Penultimate version. If citing, please refer instead to the published version in Statistics in Medicine, DOI: 10.1002/sim.7610.

Simultaneous small-sample comparisons in longitudinal or multi-endpoint trials using multiple marginal models

Philip Pallmann[1], Christian Ritz[2], Ludwig A. Hothorn[3]

[1]Medical and Pharmaceutical Statistics Research Unit, Department of Mathematics and Statistics, Lancaster University, Lancaster, UK
[2]Department of Nutrition, Exercise and Sports, University of Copenhagen, Frederiksberg C, Denmark
[3]Institute of Biostatistics, Leibniz University Hannover, Hannover, Germany

Abstract

Simultaneous inference in longitudinal, repeated-measures, and multi-endpoint designs can be onerous, especially when trying to find a reasonable joint model from which the interesting effects and covariances are estimated. A novel statistical approach known as "multiple marginal models" greatly simplifies the modelling process: the core idea is to "marginalise" the problem and fit multiple small models to different portions of the data, and then estimate the overall covariance matrix in a subsequent, separate step. Using these estimates guarantees strong control of the familywise error rate, however only asymptotically. In this paper we show how to make the approach also applicable to small-sample data problems. Specifically, we discuss the computation of adjusted p-values and simultaneous confidence bounds for comparisons of randomised treatment groups as well as for levels of a non-randomised factor such as multiple endpoints, repeated measures, or a series of points in time or space. We illustrate the practical use of the method with a data example.

Keywords: *correlated data, multiple contrast test, degrees of freedom, linear mixed-effects model*

1 Introduction

Simultaneous inference with the aim to control the familywise type I error rate (FWER) is well explored for simple randomised settings and easily available in major statistical software packages [2, 60]. Affairs can get messy, however, as soon as some element of repeatedness induces correlation between observations, as is the case with longitudinal data or multiple endpoints. In this paper we focus on what Quan *et al.* [46] called a "two-dimensional multiplicity problem": simultaneous inference for datasets with a randomised (or between-group) factor variable, such as treatment in a clinical trial, and a repeated (or within-group) factor variable, such as points in time or space, or multiple endpoints, all of which are situations where the instances are not mutually randomised—the second time point always comes after the first. Every subject is exposed to exactly one level of the randomised factor but usually multiple levels of the repeated factor. Specifically, we look into comparisons

a) between levels of the randomised factor *separately and simultaneously* at several levels of the repeated factor (e.g., comparing treatment effects at each time point), and

b) between levels of the repeated factor *separately and simultaneously* at several levels of the randomised factor (e.g., comparing differences between time points within each treatment group).

Here *separately* means that we are interested in detailed level-wise comparisons rather than just global tests, while *simultaneously* refers to our desire to control the FWER strongly over all comparisons. It is further possible to combine comparisons of both types a) and b) simultaneously under joint FWER control.

The analysis of suchlike correlated data is a bit of a statistical minefield, with flawed analyses still being omnipresent in the applied sciences [36]. For instance, separate ANOVAs per time point or endpoint followed by unadjusted *post hoc* tests can lead to frighteningly high overall type I error rates [19]. Using a Bonferroni correction is unnecessarily conservative, especially when the levels of the repeated factor are highly correlated [64]. *Ad hoc* solutions to cushion the conservatism of Bonferroni [55] have a very limited range of application, and it is unclear under what circumstances they are able to control the FWER.

Boiling data from serially repeated measurements down to a single summary measure [9, 10, 53] makes the analyses simple but often leaves many specific questions open that could be answered with more refined approaches [7, 40, 1]. Standard methods for analysing multiple endpoints [8, 50, 51, 59] compare the treatment means *simultaneously but jointly* at multiple occasions i.e., they provide only a global statement of "significance" expressed as p-value(s) and no confidence intervals. When the interest lies in *separate* level-wise comparisons, this is neither as detailed

nor as insightful as typically requested by researchers. Moreover, widespread techniques such as the multivariate analysis of variance (MANOVA) and repeated measures ANOVA rely on unrealistic—and unverifiable—assumptions e.g., multisample sphericity [22], and cannot cope with missing values properly.

Our preferred method for comparing means in repeated-measures or multi-endpoint settings are multiple contrast tests (MCTs) [20], for two reasons: first, they cover many standard procedures such as comparisons to a common control [6], between all pairs of treatments [57], or with the grand mean [41]; and second, they provide not only test decisions and adjusted p-values but also informative simultaneous confidence intervals, which are the favoured form of presenting statistical results [12]. But for MCTs to be sound in the context of correlated data, capturing dependencies across levels of the repeated variable is pivotal. Traditionally linear mixed-effects modeling [58] has been the method of choice, but this leaves us with the problem how to shape the random effects and residual covariance matrices.

Pipper *et al.* [45] introduced an asymptotic method that makes simultaneous inferences for datasets with both randomised and repeated variables stunningly easy: the idea is to fit a number of relatively simple models separately to different portions of the data (e.g., one model per time point or endpoint) and then estimate the joint covariance matrix in a data-driven manner. Pallmann *et al.* [42] showed that this method is readily applicable to longitudinal data, where it can replace much more intricate analyses that would involve specifying random effects and structured residual covariance matrices. Jensen *et al.* [26] applied the method to simultaneous inference of model-averaged derived parameters, Große Ruse *et al.* [13] used it to assess a binary composite endpoint jointly with its components, and Kitsche *et al.* [28] employed it for the joint consideration of the three basic modes of inheritance in genetic association studies. An extension to settings with more than one repeated variable, such as in longitudinal multi-endpoint studies, was described as well [25, 49]. All these publications have illustrated the wide applicability of the method, but also emphasised that it is inherently asymptotic and may therefore break down with too small sample sizes. This is potentially problematic as many real-world datasets involve smallish samples, often with less than 20 or even ten independent subjects per treatment group. In this paper we discuss modifications of the Pipper *et al.* method that make it applicable to small and even tiny datasets.

In Section 2 we outline the underlying method and our proposed small-sample variant, with a focus on the degrees of freedom to be used. A simulation study on its performance under the null and alternative hypothesis is summarised in Section 3. We illustrate the practical application to a longitudinal heart rate study in Section in 4, followed by a general discussion in Section 5.

2 Methods

Consider a clinical trial setting where several measurements $j = 1, \ldots, m$ (e.g., a longitudinal outcome or multiple endpoints) are taken from patients $i = 1, \ldots, n$ randomised to treatment arms $k = 1, \ldots, q$. The data can be summarised in a vector $\mathbf{y}$, with y_{ijk} being measurement j from patient i receiving treatment k. The number of patients randomised to arm k is denoted by n_k.

The goal of statistical inference is to estimate a set of unknown parameters $\boldsymbol{\beta}$ that are usually related to effects of covariates. While some (like age or sex) are only included as adjustments, others (like treatment) are of explicit interest as they characterise the medical intervention under study, and we wish to estimate them separately for each of the m time points or endpoints. Here the classical approach is to fit one general linear mixed-effects model with random effects and/or structured residual covariance matrix to the whole dataset [3]. Working out a "reasonably good" model can be tedious though, even with the help of an information criterion [24, 61, 62, 27, 65]. Both under- and overfitting may be detrimental to subsequent inferences [34, 43, 23, 52, 37, 14].

2.1 Multiple marginal models

A hassle-free and foolproof alternative has been suggested by Pipper *et al.* [45]: instead of devising one big model for the entire dataset, several "marginal" models are fitted to different slices of the data. For a longitudinal dataset, one would simply fit one model per time point, and for multivariate data, every endpoint would be modelled separately. These "marginal" models are typically much simpler than a joint model with all its random effects and/or residual covariances.

Assume that the data $\mathbf{y}$ are sensibly sliceable into m portions $\mathbf{y}_1, \ldots, \mathbf{y}_m$ e.g., because they contain m longitudinal measurements or m endpoints. Further assume that $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jq})$ are interesting model parameters, such as the effects of q different drugs in a pharmaceutical study, measured at the jth time point or for the jth outcome measure. In the simplest case these "marginal" models are just linear models that can be fitted with least squares, but they may also be e.g., generalised linear models (for discrete outcomes), Cox proportional hazard models (for survival outcomes), or linear mixed-effects models (when additional repeated factors are present). We fit one "marginal" model $\mathcal{M}_j$ per data portion $\mathbf{y}_j$ using some fixed and known design matrix $\mathbf{X}_j$ to estimate $\boldsymbol{\beta}_j$. Adjustment for (baseline) covariates is easily possible within the "marginal" models; in that case additional fixed-effect parameters other than $\beta_{j1}, \ldots, \beta_{jq}$ will be estimated. Note that the method is sufficiently general to allow different types of "marginal" models, different sets of (baseline) covariates for different "marginal" models, and $\mathbf{y}_j$ can also be the full dataset $\mathbf{y}$ for some or all j.

All that is left to do now is recover the correlation between the estimates $\widehat{\boldsymbol{\beta}}_1, \ldots, \widehat{\boldsymbol{\beta}}_m$, which is achieved with a little mathematical trick that involves "stacking" the score contributions (derivatives of the log-likelihood) of the parameter

estimates [45]. Here it is instructive to see that asymptotically

$$(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)\sqrt{n} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Psi_{ij} + o_P(1)$$

where

$$\Psi_{ij} = -\mathcal{I}_j^{-1} \widetilde{\Psi}_{ij}$$

and $\mathcal{I}_j^{-1}$ is the row in the inverse Fisher information matrix that corresponds to $\boldsymbol{\beta}_j$, $\widetilde{\Psi}_{ij}$ is the score function for the ith individual, and $o_P(1)$ denotes a sequence of random vectors converging to zero in probability.

We can obtain a similar representation when the $\boldsymbol{\beta}_j$, $\widehat{\boldsymbol{\beta}}_j$, and Ψ_{ij} are "stacked" over all $j = 1, \ldots, m$ so as to get longer vectors (but not matrices) $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m)$, $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_1, \ldots, \widehat{\boldsymbol{\beta}}_m)$, and $\boldsymbol{\Psi}_i = (\Psi_{i1}, \ldots, \Psi_{im})$. This yields

$$(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\sqrt{n} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{\Psi}_i + o_P(1),$$

which is the asymptotic m-variate extension of the above. The left-hand side converges (by the multivariate central limit theorem) to an m-variate normal distribution

$$(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\sqrt{n} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}).$$

As a consequence we can estimate the covariance $\boldsymbol{\Sigma}$ consistently as

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{i=1}^{n} \widehat{\boldsymbol{\Psi}}_i^T \widehat{\boldsymbol{\Psi}}_i$$

where the $\widehat{\boldsymbol{\Psi}}_i$ are obtained by plugging the parameter estimates from the m marginal models into $\boldsymbol{\Psi}_i$. This is implemented as function `mmm` in the `R` package `multcomp` [21].

2.2 Simultaneous inference

Having estimated the effect parameters $\boldsymbol{\beta}$ and covariances $\boldsymbol{\Sigma}$ as shown in 2.1, we can specify comparisons of interests as linear contrasts

$$\boldsymbol{\eta} = \mathbf{C}\boldsymbol{\beta}$$

where $\mathbf{C}$ is a $z \times mq$ coefficient matrix, and z the number of single contrasts. With the parameters in the vector $\boldsymbol{\beta}$ being ordered as

$$\boldsymbol{\beta} = (\beta_{11}, \ldots, \beta_{1q}, \ldots, \beta_{m1}, \ldots, \beta_{mq}),$$

comparing treatments separately and simultaneously for outcomes $1, \ldots, m$ requires a coefficient matrix

$$\mathbf{C}_1 = \mathbf{I}_m \otimes \mathbf{C}_0$$

where $\mathbf{I}_m$ is an identity matrix of dimension m, $\otimes$ denotes the Kronecker product, and $\mathbf{C}_0$ is a "per-outcome" coefficient matrix that defines the comparisons of treatment means carried out for one single endpoint or time point. As an example, with $q = 3$ treatments, this would result in the following matrices for many-to-one, all-pairwise, and grand-mean comparisons:

$$\mathbf{C}_0^{mto} = \begin{bmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix}, \quad \mathbf{C}_0^{ap} = \begin{bmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \\ 0 & -1 & 1 \end{bmatrix}, \quad \mathbf{C}_0^{gm} = \begin{bmatrix} 2/3 & -1/3 & -1/3 \\ -1/3 & 2/3 & -1/3 \\ -1/3 & -1/3 & 2/3 \end{bmatrix}.$$

For comparisons of outcome means simultaneously and separately for treatment arms $1, \ldots, q$ the coefficient matrix becomes

$$\mathbf{C}_2 = \mathbf{C}_0 \otimes \mathbf{I}_q,$$

and $\mathbf{C}_0$ would be the "per-group" coefficient matrix.

The computation of adjusted p-values and simultaneous CIs is described in detail in [20].

2.3 Small-sample degrees of freedom

Using the covariance $\widehat{\boldsymbol{\Sigma}}$ estimated as in 2.1 warrants strong FWER control asymptotically [45] but not for small sample sizes. Pallmann *et al.* found by simulation that the asymptotic tests are unduly liberal for sample sizes of less than 30 per treatment group [42]. This requirement is unacceptable in many real-world experiments, and thus solutions for smaller samples are needed.

We propose to replace the multivariate normal reference distribution of the asymptotic procedure by a multivariate t with some appropriate degrees of freedom (DF). The residual DF associated with the jth marginal linear model is

$$\nu_j = n_j - q_j$$

where n_j and q_j are the numbers of patients and treatment groups, respectively, from which measurements were obtained for the jth time point or endpoint. A practicable approximation to the DF for a set of outcome-wise comparisons of treatments is the minimum of the marginal models' residual DFs:

$$\nu_j^{min} = \min_j \nu_j.$$

In case of substantial sample size imbalance, this DF will get conservative. Alternatively, the average of the marginal models' DFs could be used:

$$\bar{\nu}_j = \frac{1}{m} \sum_{j=1}^{m} \nu_j.$$

For treatment-wise comparisons of outcomes these approximations are rather poor; an *ad hoc* alternative for the kth treatment group is

$$\nu_k = n_k - 1,$$

and the DF for the set of treatment-wise comparisons of outcomes may again be approximated as the minimum

$$\nu_k^{min} = \min_k \nu_k$$

or average

$$\bar{\nu}_k = \frac{1}{q} \sum_{k=1}^{q} \nu_k.$$

Non-integer values of $\bar{\nu}_j$ and $\bar{\nu}_k$ are rounded down to the nearest integer.

We investigate the performance of these DF methods via simulation in 3.1 and 3.2.

2.4 Extension: a duplex procedure

Assume we wanted to perform the comparisons defined in $\mathbf{C}_1$ and $\mathbf{C}_2$ (see 2.2) under joint FWER control, then the resulting coefficient matrix would be

$$\mathbf{C}_{12} = \begin{bmatrix} \mathbf{C}_1 \\ \mathbf{C}_2 \end{bmatrix}.$$

In such a set of comparisons between and within treatment groups the DFs ν_j and ν_k can be very different. For example, in a balanced design with $q = 4$ treatments and $n_k = 10$ patients per arm, we get $\nu_j = 36$ and $\nu_k = 9$, which leads to very different reference quantiles. Unfortunately, the most common form of multivariate t distributions [29] only allows one single DF. Straightforward choices are the minimum DF

$$\nu^{min} = \min(\nu_j^{min}, \nu_k^{min})$$

and the (weighted) average DF

$$\bar{\nu} = \frac{1}{z}(z_j \nu_j^{min} + z_k \nu_k^{min})$$

where z_j and z_k are the numbers of comparisons among treatment and outcome means, respectively, and $z_j + z_k = z$. As in 2.3, non-integer values are rounded down.

Both ν^{min} and $\bar{\nu}$ are clearly not ideal: ν^{min} will make the test procedure conservative overall i.e., the global test size will be less than α, because some elementary comparisons will be tested with too stringent a reference quantile, except in the unlikely case when $\nu_j^{min} = \nu_k^{min}$. On the other hand, $\bar{\nu}$ will allow too many rejections of H_0 for some hypotheses and too few for others, again unless $\nu_j^{min} = \nu_k^{min}$.

One possible corrective that circumvents these issues is to use a vector of comparison-specific DFs

$$\boldsymbol{\nu} = (\nu_1, \ldots, \nu_z)$$

and compare every test statistic against a quantile from its own multivariate t reference distribution [11, 16]. This approach—although an approximation only—has been shown in simulations to work reasonably well for heteroscedastic data [15]. As an alternative we can use a z-dimensional reference distribution whose marginals are univariate t distributions with DFs $\nu_1, \ldots, \nu_z$, linked by a normal or t copula [39] with the estimated correlation matrix $\widehat{\boldsymbol{\Sigma}}$. Supplementary Figure S4 displays the peculiar shapes of these copula-generated distributions.

We will investigate the effects of both approaches on the test size of our procedure in 3.3. Additionally, they may also be applied to the DFs ν_j and ν_k in 2.3 when there is imbalance, either by design (e.g., more patients randomised to the control arm) or due to dropout.

Two `R` functions `multipleDF` and `copulaDF` for calculating adjusted p-values and simultaneous CIs using comparison-specific DFs and a copula-based multivariate reference distribution, respectively, are provided in a supplementary file.

2.5 Approximate power

We can approximate the global power i.e., the probability of finding at least one elementary comparison of Gaussian means significant [17], of the "multiple marginal models" procedure with two-sided hypotheses as

$$P\left(\exists\ h : |T_h| > t^{two}_{z,1-\alpha}(\nu, \mathbf{R})\right) = 1 - \mathcal{T}_z\left(-t^{two}_{z,1-\alpha}(\nu, \mathbf{R}), t^{two}_{z,1-\alpha}(\nu, \mathbf{R}), \mathbf{R}, \nu, \boldsymbol{\delta}\right)$$

where T_h is the Wald-type test statistic for the hth contrast ($h = 1, \ldots, z$), $t^{two}_{z,1-\alpha}$ is the two-sided equicoordinate $1-\alpha$ quantile of the z-dimensional t distribution with ν DF and correlation matrix $\mathbf{R}$, and $\mathcal{T}_z(a, b, \mathbf{R}, \nu, \boldsymbol{\delta})$ is the z-variate t probability with integration bounds a and b (same in all z dimensions), correlation matrix $\mathbf{R}$, ν DF, and noncentrality vector $\boldsymbol{\delta}$.

The one-sided analogues are

$$P\left(\exists\ h : T_h > t^{one}_{z,1-\alpha}(\nu, \mathbf{R})\right) = 1 - \mathcal{T}_z\left(-\infty, t^{one}_{z,1-\alpha}(\nu, \mathbf{R}), \mathbf{R}, \nu, \boldsymbol{\delta}\right)$$

and

$$P\left(\exists\ h : T_h < -t^{one}_{z,1-\alpha}(\nu, \mathbf{R})\right) = 1 - \mathcal{T}_z\left(-t^{one}_{z,1-\alpha}(\nu, \mathbf{R}), \infty, \mathbf{R}, \nu, \boldsymbol{\delta}\right).$$

Similar to the global power is the any-pair power i.e., the probability of finding at least one elementary comparison significant for which the corresponding contrast is truly under the alternative [48]. We can approximate it as

$$P\left(\exists\ h \in \mathcal{A} : |T_h| > t^{two}_{z,1-\alpha}(\nu, \mathbf{R})\right) = 1 - \mathcal{T}_{z^*}\left(-t^{two}_{z,1-\alpha}(\nu, \mathbf{R}), t^{two}_{z,1-\alpha}(\nu, \mathbf{R}), \mathbf{R}^*, \nu, \boldsymbol{\delta}^*\right)$$

with z^* being the number of elementary hypotheses that is under the alternative, $\mathbf{R}^*$ the submatrix of $\mathbf{R}$ containing the elements of $\mathbf{R}$ that correspond to the elementary hypotheses under the alternative, and $\boldsymbol{\delta}^*$ the subvector of $\boldsymbol{\delta}$ that contains only the z^* values referring to the set $\mathcal{A} = \{h : H^h_A : \eta_h \neq \delta_h\}$, which is the subset of contrasts that are truly under the alternative, and $1 \leq z^* \leq z$. In case all contrasts are under H_0, the any-pair power is set to α. The expressions for the one-sided analogues are similar to those for the global power.

For practical purposes we suppose the global power is more relevant; it is "contaminated" with elementary type I errors from contrasts that are truly under the null, but in practice these will be indistinguishable from true effects.

We provide two `R` functions `globalPower` and `anypairPower` to compute these powers in a supplementary file.

3 Simulation study

Proof that the "multiple marginal models" method controls the FWER asymptotically was given in [45], but its small-sample behaviour is best investigated in simulations. We looked into the type I error rates and powers of the approximate procedure that uses a multivariate t reference distribution with DF as in 2.3 and 2.4. All simulations were run in `R` version 3.1.3 [47], using the add-on packages `multcomp` [21] for the "multiple marginal models" procedure and `copula` [18] for copulas.

3.1 Small-sample test size

We studied balanced settings involving $q \in \{3, 4, 5\}$ randomised treatments, $m \in \{3, 4, 5\}$ repeated measurements, and $n_k \in \{3, 4, \ldots, 20\}$ subjects per group, with repeated observations being correlated and heteroscedastic as they would be e.g., in many longitudinal experiments. Random data for each treatment arm were drawn from an m-variate normal distribution $\mathcal{N}_m(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ with mean vector $\boldsymbol{\mu} = (10, \ldots, 10)$ and joint covariance matrix $\boldsymbol{\Lambda} = \mathbf{ABA}$ where $\mathbf{B} = (\rho_{jj'})$, $j \neq j'$ is an $m \times m$ Toeplitz correlation matrix with elements $\rho_{jj'} = 1 - \frac{|j-j'|}{10}$ that is pre- and postmultiplied by $\mathbf{A} = \text{diag}(\sqrt{1}, \sqrt{2}, \ldots, \sqrt{m})$. With $m = 4$, for instance, this generates

$$\mathbf{B} = \begin{bmatrix} 1 & 0.9 & 0.8 & 0.7 \\ & 1 & 0.9 & 0.8 \\ & & 1 & 0.9 \\ & & & 1 \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Lambda} = \begin{bmatrix} 1 & 1.27 & 1.39 & 1.40 \\ & 2 & 2.20 & 2.26 \\ & & 3 & 3.12 \\ & & & 4 \end{bmatrix}.$$

We simulated 5000 datasets under H_0 and carried out many-to-one (Dunnett), all-pairwise (Tukey), and grand-mean comparisons among treatment means per measured outcome, or among outcome means per treatment group, for two-sided hypotheses using the "multiple marginal models" method

a) with a multivariate t reference distribution and ν_{min} DF (equal to $\bar{\nu}$ in this case), or

b) with the asymptotic multivariate normal reference distribution,

and checked for each of them whether the smallest adjusted p-value was below $\alpha = 0.05$.

Figure 1 shows the simulated type I error rates for $q = 3$ treatments and $m = 3$ outcomes. When comparing multiple treatments per outcome measure using the asymptotic method, the realised α can be as high as 12% with $n_k = 4$ and even 17% with $n_k = 3$. Sample sizes of at least 15–20 are necessary for the test sizes to come reasonably close to the desired 5%, as suggested by [42]. The small-sample variant using ν^{min}_j on the other hand keeps the nominal FWER very well, with only minor α inflation (around 6%) for samples as small as $n_k = 3$.

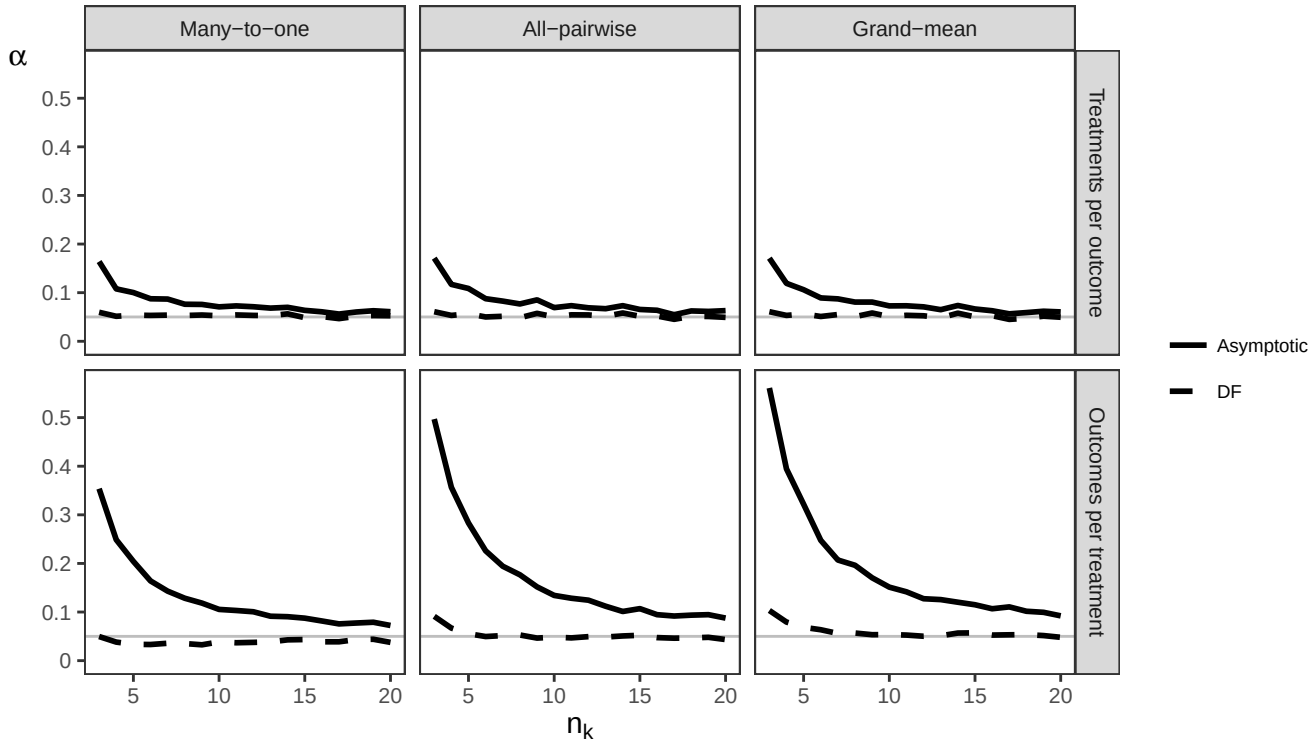

Figure 1: Simulated type I error rates for many-to-one, all-pairwise, and grand-mean comparisons among 3 Gaussian treatment means separately and simultaneously for 3 repeatedly measured outcomes (top row), or among 3 repeatedly measured Gaussian outcome means separately and simultaneously for 3 treatment groups (bottom row), with n_k independent subjects per treatment group (5000 simulation runs). The horizontal grey line is at $\alpha = 0.05$.

When comparing multiple outcome measure per treatment group, the asymptotic test procedure is essentially unusable, with type I error rates of 50% and more for $n_k = 3$. Even with seemingly reasonable sample sizes like $n_k = 20$, it is still noticeably liberal with α around 7–9%. In contrast, the small-sample version using ν_k^{min} keeps the nominal type I error level very well for $n_k > 5$ or, as in the case of the many-to-one test, is even slightly conservative with test sizes between 3.5 and 4.5%.

Further simulation results for $m, q > 3$ are shown in supplementary Figures S1 and S2. We also simulated a scenario with patient dropout (10% chance of missingness at the last time point, completely at random), but this led to essentially no worse performance in comparison to complete data (supplementary Figure S3).

3.2 Small-sample power

To assess the power advantage of the multiplicity adjustment using quantiles from a multivariate t reference distribution in comparison to Bonferroni, we simulated data in a similar way as in 3.1 but no longer under H_0. Instead we mimicked a treatment effect in one treatment group that arises for the last outcome measure only, which is the one with the largest variance. With $q = 3$ groups and $m = 3$ repeated outcomes we now have treatment means $\boldsymbol{\mu} = (10, 10, 10)$ for two outcomes and $\boldsymbol{\mu} = (10, 10, 10 + \delta)$ for the third outcome, where $\delta \in \{0, 0.2, \ldots, 4\}$. We generated 1000 datasets per δ for $n_k \in \{5, 10, 20\}$ and performed two-sided many-to-one, all-pairwise, and grand-mean comparisons of treatment means per outcome measure, or outcome means per treatment group, using a "multiple marginal models" analysis

a) with a multivariate t reference distribution and ν_{min} DF (equal to $\bar{\nu}$ in this case), or

b) with a Bonferroni-type correction that ignores any correlation between outcome measures.

Figure 2 shows the simulated power curves. We see that the adjustment based on the multivariate t distribution is, not surprisingly, more powerful than Bonferroni throughout. With $n_k = 5$ the power gain can be up to 15 percentage points for comparisons of treatment means and even more than 20 percentage points for comparisons of outcome means. This advantage is somewhat lower as n_k increases, but still up to around 10 percentage points for $n_k = 20$ (the steepness of these curves makes the vertical distances between them look negligible where in reality they are not).

3.3 Small-sample test size for the duplex procedure

We simulated 5000 datasets under H_0 as detailed in 3.1 with $q = 3$ treatment groups and $m = 3$ repeated measurements and applied the duplex procedure of 2.4 using

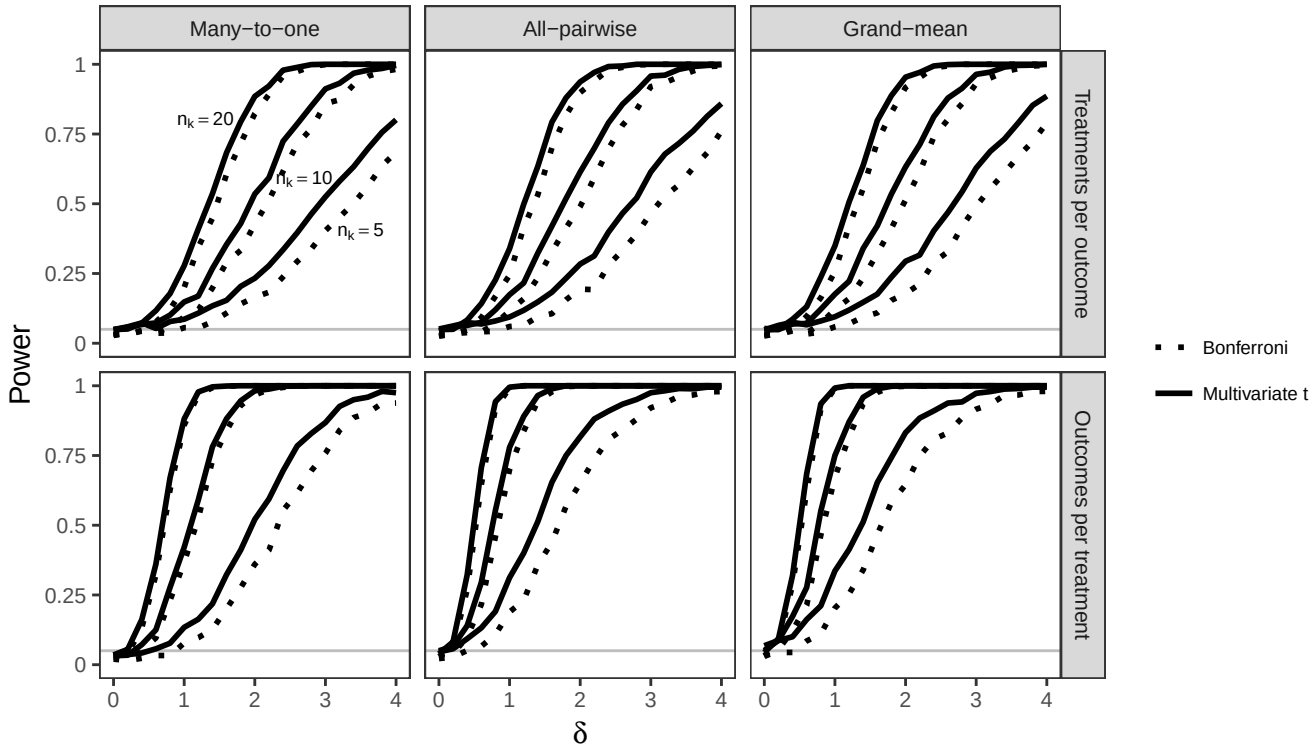

Figure 2: Simulated powers for many-to-one, all-pairwise, and grand-mean comparisons among 3 Gaussian treatment means separately and simultaneously for 3 repeatedly measured outcomes (top row), or among 3 repeatedly measured Gaussian outcome means separately and simultaneously for 3 treatment groups (bottom row), with $n_k \in \{5, 10, 20\}$ independent subjects per treatment group (1000 simulation runs). The horizontal grey line is at $\alpha = 0.05$.

a) a multivariate t reference distribution with either ν_{min} or $\bar{\nu}$,

b) separate multivariate t reference distributions providing comparison-specific critical values based on comparison-specific DFs $\nu_1, \ldots, \nu_z$, or

c) a joint multivariate reference distribution composed of univariate t distributions with comparison-specific DFs, linked by a normal or t copula (using the average $\bar{\nu}$ for the t copula itself).

Figure 3 shows that the average DF leads to clearly too many rejections of H_0, even for sample sizes of $n_k \geq 10$. This problem is attenuated with the minimum DF, at the cost of making the procedure slightly conservative. Using comparison-specific critical values from reference distributions with variable DFs is a compromise that seems to work well unless the sample size becomes too small ($n_k < 5$). In that case only the copula-based approach maintains FWER control, but it also has a tendency to be conservative, a little more so with the normal copula than with the t copula.

4 Data example: heart rates

Milliken and Johnson [38] described a placebo-controlled clinical trial that assessed the efficacy of two novel drugs, AX23 and BWW9. 24 women were randomly assigned to one of the active drug arms or control, resulting in three groups of eight subjects each. The clinical endpoint of interest was the heart rate, measured for each woman at four subsequent occasions every five minutes. Heart rate patterns over time are quite different among the three treatments (Figure 4). The rates seem to be slightly decreasing over time for control and BWW9 whereas AX23 leads to much higher rates at the second and third time point compared to the first and last. The difference between BWW9 and control is relatively constant over time. We also see that BWW9 exhibits the smallest and control the largest variability at all time points. Correlations are high between adjacent time points, especially in the control arm, and tend to decrease with greater separation in time (Figure 5), which translates to the correlations of test statistics shown in supplementary Figure S5.

One research question we can address with this dataset is: which treatments increase the heart rate compared to control at any particular time point, and by how much? This suggests comparing the treatment arms separately and simultaneously for each of the four measurement times. Another research goal could be to find out for each treatment arm whether there are relevant differences of heart rates over time, which calls for comparisons between time points separately and simultaneously for each treatment. And finally, we can combine both questions into one "claim" that we investigate whilst controlling a joint type I error rate. The sample size here is clearly too small to rely on asymptotics

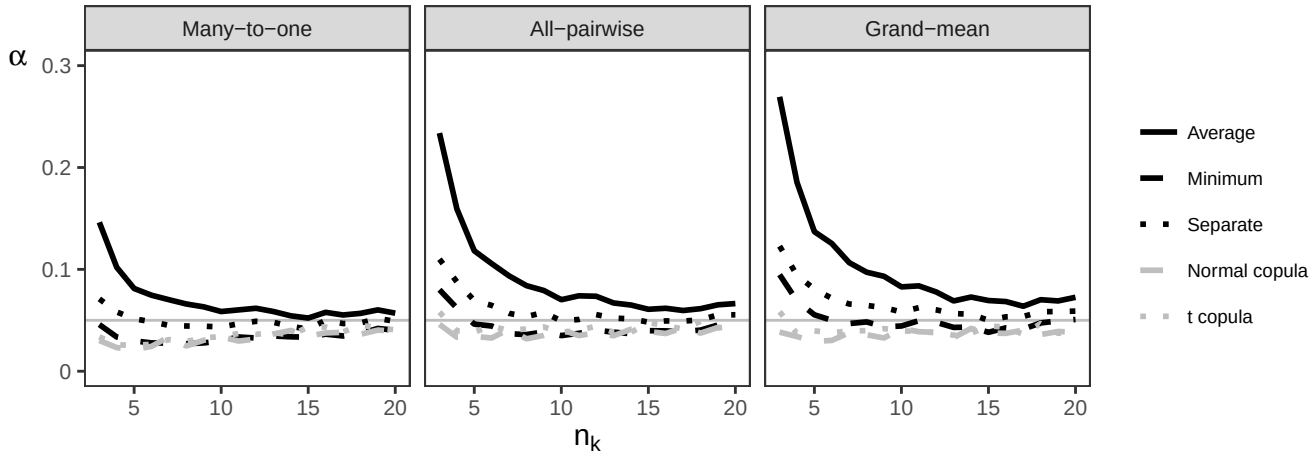

Figure 3: Simulated type I error rates for many-to-one, all-pairwise, and grand-mean comparisons among 3 Gaussian treatment means separately and simultaneously for 3 repeatedly measured outcomes and among 3 repeatedly measured Gaussian outcome means separately and simultaneously for 3 treatment groups, with n_k independent subjects per treatment group and using average, minimum, or separate comparison-specific degrees of freedom, or a copula-generated reference distribution (5000 simulation runs). The horizontal grey line is at $\alpha = 0.05$.

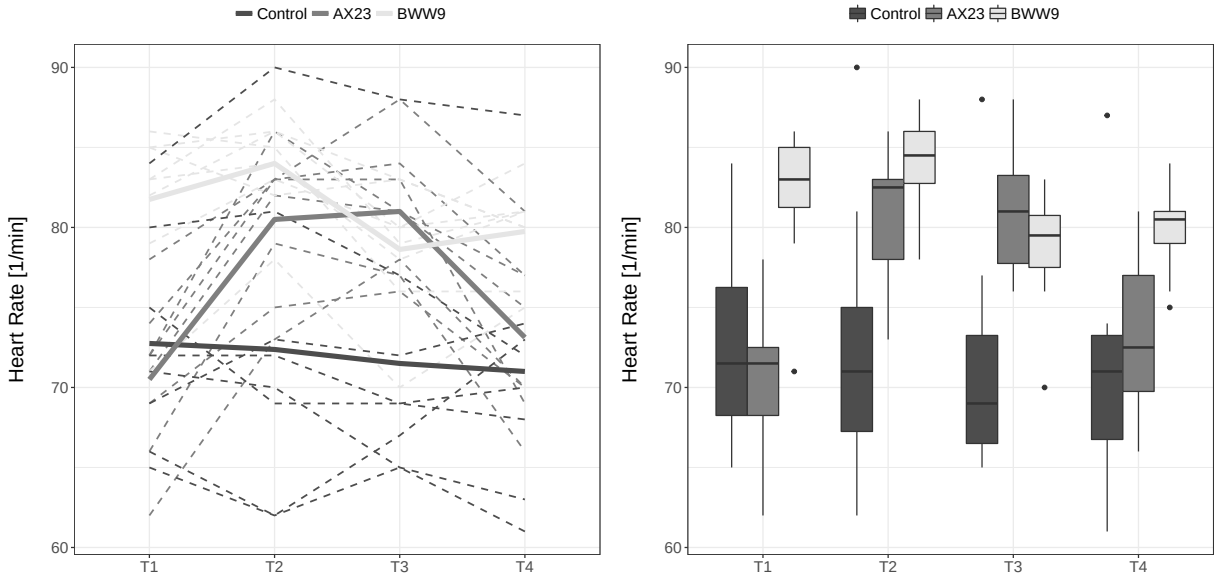

Figure 4: Heart rate data. Left: individual patient trajectories (dashed) and sample mean trajectories per treatment arm (solid) of heart rates; right: boxplots.

Figure 5: Heart rate data: correlation between time points for treatment arms AX23 (left), BWW9 (centre), and control (right).

[42], thus we make use of the small-sample procedures described in Section 2. `R` code to reproduce the analyses is

provided in a supplementary file.

4.1 Comparing multiple drugs simultaneously at multiple time points

Our first question is: when do which active treatments differ (significantly) from control, and by how much? To answer this we perform many-to-one comparisons of treatment means separately and simultaneously for each of the four time points. We use $\nu_j = 21$ as a DF approximation; as the design of the study is balanced, ν_j is the same for all m marginal models and hence also the same as ν_j^{min} and $\bar{\nu}_j$. For comparison purposes we include the asymptotic test procedure as well as a Bonferroni-corrected version that ignores all correlation between time points.

Table 1: Simultaneous inference for the heart rate data: estimated differences of heart rates, standard errors, and adjusted p-values for many-to-one comparisons of AX23 and BWW9 against control separately and simultaneously per time point.

	Estimate	SE	p(DF)	p(asym)	p(Bonf)
T1: AX23 - control	-2.250	2.762	0.897	0.901	1.000
T1: BWW9 - control	9.000	2.762	0.018	0.007	0.030
T2: AX23 - control	8.125	3.132	0.074	0.047	0.135
T2: BWW9 - control	11.625	3.132	0.007	0.001	0.010
T3: AX23 - control	9.500	2.794	0.013	0.004	0.022
T3: BWW9 - control	7.125	2.794	0.081	0.053	0.149
T4: AX23 - control	2.125	2.859	0.927	0.932	1.000
T4: BWW9 - control	8.750	2.859	0.028	0.014	0.047

The summarised results in Table 1 show that AX23 leads to a significantly increased heart rate only at T3 whereas the increase in the BWW9 group is significant at all time points except T3—this is where the mean profiles of AX23 and BWW9 cross paths (Figure 4). The simultaneous 95% CIs in Figure 6 visualise the actual magnitude of the estimated treatment effects, revealing that even for the significant comparisons the CI bounds are rather close to zero.

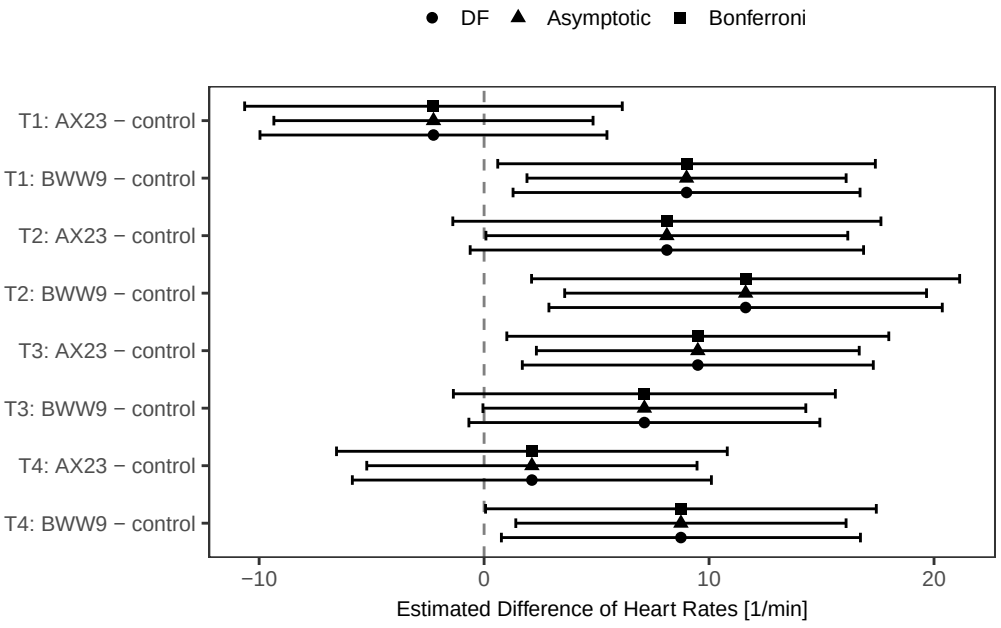

Figure 6: Heart rate data: 95% simultaneous confidence intervals for many-to-one comparisons of AX23 and BWW9 against control separately and simultaneously per time point. DF: using a multivariate t reference distribution with a small-sample approximation to the degrees of freedom; asymptotic: using a multivariate normal reference distribution; Bonferroni: using a Bonferroni adjustment ignoring correlation across time points.

Using the asymptotic normal reference distribution—which is clearly inadequate with so few patients—makes the adjusted p-values unduly small and the simultaneous CIs too narrow. By contrast, ignoring the correlation and adjusting with Bonferroni yields unnecessarily large p-values and wide simultaneous CIs.

4.2 Comparing multiple time points simultaneously for multiple drugs

Our second question is: by how much do the heart rates differ across time points within each treatment arm? To this end we carry out all-pairwise comparisons of time point means separately and simultaneously for each of the three

treatments. We approximate the DF as $\nu_k = 7$, which equals both ν_k^{min} and $\bar{\nu}_k$ due to the balanced study design. Again we include the asymptotic test procedure and a Bonferroni version for comparison purposes.

Table 2: Simultaneous inference for the heart rate data: estimated differences of heart rates, standard errors, and adjusted p-values for all-pairwise comparisons of T1, T2, T3, and T4 separately and simultaneously per treatment group.

	Estimate	SE	p(DF)	p(asym)	p(Bonf)
AX23: T2 - T1	10.000	1.417	0.002	<0.001	0.004
AX23: T3 - T1	10.500	1.066	<0.001	<0.001	<0.001
AX23: T4 - T1	2.625	1.324	0.504	0.429	1.000
AX23: T3 - T2	0.500	1.673	1.000	1.000	1.000
AX23: T4 - T2	-7.375	1.924	0.056	0.002	0.116
AX23: T4 - T3	-7.875	1.319	0.006	<0.001	0.010
BWW9: T2 - T1	2.250	1.532	0.779	0.799	1.000
BWW9: T3 - T1	-3.125	1.411	0.395	0.280	1.000
BWW9: T4 - T1	-2.000	2.217	0.976	0.988	1.000
BWW9: T3 - T2	-5.375	1.943	0.207	0.073	0.501
BWW9: T4 - T2	-4.250	1.925	0.398	0.283	1.000
BWW9: T4 - T3	1.125	1.544	0.993	0.997	1.000
Control: T2 - T1	-0.375	0.816	1.000	1.000	1.000
Control: T3 - T1	-1.250	1.009	0.884	0.914	1.000
Control: T4 - T1	-1.750	1.613	0.937	0.960	1.000
Control: T3 - T2	-0.875	0.778	0.925	0.951	1.000
Control: T4 - T2	-1.375	1.438	0.967	0.982	1.000
Control: T4 - T3	-0.500	0.891	0.999	1.000	1.000

The heart rate in the AX23 arm is significantly higher at T2 and T3 compared to T1, as well as at T3 compared to T4, and the comparison between T2 and T4 is just marginally non-significant (Table 2), reflecting the up-and-down pattern we observed in Figure 4. In particular for the comparisons of T2 and T3 against T1 the simultaneous CI bounds are far from zero (Figure 7). No significant time effects can be detected for BWW9 or control, whose mean profiles are relatively constant over time. Again the Bonferroni-adjusted CIs are needlessly wide, and those based on the asymptotic procedure too narrow.

4.3 Comparing multiple drugs and multiple time points

Finally, we can combine the many-to-one comparisons of treatment groups and the all-pairwise comparisons of time points into one set of hypotheses to be assessed under joint FWER control. The DF for comparisons among treatments is $\nu_j = 21$, and for comparisons among time points it is $\nu_k = 7$. As there are no missing values, the DF is the same for all comparisons of each type. We apply the duplex procedure with the three DF approximations discussed in 2.4.

The results of this analysis are summarised in supplementary Table S1 and supplementary Figure S6. Using the minimum DF $\nu^{min} = 7$ and also the weighted average DF $\bar{\nu} = 11$ makes the simultaneous CIs for the comparisons between treatments wider than with separate DFs where $\nu_j = 21$. For the comparisons between time points, the simultaneous CIs using ν^{min} and separate ν_k's are identical because the DFs are; the intervals using $\bar{\nu}$, however, are too narrow, but still wider than the asymptotic ones that we included for comparison.

Compared to the previous analyses in 4.1 and 4.2, the larger set of hypotheses makes the simultaneous CIs in supplementary Figure S6 wider than those in Figures 6 and 7. As a consequence, we can no longer claim that BWW9 raises the heart rate significantly at the first and last time point. This conservatism is the price we pay for controlling the FWER at 5% over comparisons of treatment arms as well as time points simultaneously, which we have done here merely for illustrative purposes. In practice one will have to weigh up whether FWER control over a large set of hypotheses is necessary and meaningful [33].

5 Discussion

The "multiple marginal models" method is extremely convenient for simultaneous inference in repeated-measures or multi-endpoint settings as it obviates the need for a complicated joint model; instead, a set of much simpler models are fitted to different portions of the data, and the covariances between parameters from the different models are subsequently recovered. To make this applicable to small-scale studies as well, we have shown in this paper that using a multivariate t reference distribution with some appropriate DF yields an approximate procedure that is able to control the FWER even for sample sizes as small as $n_k = 5$ per treatment arm. Adjusted p-values and informative simultaneous CIs are readily available in `R`.

The method allows for comparisons among treatment levels per time point or endpoint, among time points or endpoints per treatment arm, and even for comparisons of both types within the same "claim" [44] if that is of interest. Especially for this latter type of comparisons, but also when the data are notably unbalanced, we recommend using separate DFs per elementary comparison and thereby also making the critical values comparison-specific. They were shown to work well in our simulations and also by [16] and [15], but are admittedly somewhat *ad hoc*. As an alternative, a joint multivariate reference distribution with comparison-specific marginal DFs can be constructed using a normal or t copula; this ensured FWER control even for very small sample sizes ($n_k \leq 5$) in our simulations. The minimum DF is a safe and simple workaround that makes the test procedure at worst a bit conservative. Using the average DF,

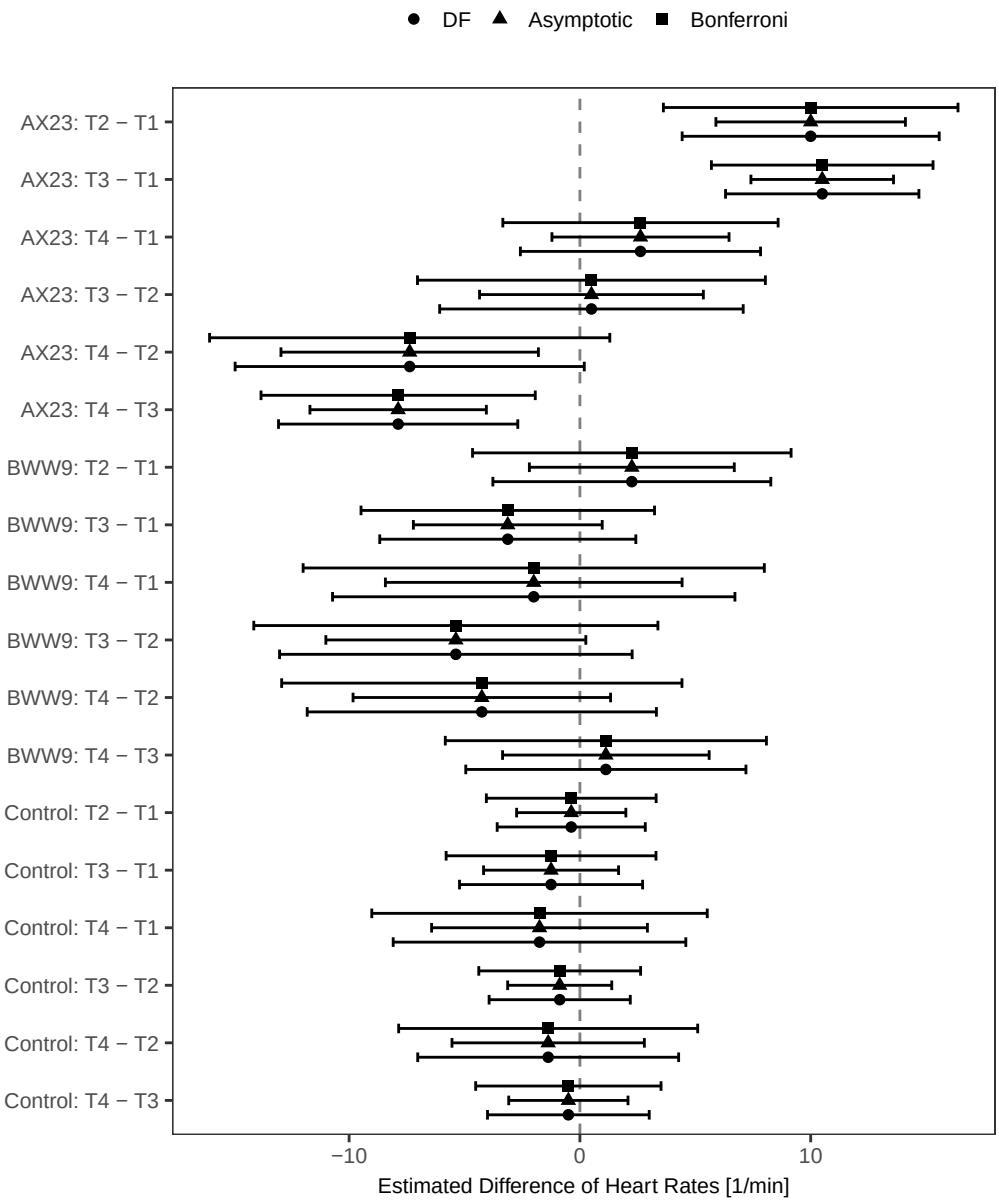

Figure 7: Heart rate data: 95% simultaneous confidence intervals for all-pairwise comparisons of T1, T2, T3, and T4 separately and simultaneously per treatment group. DF: using a multivariate t reference distribution with a small-sample approximation to the degrees of freedom; asymptotic: using a multivariate normal reference distribution; Bonferroni: using a Bonferroni adjustment ignoring correlation across time points.

however, is a poor idea as it assigns an unduly large DF to some of the contrasts, thus inflating the size of the overall maximum test. We wonder if this is also a potential problem in other procedures where DFs are routinely averaged over—as with `lsmeans` in `R` [32] using average Kenward-Roger DFs—but this shall be the object of future studies. All these considerations would probably become obsolete if one could express a multivariate t density with variable marginal DFs and arbitrary correlation in closed form, but to date this has only been accomplished for very restrictive special cases e.g., bivariate with zero correlation [54].

Our procedure will break down with multivariate data where the number of endpoints exceeds the number of independent experimental units by far. One flexible alternative, especially for such high-dimensional scenarios, is rotation tests [30]. They have recently gained some attention for applications in high-dimensional gene expression analysis [4, 63, 5, 56] but are also useful with smaller datasets in low dimensions [31]. Like the "multiple marginal models" approach, rotation tests do not require any specification of the covariance and provide adjusted p-values as well as informative simultaneous CIs. Similar to resampling methods, however, they need a large number of replications for the results to stabilise and become reliable.

We have focused on modelling continuous data here but the "multiple marginal models" method is applicable to

other data types as well. When the outcome is binary or a count variable, the marginal models are GLMs. We suppose in this case the asymptotic procedure can be tuned in a similar fashion to the small-sample approximation for continuous data, by using the residual DFs of the marginal GLMs. The same should be true when the marginal models are linear mixed-effects models, as in longitudinal multi-endpoint studies [25, 49].

Despite its great flexibility and practical usefulness, two limitations of the "multiple marginal models" procedure should not go unmentioned: first, it can handle missing data only if they are missing completely at random i.e., the probability to be missing must depend neither on the observed nor unobserved values [35], whereas missingness at random is usually sufficient with a joint mixed-effects model. In the presence of missing data, not using a joint model may entail a slight loss in efficiency, as separate "marginal" models cannot "borrow strength" from one another. And second, "multiple marginal models" are convenient for obtaining adjusted p-values and simultaneous CIs, but whenever variance components or random effects are to be estimated and interpreted, only a joint mixed-effects model will serve the purpose.

Acknowledgements

We would like to thank the Associate Editor and two anonymous referees for some very helpful suggestions, especially with regard to the simulation study.

References

[1] PS Albert. Longitudinal data analysis (repeated measures) in clinical trials. *Stat Med*, 18(13):1707–1732, 1999.

[2] F Bretz, T Hothorn, and P Westfall. *Multiple Comparisons Using R*. Chapman & Hall/CRC, Boca Raton, FL, 2010.

[3] A Cnaan, NM Laird, and P Slasor. Using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data. *Stat Med*, 16(20):2349–2380, 1997.

[4] G Dørum, L Snipen, M Solheim, and S Sæbø. Rotation testing in gene set enrichment analysis for small direct comparison experiments. *Stat Appl Genet Mol Biol*, 8(1):article 34, 2009.

[5] G Dørum, L Snipen, M Solheim, and S Sæbø. Rotation gene set testing for longitudinal expression data. *Biom J*, 56(6):1055–1075, 2014.

[6] CW Dunnett. A multiple comparison procedure for comparing several treatments with a control. *J Am Stat Assoc*, 50(272):1096–1121, 1955.

[7] BS Everitt. The analysis of repeated measures: a practical review with examples. *J R Stat Soc Ser D Statistician*, 44(1):113–135, 1995.

[8] D Follmann. Multivariate tests for multiple endpoints in clinical trials. *Stat Med*, 14(11):1163–1175, 1995.

[9] L Frison and SJ Pocock. Repeated measures in clinical trials: analysis using mean summary statistics and its implications for design. *Stat Med*, 11(13):1685–1704, 1992.

[10] L Frison and SJ Pocock. Linearly divergent treatment effects in clinical trials with repeated measures: efficient analysis using summary statistics. *Stat Med*, 16(24):2855–2872, 1997.

[11] PA Games and JF Howell. Pairwise multiple comparison procedures with unequal n's and/or variances: a Monte Carlo study. *J Educ Stat*, 1(2):113–125, 1976.

[12] MJ Gardner and DG Altman. Confidence intervals rather than p values: estimation rather than hypothesis testing. *Br Med J*, 292(6522):746–750, 1986.

[13] M Große Ruse, C Ritz, and LA Hothorn. Simultaneous inference of a binary composite endpoint and its components. *J Biopharm Stat*, 27(1):56–69, 2017.

[14] MJ Gurka, LJ Edwards, and KE Muller. Avoiding bias in mixed model inference for fixed effects. *Stat Med*, 30(22):2696–2707, 2011.

[15] M Hasler. Heteroscedasticity: multiple degrees of freedom vs. sandwich estimation. *Stat Pap*, 57(1):55–68, 2016.

[16] M Hasler and LA Hothorn. Multiple contrast tests in the presence of heteroscedasticity. *Biom J*, 50(5):793–800, 2008.

[17] AJ Hayter and W Liu. A method of power assessment for tests comparing several treatments with a control. *Commun Stat Theory Methods*, 21(7):1871–1889, 1992.

[18] M Hofert, I Kojadinovic, M Maechler, and J Yan. copula: Multivariate dependence with copulas. R package version 0.999-18, 2017.

[19] WP Hoffman, J Recknor, and C Lee. Overall type I error rate and power of multiple Dunnett's tests on rodent body weights in toxicology studies. *J Biopharm Stat*, 18(5):883–900, 2008.

[20] T Hothorn, F Bretz, and P Westfall. Simultaneous inference in general parametric models. *Biom J*, 50(3):346–363, 2008.

[21] T Hothorn, F Bretz, P Westfall, RM Heiberger, A Schuetzenmeister, and S Scheibe. multcomp: Simultaneous inference in general parametric models. R package version 1.4-7, 2017.

[22] H Huynh. Some approximate tests for repeated measurement designs. *Psychometrika*, 43(2):161–175, 1978.

[23] H Jacqmin-Gadda, S Sibillot, C Proust, JM Molina, and R Thiébaut. Robustness of the linear mixed model to misspecified error distribution. *Comput Stat Data Anal*, 51(10):5142–5154, 2007.

[24] RI Jennrich and MD Schluchter. Unbalanced repeated-measures models with structured covariance matrices. *Biometrics*, 42(4):805–820, 1986.

[25] SM Jensen, CB Pipper, and C Ritz. Evaluation of multi-outcome longitudinal studies. *Stat Med*, 34(12):1993–2003, 2015.

[26] SM Jensen and C Ritz. Simultaneous inference for model averaging of derived parameters. *Risk Anal*, 35(1):68–76, 2015.

[27] HJ Keselman, J Algina, RK Kowalchuk, and RD Wolfinger. A comparison of two approaches for selecting covariance structures in the analysis of repeated measurements. *Commun Stat Simul Comput*, 27(3):591–604, 1998.

[28] A Kitsche, C Ritz, and LA Hothorn. A general framework for the evaluation of genetic association studies using multiple marginal models. *Hum Hered*, 81(3):150–172, 2016.

[29] S Kotz and S Nadarajah. *Multivariate t Distributions and Their Applications*. Cambridge University Press, Cambridge, UK, 2004.

[30] Ø Langsrud. Rotation tests. *Stat Comput*, 15(1):53–60, 2005.

[31] Ø Langsrud, K Jørgensen, R Ofstad, and T Næs. Analyzing designed experiments with multiple responses. *J Appl Stat*, 34(10):1275–1296, 2007.

[32] RV Lenth. Least-squares means: the R package lsmeans. *J Stat Softw*, 69(1):1–33, 2016.

[33] G Li, M Taljaard, ER van den Heuvel, MAH Levine, DJ Cook, GA Wells, PJ Devereaux, and L Thabane. An introduction to multiplicity issues in clinical trials: the what, why, when and how. *Int J Epidemiol*, 46(2):746–755, 2007.

[34] RC Littell, J Pendergast, and R Natarajan. Modelling covariance structure in the analysis of repeated measures data. *Stat Med*, 19(13):1793–1819, 2000.

[35] RJA Little and DB Rubin. *Statistical Analysis with Missing Data*. John Wiley & Sons, New York, NY, 2002.

[36] C Liu, TP Cripe, and MO Kim. Statistical issues in longitudinal data analysis for treatment efficacy studies in the biomedical sciences. *Mol Ther*, 18(9):1724–1730, 2010.

[37] K Lu and DV Mehrotra. Specification of covariance structure in longitudinal data analysis for randomized clinical trials. *Stat Med*, 29(4):474–488, 2010.

[38] GA Milliken and DE Johnson. *Analysis of Messy Data. Volume I: Designed Experiments*. Chapman & Hall, London, UK, 1992.

[39] RB Nelsen. *An Introduction to Copulas. Second Edition*. Springer, New York, NY, 2006.

[40] RZ Omar, EM Wright, RM Turner, and SG Thompson. Analysing repeated measurements data: a practical comparison of methods. *Stat Med*, 18(13):1587–1603, 1999.

[41] P Pallmann and LA Hothorn. Analysis of means: a generalized approach using R. *J Appl Stat*, 43(8):1541–1560, 2016.

[42] P Pallmann, M Pretorius, and C Ritz. Simultaneous comparisons of treatments at multiple time points: combined marginal models versus joint modeling. *Stat Methods Med Res*, 2015.

[43] T Park, JK Park, and CS Davis. Effects of covariance model assumptions on hypothesis tests for repeated measurements: analysis of ovarian hormone data and pituitary-pteryomaxillary distance data. *Stat Med*, 20(16):2441–2453, 2001.

[44] A Phillips, C Fletcher, G Atkinson, E Channon, A Douiri, T Jaki, J Maca, D Morgan, JH Roger, and P Terrill. Multiplicity: discussion points from the Statisticians in the Pharmaceutical Industry multiplicity expert group. *Pharm Stat*, 12(5):255–259, 2013.

[45] CB Pipper, C Ritz, and H Bisgaard. A versatile method for confirmatory evaluation of the effects of a covariate in multiple models. *J R Stat Soc Ser C Appl Stat*, 61(2):315–326, 2012.

[46] H Quan, X Luo, and T Capizzi. Multiplicity adjustment for multiple endpoints in clinical trials with multiple doses of an active treatment. *Stat Med*, 24(14):2151–2170, 2005.

[47] R Core Team. R: a language and environment for statistical computing, 2015.

[48] PH Ramsey. Power differences between pairwise multiple comparisons. *J Am Stat Assoc*, 73(363):479–485, 1978.

[49] C Ritz, R Pilmann Laursen, and C Trab Damsgaard. Simultaneous inference for multilevel linear mixed models—with an application to a large-scale school meal study. *J R Stat Soc Ser C Appl Stat*, 66(2):295–311, 2017.

[50] AJ Sankoh, MF Huque, and SD Dubey. Some comments on frequently used multiple endpoint adjustment methods in clinical trials. *Stat Med*, 16(22):2529–2542, 1997.

[51] AJ Sankoh, MF Huque, HK Russell, and RB D'Agostino Sr. Global two-group multiple endpoint adjustment methods applied to clinical trials. *Drug Inf J*, 33(1):119–140, 1999.

[52] H Schielzeth and W Forstmeier. Conclusions beyond support: overconfident estimates in mixed models. *Behav Ecol*, 20(2):416–420, 2009.

[53] S Senn, L Stevens, and N Chaturvedi. Repeated measures in clinical trials: simple strategies for analysis using summary measures. *Stat Med*, 19(6):861–877, 2000.

[54] WT Shaw and KTA Lee. Bivariate Student t distributions with variable marginal degrees of freedom and independence. *J Multivar Anal*, 99(6):1276–1287, 2008.

[55] Q Shi, ES Pavey, and RE Carter. Bonferroni-based correction factor for multiple, correlated endpoints. *Pharm Stat*, 11(4):300–309, 2012.

[56] A Solari, L Finos, and JJ Goeman. Rotation-based multiple testing in the multivariate linear model. *Biometrics*, 70(4):954–961, 2014.

[57] JW Tukey. The problem of multiple comparisons. In HI Braun, editor, *The Collected Works of John W. Tukey, Volume VIII*, pages 1–300. Chapman & Hall, New York, NY, 1994.

[58] G Verbeke and G Molenberghs. *Linear Mixed Models for Longitudinal Data*. Springer, New York, NY, 2000.

[59] G Wassmer, P Reitmeir, M Kieser, and W Lehmacher. Procedures for testing multiple endpoints in clinical trials: an overview. *J Stat Plan Inference*, 82(1-2):69–81, 1999.

[60] PH Westfall, RD Tobias, and RD Wolfinger. *Multiple Comparisons and Multiple Tests Using SAS. Second Edition*. SAS Institute, Cary, NC, 2011.

[61] R Wolfinger. Covariance structure selection in general mixed models. *Commun Stat Simul Comput*, 22(4):1079–1106, 1993.

[62] RD Wolfinger. Heterogeneous variance: covariance structures for repeated measures. *J Agric Biol Environ Stat*, 1(2):205–230, 1996.

[63] D Wu, E Lim, F Vaillant, ML Asselin-Labat, JE Visvader, and GK Smyth. ROAST: rotation gene set tests for complex microarray experiments. *Bioinformatics*, 26(17):2176–2182, 2010.

[64] FB Yoon, GM Fitzmaurice, SR Lipsitz, NJ Horton, NM Laird, and SLT Normand. Alternative methods for testing treatment effects on the basis of multiple outcomes: Simulation and case study. *Stat Med*, 30(16):1917–1932, 2011.

[65] B Zou, B Jin, GG Koch, H Zhou, SE Borst, S Menon, and JJ Shuster. On model selections for repeated measurement data in clinical studies. *Stat Med*, 34(10):1621–1633, 2015.

Supporting information

Additional Supporting Information may be found online in the supporting information tab for this article.