Penultimate version. If citing, please refer instead to the published version in Statistical Methods in Medical Research, DOI: 10.1177/0962280215603743.

Simultaneous comparisons of treatments at multiple time points: combined marginal models versus joint modeling

Philip Pallmann[1*], Mias Pretorius[2], Christian Ritz[3]

[1]Institute of Biostatistics, Leibniz University Hannover, Germany
[2]Department of Anesthesiology, Vanderbilt University Medical School, Nashville, TN
[3]Department of Nutrition, Exercise and Sports, University of Copenhagen, Denmark

Correspondence to: Philip Pallmann, Institute of Biostatistics, Leibniz University Hannover, Herrenhaeuser Strasse 2, 30419 Hannover, Germany. Phone: +49 511 762 5821. Fax: +49 511 762 4966. E-mail: pallmann@biostat.uni-hannover.de

Abstract

We discuss several aspects of multiple inference in longitudinal settings, focusing on many-to-one and all-pairwise comparisons of a) treatment groups simultaneously at several points in time, or b) time points simultaneously for several treatments. We assume a continuous endpoint that is measured repeatedly over time and contrast two basic modeling strategies: fitting a joint model across all occasions (with random effects and/or some residual covariance structure to account for heteroscedasticity and serial dependence), and a novel approach combining a set of simple marginal i.e., occasion-specific models. Upon parameter and covariance estimation with either modeling approach, we employ a variant of multiple contrast tests (MCTs) that acknowledges correlation between time points and test statistics. This method provides simultaneous confidence intervals (SCIs) and adjusted p-values for elementary hypotheses as well as a global test decision. We compare via simulation the powers of MCTs based on a joint model and multiple marginal models, respectively, and quantify the benefit of incorporating longitudinal correlation i.e., the advantage over Bonferroni. Practical application is illustrated with data from a clinical trial on bradykinin receptor antagonism.

Keywords: *longitudinal data, repeated measurements, generalized least squares, linear mixed-effects model, AICc*

1 Introduction

Trials involving longitudinally repeated measurements may pursue different goals. If subject-matter interest centers on an outcome's development in the course of time, linear and nonlinear mixed-effects models[1,2] are often the first choice for analysis, and inference is commonly focused on a summarizing parameter (e.g., the slope). Or to keep things simple, using a summary measure like the area under the curve (AUC) may be conceivable as well. By contrast, if a researcher is interested in simultaneous inferences for (a few) clearly specified points in time, these approaches are of limited avail.
For such an endeavor to make sense, the measurement occasions should preferably not be picked at haphazard but rather have an inherent medical relevance. In a clinical application, some endpoint could be assessed pre- and post-surgery and after one month of convalescence. Similar scenarios with meaningfully defined measurement times arise in many related life sciences such as pharmacology, toxicology, and more.

If one derives a family of statistical hypotheses that represent the research questions well, simultaneous inference is usually performed upon condition that the familywise error rate (FWER) is controlled at some predefined level α. In the case of a longitudinal setting, such a family – or "claim"[3] – may embrace individual hypotheses of two basic types: i) comparisons of estimated treatment means, separately and simultaneously at several points in time; and ii) comparisons of estimated means of time points, separately and simultaneously for several treatment groups. For instance, one may wish to investigate which treatments are better than control at certain occasions, leading to a set of many-to-one comparisons of treatments[4]. Just as well, one could be interested in the measurement occasion associated with the highest value of the outcome in each treatment group, which calls for all-pairwise comparisons of time points[5].

Traditional methodology for inference across multiple occasions includes repeated measures ANOVA, which comes with harsh assumptions such as constant variance over time and sphericity (i.e., equicorrelation) that are rarely met in longitudinal trials, and multivariate ANOVA. Both of them cannot cope with missing values properly, and omitting all individuals with incomplete observations is undesirable. In addition, the associated tests are geared to global inference and therefore unrewarding for e.g., comparisons of several treatments to a common control at multiple occasions.

This situation rather demands many-to-one tests applied to each of several points in time, which naturally acknowledges the multiplicity of treatment groups but leaves open how multiplicity of occasions should be accounted for. All too often practical researchers turn a blind eye to this problem (thus inflating the overall type I error rate), or they apply a simple but conservative Bonferroni adjustment, thereby impairing the power to detect effects. In recent years, statisticians have sought ways out of this dilemma; however, many of their solutions are crude and clumsy, or limited to specialized applications[6,7].

Multiple contrast tests (MCT) as outlined by Mukerjee et al.[8] and Bretz et al.[9] and generalized in Hothorn et al.[10] offer a versatile framework that can be the key to precise yet powerful comparisons in longitudinal designs. They capture dependencies among test statistics and allow for localized inferences (in terms of multiplicity-adjusted p-values and compatible simultaneous confidence intervals (SCIs)) as well as a global statement of significance for the entire "claim". When operating with MCTs in longitudinal contexts, we must address the correlation among repeated measurements to make tests considerably sharper than Bonferroni.

Recently, MCTs were particularized for several problems related to ours. Hasler and Hothorn[11,12] developed MCTs for multiple endpoints, which can be applied to our longitudinal setting directly by treating occasions as endpoints. However, this is nonsatisfying from several different points of view. First, it only allows for comparisons among treatments per time point but not among time points per treatment. Second, the underlying model makes the procedure fairly inflexible as regards dealing with covariates and missing values. And third, covariances of endpoints are always modeled as heteroscedastic (variances) and unstructured (correlation), either for all treatment groups or even separately for each treatment[13]; in a scenario involving a mere four treatments and five occasions, say, there are 60 covariance parameters to estimate! This may be undesirable (if not unfeasible) especially with small sample sizes, and sparser modeling should be aspired to.

Hasler[14] discussed MCTs for comparing means at subsequent points in time, but his method is limited to comparing occasions in one treatment group only and raises similar difficulties as the multi-endpoint MCTs.

The goal of this paper is to establish a much more versatile MCT approach on the basis of models that cope with typical challenges of longitudinal data such as correlated observations, heterogeneous variances over time, and missing values (e.g., due to dropout). In particular, we study the applicability of joint models (focusing on extended linear models and conditional independence models embracing all occasions) in contrast to a combination of occasion-specific marginal linear models.

Section 2 describes different approaches to model building that pave the way for flexible simultaneous inference in longitudinal settings using MCTs. Simulation results for type I error and power are summarized in Section 3. We analyze a clinical data example in Section 4. A discussion in Section 5 concludes the paper.

2 Methods

We strive to compare several treatments at multiple occasions, or several occasions within multiple treatment groups, under strong FWER control, with the measurements being mutually independent in respect of treatments but serially correlated in respect of time points. Our MCT strategy crucially depends on finding valid estimates for the mean parameters and their covariances. We will deal with two basic types of models for multiple correlated measurements over time:

1. A well-balanced joint model over all measurement times. By well-balanced, we mean that the complexity of the fitted model covariance structure is governed by whether the data basis is meager or rich. An AIC-like criterion may assist in choosing a particular structure from several candidates. The joint model will typically be a linear mixed-effects model (LMM) with random subject effects or an extended linear model that allows to impose some (parsimonious) pattern on the residual covariance matrix (Section 2.1). Variations of this joint modeling strategy have become a standard tool for many applied statisticians.
2. A multiple marginal models approach. Repeated measurements across a number of occasions are treated as multivariate correlated endpoints, but univariate linear models are fitted, one for each occasion. Subsequently, these univariate model fits are combined to allow to devise covariances between estimates from the different models, and carry out inference across occasions and treatments (Section 2.2). We show that this technique performs comparable to joint modeling but has major practical advantages e.g., one need not bother about the structure of random effects and/or residual covariance.

Then according to the elementary hypotheses of interest we calculate a collection of test statistics whose (asymptotic) joint distribution we approximate as multivariate normal in order to compute adjusted p-values and SCIs (Section 2.3). The problem of missing values is discussed in Section 2.4.

2.1 Joint model

We define a random variable Y_{jki} denoting some Gaussian outcome measured from individuals $i = 1, \ldots, n_k$ in treatment groups $k = 1, \ldots, q$ at occasions $j = 1, \ldots, m$. Consequently, there are $n = \sum_{k=1}^{q} n_k$ independent units, and the total number of observations is N, which equals mn if all subjects are measured at all occasions (i.e., no missing values). Further, we index the combinations of occasions and treatment groups with $b = 1, \ldots, s$ where $s = mq$.
We fit to the data an LMM of the general form[15]

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon}, \quad \mathbf{b} \sim \mathcal{N}(\mathbf{0}, \mathbf{D}), \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{R}), \quad Cov(\mathbf{b}, \boldsymbol{\epsilon}) = \mathbf{0}$$

where $\mathbf{y}$ is a column vector of the realizations of Y_{jki}, $\mathbf{X}$ is an $N \times s$ design matrix of fixed effects, $\boldsymbol{\beta}$ is a column vector containing the parameters of interest, $\mathbf{Z}$ is a design matrix of random effects, $\mathbf{b}$ is a column vector of random effects parameters, and $\boldsymbol{\epsilon}$ is a column vector of residuals. The error components $\mathbf{b}$ and $\boldsymbol{\epsilon}$ are assumed multivariate normal with mean $\mathbf{0}$ and covariance matrices $\mathbf{D}$ and $\mathbf{R}$, respectively, and independent of one another.
We parameterize our model in a pseudo one-way layout (also known as *cell means model*) which contains one mean parameter for every interaction effect of treatment and time point; this requires a block-diagonal fixed-effects design matrix of the form

$$\mathbf{X} = \begin{bmatrix} \mathbf{1}_{n_1} & & \\ & \ddots & \\ & & \mathbf{1}_{n_q} \end{bmatrix}$$

where $\mathbf{1}_{n_k}$ is a column vector of length n_k containing ones only. The effect parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_s)^T$ are estimated by generalized least squares as

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{X})^{-1} \; \mathbf{X}^T \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{y}$$

with variance

$$Var(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^T \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{X})^{-1}$$

where $\hat{\boldsymbol{\Sigma}}$ denotes the REML solution to the marginal covariance matrix

$$\boldsymbol{\Sigma} = \mathbf{Z}\mathbf{D}\mathbf{Z}^T + \mathbf{R}.$$

Notice that, despite the name *cell means model*, $\hat{\boldsymbol{\beta}}$ equals the arithmetic cell means $(\bar{y}_1, \ldots, \bar{y}_s)^T$ only when there are no missing values. For details on LMM theory see e.g., McCulloch and Searle[16]; guidelines for practical model building are provided e.g., by Cnaan et al.[17] and Cheng et al.[18].
In principle, one can model any combinations of random effects and residual covariance structures that appear expedient (and can be fitted by software), but we want to focus on two important special cases that simplify matters in practice:

- The *extended linear model* (ELM) sets $\mathbf{Z} = \mathbf{0}$ so that the random-effects part $\mathbf{Zb}$ of the model is effectively dropped, and thus also $\mathbf{D} = \mathbf{0}$. In consequence, all longitudinal association and heteroscedasticity must be captured by the error covariance matrix $\mathbf{R}$.
- The *conditional independence model* (CIM) sets $\mathbf{R} = \sigma^2\mathbf{I}$, thereby requiring that the random-effects part $\mathbf{Zb}$ is sufficient to acknowledge within-subject correlation and heterogeneous variances. This leads to assuming conditional independence of the residuals given the random effects.

Extended linear model

The complete covariance matrix $\boldsymbol{\Sigma} = \mathbf{R}$ is a block-diagonal composition of partial matrices $\boldsymbol{\Sigma}_k$ that contain the covariances of time points for the kth treatment and may or may not be the same for different treatment groups:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_1 & & \\ & \ddots & \\ & & \boldsymbol{\Sigma}_q \end{bmatrix}.$$

One key issue is how to specify the $\boldsymbol{\Sigma}_k$. We desire to work with a covariance model that balances elaborateness and parsimony in a reasonable way. For instance, a model assuming equicorrelation and homogeneous variances at all occasions is likely to be too simplistic. On the contrary, there is often no need to estimate a full unstructured covariance matrix and variances that differ between treatments. So we can either opt for a sound tradeoff (e.g., assuming AR(1) correlation and heteroscedasticity over time), or we bring model selection into play, which hopefully comes to a sparse compromise that captures the main characteristics of the data. To this end, we assemble a set of "plausible" candidate models of varying complexity i.e., different variance patterns and correlation structures of the residuals. Then we employ a selection criterion to detect the most appropriate model among those in the candidate set given the data at hand. The widespread AIC[19] chooses the model that provides the best approximation to the unknown truth in an information-theoretic sense[20]. However, AIC may be unreliable when the sample size is small relative to the number of estimated model parameters; therefore we resort to the second-order AICc[21,22]

$$AICc = AIC + \frac{2K(K+1)}{n-K-1}$$

where K is the number of model parameters and n the sample size. According to Burnham and Anderson[23], AICc's small-sample bias adjustment has a noticeable impact for $\frac{n}{K} < 40$ i.e., nearly all cases relevant to our application. AIC and AICc are asymptotically equal.
The first component of our covariance structure to be selected is the elements on the diagonal of the matrix. Reasonable options for modeling homo-/heteroscedasticity are:

i) variances are equal across treatments and time points ($\sigma^2_{kj} = \sigma^2$),
ii) variances vary over time but are constant across treatments ($\sigma^2_{kj} = \sigma^2_j$),
iii) variances vary over time and between treatments (σ^2_{kj}).

The other component is the off-diagonal entries. Jennrich and Schluchter[24] and Wolfinger[25] discuss a multitude of choices of correlation patterns with repeated measurements. Lu and Mehrotra[26] propose to always use an unstructured matrix but admit it may lead to convergence problems; in addition, Littell et al.[27] point out that this strategy makes standard error estimates unstable.
We shall focus our attention on alternatives whose complexity lies somewhere between the simplistic compound symmetry (CS) and the excessive unstructured (UN) pattern. A straightforward model for longitudinal data is AR(1), which applies to equi-spaced occasions and assumes that correlation drops exponentially with distance in time. If the gaps between adjacent time points vary, there is a continuous

version of AR(1) available. Further flexibility can be achieved e.g., with antedependence or moving average (MA) models.
Note that the candidate set encompasses models with different covariance parts but identical specification of the mean, therefore it is legitimate to compare REML-based AICc values.

Conditional independence model

Instead of "dumping" the serial dependence structure in the residual covariance matrix $\mathbf{R}$, we can account for the repeatedness of measurements by introducing random effects. With "simple" random subject effects, the random-effects design matrix (for complete and balanced data) is

$$\mathbf{Z} = \mathbf{1}_q \otimes \mathbf{I}_n$$

and has dimensions $N \times n$. This induces a CS correlation for the fixed-effect parameters, which is rarely in harmony with serially repeated measurements, and such an underspecification can lead to overoptimistic inferences[28,29].
If we allow for occasion-specific random subject effects, the design matrix becomes

$$\mathbf{Z} = \begin{bmatrix} \mathbf{1}_q \otimes \mathbf{e}_1 \\ \vdots \\ \mathbf{1}_q \otimes \mathbf{e}_m \end{bmatrix}$$

where $\mathbf{e}_j, j = 1, \ldots, m$ is the jth unit vector of size m, and the dimension of $\mathbf{Z}$ grows to $N \times N$.
Occasion-specific random effects bring about an unstructured correlation matrix that is much more likely to capture the dependency among points in time. If the random effects are both occasion- and treatment-specific, UN correlations are allowed to differ between treatment groups; this specification appears very flexible at first sight but is often hard to estimate from real data. If the measurement occasions are equidistant or otherwise known (apart from just their stochastic order), modeling random slopes for the subjects may be a sparse alternative. Yet another strategy would be to model "simple" random effects and additionally impose some structure on $\mathbf{R}$, which means abandoning the conditional independence assumption.
If unsure about finding a suitable correlation pattern, we may again employ AICc model selection, and again the candidate set encompasses models that differ only in their random part whereas the fixed portion remains unchanged and thus REML is appropriate for AICc selection purposes.
General advice on specifying the covariances in an LMM can be found in Wolfinger[30], Keselman et al.[31], and Gurka[32]. Having settled for a random-effects and/or residual covariance structure, we use the chosen model to estimate $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ for simultaneous inference as described in Section 2.3.

2.2 Multiple marginal models

In contrast to basing multiple inferences on a joint model that embraces all time points (as done in Section 2.1), we may also fit m separate linear models for occasions $j = 1, \ldots, m$. Following the approach of Pipper et al.[33], one can combine these m marginal models and determine a joint matrix of correlations between the respective score contributions of the time points from the different models. A big advantage of this method (which is implemented in the `R` add-on package `multcomp`[34]) is that one does not have to bother about how to shape the random effects and residual covariances in advance.
The parameters of interest are the treatment effects $\boldsymbol{\beta}_j$; they are estimated by the jth marginal model, and their correlation is obtained via "stacking" the score contributions (derivatives of the log-likelihood) of the ordinary least squares estimates $\hat{\boldsymbol{\beta}}_j$. Asymptotically

$$(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)\sqrt{n} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} -\mathbf{I}_j^{-1} \tilde{\Psi}_{ij} + o_P(1)$$

where $\mathbf{I}_j^{-1}$ is the row in the inverse Fisher information matrix that corresponds to $\boldsymbol{\beta}_j$, $\tilde{\Psi}_{ij}$ is the score function for the ith of measurements $i = 1, \ldots, n$, and $o_P(1)$ denotes a sequence of random vectors

converging to zero in probability. Now the idea is to "stack" the $\boldsymbol{\beta}_j$, $\hat{\boldsymbol{\beta}}_j$, and score components over all j so as to get

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

which is the m-variate asymptotic version of the above. According to the multivariate central limit theorem, the left side converges in distribution to m-variate normality:

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

A consistent estimator of $\boldsymbol{\Sigma}$ is

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{i=1}^{n} \hat{\boldsymbol{\Psi}}_i' \hat{\boldsymbol{\Psi}}_i$$

where the $\hat{\boldsymbol{\Psi}}_i$ are obtained by plugging in the parameter estimates from the m marginal models. Along with $\hat{\boldsymbol{\beta}}$ we have what we need to continue with simultaneous inference.
In practice, the jth marginal linear model has the form

$$\mathbf{y}^{(j)} = \mathbf{X}^{(j)} \boldsymbol{\beta}^{(j)} + \boldsymbol{\epsilon}^{(j)}$$

where the superscript index signalizes belonging to occasion j. The design matrix $\mathbf{X}^{(j)}$ needs to be arranged such that there is one parameter for each of the treatments' means.

2.3 Simultaneous inference

Comparisons of interest $h = 1, \ldots, z$ are specified in a $z \times s$ coefficient matrix

$$\mathbf{C} = (\mathbf{c}_1, \ldots, \mathbf{c}_z)' = (c_{hb})$$

where $\sum_{b=1}^{s} c_{hb} = 0$. The resulting set of contrasts

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

may comprise comparisons of treatments per time point, or comparisons of time points per group. Even more generally, all sorts of comparisons can be realized, but it is hard to see where e.g., comparing treatment A at occasion 1 versus treatment B at occasion 2 could be of relevant use.
The asymptotic joint reference distribution for the vector of contrast test statistics under H_0 is z-variate normal with a correlation matrix that incorporates the longitudinal dependency of time points. For details on the computation of adjusted p-values and compatible simultaneous confidence limits around the estimated contrasts

$$\hat{\eta}_h = \sum_{b=1}^{s} c_{hb} \hat{\beta}_b$$

we refer to Hothorn et al.[10] and Bretz et al.[35].
In the case of small sample sizes the null distribution may be approximated as z-dimensional t with a reasonable choice for the degrees of freedom e.g., based on the Kenward-Roger method[36] or a recent suggestion using effective sample sizes[37].

2.4 Missing values

Missing values are the rule in longitudinal data settings and not some curious exception; thus any method that cannot cope with missings properly is virtually useless. Data points may be missing intermittently (e.g., due to failure of recording values because of technical errors) or there can be dropout (e.g., due to death or loss of follow-up in clinical studies).
The joint LMMs make use of all data ("available-case analysis"), and they yield valid results under the assumption of MAR (missing at random i.e., the probability of a data point to be missing may depend on observed but not on unobserved values) due to their specifying a joint likelihood[38]. Fitting multiple occasion-specific models also uses the entire data but requires the stricter MCAR assumption (missing

completely at random i.e., the probability to be missing may neither depend on observed nor on unobserved values).
Note that the parameter estimates from joint and multiple separate models disagree when data are unbalanced e.g., due to dropout. With the joint modeling approach, the estimates of $\boldsymbol{\beta}$ depend on $\hat{\boldsymbol{\Sigma}}$, the REML solution for the covariance matrix. When there are missing values, information is "borrowed" from adjoining measurements, which makes the estimation of $\boldsymbol{\beta}$ depend on how well the error (co-)variances or random effects are specified. By contrast, the estimator $\hat{\boldsymbol{\beta}}$ with the multiple marginal models approach is always the arithmetic mean of the available values i.e., we cannot "borrow strength" across occasions.

3 Simulation study

We investigate our methods' finite-sample performances under the null and under various configurations of the alternative via simulation. In Section 3.1 we assess how well the MCT procedures preserve the nominal FWER in the presence of moderate to large sample sizes. In Section 3.2 we compare the powers of MCTs based on joint and combined marginal models as well as simple Bonferroni-type adjustments. All simultations were executed in `R` version 3.1.3[39].

3.1 Type I error

We consider balanced layouts involving $q = \{3, 4, 5\}$ treatment groups, $m = \{3, 4, 5\}$ time points, and $n_k = \{10, 20, \ldots, 120\}$ subjects per group with longitudinal observations being correlated and heteroscedastic over time. Simulation data for each treatment are 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 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 yields

$$\boldsymbol{\Lambda} = \begin{bmatrix} 1 & 1.27 & 1.39 & 1.40 \\ & 2 & 2.20 & 2.26 \\ & & 3 & 3.12 \\ & & & 4 \end{bmatrix}.$$

We simulate 5000 datasets under H_0 and carry out many-to-one or all-pairwise comparisons among treatments per measurement occasion, or among occasions per treatment, for two-sided hypotheses using parameter and covariance estimates obtained from

a) a joint ELM assuming AR(1) correlation and heterogeneous variances across time,
b) a joint CIM with occasion-related random subject effects,
c) the combination of marginal occasion-specific linear models,

and check for each of them whether the minimum adjusted p-value is less than the nominal $\alpha = 0.05$ bound.
For comparisons of treatments simultaneously and separately at multiple occasions (top row of Figure 1), the simulated type I error rates of all methods are slightly inflated to around 7-8% with $n_k = 10$ but level off at the nominal 5% once the sample sizes exceed 20 or 30. This holds true for both many-to-one and all-pairwise comparisons. Varying the number of treatment groups seems to have no influence on the achieved test sizes whatsoever. However, when the number of occasions increases (which raises the overall heteroscedasticity in our simulation setting), the CIM-based tests become conservative, probably indicating that the simple random-effects structure is not adequate.
The simulations of treatment-wise comparisons among occasions (bottom row of Figure 1) give a slightly different picture; in particular the choice of comparisons (many-to-one or all-pairwise) can make a difference. Many-to-one (but not all-pairwise) tests relying on the CIM turn conservative with type I error rates around 4% or even lower as the number of occasions rises. In the approach with multiple occasion-specific models the realized α for all-pairwise comparisons skyrockets to 14-23% for $n_k = 10$ and requires sample sizes of 50 and more to come close to the nominal test size. In contrast, comparisons based on the ELM generally perform well except for the slight inflation of α when n_k is relatively small.

All procedures investigated here are asymptotic in nature, and hence discernibly anticonservative unless n_k exceeds at least 20-30. Their small-sample performances may be improved by using a multivariate t reference distribution with some reasonable approximation to the degrees of freedom (simulation results not shown).

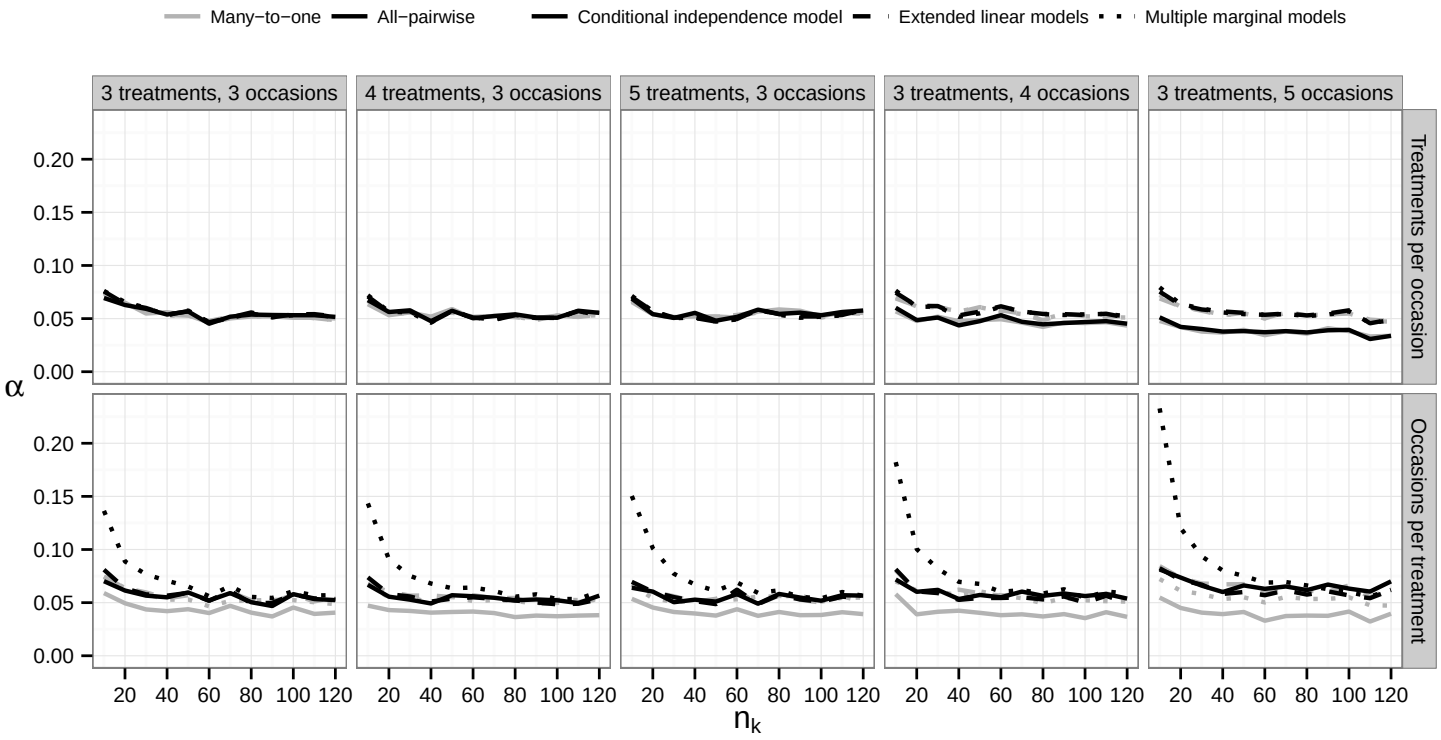

Figure 1: Simulated type I error levels for many-to-one and all-pairwise comparisons involving $q = \{3, 4, 5\}$ treatment groups, $m = \{3, 4, 5\}$ time points, and n_k independent subjects per treatment group (5000 simulation runs). Top row: comparisons of treatments within occasions; bottom row: comparisons of occasions within treatment groups.

3.2 Power

Having substantiated that the test procedures based on either joint or multiple marginal models keep the α level as long as n_k does not get too small, one may now wonder if the methods differ in terms of statistical power. Another relevant question could be whether approximating the joint distribution of test statistics as multivariate t is actually worth the effort. Or asked differently, how much worse (in terms of power) is a simple foolproof solution like calculating a bunch of single t-tests followed by a Bonferroni adjustment?

We focus our power investigations on many-to-one and all-pairwise comparisons of $q = \{3, 4, 5\}$ groups simultaneously for $m = \{3, 4, 5\}$ time points and $n_k = 100$. Simulation data are drawn similar as in Section 3.1, but now we mimic a treatment effect in one non-control group that arises only at the last time point. Thus for one of the treatments the mean vector is now $\boldsymbol{\mu} = (\mu, \ldots, \mu, \mu + \delta)$ with non-centralities $\delta = \{0, 0.1, \ldots, 1.5\}$. This leads to a scenario with *exactly one* many-to-one comparison being under H_A; the number of all-pairwise tests under the alternative is $q - 1$ (when comparing treatments per occasion) or $m - 1$ (when comparing occasions per treatment).

We generate 1000 datasets for each combination of parameter values and evaluate them fivefold:

1. Calculate standard t-tests for all single comparisons and adjust the resulting p-values with Bonferroni. This means turning a blind eye to any correlations.
2. Perform an MCT within each time point (or within each treatment arm) and adjust via Bonferroni for the multiplicity of time points (or treatment arms). This approach incorporates the portion of correlation that originates from multiple test statistics being built with overlapping subsets of $\hat{\boldsymbol{\beta}}$, but ignores the correlation among time points.

3. Build the longitudinal MCT on a joint CIM with occasion-depending random subject effects.
4. Base the longitudinal MCT on a joint ELM with variance heterogeneity and AR(1) correlation on the residual covariance matrix.
5. Fit multiple occasion-specific marginal models and combine them as detailed in Section 2.2 to obtain joint covariance estimates for use in a longitudinal MCT.

Strategies 3 to 5 account for correlations among both time points and test statistics.
The empirical curves of global power i.e., the probability of rejecting at least one elementary H_0, are shown in Figure 2 for many-to-one comparisons (the results for all-pairwise contrasts are very similar, therefore we will not present and discuss them separately). All three methods that acknowledge dependence of repeated measurements have clearly superior power compared to the Bonferroni-corrected t-tests and MCTs. This may not strike the eye at first sight, but the *vertical* distances between the gray and black curves in the top row of Figure 2 indicate a power advantage of 7-8 percentage points at the steepest point near $\delta = 0.7$ when comparisons are carried out between treatments at multiple occasions. For comparisons of occasions within each treatment arm (bottom row of Figure 2), the superiority of the joint or multiple marginal models is evident beyond any doubt. By contrast, the power advantage of (Bonferroni-corrected) single MCTs over (Bonferroni-corrected) single t-tests is marginal (around 1%) throughout the simulation settings.
The power curves for joint and marginal models-based MCTs are almost indistinguishable in the majority of cases; only the CIM-based comparisons of treatments per occasion prove again somewhat conservative as the number of occasions increases.
The power considerations up to this point assumed a fixed correlation $\rho_1 = 0.90$ for adjoining measurements, then $\rho_2 = 0.80$, etc. Now we want to shed light upon the actual impact of longitudinal correlation on the powers of joint and marginal model-based MCTs. For $q = 3$ treatments and $m = 3$ occasions we choose values for the elements on the first and second off-diagonal of $\mathbf{B}$ (denoted by ρ_1 and ρ_2) from $\boldsymbol{\rho} = (\rho_1, \rho_2) \in \{(0.50, 0.20), (0.80, 0.50), (0.90, 0.80), (0.95, 0.90)\}$.
The corresponding power curves are displayed in Figure 3 for many-to-one comparisons (the results for all-pairwise contrasts are again very similar). We recognize that the power advantage of both joint and multiple marginal models-based MCTs over Bonferroni-type adjustments increases with the correlation among occasions. The power gain when accounting for longitudinal dependence in comparisons between treatments at multiple occasions (upper row of Figure 3) is actually negligible unless $\rho_1 \geq 0.9$. On the contrary, treatment-wise comparisons among occasions (bottom row of Figure 3) benefit substantially already for much weaker longitudinal dependence.

4 Illustration: bradykinin receptor antagonism

Cardiopulmonary bypass (CPB) puts cardiac surgery patients in jeopardy of postoperative bleeding, which in turn may require transfusion of blood products. This bleeding is often caused by fibrinolysis i.e., fibrin in blood clots gets degraded to little protein fragments called D-dimers. Consequently, D-dimers are commonly used as a biomarker for fibrinolysis. Researchers have been seeking strategies to prevent fibrinolytic degradation for it is, of course, desirable to avoid blood transfusion during and after surgical intervention.
It is known that CPB promotes fibrinolysis via a peptide called bradykinin and the associated bradykinin B_2 receptor. Therefore Balaguer et al.[40] investigated whether bradykinin B_2 receptor antagonism can reduce fibrinolysis. They conducted a randomized, double-blind trial at Vanderbilt University Medical School, Nashville, TN between 2007 and 2012 where 115 patients about to undergo cardiac surgery with the aid of a heart-lung machine ("on-pump") were randomized to one of three intravenous treatments:

- HOE 140, a specific bradykinin B_2 receptor antagonist,
- ϵ-aminocaproic acid (EACA), a well-known antifibrinolytic drug,
- normal saline (placebo).

One of their secondary endpoints was the concentration of D-dimer in blood samples taken at five selected time points:

- prior to surgical incision (baseline)
- after 30 minutes "on-pump"
- after 60 minutes "on-pump"
- after separation from the heart-lung machine (post-bypass)

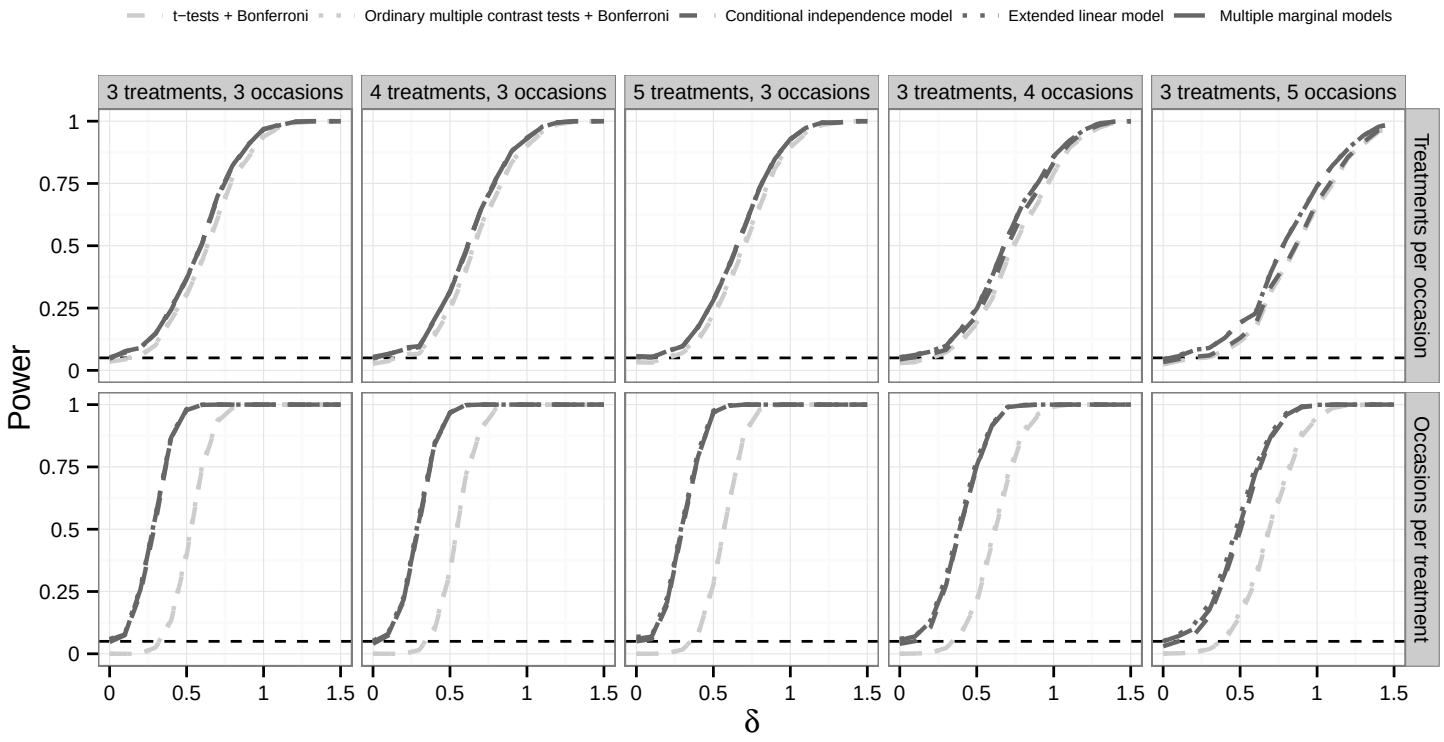

Figure 2: Simulated powers for many-to-one comparisons involving $q = \{3, 4, 5\}$ treatment groups, $m = \{3, 4, 5\}$ time points, and $n_k = 100$ independent subjects per treatment group, with exactly one comparison under the alternative (1000 runs). Top row: comparisons of treatments within occasions; bottom row: comparisons of occasions within treatment groups.

- on the first postoperative day (POD1).

Note that these time points were not picked haphazardly but are closely associated with relevant medical steps (e.g., initiation of anesthesia, separation from CPB).

The goal of this investigation was to quantify fibrinolysis (as measured via D-dimer concentrations) at distinctive occasions over the course of CPB until the day after under each of three treatments. The time intervals between measurement occasions are obviously unequal; nonetheless, observations from the same patient are for sure correlated.

Assuming multivariate normality of the natural logarithms of D-dimer concentrations in each treatment arm, we generated artificial data based on sample sizes, means, variances, and covariances of the (log-transformed) original data. Figure 4 displays the simulated bradykinin dataset with 38 patients in the HOE 140 arm and 37 patients in the placebo and EACA arms. The simulated dataset is available online from GitHub; we provide an R script for downloading and analyzing the bradykinin data as supplementary material.

4.1 Comparing treatments simultaneously at multiple time points

One relevant research question in this context is: when do which active drugs (HOE 140, EACA) reduce D-dimer concentrations (thus: reduce fibrinolysis) compared to placebo? We want to answer this question separately and simultaneously for each measurement occasion (except baseline) while controlling a common FWER. So we are out for many-to-one comparisons of drugs (HOE 140 vs. placebo and EACA vs. placebo) at each of four occasions. This we can achieve with longitudinal MCTs using estimates from

- an AICc-selected ELM,
- an AICc-selected CIM, or
- a combination of occasion-specific marginal models.

For comparison, we also compute "ordinary" MCTs for each of the four time points, followed by a Bonferroni adjustment i.e., multiplying the unadjusted p-values by four).

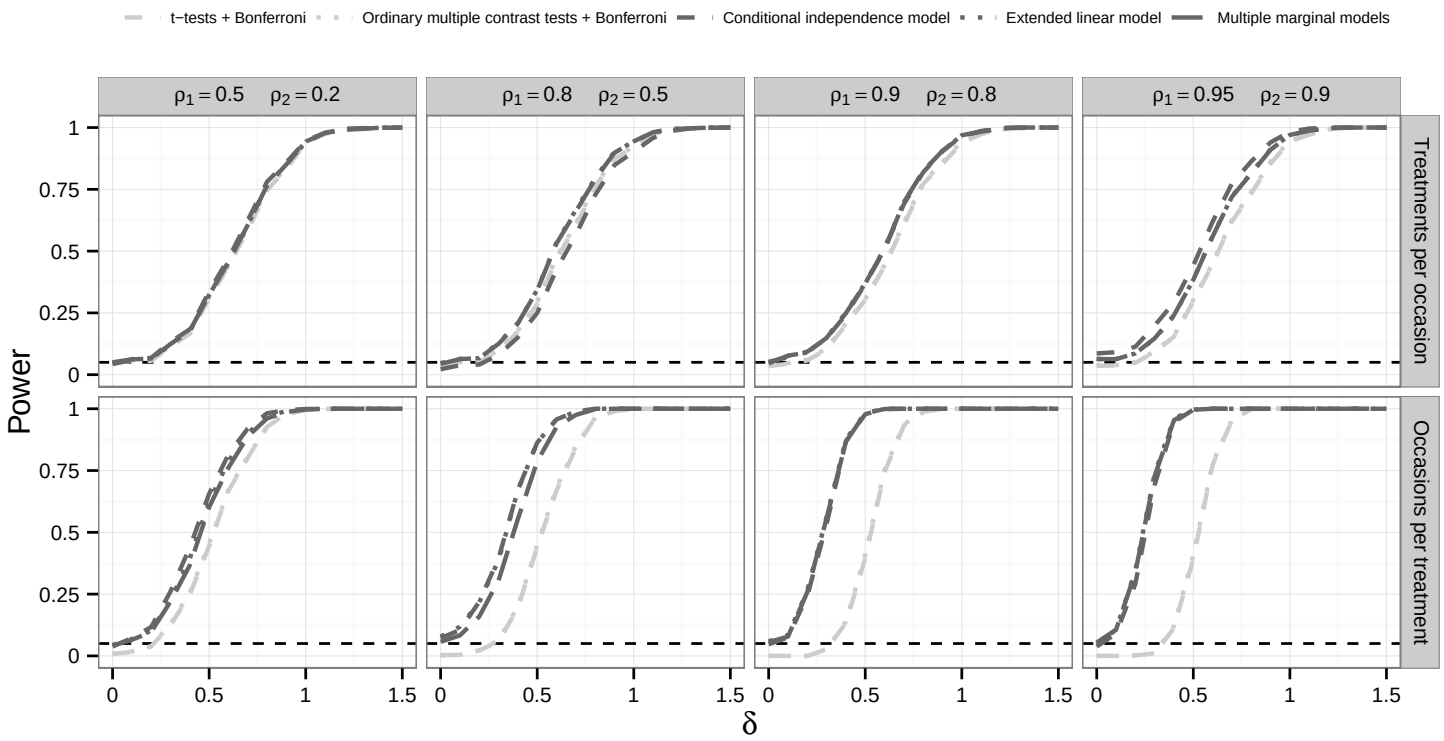

Figure 3: Simulated powers for many-to-one comparisons of $q = 3$ treatment groups at $m = 3$ time points with $n_k = 100$ independent subjects per treatment group under various longitudinal correlations (1000 runs). Top row: comparisons of treatments within occasions; bottom row: comparisons of occasions within treatment groups.

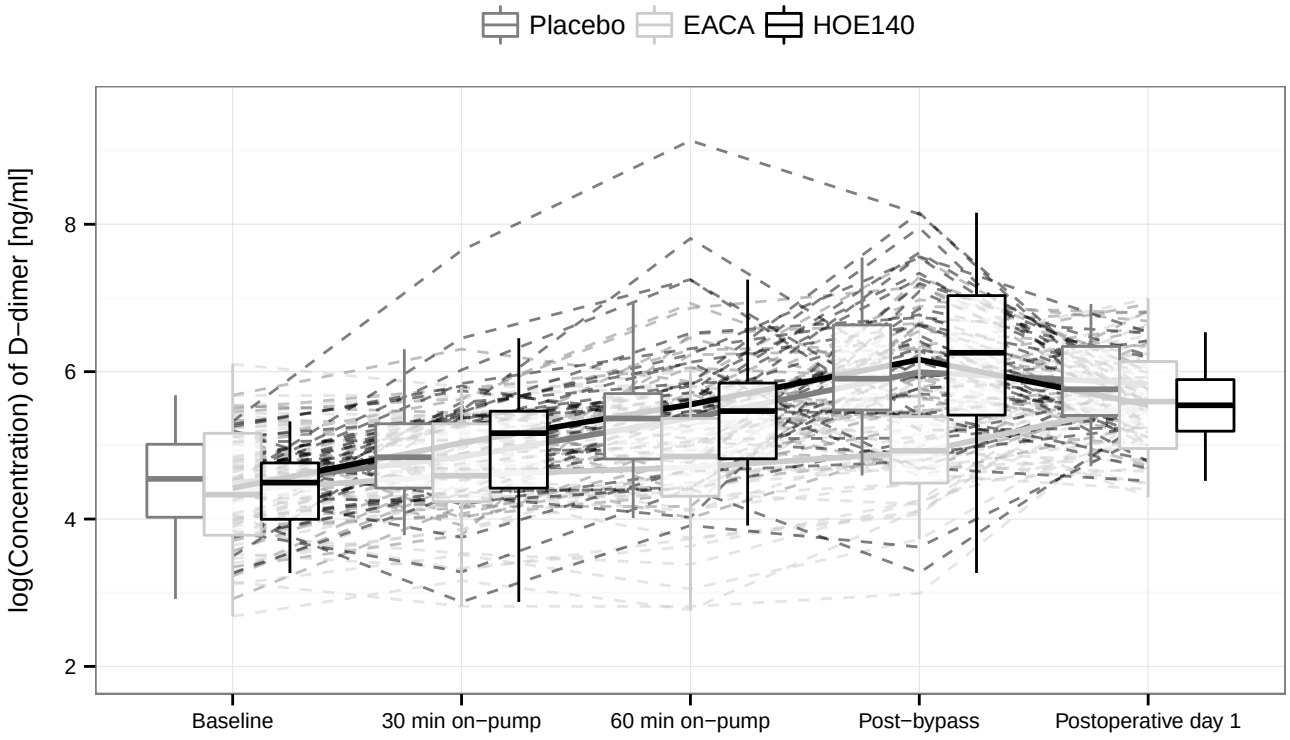

Figure 4: Bradykinin data: individual patient trajectories (dotted), sample mean trajectories per treatment arm (solid), and boxplots of log-concentrations of D-dimer.

We fit various candidate models for AICc to choose from. The set of ELMs involves all combinations of four variance structures (constant, heterogeneous over time, heterogeneous across treatment arms, heterogeneous both over time and across treatments) and three correlation patterns (CS, AR(1), UN) i.e., a total of twelve candidate models. AICc selects the most complex model with unstructured correlations and heteroscedasticity between occasions and treatments.
The candidate set of CIMs comprises three models with random patient effects being unstratified, occasion-specific, or both occasion- and drug-specific. Here AICc does not pick the most complex alternative but rather the model with occasion-specific random effects, which implies an unstructured correlation matrix that is the same for all three drugs and is therefore similar to that of the selected ELM. This similarity becomes apparent in the resulting correlation matrices of test statistics (Figures 5 and 7). Notice that the most complex random-effects structure would correspond to unstructured correlation matrices differing between treatment arms i.e., a configuration even more complicated than those of all ELMs in the candidate set.
Our subsequent inferences build upon parameter and covariance estimates from the AICc-selected joint models and the combined marginal models, respectively, as detailed in Section 2.3. We find that neither drug affects D-dimer concentrations in the initial phase (after 30 minutes) but EACA is superior towards the end of the surgical procedure: it reduces D-dimer significantly compared to placebo after 60 minutes and when the patient is separated from the heart-lung machine (Table 1). This effect cannot be shown for HOE 140, which seems to even slightly increase D-dimer levels during CBP. After one day the beneficial effect of EACA vanishes, and we observe only a minor reduction of D-dimer log-concentrations that is similar (but non-significant) for both treatments.
The estimated standard errors reveal that there is variance heterogeneity over time: variability of D-dimer rises over the course of the surgical intervention but declines and stabilizes the day after. The ELM approach uses standard error estimates that additionally differ between comparisons of drugs: they are distinctly higher for HOE 140 than for EACA during surgery (i.e., after 30 and 60 minutes and post-bypass) but a little lower at baseline and after one day.
The adjusted p-values do not differ much between the three modeling strategies that acknowledge temporal correlation. In contrast, occasion-specific MCTs followed by a Bonferroni adjustment have drastically increased p-values (e.g., that for EACA versus placebo after 60 minutes doubles from 0.01 to 0.02) indicating that there is a lot of power to gain by exploiting the dependence of occasions. So we conclude it is essential to incorporate some, at least half-decent estimate of the correlation over time whereas the choice of modeling approach is of secondary importance here. This becomes also apparent from the 95% SCIs displayed in Figure 6. Intervals obtained from CIM or multiple marginal models analyses are about the same width whereas those from ELM analysis can be a little wider or narrower due to their using comparison-specific standard error estimates.

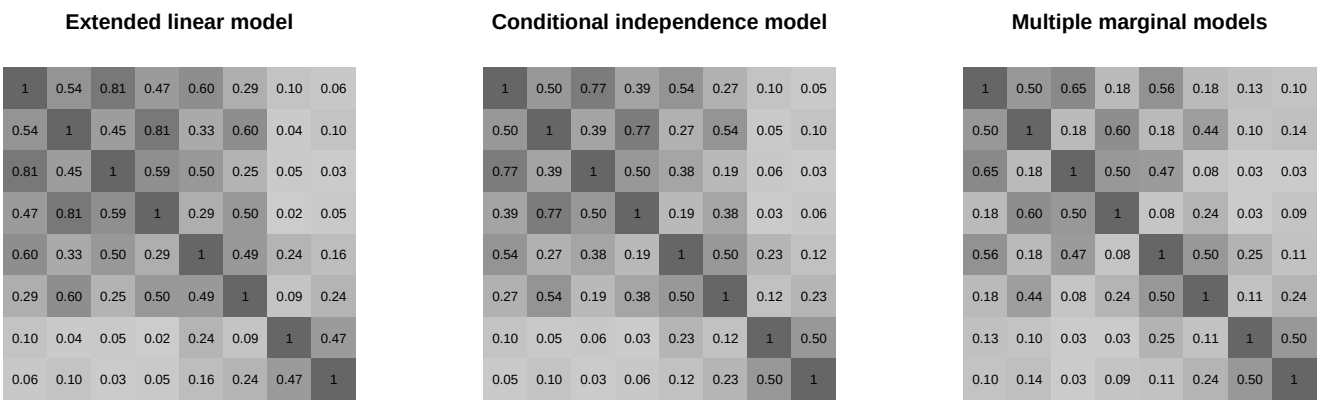

Figure 5: Bradykinin data: correlation matrices of test statistics for many-to-one comparisons of treatment arms per occasion.

Table 1: Simultaneous inference for the bradykinin data: estimated differences of D-dimer log-concentrations, standard errors (SE), and adjusted p-values for occasion-wise many-to-one comparisons of HOE 140 and EACA against placebo. Bon: Bonferroni; CIM: conditional independence model; ELM: extended linear model; MMM: multiple marginal models.

	Estimate	SE(ELM)	SE(CIM, MMM)	p(ELM)	p(CIM)	p(MMM)	p(Bon)
30 min on-pump: EACA - Placebo	-0.246	0.170	0.176	0.607	0.656	0.671	1.000
30 min on-pump: HOE140 - Placebo	0.191	0.213	0.174	0.933	0.856	0.865	1.000
60 min on-pump: EACA - Placebo	-0.668	0.208	0.208	0.009	0.009	0.010	0.020
60 min on-pump: HOE140 - Placebo	0.186	0.250	0.207	0.973	0.939	0.944	1.000
Post-bypass: EACA - Placebo	-1.088	0.176	0.211	<0.001	<0.001	<0.001	<0.001
Post-bypass: HOE140 - Placebo	0.169	0.249	0.210	0.984	0.965	0.968	1.000
Postoperative day 1: EACA - Placebo	-0.259	0.155	0.144	0.438	0.369	0.382	1.000
Postoperative day 1: HOE140 - Placebo	-0.290	0.127	0.143	0.133	0.240	0.250	0.619

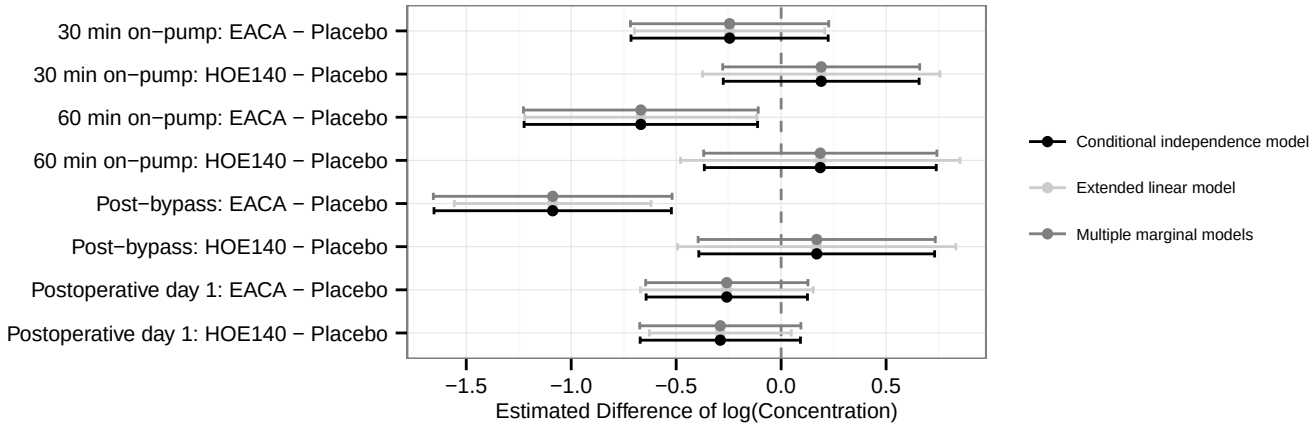

Figure 6: Bradykinin data: 95% simultaneous confidence intervals based on AICc-selected joint models and combined marginal models.

4.2 Comparing time points simultaneously for multiple treatments

Now suppose we were to gauge the D-dimer levels at various time points in comparison to baseline separately and simultaneously for the three treatment arms. As our power simulations have shown, here it can be way more painful not to incorporate serial correlation than it is with comparisons of treatments at several points in time.

We use the AICc-selected joint models from Section 4.1 as well as the multiple models approach. Again there is considerable discrepancy between pooled-treatment and treatment-specific standard error estimates, and they bring about p-values that differ quite a bit between the modeling strategies (Table 2). The widths of the corresponding 95% SCIs shown in Figure 8 also vary according to the standard error estimates involved, but the discrepancies appear less drastic than when looking at the adjusted p-values. The latter are probably hardly meaningful anyway when the actual task is not to test point-zero hypotheses but to quantify the uncertainty of estimated differences, which is much better done by SCIs. The log-concentrations of D-dimer raise quickly above baseline as the surgical procedure takes its course in patients treated with either HOE 140 or placebo. However, the situation is quite different for the EACA arm: D-dimer levels remain rather close to the baseline value until the patient is separated from the heart-lung machine. So it makes good sense to do the comparisons versus baseline separately (but simultaneously) for the three treatment arms.

5 Discussion

If simultaneous inference at a few prudently chosen points in time, or between such time points, is a promising means to scientific insight, we propose an MCT-type procedure that may be built upon either a joint model for all measurement occasions, or separate occasion-specific models. Both MCTs based on joint and marginal models lead to SCIs, adjusted local p-values, and a global decision. The question remains which modeling approach should be used. Simply put, both of them come with assets and

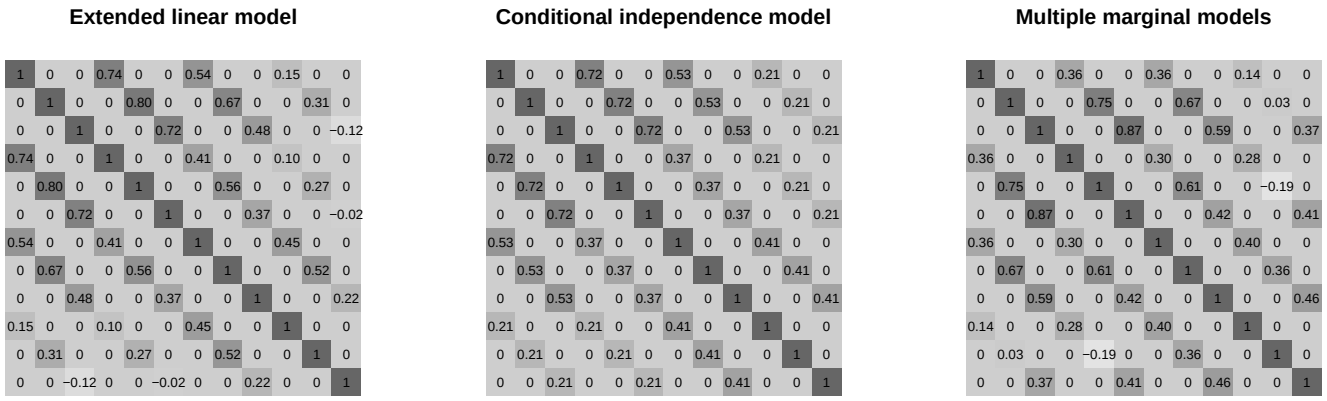

Figure 7: Bradykinin data: correlation matrices of test statistics for many-to-one comparisons of occasions per treatment arm.

Table 2: Simultaneous inference for the bradykinin data: estimated differences of D-dimer log-concentrations, standard errors (SE), and adjusted p-values for treatment-wise many-to-one comparisons of occasions against baseline. CIM: conditional independence model; ELM: extended linear model; MMM: multiple marginal models.

	Estimate	SE(ELM)	SE(CIM)	SE(MMM)	p(ELM)	p(CIM)	p(MMM)
Placebo: 30 min on-pump - Baseline	0.361	0.100	0.098	0.094	0.004	0.003	0.001
Placebo: 60 min on-pump - Baseline	0.886	0.149	0.138	0.161	<0.001	<0.001	<0.001
Placebo: Post-bypass - Baseline	1.511	0.149	0.155	0.164	<0.001	<0.001	<0.001
Placebo: POD1 - Baseline	1.354	0.143	0.139	0.124	<0.001	<0.001	<0.001
EACA: 30 min on-pump - Baseline	0.210	0.081	0.098	0.059	0.097	0.289	0.004
EACA: 60 min on-pump - Baseline	0.312	0.107	0.138	0.081	0.039	0.220	0.001
EACA: Post-bypass - Baseline	0.517	0.117	0.155	0.128	<0.001	0.010	0.001
EACA: POD1 - Baseline	1.189	0.151	0.139	0.146	<0.001	<0.001	<0.001
HOE140: 30 min on-pump - Baseline	0.615	0.113	0.097	0.129	<0.001	<0.001	<0.001
HOE140: 60 min on-pump - Baseline	1.135	0.150	0.136	0.161	<0.001	<0.001	<0.001
HOE140: Post-bypass - Baseline	1.744	0.186	0.153	0.159	<0.001	<0.001	<0.001
HOE140: POD1 - Baseline	1.127	0.118	0.137	0.141	<0.001	<0.001	<0.001

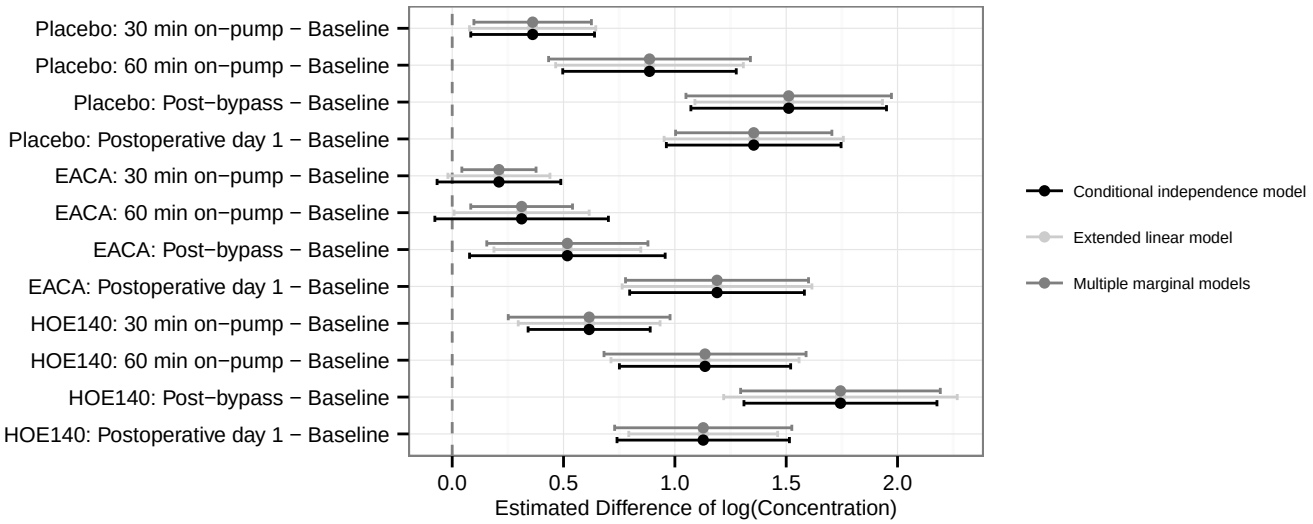

Figure 8: Bradykinin data: 95% simultaneous confidence intervals based on AICc-selected joint models and combined marginal models.

drawbacks.

Combining multiple marginal models saves the trouble of devising a suitable structure for the random

effects and/or error covariance matrix, which can be a fairly intricate affair with joint modeling. In practice, the decision which type of joint model to fit (ELM, CIM, or some model with both non-zero elements of $\mathbf{D}$ and structure in $\mathbf{R}$) will be guided by characteristics of the data-generating process. Failure to capture these dependencies adequately may affect the joint model-based tests; in addition, numerical convergence problems may prevent fitting a more complex (and more appropriate) model. Both these problems can be evaded by combining simple univariate occasion-specific models.
An evident virtue of the joint modeling strategy is that it can handle missing data and dropout rather straightforwardly assuming MAR, and "borrow strength" from adjoining occasions, whereas the multiple marginal models approach requires the harsher MCAR condition, and no information can be "borrowed".
The simulations in Sections 3.1 and 3.2 show that the test procedure based on multiple occasion-specific models performs very well in most situations, especially for comparisons of treatments at multiple occasions, and has power equivalent to the tests based on joint models. However, one should exercise care when comparing occasions within treatment arms: here the marginal models approach may produce too many rejections of H_0 with small sample sizes. Comparisons among repeated occasions are generally more intricate than comparisons among randomized treatment groups.
The power simulations of Section 3.2 clearly suggest that taking dependencies of time points into account does pay off. In particular with high correlation the longitudinal MCT procedures have distinctly superior power compared to simple Bonferroni adjustments.
A possible alternative to parametric simultaneous inference in longitudinal designs are distribution-free techniques. Nonparametric MCTs in the context of repeated measures were treated in Konietschke et al.[41].
All methodology described in this paper concerns "small-scale" scenarios where the dimensionality of the problem (i.e., the number of treatments and occasions) is smaller than the number of experimental units ($d < n$). The high-dimensional case ($d > n$) creates extra challenges such as singular covariance matrices. Strategies for $d \gg n$ problems with small n are discussed in Konietschke et al.[42].
We are aware that strong error control over multiple endpoints is a controversial issue[43]. In our opinion, it can be meaningful to look at several time points simultaneously, especially if they are of particular medical or biological relevance. So if there is a claim à la "find at least one treatment that is different from control at least at one occasion", it is reasonable to control the FWER for the entire claim. In no way is our procedure meant to be an invitation to a "fishing trip for significances", nor is it designed for analyzing excessive numbers of time points separately. It gives interpretable results for datasets with a handful of well-chosen occasions (like in the bradykinin example of Section 4), but as soon as there are, say, ten or more time points, one should certainly resort to modeling the longitudinal evolution and making inference for a summary measure (e.g., compare the slope parameters).
This paper focused on many-to-one and all-pairwise comparisons, but the techniques presented are applicable to multiple contrasts of any kind e.g., comparisons with the grand mean, Williams trend tests[44–46], or user-defined contrasts. Similarly, we laid the focus on inference in longitudinal designs, but there is no obstacle why one should not use our methods for other repeated measurement problems where the "occasions" are not points in time but rather e.g., spatial points, or experimental conditions.
Another extension worth looking at is discrete data; so far we have limited our considerations to continuous endpoints, but in biomedical research the outcome is often counts or proportions. The approach combining multiple marginal models seamlessly extends to discrete data by fitting occasion-specific generalized linear models (GLMs); this will be the topic of further investigations.

Funding

The research of Mias Pretorius was supported by the National Institute of Health [grant number R01HL085740]. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Acknowledgements

We are grateful to the editor and an anonymous reviewer for several helpful suggestions. We would also like to thank Ludwig Hothorn for his comments on an earlier version of the manuscript.

References

1. Verbeke G, Molenberghs G. *Linear mixed models for longitudinal data.* New York: Springer, 2000.
2. Davidian M, Giltinan DM. *Nonlinear models for repeated measurement data.* Boca Raton: Chapman & Hall/CRC, 1995.
3. Phillips A, Fletcher C, Atkinson G, et al. Multiplicity: discussion points from the statisticians in the pharmaceutical industry multiplicity expert group. *Pharm Stat* 2013; **12**: 255-259.
4. Dunnett CW. A multiple comparison procedure for comparing several treatments with a control. *J Am Stat Assoc* 1955; **50**: 1096-1121.
5. Tukey JW. The problem of multiple comparisons. Unpublished manuscript reprinted in: Braun HI (editor) *The collected works of John W. Tukey, Volume VIII.* New York: Chapman & Hall, 1994.
6. Hoffman WP, Recknor J, Lee C. Overall type I error rate and power of multiple Dunnett's tests on rodent body weights in toxicology studies. *J Biopharm Stat* 2008; **18**: 883-900.
7. Liu C, Cripe TP, Kim MO. Statistical issues in longitudinal data analysis for treatment efficacy studies in the biomedical sciences. *Mol Ther* 2010; **18**: 1724-1730.
8. Mukerjee H, Robertson T, Wright FT. Comparison of several treatments with a control using multiple contrasts. *J Am Stat Assoc* 1987; **82**: 902-910.
9. Bretz F, Genz A, Hothorn LA. On the numerical availability of multiple comparison procedures. *Biometrical J* 2001; **43**: 645-656.
10. Hothorn T, Bretz F, Westfall P. Simultaneous inference in general parametric models. *Biometrical J* 2008; **50**: 346-363.
11. Hasler M, Hothorn LA. A Dunnett-type procedure for multiple endpoints. *Int J Biostat* 2011; **7**: article 3.
12. Hasler M, Hothorn LA. A multivariate Williams-type trend procedure. *Stat Biopharm Res* 2012; **4**: 57-65.
13. Hasler M. Multiple contrast tests for multiple endpoints in the presence of heteroscedasticity. *Int J Biostat* 2014; **10**: 17-28.
14. Hasler M. Multiple contrasts for repeated measures. *Int J Biostat* 2013; **9**: 1-13.
15. Laird NM, Ware JH. Random-effects models for longitudinal data. *Biometrics* 1982; **38**: 963-974.
16. McCulloch CE, Searle SR. *Generalized, linear, and mixed models.* New York: John Wiley & Sons, 2001.
17. Cnaan A, Laird NM, Slasor P. Using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data. *Stat Med* 1997; **16**: 2349-2380.
18. Cheng J, Edwards LJ, Maldonado-Molina MM, et al. Real longitudinal data analysis for real people: building a good enough mixed model. *Stat Med* 2010; **29**: 504-520.
19. Akaike H. A new look at the statistical model identification. *IEEE T Automat Contr* 1974; **19**: 716-723.
20. Kullback S, Leibler RA. On information and sufficiency. *Ann Math Stat* 1951; **22**: 79-86.
21. Sugiura N. Further analysis of the data by Akaike's information criterion and the finite corrections. *Commun Stat-Theor M* 1978; **7**: 13-26.
22. Hurvich CM, Tsai CL. Regression and time series model selection in small samples. *Biometrika* 1989; **76**: 297-307.
23. Burnham KP, Anderson DR. *Model selection and multimodel inference: a practical information-theoretic approach.* 2nd ed. New York: Springer, 2002.
24. Jennrich RI, Schluchter MD. Unbalanced repeated-measures models with structured covariance matrices. *Biometrics* 1986; **42**: 805-820.
25. Wolfinger RD. Heterogeneous variance: covariance structures for repeated measures. *J Agr Biol Envir St* 1996; **1**: 205-230.
26. Lu K, Mehrotra DV. Specification of covariance structure in longitudinal data analysis for randomized clinical trials. *Stat Med* 2010; **29**: 474-488.
27. Littell RC, Pendergast J, Natarajan R. Modelling covariance structure in the analysis of repeated measures data. *Stat Med* 2000; **19**: 1793-1819.
28. Schielzeth H, Forstmeier W. Conclusions beyond support: overconfident estimates in mixed models. *Behav Ecol* 2009; **20**: 416-420.
29. Gurka MJ, Edwards LJ, Muller KE. Avoiding bias in mixed model inference for fixed effects. *Stat Med* 2011; **30**: 2696-2707.

30. Wolfinger R. Covariance structure selection in general mixed models. *Commun Stat-Simul C* 1993; **22**: 1079-1106.
31. Keselman HJ, Algina J, Kowalchuk RK, et al. A comparison of two approaches for selecting covariance structures in the analysis of repeated measurements. *Commun Stat-Simul C* 1998; **27**: 591-604.
32. Gurka MJ. Selecting the best linear mixed model under REML. *Am Stat* 2006; **60**: 19-26.
33. Pipper CB, Ritz C, Bisgaard H. A versatile method for confirmatory evaluation of the effects of a covariate in multiple models. *J Roy Stat Soc C-App* 2012; **61**: 315-326.
Hothorn T, Bretz F, Westfall P, et al. multcomp: simultaneous inference in general parametric models, R package version 1.4-0, 2015.
Bretz F, Hothorn T, Westfall P. *Multiple comparisons using R.* Boca Raton: Chapman & Hall/CRC, 2010.
34. Kenward MG, Roger JH. Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics* 1997; **53**: 983-997.
35. Faes C, Molenberghs G, Aerts M, et al. The effective sample size and an alternative small-sample degrees-of-freedom method. *Am Stat* 2009; **63**: 389-399.
36. Molenberghs G, Kenward MG. *Missing data in clinical studies.* Chichester: John Wiley & Sons, 2007.
37. R Core Team. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2015.
38. Balaguer JM, Yu C, Byrne JG, et al. Contribution of endogeneous bradykinin to fibrinolysis, inflammation, and blood product transfusion following cardiac surgery: a randomized clinical trial. *Clin Pharmacol Ther* 2013; **93**: 326-334.
39. Konietschke F, Bathke AC, Hothorn LA, Brunner E. Testing and estimation of purely nonparametric effects in repeated measures designs. *Comput Stat Data Anal* 2010; **54**: 1895-1905.
40. Konietschke F, Gel YR, Brunner E. On multiple contrast tests and simultaneous confidence intervals in high-dimensional repeated measures designs. In: Ahmed SE (editor) *Perspectives on big data analysis: methodologies and applications.* Montréal: Centre de Recherches Mathématiques, 2014.
41. Stone A, Chuang-Stein C. Strong control over multiple endpoints: are we adding value to the assessment of medicines? *Pharm Stat* 2013; **12**: 189-191.
42. Williams DA. A test for differences between treatment means when several dose levels are compared with a zero dose control. *Biometrics* 1971; **27**: 103-117.
43. Williams DA. The comparison of several dose levels with a zero dose control. *Biometrics* 1972; **28**: 519-531.
44. Bretz F. An extension of the Williams trend test to general unbalanced linear models. *Comput Stat Data Anal* 2006; **50**: 1735-1748.