that we have with the Wellek-Welch test. Our primary concern is with the interpretation and setting of the associated equivalence margin ε^2 . Although ε^2 can be thought of as a ratio of between-group to within-group variability, practically speaking, this scale is ambiguous and not at the scale in which scientists and researchers think about defining equivalency between groups. Wellek (2010) recommends setting $\varepsilon = 0.36/\sqrt{2} \approx 0.25$ or $\varepsilon = 0.74/\sqrt{2} \approx 0.5$ depending on whether the equivalence criterion is intended to be rather strict (i.e., conservative) or comparatively weak (i.e., liberal), however, we view this as a poor recommendation in practice in that the equivalence margin should be determined by subject matter experts rather than a binary rule of thumb prescribed by statisticians. Moreover, the scale at which to set the equivalence margin should be rooted in the science of the problem, and thus be directly interpretable and understandable by scientists in the field. For example, if there were three groups of scales used to measure weights, then it makes sense to be able to say something like the averages of any two groups should not differ by more than, say, ± 5 lbs to be equivalent, but this type of simple interpretation is lost to scientists in the use of ε^2 . Of lesser concern, but equally problematic, the Wellek-Welch test evaluates equivalency amongst all of the groups simultaneously given that it is an omnibus test for equivalence. However, it may not be the intent of the scientist or researcher to test equivalency of all the $c = k!/(2(k-2)!)$ pairwise comparisons, but perhaps only some of the c pairwise comparisons. For example, if there are three different drugs A , B , and C , then it might be of interest to only show that A and B are equivalent, while also showing that A and C are equivalent, but the comparison of B and C may not be of interest. In this example, the Wellek-Welch test would test all 3 pairwise comparisons and may find that the groups are not equivalent merely because B and C are not equivalent, even though A and B , and A and C , may be equivalent, which answers the original question of interest. However, this is not the fault of the Wellek-Welch test as it is not particularly suited for this situation. Rather, when all pairwise comparisons are not of interest, one could use a multiple hypothesis testing procedure, but this typically comes with a loss of statistical power for fixed n_k as the number of comparisons increases. This will not be an issue

for our Bayesian approach as it will be able to test any subset of all pairwise comparisons within a single hypothesis test. And so, rather than use the Wellek-Welch test or a multiple hypothesis testing procedure, we propose using a Bayesian approach to equivalence testing that would allow for simultaneous equivalence testing of any c pairwise comparisons of interest for the k multiple groups without any of the aforementioned issues that were raised. The details of this Bayesian multigroup equivalence testing procedure are described in what follows in the next section.

2.2 | A Bayesian Approach to Equivalence Testing

This section introduces our Bayesian model for the data, and describes three different potential approaches to testing equivalency amongst multiple groups.

2.2.1 | The Model Setup

For all $i = 1, \dots, k$ groups, let us assume a reasonably non-informative prior of the following form:

$$p(\mu_i, \sigma_i^2) \propto (\sigma_i^2)^{-1}. \quad (7)$$

The choice of prior in (7) leads to a well-known closed-form joint posterior distribution for μ_i and σ_i^2 (for example, see Gelman et al. (2013), p. 64), i.e.,

$$p(\mu_i, \sigma_i^2 | x_{i1}, \dots, x_{in_i}) = \sigma_i^{-n-2} \exp \left\{ -\frac{1}{2\sigma_i^2} \times \left(\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 + n_i(\bar{x}_i - \mu_i)^2 \right) \right\}. \quad (8)$$

Furthermore, as applicable to all $i = 1, \dots, k$ groups, the conditional posterior distribution of μ_i is

$$p(\mu_i | \sigma_i^2, x_{i1}, \dots, x_{in_i}) \propto \exp \left\{ -\frac{n_i}{2\sigma_i^2} (\bar{x}_i - \mu_i)^2 \right\}, \quad (9)$$

and the marginal posterior distribution of σ_i^2 is

$$\begin{aligned} p(\sigma_i^2 | x_{i1}, \dots, x_{in_i}) &= \int_{-\infty}^{\infty} p(\mu_i, \sigma_i^2 | x_{i1}, \dots, x_{in_i}) d\mu_i \\ &= \int_{-\infty}^{\infty} \exp \left\{ -\left[n_i(\bar{x}_i - \mu_i)^2 + \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 \right] / 2\sigma_i^2 \right\} \frac{d\mu_i}{\sigma_i^{n_i+2}} \\ &\propto \exp \left\{ -\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 / 2\sigma_i^2 \right\} \sigma_i^{-n_i-1}. \end{aligned} \quad (10)$$

Both (9) and (10) can be recognized as the unnormalized forms of the normal distribution and the inverse-gamma distribution, respectively, i.e.,

$$\mu_i | \sigma_i^2, x_{i1}, \dots, x_{in_i} \sim N \left(\bar{x}_i, \frac{\sigma_i^2}{n_i} \right) \quad (11)$$

and

$$\sigma_i^2 | x_{i1}, \dots, x_{in_i} \sim \text{Inverse-Gamma} \left(n_i/2 - 1/2, \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 / 2 \right). \quad (12)$$

Note that a non-informative prior need not be chosen for the equivalence testing procedure in the following sections to work, and that a more informative prior may be placed on either the mean or variance (or both) if prior information is available. However, the closed form equations and distributions in the following sections may no longer hold if any general form of prior is used, instead, Markov chain Monte Carlo methods will be necessary to sample from the distributions of interest in order to carry out the equivalence testing procedures. In the special case of a normal inverse-gamma prior (i.e., the conjugate prior) all closed form distributions will be obtainable but will contain updated posterior parameters in relationship to the ones derived in the following sections. Furthermore, we choose to use a non-informative prior when developing the multigroup equivalence testing procedure owing to the fact that most criticisms of Bayesian methods stem from the subjectivity induced by the choice of prior, and we also hope that more regulated industries (e.g., the pharmaceutical industry since equivalence testing is highly prevalent there) might adopt a Bayesian approach given that concerns over prior influence could be easily mitigated.

2.2.2 | Equivalence Based on Mean Differences

There is no universally accepted definition for what it means for two or more groups to be equivalent. Much like the equivalence margin, the definition of the metric of interest for showing equivalence is debatable and should be rooted in the field of application and based on subject matter

expertise. Here we develop a multigroup equivalence testing procedure based on perhaps the most widely used definition of equivalence which is equivalence based on mean differences between groups. Now, let us define Δ to be a $c \times 1$ matrix of c pairwise group comparisons. Without loss of generality, and for clarity sake, let us consider the case of $k = 3$ independent groups where it is of interest to consider all pairwise comparisons of the groups. Then, we can write Δ as follows:

$$\Delta := \begin{bmatrix} \delta_{12} \\ \delta_{13} \\ \delta_{23} \end{bmatrix} = \begin{bmatrix} \mu_1 - \mu_2 \\ \mu_1 - \mu_3 \\ \mu_2 - \mu_3 \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} =: C\mu. \quad (13)$$

Here we have that the conditional posterior distribution of Δ is

$$\Delta | \sigma_1^2, \sigma_2^2, \sigma_3^2, x_{11}, \dots, x_{1n_1}, x_{21}, \dots, x_{2n_2}, x_{31}, \dots, x_{3n_3} \sim N_3 \left(C\mathbb{E}(\mu), C\mathbb{Cov}(\mu)C^T \right) \quad (14)$$

where

$$\mathbb{E}(\mu) = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \bar{x}_3 \end{bmatrix} \quad \text{and} \quad \mathbb{Cov}(\mu) = \begin{bmatrix} \sigma_1^2/n_1 & 0 & 0 \\ 0 & \sigma_2^2/n_2 & 0 \\ 0 & 0 & \sigma_3^2/n_3 \end{bmatrix}. \quad (15)$$

In general, for any number of k independent groups and c pairwise comparisons, we have that

$$\Delta | \sigma_1^2, \dots, \sigma_k^2, x_{11}, \dots, x_{1n_1}, \dots, x_{k1}, \dots, x_{kn_k} \sim N_c \left(C\mathbb{E}(\mu), C\mathbb{Cov}(\mu)C^T \right), \quad (16)$$

with

$$\mathbb{E}(\mu) = \begin{bmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_k \end{bmatrix} \quad \text{and} \quad \mathbb{Cov}(\mu) = \begin{bmatrix} \sigma_1^2/n_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma_k^2/n_k \end{bmatrix}, \quad (17)$$

where $\mathbb{Cov}(\mu)$ is a diagonal covariance matrix.

Returning to the general multigroup equivalence testing hypotheses in (1), we would like to test the following null and alternative hypothesis

$$H_0 : \Delta \notin (-\theta_L, \theta_U) \quad \text{versus} \quad H_A : \Delta \in (-\theta_L, \theta_U), \quad (18)$$

where θ_L and θ_U are vectors of length c containing the lower and upper equivalence margins, respectively, for the average differences between groups. For example, $-\theta_L = (-\theta_{12L}, -\theta_{13L}, -\theta_{23L})$ would contain the lower equivalence margins for the comparisons of groups 1 and 2, groups 1 and 3, and groups 2 and 3, and similarly defined for the upper equivalence margins. Now, one way to test the hypotheses in (18) would be to calculate all $100(1 - \alpha)\%$ credible (or highest posterior density) intervals associated with the c pairwise differences δ_{ij} , and to reject H_0 (i.e., conclude equivalency of all of the groups) if the $100(1 - \alpha)\%$ credible intervals for all of the δ_{ij} 's are entirely contained within their corresponding equivalence margin $(-\theta_{ijL}, \theta_{ijU})$. Here α would be prespecified and would be chosen to reflect the desired level of probabilistic certainty for the equivalency conclusion to be true. An alternative approach would be to define the probability q such that

$$q = \Pr(\Delta \in (-\theta_L, \theta_U)) \quad (19)$$

which is equivalent to calculating

$$q = \Pr(-\theta_{ijL} < \delta_{ij} < \theta_{ijU}) \quad (20)$$

for all c pairwise comparisons of interests jointly. For a prespecified value of q_0 , one could accept H_A if $q \geq q_0$. Here, we note that calculation of the posterior probability q proceeds very easily by calculation the proportion of posterior draws of μ_1, μ_2 and μ_3 , that satisfy the inequalities in (20). Of the two proposed strategies for testing multigroup equivalency, we prefer the latter strategy as we favor the simple probabilistic framework and interpretation that comes with the definition of q .

As an interesting aside, even though the purpose of the current paper is for establishing a strategy for multigroup testing, in the case of $k = 2$ groups, all of the aforementioned methodology can still be used, and our testing strategy reduces to the well-known region of practical equivalence (ROPE) procedure of Kruschke (2018) when a $100(1 - \alpha)\%$ highest posterior density interval is used.

2.2.3 | Equivalence Based on Standard Deviations

An advantage of taking a Bayesian approach to equivalence testing is that the equivalence margin can be defined as fairly complicated metrics or decision rules that might not be so accessible to evaluate otherwise. Often of interest is to also incorporate some sort of comparison amongst the

standard deviations (or variances) of the different groups when declaring equivalency. For example, it might be of interest to test, for all c pairwise comparisons of interests, the following null and alternative hypothesis

$$H_0 : \frac{\sigma_i}{\sigma_j} \leq \gamma_{ijL} \quad \text{or} \quad \frac{\sigma_i}{\sigma_j} \geq \gamma_{ijU} \quad \text{versus} \quad H_A : \gamma_{ijL} < \frac{\sigma_i}{\sigma_j} < \gamma_{ijU}. \quad (21)$$

Here, γ_{ijL} and γ_{ijU} are the lower and upper equivalence margins for the ratio of the standard deviation of group i compared to group j . If, for example, we want to declare that the groups are all equivalent if no group has a standard deviation two times larger than any other, then one could set $\gamma_{ijL} = 0.5$ and $\gamma_{ijU} = 2$ to test this equivalence criterion. On the other hand, there are examples (see FDA 2022, for an example where for any test group compared to the reference group, the ratio of the test to reference standard deviations cannot exceed 2.5) where it may be of interest to treat one of the groups as a reference group in which to compare all of the other groups' standard deviations against. If so, the set of hypotheses can be simplified to

$$H_0 : \frac{\sigma_i}{\sigma_j} \geq \gamma_{ijU} \quad \text{versus} \quad H_A : \frac{\sigma_i}{\sigma_j} < \gamma_{ijU} \quad (22)$$

where group j is the reference group and the equivalence margin for each comparison is now $(0, \gamma_{ijU})$. Testing either set of hypotheses proceeds analogously to the methods outlined in Section 2.2.2 where either $100(1 - \alpha)\%$ credible intervals, or the corresponding probability calculation $q = \Pr(\gamma_{ijL} < \frac{\sigma_i}{\sigma_j} < \gamma_{ijU})$, can be used to determine whether to reject the null hypothesis (or not) and to declare equivalency. Note, that the ratio of standard deviations in (22) does not put a bound on the size of the standard for the reference group. If desired, one could instead use (21) and set the lower equivalence margin to be $\gamma_{ijL} = 1/\gamma_{ijU}$ to obtain a symmetric interval that ensures a practical bound on the reference group standard deviation.

Testing the ratio of standard deviations (or variances) seems like a sensible choice for equivalence testing, however, any complex, yet meaningful, transformation of the parameters of interest can be used. For example, for any group i and j , equivalence could be declared based on the difference in standard deviations (i.e., $\sigma_i - \sigma_j$), or the percent change in standard deviations (i.e., $(\sigma_i - \sigma_j)/\sigma_j \times 100$), or any other meaningful quantity of interest. Once again, the advantage of the Bayesian approach is that probabilistic calculations that may seem daunting (like q) are easily computed for any quantity of interest due to the fact that we can calculate q using sample based methods using samples from the appropriate posterior distributions of interest. For example, given our Bayesian model in Section 2.2.1, it would be easy to calculate the probability $q = \Pr(0 < (\sigma_i - \sigma_j)/\sigma_j \times 100 < 10\%)$ by using a Monte Carlo approximation by first obtaining posterior samples of σ_i and σ_j from (12), and then directly plugging them into $(\sigma_i - \sigma_j)/\sigma_j \times 100$, and then calculating the proportion of times that the quantity was between 0 than 10%.

Lastly, equivalence testing need not be carried out on a single equivalence metric. In fact, it may be of interest to judge equivalency based jointly on, say, average differences and ratio of standard deviations. For example, we may want to judge multigroup equivalency based on the following probability

$$q = \Pr\left(-\theta_{ijL} < \delta_{ij} < \theta_{ijU} \quad \text{and} \quad \gamma_{ijL} < \frac{\sigma_i}{\sigma_j} < \gamma_{ijU}\right). \quad (23)$$

Yet again, under a Bayesian modeling approach the probability in (23) is easily calculated based on Monte Carlo approximations by sampling appropriately from the posterior distributions of the mean differences and variances.

2.2.4 | Equivalence Based on Future Differences

Irrespective of $k = 2$ or $k > 2$ groups, nearly all of the equivalence testing procedures in the statistical literature are based on judging equivalence using parameters estimated from the data. For example, parameter estimates of the averages of each group can be obtained to construct average differences in which to judge equivalency. However, scientists and researchers are not always interested in judging equivalency based on average behaviors. In fact, a more natural inclination might be to want to ensure that any given difference (i.e., not just the average difference) is considered equivalent amongst the groups. Using the framework of Bayesian statistics, we propose a novel testing strategy that uses the posterior predictive distribution to test equivalency of multiple groups based on ensuring that future differences between groups will fall within a prespecified equivalence margin. The key idea is to think about potential future observations, x_i^* , $i = 1, \dots, k$, and to test for equivalency on the collection of all possible sets of future observations, where one cares about the difference that would be calculated on the actual future observations obtained (e.g., $x_1^* - x_2^*$). We can use the posterior predictive distribution to describe the collection of all possible sets of future observations.

Now, let us define D^* to be a $c \times 1$ matrix of c pairwise group comparisons. Note that D^* differs from Δ in that D^* represents the future differences between groups of interest that we have yet to see, whereas Δ represented the average difference between the groups of interest. Let us also introduce the notation X here to denote the collection of all observations currently observed, i.e., $X = x_{11}, \dots, x_{1n_1}, \dots, x_{k1}, \dots, x_{kn_k}$,

then the posterior predictive distribution of D^* can be calculated as

$$\begin{aligned} p(D^*|X) &= \int_{\mathbb{R}^c} \int_0^\infty \cdots \int_0^\infty p(D^*|\Delta, \sigma_1^2, \dots, \sigma_k^2, X) p(\Delta, \sigma_1^2, \dots, \sigma_k^2|X) d\Delta d\sigma_1^2 \cdots d\sigma_k^2 \\ &= \int_{\mathbb{R}^c} \int_0^\infty \cdots \int_0^\infty p(D^*|\Delta, \sigma_1^2, \dots, \sigma_k^2, X) p(\Delta|\sigma_1^2, \dots, \sigma_k^2, X) p(\sigma_1^2, \dots, \sigma_k^2|X) d\Delta d\sigma_1^2 \cdots d\sigma_k^2. \end{aligned} \quad (24)$$

While carrying out the integration directly is not trivial, Monte Carlo integration is easily performed in this case in order to obtain samples of the posterior predictive distribution. To make this explicit, without loss of generality, let us consider the case of $k = 3$ independent groups where it is of interest to consider all pairwise group comparisons. Then, we can write D^* as follows:

$$D^* := \begin{bmatrix} d_{12}^* \\ d_{13}^* \\ d_{23}^* \end{bmatrix} = \begin{bmatrix} x_1^* - x_2^* \\ x_1^* - x_3^* \\ x_2^* - x_3^* \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} x_1^* \\ x_2^* \\ x_3^* \end{bmatrix} =: Cx^*. \quad (25)$$

Under this definition of D^* , we have that the conditional predictive distribution of D^* is the following:

$$D^*|\Delta, \sigma_1^2, \dots, \sigma_k^2, x_{11}, \dots, x_{1n_1}, \dots, x_{31}, \dots, x_{3n_3} \sim N_3 \left(C\mathbb{E}(x^*), C\mathbb{Cov}(x^*)C^T \right), \quad (26)$$

with

$$\mathbb{E}(x^*) = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} \quad \text{and} \quad \mathbb{Cov}(x^*) = \begin{bmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{bmatrix}. \quad (27)$$

In general, for any number of groups k and pairwise comparisons c , we have the following:

$$D^*|\Delta, \sigma_1^2, \dots, \sigma_k^2, x_{11}, \dots, x_{1n_1}, \dots, x_{k1}, \dots, x_{kn_k} \sim N_c \left(C\mathbb{E}(x^*), C\mathbb{Cov}(x^*)C^T \right), \quad (28)$$

with

$$\mathbb{E}(x^*) = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_k \end{bmatrix} \quad \text{and} \quad \mathbb{Cov}(x^*) = \begin{bmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma_k^2 \end{bmatrix}, \quad (29)$$

where $\mathbb{Cov}(x^*)$ is a diagonal covariance matrix. Given the independence of the k groups, we can proceed to obtain posterior predictive samples of D^* by first independently sampling $\sigma_1^2, \dots, \sigma_k^2$ from the posterior distribution in (12), next taking a joint sample of Δ from the posterior distribution in (14), and then finally sampling D^* in (28) conditional on the obtained posterior samples of $\Delta, \sigma_1^2, \dots, \sigma_k^2$.

Similar to the case of Δ , the multigroup equivalence testing hypotheses can be represented by the following null and alternative hypothesis

$$H_0 : D^* \notin (-\theta_L, \theta_U) \quad \text{versus} \quad H_A : D^* \in (-\theta_L, \theta_U), \quad (30)$$

where now $-\theta_L$ and θ_U are lower and upper equivalence margins for future differences between the groups. Note that equivalence margins for future differences would likely need to be set wider than those of average differences. Likewise, testing of the hypotheses in (30) proceeds analogously to the case of Δ , i.e., via $100(1 - \alpha)\%$ credible intervals or probabilistic calculations of $q = \Pr(D^* \in (-\theta_L, \theta_U))$.

3 | ILLUSTRATIVE EXAMPLES

3.1 | Simulation Studies

Both the Wellek-Welch test and TOST procedure (adjusted for multiple comparisons) can be used to test equivalence of multiple groups, yet both come with inherent issues as discussed in Section 2.1. In this section we highlight some of these issues and show why our Bayesian approach to multigroup equivalence testing is a preferable choice. In the case of the Wellek-Welch test, it is easy to see that the test suffers from 1) having to test all pairwise group comparisons (although not all may be of interest), as well as 2) a lack of direct interpretability of ε^2 as compared to the equivalence margin in, say, the TOST procedure. Beyond these issues, what we have found empirically is that the Wellek-Welch test requires about double the signal-to-noise ratio (SNR) (i.e., $SNR = \mu/\sigma$) to come to the correct conclusion of multigroup equivalence as compared to the Bayesian approach. We illustrate this point with a small simulation study. For varying sample sizes (from 10 to 100) we simulate data for three groups from the exact same normal distribution (a $N(\mu = 0, \sigma = 3.5)$, i.e., all groups are equivalent) and apply the Wellek-Welch test with $\varepsilon = 0.25$ and $\varepsilon = 0.5$, and the Bayesian multigroup equivalence testing procedure with the equivalence margins set to $-\theta_L = (-5, -5, -5)$ and $\theta_U = (5, 5, 5)$

for comparing average differences. To make the testing procedures comparable, we set $q_0 = 1 - \alpha$ where $\alpha = 0.05$ is the associated α -level for the Wellek-Welch test. We ran the simulations 100 times and the results are summarized in Figure 1.

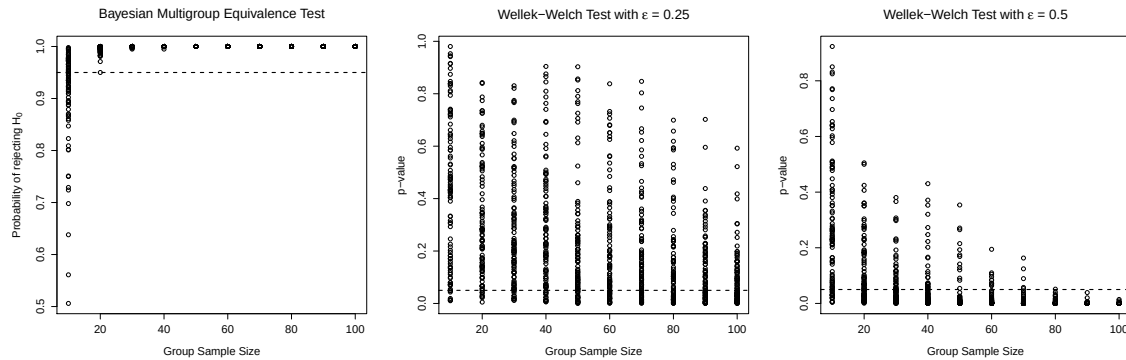


Figure 1 Simulation results for the different equivalence testing procedures when all of the groups are simulated from a $N(0, 3.5)$ distribution. The dashed lines represent either q_0 or α .

In this scenario, two things are apparent. First, for small group sample sizes (i.e., around 20), the Bayesian multigroup equivalence testing procedure is capable of coming to the correct conclusion with very high certainty, while the Wellek-Welch test is not. For $\varepsilon = 0.25$, it often fails to conclude that the groups are all equivalent, whereas for $\varepsilon = 0.5$, it takes a much larger sample size (around 60 to 70) to get the right conclusion, as compared to our Bayesian approach. Secondly, the case of $\varepsilon = 0.25$ may be too strict of a criterion for even moderate to large sample sizes. To investigate when the Bayesian approach might given comparable answers to the Wellek-Welch test, we ran further simulations and found empirically that it required having the SNR to be cut in roughly half for the results of the Bayesian multigroup equivalence testing procedure to mirror the results of the Wellek-Welch test with $\varepsilon = 0.5$. For example, running another 100 simulations as before, we changed the standard deviation of all the groups to be $\sigma = 7$ (i.e., halving the SNR), and we found that the Bayesian multigroup equivalence testing procedure declared equivalence amongst all of the groups at roughly the same rate as the Wellek-Welch test (Figure 2). Interestingly, when groups should be declared not equivalent based on average differences, the Bayesian multigroup equivalence testing procedure and the Wellek-Welch test tended to exhibit the same performance in decision making, which shows that the Bayesian approach has more power for an equivalent Type I error rate.

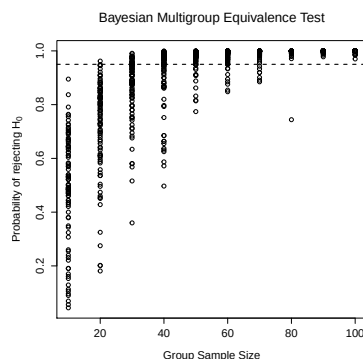


Figure 2 Simulation results for the Bayesian multigroup equivalence testing procedure when all of the groups are simulated from a $N(0, 7)$ distribution. The dashed line corresponds to q_0 .

In a more traditional power like analysis, we investigated the probability of rejecting the null hypothesis H_0 (i.e., declaring equivalence of all the groups) using the Bayesian multigroup equivalence testing procedure. To carry out this assessment, we ran a simulation study based on testing equivalence of three groups using the average difference as our equivalence criterion. Here we varied the group sample sizes and the equivalence margins as a function of the standard deviations. We set all three groups to come from the exact same normal distribution, however, we allowed

the mean of the first group to vary as a function of the standard deviation in order to assess the probability of rejecting H_0 . The simulations were run 1000 times and are summarized in Figure 3.

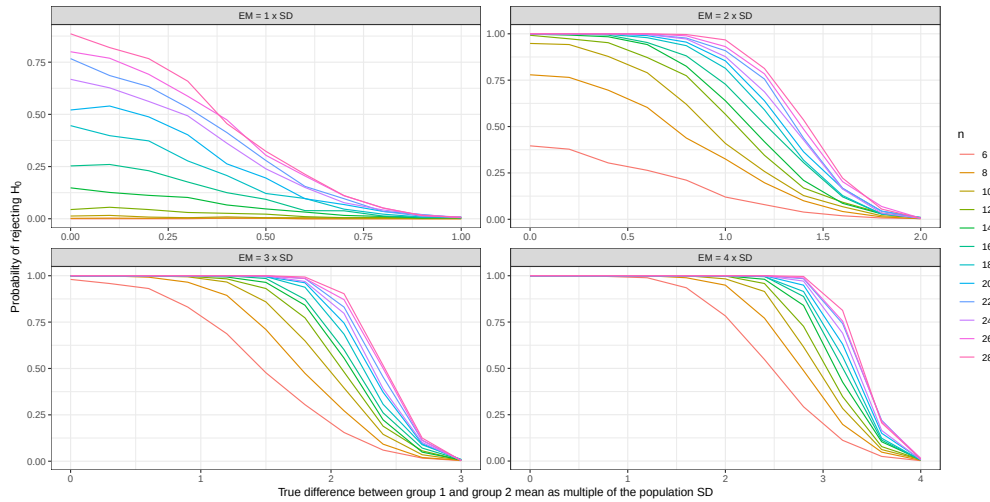


Figure 3 Probability of rejecting the null hypothesis as function of the equivalence margin (EM), average difference, and sample size.

As expected, the probability of rejecting H_0 (i.e., declaring multigroup equivalence) decreases as the true average difference approaches the equivalence margin. Likewise, as the equivalence margin is set wider, the probability to declare multigroup equivalency is also easier with much smaller sample sizes. None of this is too surprising, but the simulation results are reassuring that the Bayesian multigroup equivalence testing procedure produces similar power analysis results compared to the traditional TOST procedure (for example, Anderson-Cook & Borror 2016). Lastly, we reran this same power analysis shown in Figure 3 but instead using the TOST procedure under a Bonferroni correction for multiple hypothesis testing, as well as the Wellek-Welch test with ε fixed at 0.5. Here the TOST procedure would declare multigroup equivalence if all of the individual pairwise tests rejected the null hypothesis. Note that having multigroup equivalency be established based on all pairwise TOST tests rejecting the null hypothesis constitutes an intersection-union test, and does not actually require multiplicity correction as pointed out in Berger and Hsu (1996). However, the intent of this simulation is to showcase the impact of multiplicity correction on equivalence testing, as well as its comparison to the Bayesian multigroup equivalence testing approach, which is why we do not formulate the problem as an intersection-union test. Likewise, as pointed in Chapter 7.1 of Wellek (2010), in the case of multigroup equivalence testing, formulating the problem as an intersection-union test leads to test that may be markedly over conservative, and so, one way to potentially remedy this issue would be to use a multiplicity correction procedure. The comparison of these simulation results for the case of setting the equivalence margin to 2 times the standard deviation are presented in Figure 4.

In all cases, the Bayesian multigroup equivalence testing procedure obtains higher statistical power than the TOST procedure and the Wellek-Welch test. The TOST procedure incurs a loss in power as a result of needing to correct for multiple hypothesis testing. Figure 4 demonstrates the clear benefit of using the Bayesian multigroup equivalence testing procedure as opposed to the corrected TOST procedure and the Wellek-Welch test. On the other hand, one could ignore correcting the TOST for multiple comparisons which would lead to an increase in power, however, this may come at the expense of inflating the family-wise error rate if the problem is not formulated as intersection-union test. It would be remiss however to not note that a potential limitation of this simulation study is that its a bit of an “apples to oranges” comparison comparing the Bayesian multigroup equivalence testing procedure, and TOST under multiplicity correction, to the Wellek-Welch test since there is not a direct translation between an equivalence margin of $-\theta_L = (-5, 5, 5)$ and $\theta_U = (5, 5, 5)$ to either $\varepsilon = 0.25$ or $\varepsilon = 0.5$. However, $\varepsilon = 0.25$ or $\varepsilon = 0.5$ are the default values suggested in Wellek (2010) and thus the most reasonable values to consider and compare to.

3.2 | Preclinical Toxicology Study

In the initial phase of preclinical studies, numerous short duration pilot experiments are performed to evaluate the potential of a candidate drug to move forward to more extensive preclinical studies. Part of this process usually entails running a preclinical toxicology study, where it may be of importance to demonstrate that there is no difference (i.e., equivalency) in toxicity across different dosage levels of the candidate drug in animals. The following simulated example is based on characteristics and considerations encountered in real preclinical toxicology studies. More

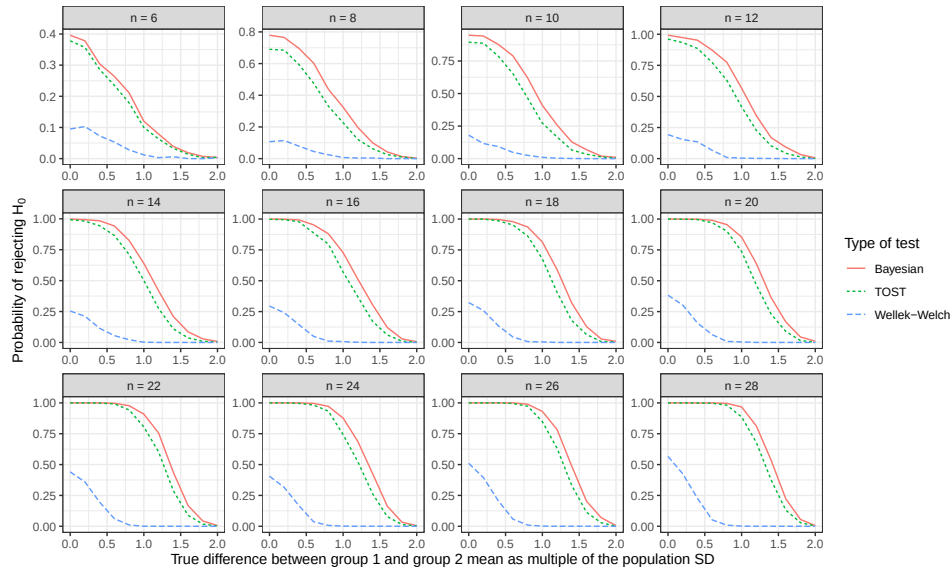


Figure 4 Comparison of the probability of rejecting the null hypothesis under the Bayesian multigroup equivalence testing procedure, the TOST procedure correcting for multiple hypothesis testing, and the Wellek-Welch test with $\varepsilon = 0.5$. The simulations correspond to the $EM = 2 \times SD$ scenario in Figure 3.

specifically, we would like to test the equivalence hypothesis that there are no differences in toxicity across four different dosage levels of the same drug administered in a study of rats. Typically a surrogate outcome is used as a proxy for toxicity, and in this case, we are looking at change in baseline body weight (i.e., terminal body weight minus baseline body weight) as the surrogate outcome to use for equivalence testing. To mirror real sample sizes used in pilot preclinical toxicology studies, we choose the sample sizes of our four groups based on the real life examples given in Pourmohamad and Wang (2022), i.e., 8, 12, 12, and 12 rats in a control, low dose, medium dose, and high dose groups, respectively. Here, the control dose group corresponds to rats that are given no dose of the drug, and so of interest would be only comparing the low, medium, and high dose group against the control dose group for equivalence (i.e., $c = 3$) rather than all potential pairwise comparisons (i.e., $c = 6$). Let $j = 0$ represent the control dose group and $i = 1, 2, 3$ represent the low, medium, and high dose groups, respectively. Based on prior experience and discussions with preclinical toxicologists, a 10 gram window seemed reasonable for setting the equivalence margin on the average difference for change in baseline body weight, i.e., $-\theta_{i0L} = -5$ and $\theta_{i0U} = 5$ for all i . Similarly, if one wanted to investigate the variability in those outcomes, then it was decided that the variability in the low, medium, and high dose group should not be more than 2 times larger than that of the control dose group (i.e., $\sigma_i/\sigma_0 < 2$ for all i) to be declared equivalent. Formalizing this all into two probabilities of interest, we have that

$$q_1 = \Pr(-5 < \delta_{10}, \delta_{20}, \delta_{30} < 5) \quad \text{and} \quad q_2 = \Pr\left(-5 < \delta_{10}, \delta_{20}, \delta_{30} < 5 \text{ and } 0 < \frac{\sigma_1}{\sigma_0}, \frac{\sigma_2}{\sigma_0}, \frac{\sigma_3}{\sigma_0} < 2\right). \quad (31)$$

For a prespecified q_0 , we can declare equivalence of the multiple dose groups based on average differences if $q_1 \geq q_0$, or based on average differences and variability if $q_2 \geq q_0$. Given the exploratory nature of pilot studies, it would be reasonable to set q_0 slightly lower than in a confirmatory study at, say, $q_0 = 0.8$.

For our simulation study, we considered six different scenarios based on varying the group sample sizes and the standard deviation of the high dose group. In general, data was generated from the same normal distribution for all of the groups (i.e., $N(\mu = 0, \sigma = 2)$), with the exception that in half of the scenarios, the high dose group had either a similar standard deviation of 2, or a larger standard deviation of 5. We ran simulations for each of the six scenarios 1000 times and the results of these simulations are captured in Table 1.

From Table 1, we see that there is a very high chance (i.e., $\Pr(q_1 \geq q_0)$ is close to 1) of making the correct decision in declaring the groups equivalent based on testing the average difference across all six scenarios. Likewise, when the standard deviation of the high dose group is set to be much larger than the other dose groups, the correct decision that the groups are not all equivalent is declared very consistently (i.e., $\Pr(q_2 \geq q_0)$ is very close to 0). On the other hand, when all of the standard deviations of all of the dose groups are the same, it is a much harder criterion to declare multigroup equivalency using q_2 with smaller sample sizes. In this case, the sample size needs to almost double to achieve high certainty in q_2 being true, which would be rare in a pilot toxicology study as sample sizes rarely exceed 12 animals per group.

Similarly, we investigated the use of the posterior predictive distribution for establishing multigroup equivalence based on future differences in a simulation setup. The simulations proceeded analogously to before, but now the equivalence margins were varied to reflect the fact that they

Sample Sizes	Std. Devs	$P(q_1 \geq q_0)$	$\mathbb{E}(q_1)$	$SD(q_1)$	$P(q_2 \geq q_0)$	$\mathbb{E}(q_2)$	$SD(q_2)$
8, 12, 12, 12	2, 2, 2, 5	0.981	0.962	0.052	0.032	0.276	0.238
8, 12, 12, 12	2, 2, 2, 2	1	0.996	0.007	0.576	0.768	0.228
20, 20, 20, 20	2, 2, 2, 5	0.999	0.997	0.010	0.029	0.237	0.234
20, 20, 20, 20	2, 2, 2, 2	1	0.999	0.000	0.915	0.937	0.104
30, 30, 30, 30	2, 2, 2, 5	1	0.999	0.000	0.016	0.178	0.201
30, 30, 30, 30	2, 2, 2, 2	1	1	0	0.991	0.983	0.048

Table 1 Summary of the 1000 simulation results for the probability, average, and standard deviation of q_1 and q_2 for each of the six different scenarios.

need to be set wider compared to those for average differences. With a mean of 0 and standard deviation of 2 for all of the groups, it would be expected that almost all of the future differences would lie between -14 to 14, however, in practice the equivalence margins would be set based on scientific judgement and rationale. Here we would declare equivalence of the multiple dose groups if $q^* \geq q_0$, where $q_0 = 0.8$ and

$$q^* = \Pr(-\theta_L < d_{10}^*, d_{20}^*, d_{30}^* < \theta_U). \quad (32)$$

The results of the 1000 simulations for the three different scenarios are presented in Table 2.

Sample Sizes	Std. Deviations	$-\theta_L$	θ_U	$\Pr(q^* \geq q_0)$	$\mathbb{E}(q^*)$	$SD(q^*)$
8, 12, 12, 12	2, 2, 2, 2	-10	10	1	0.969	0.023
8, 12, 12, 12	2, 2, 2, 5	-10	10	0.670	0.832	0.081
8, 12, 12, 12	2, 2, 2, 5	-14	14	0.989	0.940	0.048

Table 2 Summary of the 1000 simulation results for the probability, average, and standard deviation of q^* for each of the three different scenarios.

When the variability is the same amongst all the dose groups, it is fairly easy to declare equivalency with a modestly wide range for the equivalence margin (i.e., $-\theta_L = -10$ and $\theta_U = 10$). However, when the variability in the high dose group increases substantially, the modest equivalence margin is no longer wide enough to come to the same conclusion. The third scenario shows that when the equivalence margins increase to -14 and 14, the equivalency of the groups is again established. This last simulation highlights the importance of understanding and quantifying the variability in the equivalence margin. Intuitively, the equivalence margin for a future difference needs to be set larger than that of an average difference, however, the size of the equivalence margin must be reflective of the actual scientific problem at hand.

Turning to a real world example, the following historical data was taken from a real preclinical toxicology experiment in rats (i.e., what follows is a retrospective analysis of data taken from a pilot preclinical toxicology study). Due to concerns of confidentiality, all identifiers for the candidate drugs and study have been removed from the data set. Rats were randomly allocated to one of four dose groups in the study: control, low dose, medium dose, or high dose, and on day one of the study their baseline measurements were recorded pre-dose. Rats in the control dose group received a sham treatment that contained no dose of the drug material under study. After receiving their first dose, rats were followed on study for a predetermined number of days before being euthanized on the last day of study, and their final data collected. Body weight measurements were collected prior to dosing and at the end of study, and these body weight measurements were then used to calculate the body weight gain of each rat over the course of the study, i.e., body weight gain = (final weight) - (initial weight). This change in baseline body weight was considered an important outcome for the study as it served as a surrogate endpoint for toxicity. Here no weight gain, as compared to the control group, was considered a positive finding. The measured change in baseline body weight (in grams), by group, is shown in Figure 5 and summarized in Table 3.

The data set consisted of 44 rats of which 8 were in the control, 12 in the low dose group, 12 in the medium dose group, and 12 in the high dose group. It is of scientific interest to show that the change in baseline body weight of the three different dose groups was equivalent to the control group. Here we can use the Bayesian multigroup equivalence testing procedure to test multigroup equivalence based on average differences. Given that this is a retrospective analysis, had the Bayesian multigroup equivalence testing procedure been used at the time, an appropriate equivalence margin would have needed to be elicited for the study based on scientific expertise. Let us again assume that a 10 gram window would be reasonable for setting the equivalence margin on the average difference for change in baseline body weight, i.e., $-\theta_{i0L} = -5$ and $\theta_{i0U} = 5$ for all i , where $j = 0$ denotes the control group and $i = 1, 2$, and 3 denotes the low, medium, and high dose groups. Here we would declare equivalence of the multiple dose groups if $q \geq q_0$, where $q_0 = 0.8$ and $q = \Pr(-\theta_L < \delta_{10}, \delta_{20}, \delta_{30} < \theta_U)$. Calculating q , we have that $q = 0.8373$ and that we would

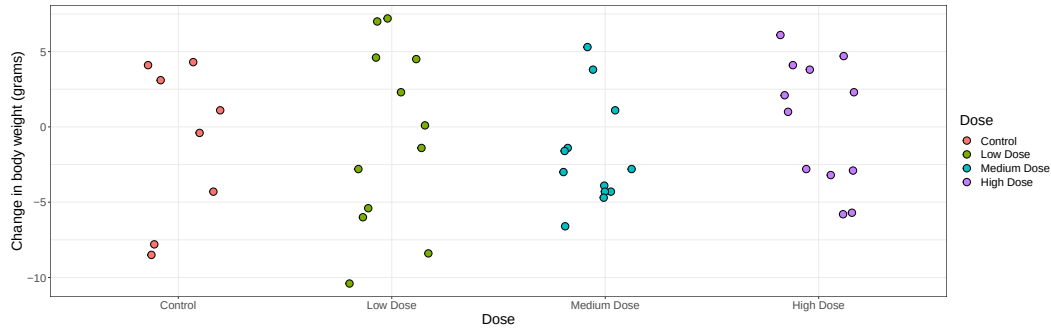


Figure 5 Weight gain data for 44 rats assigned to either the control, lose dose, medium dose, or high dose group.

Dose Group	Sample Size	Average	Std. Deviation
Control	8	-1.05	5.20
Low	12	-0.73	5.99
Medium	12	-1.87	3.58
High	12	0.31	4.19

Table 3 Summary of the change in baseline body weight data for the different dose groups in the study. Empirically, the medium dose group had the largest average change in body weight.

declare equivalency of the multiple dose groups, implying that there is likely no toxicological effect present. From Figure 5 one might expect that multigroup equivalency might be declared, however, we can also plot the posterior distributions of the pairwise comparisons to get a better idea (Figure 6).

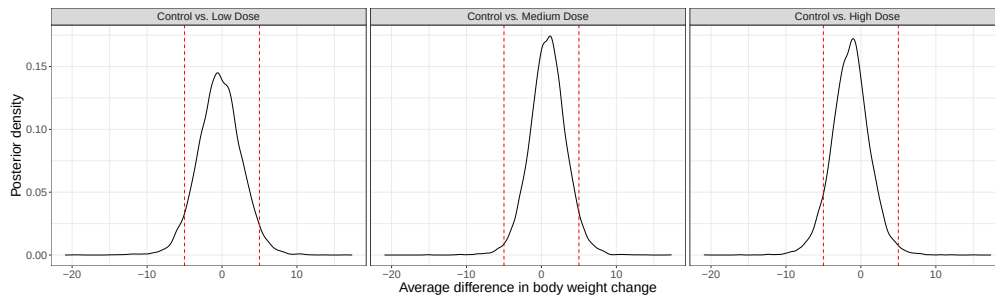


Figure 6 Posterior distributions for the pairwise comparisons δ_{10} , δ_{20} , and δ_{30} . The red dashed lines correspond to the respective equivalence margins.

Indeed, for all of the pairwise comparisons we see that the majority of their posterior densities lie within the region defined by the equivalence margins. Had we not declared equivalence, these posterior density plots would be useful in helping to identify the potential comparisons that are driving the non-equivalence. Alternatively, we could have instead investigated the multigroup equivalence testing hypothesis using equivalence based on future differences as our evaluation criterion. For example, given that we have small sample sizes in the low and medium dose groups (i.e., potentially higher uncertainty in our variance estimates), and that we know that our equivalence margins in general should be wider for future differences, we could set our equivalence margins to be slightly wider at $-\theta_{i0L} = -15$ and $\theta_{i0U} = 15$ for all i , where again $j = 0$ denotes the control group and $i = 1, 2$, and 3 denotes the low, medium, and high dose groups. Here we would declare equivalence of the multiple dose groups if $q^* \geq q_0$, where $q_0 = 0.8$ and $q^* = \Pr(-\theta_L < d_{10}^*, d_{20}^*, d_{30}^* < \theta_U)$. Calculating q^* , we have that $q = 0.7463$ and therefore we do not declare equivalency of the multiple dose groups, implying that there may be a toxicological effect present based on this equivalence criterion. Note this time though that there was only a slightly lower probability in favor of multigroup equivalency using the future differences ($q^* = 0.7463$) as opposed to the average differences (i.e., $q = 0.8373$). From Figure 7, this last point is clear as there is a lower chance of multigroup equivalency based on using

the future differences since a slightly larger proportion of the posterior predictive distributions are outside of the equivalence margin as opposed to the proportion of the posterior distributions in Figure 6.

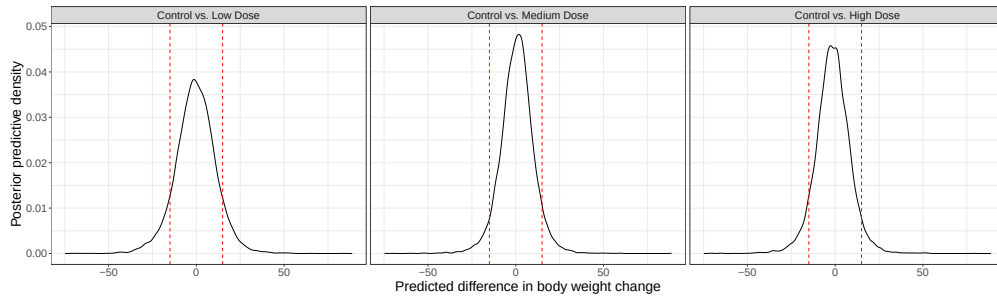


Figure 7 Posterior predictive distributions for the pairwise comparisons d_{10}^* , d_{20}^* , and d_{30}^* . The red dashed lines correspond to the respective equivalence margins.

4 | DISCUSSION

Using a Bayesian approach allows for an elegant solution to the problem of testing for multigroup equivalence. This new test procedure uses a scientifically meaningful equivalence margin to define comparability, and it is a more powerful test than using multiple pairwise comparisons or than the Wellek-Welch test. A significant contribution, the Bayesian multigroup equivalence testing procedure avoids the need to do all pairwise comparisons, but rather is capable of carrying out only the pairwise comparisons of interest in a single hypothesis test while remaining highly customizable to the definition of equivalence (e.g., based on average differences, ratios, some combination of the two, and more). This approach also allows for a new definition of comparability based on future observations.

There are several ways this work could be extended. The paper focused on the case of independent groups, however, the equivalence testing framework is amenable to the very common case of dependent groups. For example, when comparing measurements from different devices or assays, the samples that get used by each device or assay may be the same samples and so clearly the groups of measurements are no longer independent. The good thing about our Bayesian approach is that the framework allows for capturing this dependence structure when specifying the data likelihood, however, the closed form solutions that we obtained in Section 2.2 may no longer hold. However, under a Bayesian modeling framework any type of evaluation criterion can be evaluated as long as posterior samples (or posterior predictive samples) for the parameters of interest can be obtained, which is what makes the Bayesian approach so flexible and desirable. Similarly, we explored the case of sampling directly from the posterior distribution of the average difference Δ , however, it might be of interest to obtain posterior samples directly of the group means so that one could create an evaluation criterion based on a different scientifically meaningful function of the groups means (e.g., $\Delta = \mu_i/\mu_j$). In our application we found it more natural to work with the joint posterior distribution of the average differences, however the intermediate step of sampling from the posterior distribution of the individual group averages can easily be carried out using the conditional posterior distribution in (9).

Lastly, there are active areas of research in bioequivalence studies that may benefit from using our Bayesian testing approach. For example, three-arm clinical endpoint bioequivalence studies (Yang & Sun 2019) are used to test equivalency between a generic drug (test group) and an innovator drug (reference group). A third group comprised of a placebo is also included as part of the equivalency comparison. For overall bioequivalence to be established (i.e., to test equivalency), multiple hypothesis tests are carried out to tests two superiority tests (test group versus placebo, and reference group versus placebo) and one equivalence test (test versus reference group). Traditional methods struggle under this scenario due to the constraints of multiple hypothesis testing (i.e., lower power or the need to increase sample sizes), however, under our Bayesian framework, superiority tests can be redefined as equivalence tests with an equivalence margin defined from 0 to ∞ , which then would allow us to conduct all three tests simultaneously using our Bayesian multigroup testing procedure.

Acknowledgments

We are grateful for the helpful comments and suggestions from three anonymous reviewers and an associate editor.

Conflict of interest

The authors declare no potential conflict of interests.

References

- Anderson-Cook, C. M., & Borror, C. M. (2016). The difference between "equivalent" and "not different". *Quality Engineering*, 28(3), 249–262.
- Berger, R. L. (1982). Multiparameter hypothesis testing and acceptance sampling. *Technometrics*, 24(4), 295–300.
- Berger, R. L., & Hsu, J. C. (1996). Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science*, 11(4), 283–319.
- FDA. (2022). *Statistical approaches to establishing bioequivalence: Guidance for industry* (Tech. Rep.). U.S. Department of Health and Human Services, FDA. Center of Drug Evaluation and Research (CDER), Rockville, MD.
- Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., & Rubin, D. (2013). *Bayesian data analysis, third edition*. Taylor & Francis. Retrieved from <http://books.google.com/books?id=ZXL6AQAQBAJ>
- Ghosh, P., & Gönen, M. (2008). Bayesian modeling of multivariate average bioequivalence. *Statistics in Medicine*, 27(13), 2402–2419.
- Hoyda, J. J., Counsell, A., & Cribbie, R. A. (2019). Traditional and Bayesian approaches for testing mean equivalence and a lack of association. *The Quantitative Methods for Psychology*, 15(1), 12–24.
- Koh, A., & Cribbie, R. (2013). Robust tests of equivalence for k independent groups. *British Journal of Mathematical and Statistical Psychology*, 66, 426–434.
- Kruschke, J. K. (2018). Rejecting or accepting parameter values in Bayesian estimation. *Advances in Methods and Practices in Psychological Science*, 1(2), 270–280.
- Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, 1(2), 270–280.
- Lauzon, C., & Caffo, B. (2009). Easy multiplicity control in equivalence testing using two one-sided tests. *The American Statistician*, 63(2), 147–154.
- Liao, J. G., Midya, V., & Berg, A. (2019). Connecting and contrasting the Bayes factor and a modified ROPE procedure for testing interval null hypotheses. *The American Statistician*, 75(3), 256–264.
- Miller, R. G. (1981). *Simultaneous statistical inference* (2nd ed.). New York, NY: Springer.
- Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods*, 16(34), 406–419.
- Pallmann, P., & Jaki, T. (2016). Simultaneous confidence regions for multivariate bioequivalence. *Statistics in Medicine*, 36(39), 4585–4603.
- Parasuraman, S. (2011). Toxicological screening. *Journal of Pharmacology & Pharmacotherapeutics*, 2(2), 74–79.
- Pourmohamad, T., & Wang, C. (2022). Sequential Bayes factors for sample size reduction in preclinical experiments with binary outcomes. *Statistics in Biopharmaceutical Research*, 1–11.
- Quan, H., Bolognese, J., & Yuan, W. (2001). Assessment of equivalence on multiple endpoints. *Statistics in Medicine*, 20(21), 3159–3173.
- Richter, S. J., & Richter, C. (2002). A method for determining equivalence in industrial applications. *Quality Engineering*, 14(3), 375–380.
- Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, 56, 356–374.
- Schuurmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15, 657–680.
- van Ravenzwaaij, D., Monden, R., Tendeiro, J. N., & Ioannidis, J. P. A. (2019). Bayes factors for superiority, non-inferiority, and equivalence designs. *BMC Medical Research Methodology*, 19(71), 1–12.
- Wellek, S. (2010). *Testing statistical hypotheses of equivalence and noninferiority* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC.
- Yang, A., & Sun, W. (2019). An exact method for power calculation for a three-arm clinical endpoint bioequivalence study. *International Journal of Statistics and Probability*, 1(8), 25–39.

How to cite this article: Pourmohamad, T. and Lee, H. K. H. (2023), Equivalence Testing for Multiple Groups, *Stat*, 2023;00:1–14.