Multiple Contrast Tests with Repeated and Multiple Endpoints—with Biological Applications

Von der Naturwissenschaftlichen Fakultät der

Gottfried Wilhelm Leibniz Universität Hannover

zur Erlangung des Grades

Doktor der Naturwissenschaften (Dr. rer. nat.)

genehmigte Dissertation

von

M.Sc. Philip Steffen Pallmann

geboren am 01.03.1987 in Geislingen an der Steige

2016

Referent:	Prof. Dr. Ludwig A. Hothorn
Korreferent:	Prof. Christian Ritz, PhD
Korreferent:	Apl. Prof. Dr. Siegfried Kropf
Tag der Promotion:	11.04.2016

Abstract

The primary objective of this thesis is to develop multiple hypothesis tests and simultaneous confidence intervals for comparisons of Gaussian means in longitudinal and similar scenarios with correlated outcomes. One might wish to investigate differences between treatment means separately at several specific points in time, or between means of several time points separately for multiple treatment groups, whilst controlling a common type I error rate for the entire set of comparisons. Global tests such as repeated measures analysis of variance are clearly unsatisfying in this situation. We describe flexible procedures for approximate simultaneous inference within the framework of multiple contrast tests (MCTs) where the different time points are treated as factor levels.

Two possible modeling strategies for the repeated measures are compared: a joint model covering the entire span of measurement occasions (which is for Gaussian data typically a linear mixed-effects model), or a novel approach for combining occasion-specific marginal models i.e., one linear model per time point. Both strategies come with assets and drawbacks: fitting a joint model requires a decision about how to model random effects and/or error covariance structures, which is in many cases not trivial, but model selection techniques can help out; on the other hand, covariance estimation in the combination of marginal models is entirely data-driven, but the method comes with harsher assumptions in the case of missing values.

Asymptotic procedures are fairly straightforward to contrive whereas a challenge with small samples is that the degrees of freedom (DFs) to be used for computing the multivariate t reference quantile are unclear, especially for general unbalanced designs. We explore a number of DF approximations via simulation; well-known adjustments such as Kenward-Roger or Pinheiro-Bates turn out to perform well in many situations. Moreover, we quantify the power advantage of our longitudinal MCTs over Bonferroni under various configurations.

Rather than comparing only treatment groups per time point or only time points per treatment group, we can also unite these hypotheses in one single family that embraces both comparisons among treatments and among occasions under control of a common type I error rate. This adds complexity to the problem, and especially finding proper DFs is intricate. The use of different comparison-specific DFs appears to be a viable solution.

We extend our considerations to discrete outcomes such as proportions and counts. Again we contrast the multi-model approach (now combining several generalized linear models) with a single joint model, which is now fitted by generalized estimating equations. As the subsequent inferences are asymptotic in nature, we gauge minimum sample sizes required to ensure reasonable control of the type I error rate for binomial and Poisson outcomes, respectively.

Application of the proposed methods is illustrated with six real datasets from medical, toxicological, and horticultural research. We point out strengths and limitations of our longitudinal MCTs, discuss their interpretation in the presence of missing data, and outline a few extensions and alternatives.

Keywords: longitudinal data, linear mixed-effects model, multiple hypothesis tests, simultaneous confidence intervals, degrees of freedom, missing values

Zusammenfassung

Kern dieser Arbeit ist die Ausarbeitung multipler Hypothesentests und simultaner Konfidenzintervalle für Mittelwertsvergleiche gaußverteilter Daten in longitudinalen und ähnlichen Szenarien mit korrelierten Messungen. So könnte man sich beispielsweise für Differenzen zwischen Behandlungsmittelwerten getrennt an mehreren spezifischen Zeitpunkten interessieren, oder für Differenzen zwischen Mittelwerten mehrerer solcher Zeitpunkte getrennt für mehrere Behandlungsgruppen, unter Kontrolle einer gemeinsamen Fehlerrate erster Art für die Gesamtheit der Vergleiche. Globaltests wie etwa die Varianzanalyse für wiederholte Messungen sind in dem Fall ganz klar unbefriedigend. Wir beschreiben flexible Prozeduren für näherungsweise simultane Inferenz auf Basis multipler Kontrasttests, wobei die verschiedenen Zeitpunkte als Faktorstufen aufgefasst werden.

Zwei mögliche Modellierungsansätze für die wiederholten Messungen werden gegenübergestellt: ein gemeinsames Modell, das sämtliche Messzeitpunkte abdeckt (für gaußverteilte Daten typischerweise ein lineares gemischtes Modell), oder ein neuartiger Ansatz, mit dem sich zeitpunkt-spezifische marginale Modelle, d.h. ein lineares Modell pro Zeitpunkt, kombinieren lassen. Beide Strategien haben ihre Vor- und Nachteile: ein gemeinsames Modell anzupassen erfordert eine Entscheidung darüber, wie zufällige Effekte und/oder Kovarianzstrukturen der Residuen zu modellieren sind, was sich in vielen Fällen als nicht trivial erweist, wenngleich Modellselektionstechniken hilfreich sein können; demgegenüber ist die Kovarianzschätzung bei der Kombination marginaler Modelle vollständig datenabhängig, jedoch erfordert die Methode strengere Annahmen im Fall fehlender Werte.

Asymptotische Verfahren lassen sich relativ leicht aufstellen, wohingegen bei kleinen Fallzahlen eine Herausforderung darin besteht, dass die Freiheitsgrade für die Berechnung von Vergleichsquantilen aus der multivariaten t-Verteilung unklar sind, insbesondere für allgemeine unbalanzierte Anlagen. Wir untersuchen einige Näherungen mittels Simulation; dabei zeigt sich, dass sich etablierte Methoden wie Kenward-Roger oder Pinheiro-Bates in vielen Situationen günstig verhalten. Darüber hinaus quantifizieren wir den Gütevorteil unserer longitudinalen multiplen Kontrasttests im Vergleich zu Bonferroni unter verschiedenen Konfigurationen.

Anstatt nur Behandlungsgruppen pro Zeitpunkt oder Zeitpunkte pro Behandlungsgruppe zu vergleichen, können wir diese Hypothesen genauso gut in einer einzigen Testfamilie vereinigen, die sowohl Vergleiche zwischen Behandlungen als auch zwischen Zeitpunkten umfasst unter Kontrolle einer gemeinsamen Fehlerrate erster Art. Damit wird das Problem noch einmal komplizierter, insbesondere die Bestimmung passender Freiheitsgrade. Eine gute Lösung ergibt sich, wenn man verschiedene vergleichsspezifische Freiheitsgrade verwendet.

Wir erweitern unsere Abhandlung auf diskrete Endpunkte wie Proportionen und Zähldaten. Wiederum stellen wir den Ansatz mit mehreren Modellen (wobei wir jetzt mehrere generalisierte lineare Modelle kombinieren) dem einzelnen Gesamtmodell, welches nun mit Hilfe von verallgemeinerten Schätzgleichungen angepasst wird, gegenüber. Da die darauf basierenden Inferenzen asymptotischer Natur sind, schätzen wir die benötigte Mindestfallzahl ab, die eine vertretbare Kontrolle der Fehlerrate erster Art bei binomialen beziehungsweise Poisson-Daten gewährleistet.

Die Anwendung der vorgeschlagenen Methodik wird anhand von sechs realen Datensätzen aus der medizinischen, toxikologischen und gartenbaulichen Forschung veranschaulicht. Wir zeigen Stärken und Schwächen unserer longitudinalen multiplen Kontrasttests auf, diskutieren deren Interpretation bei fehlenden Werten, und umreißen einige Erweiterungen und Alternativen.

Schlagworte: longitudinale Daten, lineares gemischtes Modell, multiple Hypothesentests, simultane Konfidenzintervalle, Freiheitsgrade, fehlende Werte

Contents

List of Figures

List of Tables

General Notation

This list contains symbols that are recurrently used, but it is not exhaustive. A few symbols have different meanings in different contexts e.g., β as linear model parameter and type II error rate.

$\otimes$	Kronecker product
$\mathbf{1}$	column vector of ones
α	type I error rate
b	random-effect parameter
β	fixed-effect parameter / type II error rate
c	contrast coefficient
$\mathbf{C}$	contrast matrix
$\mathbf{\Gamma}$	correlation matrix
δ	contrast margin
Δ	difference
$\mathbf{D}$	random-effects covariance matrix
$\mathbf{e}$	unit vector
ϵ	residual error
η	contrast
h	index for contrasts
H_0	null hypothesis
H_A	alternative hypothesis
i	index for independent subject
$\mathbf{I}$	identity matrix
$\mathcal{I}$	expected Fisher information
j	index for repeated measurements
$\mathbf{J}$	matrix of ones
$\mathcal{J}$	observed Fisher information
k	index for treatments

λ	Poisson parameter
$\mathbf{\Lambda}$	correlation matrix
m	number of repeated measurements
μ	mean
n	sample size
$\tilde{n}$	effective sample size
N	overall sample size
$\mathcal{N}$	(uni- or multivariate) normal distribution
ν	degrees of freedom
p	p-value
π	binomial proportion
ϕ	variance inflation factor
q	number of treatments
$\mathbf{R}$	residual covariance matrix / correlation matrix
$\mathbf{R}(\alpha)$	working correlation matrix
ρ	correlation
s^2	variance estimator
σ^2	variance
$\mathbf{\Sigma}$	covariance matrix
t	critical value from a t-distribution
T	t-type test statistic
$\mathcal{T}$	(uni- or multivariate) t-distribution
x	dependent variable
$\mathbf{X}$	fixed-effects design matrix
y	independent (outcome) variable
Y	(outcome) random variable
z	number of contrasts
Z	z-type test statistic
$\mathbf{Z}$	random-effects design matrix

List of Abbreviations

AIC	Akaike information criterion
AICc	second-order / small-sample Akaike information criterion
ANOM	analysis of means
ANOVA	analysis of variance
AUC	area under the (response-time) curve
BIC	Bayesian information criterion
CDF	cumulative distribution function
CI	confidence interval
CIM	conditional independence model
CPB	cardiopulmonary bypass
DF	degrees of freedom
EACA	ϵ-aminocaproic acid
ELM	extended linear model
FWER	familywise error rate
GEE	generalized estimating equations
GLM	generalized linear model
GLMM	generalized linear mixed-effects model
GLS	generalized least squares
$HgCl_2$	mercuric chloride
ICH	International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use
KR	Kenward-Roger (degrees of freedom)
LMM	linear mixed-effects model
LVCF	last value carried forward

MANOVA	multivariate analysis of variance
MAR	missing at random
MCAR	missing completely at random
MCP	multiple comparison procedure
MCT	multiple contrast test
MI	multiple imputation
ML	maximum likelihood
MMM	multiple marginal models
MNAR	missing not at random
OLS	ordinary least squares
PB	Pinheiro-Bates (degrees of freedom)
PDF	probability density function
REML	restricted maximum likelihood
SCI	simultaneous confidence interval
SE	standard error
SI	single imputation
UIT	union intersection test

Logical validity is not a guarantee of truth.

David Foster Wallace, Infinite Jest

1 Introduction

Simultaneous comparisons of means are frequently desired in all fields of experimental research, typically with the aim to control the familywise type I error rate (FWER) at a preset level α. But what embraces the "family" i.e., the set of comparisons whose type I errors deserve to be jointly controlled, is often a controversial issue (e.g., Stone and Chuang-Stein 2013; Wason et al. 2014), especially if multiple levels of several factors are to be compared.

Adjusting for multiplicity of comparisons among more than two treatment groups is widespread, as can be seen from the popularity of the tests for all-pairwise (Tukey 1953) 1953 and many-to-one comparisons (Dunnett 1955). By contrast, additional sources of multiplicity, such as multiple outcomes or repeated measures, are often swept under the carpet. We believe that this practice is inconsistent because the multiplicity arising from e.g., several observational time points is not very different from the multiplicity arising from several group comparisons in that it can equally invalidate a conclusion if not adjusted for.

The presence of non-randomized factors further complicates matters. Time points in a longitudinal experiment are an example of a repeated factor that is by its very nature unrandomizable and gives rise to stochastic dependence among measurements. The defining property of longitudinal data is that each experimental unit is measured repeatedly over time so that values from the same individual are positively correlated (Fitzmaurice et al. 2011). This runs contrary to the assumption of uncorrelated samples, which is the basis of most statistical methods for simultaneous inference.

There is obviously a gap to fill between the complex experimental designs that are carried out in practice and lead to correlated measurements, and the widely-used simplistic statistical analyses that ignore dependencies in the data. This becomes apparent with experimental setups that involve

a) at least one *randomized* factor (or *between-subjects* variable) whose different levels are randomly assigned to *different* experimental units,

b) at least one *repeated* factor (or *within-subjects* variable) whose different levels are investigated using the *same* experimental unit,

and every subject is exposed to exactly one level of the randomized factor but multiple levels of the repeated factor. Quan et al. (2005) called this a "two-dimensional multiplicity problem" (for the case of one randomized and one repeated factor), but in principle it can be extended to higher dimensions e.g., if there is also a spatial component to the experimental design. This thesis is mainly—but not exclusively—concerned with *longitudinally* repeated measurements in controlled experiments, so the levels of the repeated factor will often be referred to as *time points* or *occasions*, and those of the randomized factor will usually be called *treatments*.

Imagine a classical repeated measures situation where the observed outcomes are correlated between occasions (repeated measurements) but uncorrelated across treatments (due to randomization). Widespread statistical techniques such as multivariate analysis of variance (MANOVA) and repeated measures ANOVA compare the treatment means *simultaneously and jointly* at multiple occasions i.e., they provide only some global statement of "significance" expressed as p-value(s) and rely on unrealistic assumptions such as multisample sphericity (Huynh 1978). Other well-known procedures for multiple endpoints, like the T^2 test (Hotelling 1931), the ordinary and generalized least squares (OLS and GLS) methods by O'Brien (1984), and the approximate likelihood ratio test of Tang et al. (1989), are also just global procedures, and many of them are limited to comparisons of two groups only (see e.g., Sankoh et al. 1997, 1999; Wassmer et al. 1999). So if interest lies in analyzing the occasions *simultaneously but separately*, all this is neither as detailed nor as insightful as typically requested by researchers.

In practice questions like "Which treatment is better than control?" or "Can we show a dose-related trend?" may be asked for each of several points in time. Thus the goal is to study the effects of treatments *separately and simultaneously* at several occasions, which enables us to localize interesting effects rather than just claim their existence. Similarly, questions like "When is the treatment's effect superior to baseline?" or "At which point in time does the treatment cause the highest effect?" may be asked for several treatments, suggesting to assess the effects of occasions *separately and simultaneously* within several treatment groups. Here *separately* means that we want inference for each and every time point rather than for the whole time span in its entirety, and *simultaneously* means that the FWER is to be controlled at level α for the whole set of elementary comparisons.

On top of that, both types of questions can be combined into one test family. Researchers may want to assess differences among treatments as well as among occasions, with the overall FWER to be bounded by α. Then the multiple inference problem is not only two-*dimensional* but also two-*directional* in the sense that comparisons are carried out across randomized treatments and also across dependent occasions.

For any such research questions to be meaningful, the number of time points should be smallish (i.e., probably not more than five or six) as vast amounts of single test results rarely lead to enlightening conclusions. If there are lots of serially repeated measurements, it is certainly wiser to resort to a summary measure of the evolution over time, such as the area under the curve (AUC) or the slope of a linear model fit. Moreover, the measurement occasions should preferably not be picked at haphazard but rather have inherent medical or biological 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 points in time arise in many life sciences from agriculture (e.g., growth stages of crops) to zoology (e.g., larval instars of insects).

Awareness of multiplicity issues in longitudinal designs has risen in the past few years, but flawed solutions are still omnipresent in the applied sciences, as a recent survey by Liu et al. (2010) shows. Comparing treatments with individual Dunnett or Tukey tests per occasion without any further adjustment seriously inflates the overall rate of type I errors (e.g., Hoffman et al. 2008) and thus invalidates the claim made at an alleged α level. On the other hand, an overall Tukey test for all combinations of treatments and occasions will blur the results of the target effects by lots of uninteresting comparisons. Correcting for the multiplicity of occasions via Bonferroni is unnecessarily conservative, especially when occasions are highly correlated. *Ad hoc* solutions to cushion the conservatism of Bonferroni (e.g., Shi et al. (2012) proposed a simple correction factor depending on intraclass correlation) often have a very limited range of application, or it is unclear in what situations they are able to control the FWER.

So we see there is want for generally applicable methodology that takes the characteristics of repeated measurement data into account and produces valid but not unnecessarily conservative results. The prodecure must cope with three major challenges of longitudinal data:

1. Measurements from the same individuals are positively correlated i.e., adjoining occasions are likely to behave in a similar way, and correlation tends to decrease with greater separation in time.
2. Variability is likely to change over the course of the study (heteroskedasticity).
3. Missing values are the rule rather than the exception, and data may be incomplete due to individuals dropping out of the study.

As a consequence, longitudinal data require a thoughtful analysis using methods that are geared to handling correlated parameters. The principal goal of this thesis is to develop and characterize a framework for multiple inference in longitudinal data settings with either Gaussian or discrete endpoints, especially focusing on how to account for heteroscedastic errors and incorporate dependency structures of repeated measurements.

The methods described and discussed in this work are an application of multiple contrast tests (MCTs) as described in Hothorn et al. (2008). MCT procedures already exist for related scenarios such as multiple endpoints (Hasler and Hothorn 2011) and repeated measurements (Hasler 2013); however, these are not based on a proper modeling strategy for the data and hence cannot easily include covariates or missing values, which makes them inflexible in practice. In addition, the repeated measures MCTs only allow for single-group comparisons of time points.

We believe that a sound procedure for multiple inference in longitudinal settings must have its foundation in a convincing model of the data. To estimate effect sizes and (co-)variances, we consider linear mixed-effects models (LMMs) for Gaussian endpoints (Verbeke and Molenberghs 2000), generalized estimating equations (GEEs) for discrete outcomes (Hardin and Hilbe 2013), and alternatively a novel approach combining separate occasion-specific models (Pipper et al. 2012).

In addition to *localizing* effects of interest, we usually want to *quantify* them, which we cannot achieve with the p-value(s) produced by a global test. Hence we think that the presentation of adjusted p-values should be replaced—or at least accompanied—by simultaneous confidence intervals (SCIs) that give a measure of uncertainty about the effect estimates. We subscribe to the opinion of Cochran and Cox (1957, p. 5), who wrote:

> In many experiments it seems obvious that the different treatments must have produced some difference, however small, in effect. Thus the hypothesis that there is *no* difference is unrealistic: the real problem is to obtain estimates of the sizes of the differences.

Regulatory bodies have not long ago started encouraging to put this into practice. The International Conference on Harmonisation (ICH) guideline E9 (1998, p. 25–26), which is relevant for Europe, the United States, and Japan, declares that “estimates of treatment effects should be accompanied by confidence intervals, whenever possible” and emphasizes “the need to provide statistical estimates of the size of treatment effects together with confidence intervals (in addition to significance tests)”. The longitudinal MCT procedures that we propose are particularly suited to meet these requirements as they do not only yield adjusted p-values for single comparisons but also SCIs that are compatible i.e., reject the null hypothesis if and only if the corresponding test does.

The remainder of this thesis is organized as follows. Chapter 2 introduces six example datasets from medicine, toxicology, and horticulture to further motivate the problem under study. In Chapter 3 we review some relevant statistical concepts and techniques that will be used as building blocks in later chapters. Chapter 4 introduces longitudinal MCTs for Gaussian outcomes along with extensive simulations of their size and power, applications to example data, and approximate power calculations. Chapter 5 extends the longitudinal MCTs to data with discrete endpoints, assesses finite-sample properties of these asymptotic procedures, and illustrates their application to real data. Chapter 6 spotlights some related ideas, alternatives, and possible extensions. Strengths and limitations of the proposed methods are discussed in Chapter 7, followed by a concise summary of the main results and conclusions in Chapter 8.

2 Example Data

The application of novel and modified statistical procedures to real-world data is an integral part of this thesis. Our six data examples arise from various research disciplines but have in common that the underlying experimental questions can be expressed as multiple comparisons among levels of a randomized factor *separately and simultaneously* at several levels of a repeated factor, or multiple comparisons among levels of a repeated factor *separately and simultaneously* for several levels of a randomized factor.

We introduce in 2.1 a clinical trial on bradykinin receptor antagonism as quantified by D-dimer concentrations at five characteristic and medically important points in time. In 2.2 we present a toxicological study measuring the body weights of rats that were force-fed with mercuric chloride over the course of two years, and there are substantial dropout rates at later occasions. Both the bradykinin and the mercuric chloride data have Gaussian endpoints and fairly large sample sizes. The problem of small samples occurs in 2.3 with a placebo-controlled trial of two novel drugs where the patients' heart rates are recorded over time.

Our fourth data example in 2.4 is an entomological study where the repeated measurements were taken from multiple parts of the plant. Another count dataset will be introduced in 2.5: it is a clinical trial on the efficacy of progabide in the treatment of epilepsy, measured as the number of seizures i.e., a discrete rate rather than a continuous endpoint. The final example dataset has an endpoint that is a proportion: the efficacy of a biopesticide on the mortality of different developmental stages of an insect is investigated; these data are presented in 2.6.

Statistical analyses of all six datasets using the methodology developed in this thesis are going to be shown in detail in chapters 4 (for the Gaussian outcomes) and 5 (for the discrete outcomes).

2.1 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 is degraded to little protein fragments called D-dimers. Consequently, D-dimers are commonly used as a biomarker for fibrinolysis.

It is clearly desirable to avoid blood transfusion during and after surgical intervention, therefore researchers have been seeking strategies to prevent fibrinolytic degradation. It is known that CPB promotes fibrinolysis via a peptide called bradykinin and the associated bradykinin B_2 receptor. Balaguer et al. (2013) 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 (ClinicalTrials.gov identifier: NCT00223704). 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).

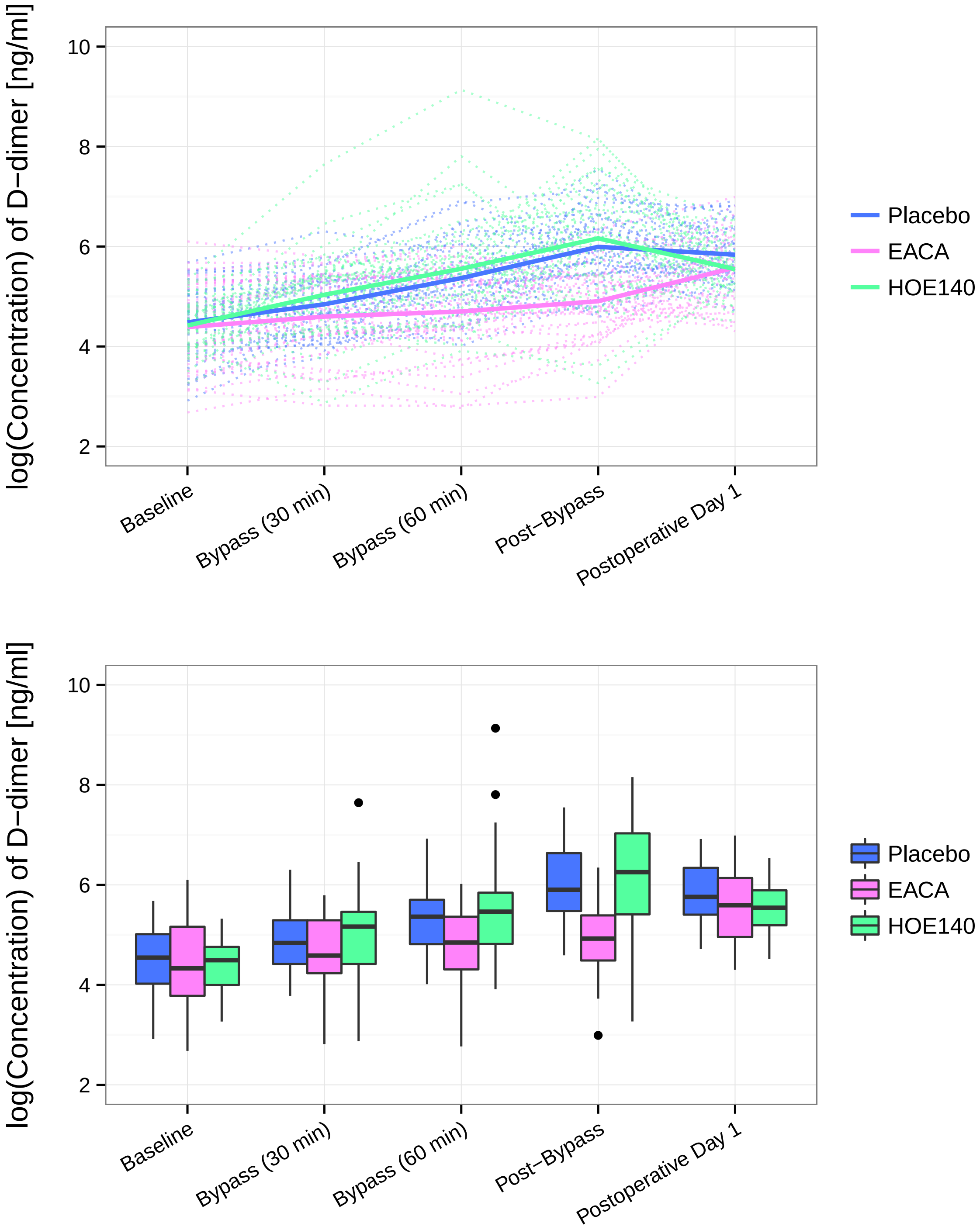


Figure 1: Bradykinin data. Top: individual patient trajectories (dotted) and sample mean trajectories per treatment arm (solid) of log-concentrations of D-dimer; bottom: boxplots.

One of the 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),
- on the first postoperative day.

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

The raw data were kindly provided by Dr. Mias Pretorius, Division of Clinical Pharmacology and Department of Anesthesiology, Vanderbilt University Medical School, Nashville, TN upon condition that artificial data be generated for use in any publication. So to avoid copyright infringements, we create and evaluate a fake dataset based on sample sizes, means, variances, and covariances of the (log-transformed) original values; these summary statistics are listed in Table 17 in Appendix C.

Prior to drawing data, we exclude patients 92 and 93 (both randomized to HOE 140) for whom no D-dimer measurements are available at all; we further exclude patient 99 (randomized to placebo) who was measured with implausible D-dimer concentrations of zero at the first two time points. Thus we are left with 38 patients in the HOE 140 group and 37 patients in the placebo and EACA arms.

Assuming that the natural logarithms of D-dimer concentrations in each treatment arm are multivariate normal, we draw random variates from five-dimensional normal distributions $\mathcal{N}_5(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with mean vector $\boldsymbol{\mu}$ and a 5×5 covariance matrix $\boldsymbol{\Sigma}$ of time points using `R` package `mvtnorm` (Genz and Bretz 2009; Genz et al. 2014). Figure 1 displays the simulated bradykinin dataset; a full table is presented in Table 18 in Appendix C.

One relevant research question for these data could be: when do which active treatments (HOE 140, EACA) reduce D-dimer concentrations (thus: reduce fibrinolysis) compared to control. We want to ask the question for treatment differences separately and simultaneously for each time point while controlling a common FWER. Another objective could be to assess the differences to baseline separately and simultaneously for each treatment arm. The analysis of bradykinin dataset with methods proposed in this thesis will be presented in 4.4.1.

2.2 Mercuric Chloride

Mercuric chloride ($HgCl_2$) was used to tan leather, disinfect seeds, embalm bodies, preserve wood, etch steel, and for many other purposes, but has fallen into disrepute after the toxicity of mercury compounds became known to the public. A long-term bioassay of the National Toxicology Program (1993, study number C60173), beginning in March 1983, investigated toxic effects of $HgCl_2$ in rodents. As part of this study, 180 female F344 rats were randomized to three dose groups: 0, 2.5, or 5 mg $HgCl_2$ per kg body weight, administered in 5 ml/kg deionized water by gavage five days a week. The animals were caged together in sets of five receiving the same treatment. Among other endpoints, their body weights were monitored over the course of two years. The study was conducted at the International Research and Development Corporation (IRDC) in

Mattawan, MI. Raw data are available from the National Toxicology Program's web database (`http://tools.niehs.nih.gov/ntp_tox`); the complete dataset is shown in Table 19 in Appendix C.

We evaluate the body weights at three points in time: when the study is halfway through (after 53 weeks), at the interim analysis after 65 weeks (when 30 animals are sacrificed to assess safety endpoints), and at the end of the study after 105 weeks (Figure 2, top and middle). Different research questions are imaginable here e.g., one could be interested in estimating the effects (with 95% confidence intervals) of low and high dosage in comparison to the vehicle control separately and simultaneously for each measurement time. A corresponding statistical analysis will be presented in 4.4.2.

One complication with this dataset is missing values (Figure 2, bottom). Few animal drop out prior to the interim analysis in week 65: only two were found moribund or dead, one in each of the active dose groups. At the end of the second year, however, about half of the data points are missing in each treatment group: ten animals per dose group were sacrificed for interim analyses, and many more found moribund or dead.

2.3 Heart Rates

The efficacy of two novel drugs, AX23 and BWW9, was assessed in a placebo-controlled clinical trial described by Milliken and Johnson (1992). 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. Complete sets of observations are available for all subjects. Table 20 in Appendix C shows the full dataset.

Plotting these data (Figure 3) reveals that heart rate patterns over time are quite different among the three treatments. The rates seem to be slightly decreasing over time for control and BWW9 whereas AX23 leads to 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 observe that BWW9 has the smallest and control the largest variance at all time points.

One research question we can answer with this dataset is: which treatments increase the heart rate compared to control at any particular occasion? This suggests comparing the treatment arms separately and simultaneously for each of the four measurement times. Another reasonable goal is to find out for each treatment arm whether there are relevant differences of heart rates over time, leading to comparisons between time points separately and simultaneously for each treatment. We will analyze the data in 4.4.3.

2.4 Greenhouse Whiteflies

The whitefly *Trialeurodes vaporariorum* is a common greenhouse pest and feared by growers because of its ability to cause dramatic economic loss. Since spraying chemicals is undesirable in a greenhouse environment, horticulturists have been searching for alternatives for efficient control of whiteflies. One solution that has arisen in the context of integrated pest management is the application of antagonists such as predatory bugs, but

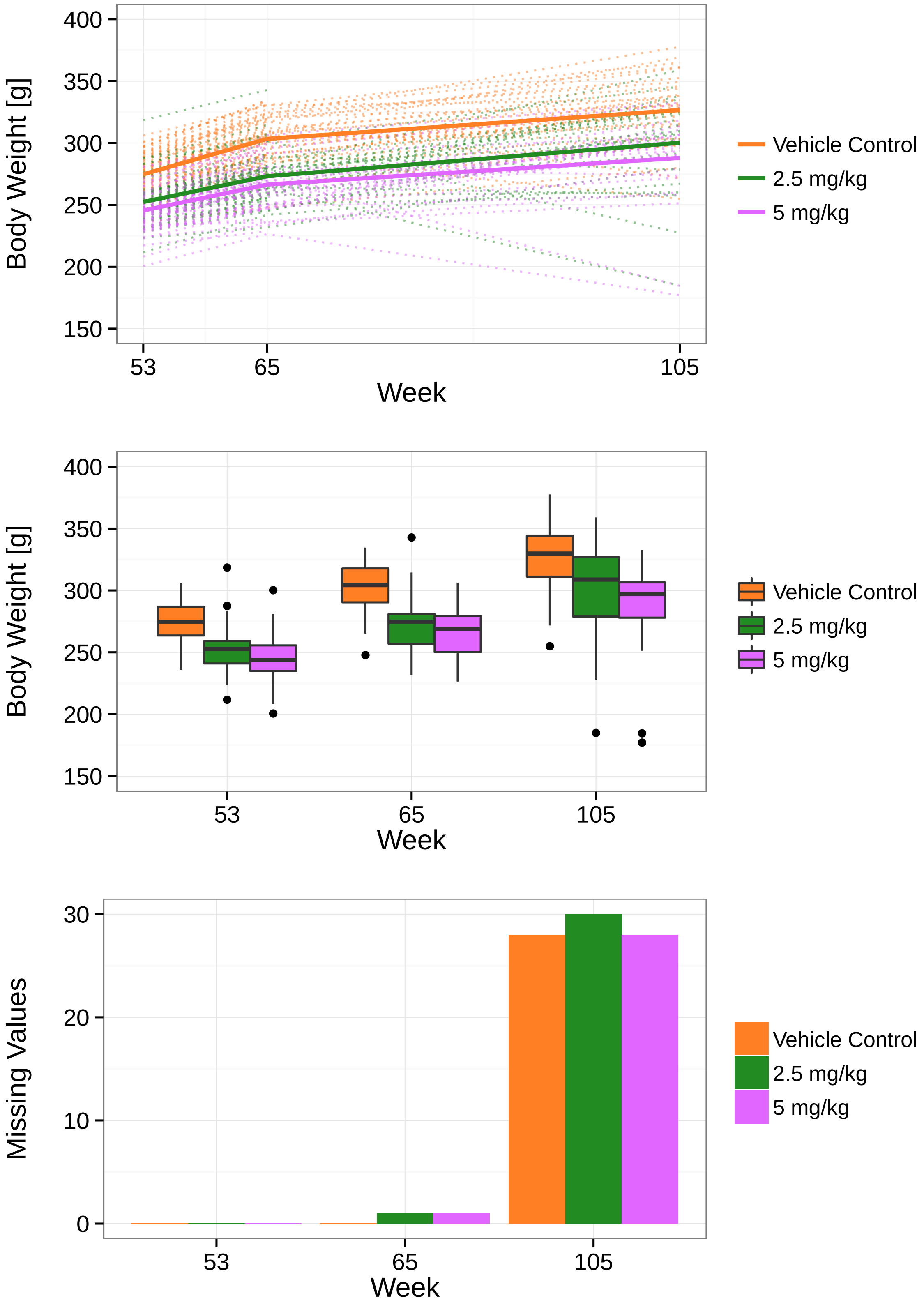


Figure 2: Mercuric chloride data. Top: individual rat trajectories (dotted) and sample mean trajectories per treatment group (solid) of body weights; middle: boxplots; bottom: numbers of missing values.

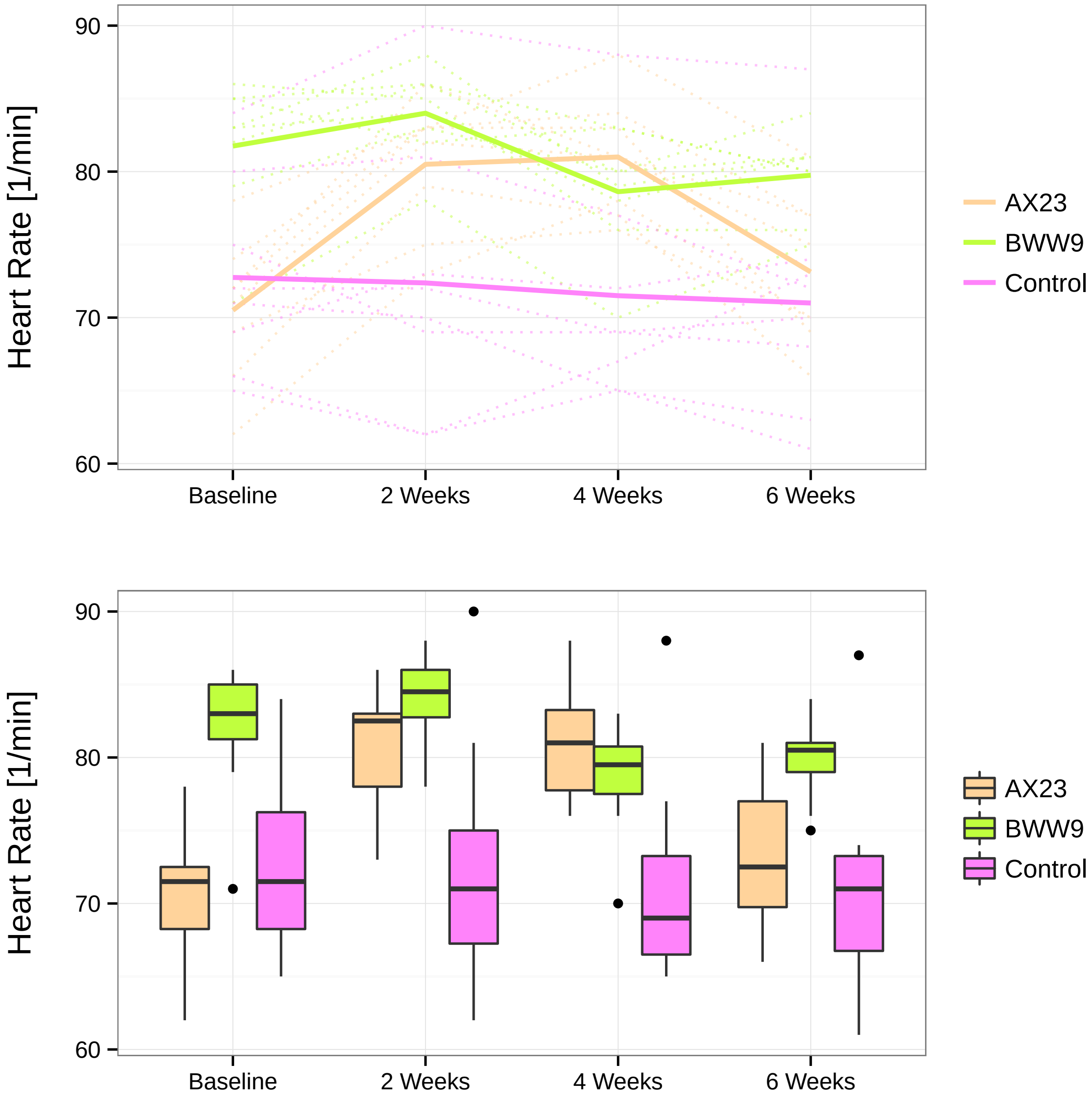


Figure 3: Heart rate data. Top: individual patient trajectories (dotted) and sample mean trajectories per treatment arm (solid) of heart rates; bottom: boxplots.

their effectiveness depends, among other factors, on the microclimatic environment. It is also important that the predators can find their prey, who might prefer different parts of the plant than their enemies.

An experiment was set up to study the preferences of whiteflies in the presence of the predator bug *Macrolophus pygmaeus* either in a conventional glasshouse or in two types of foil tunnels equipped with potted plants. The numbers of whitefly larvae per plant at midsummer were counted seperately for the three parts of each plant (top, middle, bottom), and we consider each of these plant samples as an independent experimental unit with three repeated measurements. There were 84 samples from the glasshouse and 42 from each of the two foil tunnels.

The data are graphed in Figure 4. We find by visual inspection that whiteflies obviously prefer the lower regions of the plant and avoid staying at the top. In fact, there are hardly any whiteflies at the top of most plants. Especially the conditions in foil tunnel 1 seem to reduce the number of whiteflies at the bottom and middle of the plant. Furthermore, we see that a few plants are heavily infested with whiteflies whereas others are practically pest-free. The data are clearly skewed to the right.

Unlike with the other datasets introduced in this chapter, the outcome here is measured repeatedly at multiple parts of the plant instead of multiple occasion i.e., spatial rather than temporal replication, but the methods and principles discussed in this thesis apply just as well. A few observations are missing from one of the foil tunnels (indeed, three plants could not be sampled at all).

We want to assess how the microclimatic environment influences the whiteflies' preference for the top, middle, or bottom parts of the plant. To this end we compare whitefly counts between the three parts of the plant separately and simultaneously for the glasshouse and both foil tunnels. Additionally, the question whether any environment provides better conditions for whitefly control than the others can be investigated with comparisons of environments separately and simultaneously per part of the plant. The corresponding analyses using statistical methods established in this thesis will be undertaken in 5.4.1.

These data are a subset of a larger series of experiments conducted by Elias Böckmann at the Institute of Horticultural Production Systems, Department Phytomedicine, Leibniz University Hannover in 2013. Table 22 in Appendix C shows the raw data in full.

2.5 Epileptic Seizures

Thall and Vail (1990) present data from a placebo-controlled clinical trial on the efficacy of progabide in the treatment of epilepsy, with 31 patients randomized to progabide and 28 to placebo. The outcome of interest is the number of seizures, recorded in two-week intervals two, four, six, and eight weeks after randomization. Their paper also gives a baseline value, which is the number of seizures in the eight-week interval before randomization; we divided this by four to make it intuitively comparable to the two-week intervals after randomization.

Figure 5 shows that the mean numbers of seizures are very similar in both arms, and they are also more or less constant over time. The data are heavily skewed to the right: most patients experience up to 20 or 30 seizures during two weeks, but there are a few exceptions, for example one patient in the progabide arm with 102 seizures in the first interval after randomization. This person's individual profile is way higher than any other patient's profile. Generally, individuals with relatively many (few) seizures tend to have similarly many (few) seizures in all intervals, so the numbers of seizures in a patient are positively correlated over time.

One question to be answered with these data is whether progabide is capable of reducing the rate of seizures during any two-week interval after initiation of the treatment. The analysis using comparisons separately and simultaneously for each time point will be presented in 5.4.2. The full dataset is given in Table 21 in Appendix C.

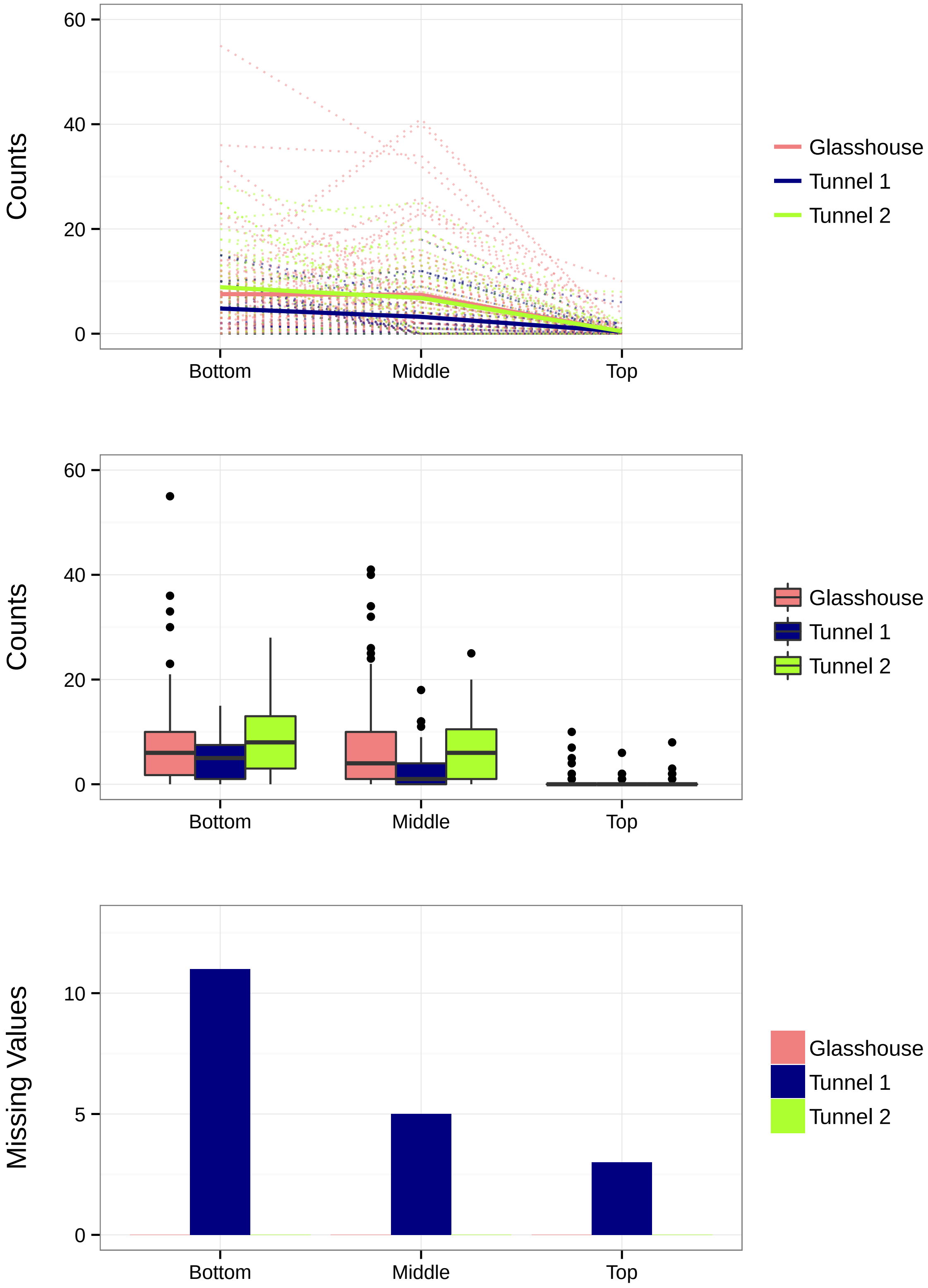


Figure 4: Greenhouse data. Top: individual sample trajectories (dotted) and mean trajectories per environment (solid) of whitefly counts at each of three different parts of the plant; middle: boxplots; bottom: numbers of missing values.

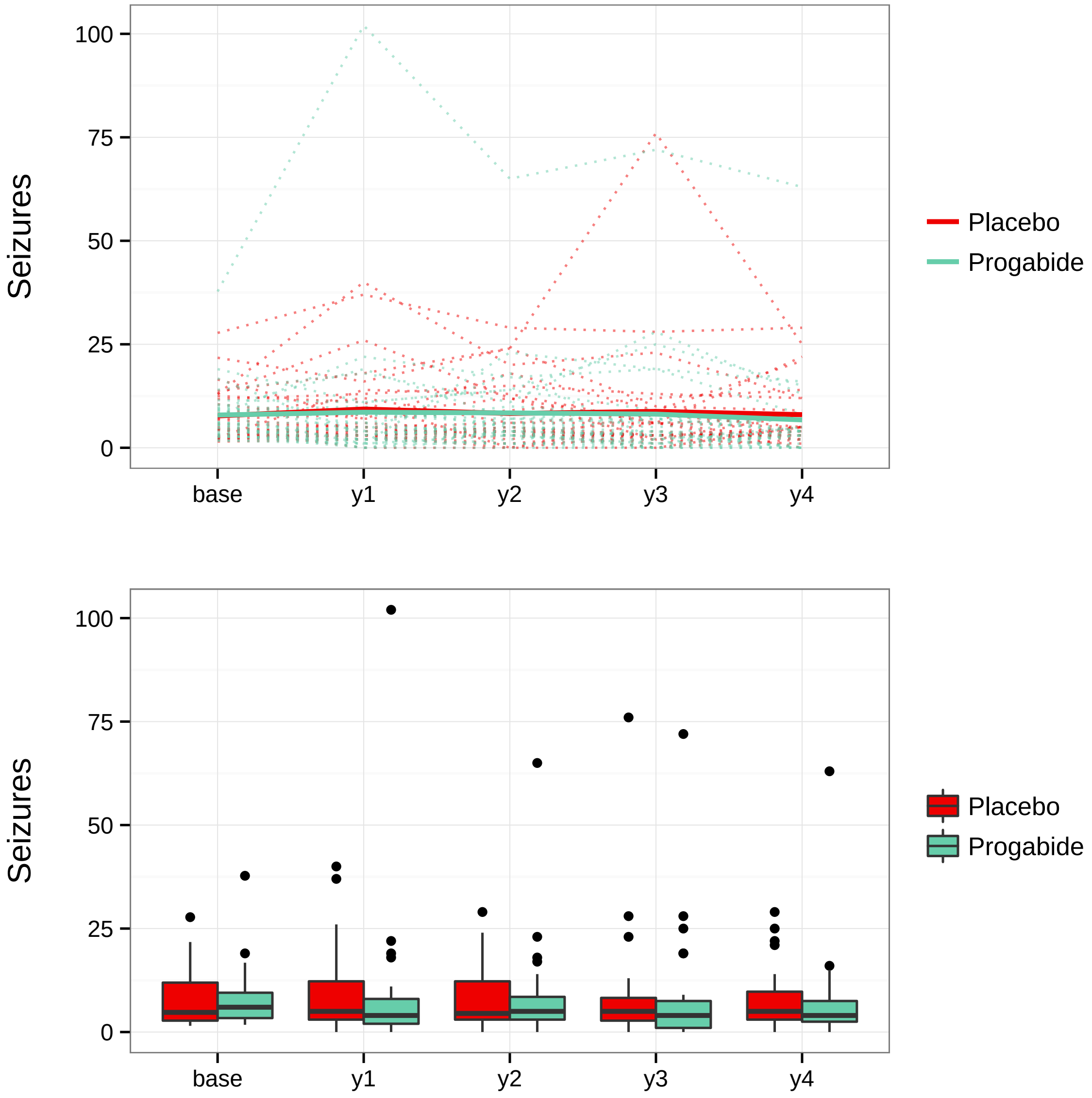


Figure 5: Epileptic seizure data. Top: individual patient trajectories (dotted) and sample mean trajectories per treatment arm (solid) of seizures per two weeks; bottom: boxplots.

2.6 Azadirachtin

Seeds of the neem tree (*Azadirachta indica*) contain a secondary metabolite called azadirachtin that is known to disrupt the development of insects and act as an antifeedant. On the other hand it is non-toxic to mammals; therefore azadirachtin has become increasingly popular as a biological insecticide for organic farming and integrated pest management.

Neem extracts can be applied either as a spray on the plant surface or as a drenching solution to be taken up via the roots. The disadvantage of spraying is the limited duration of effect due to azadirachtin degrading very fast under the influence of ultraviolet radiation. Soil application of azadirachtin is particularly appealing for the purpose of plant protection because it is systemically distributed all over the plant and quickly translocated to the insect feeding sites.

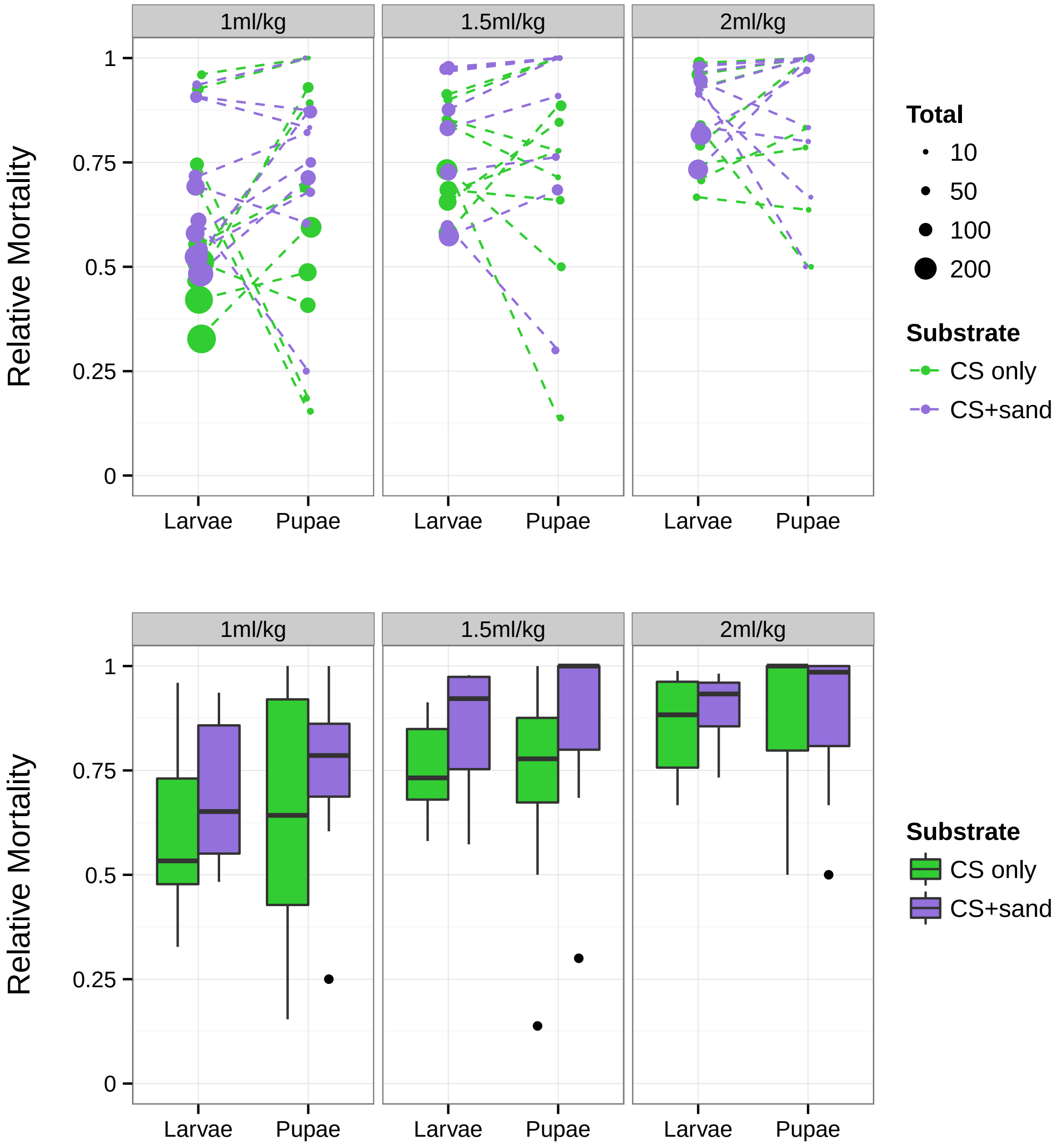


Figure 6: Azadirachtin data. Top: individual pot trajectories of relative larval and pupal mortalities with different substrates; bottom: boxplots.

We consider an investigation of the effect of soil-applied azadirachtin on larval and pupal mortality of the greenhouse whitefly *Trialeurodes vaporariorum*. 60 tomato plants in single pots were randomly arranged on a greenhouse table. Five female whitefly adults were placed on one well-developed leaf per plant, given the opportunity to lay eggs so that a population is initiated, and then removed again after one day. Then they were treated with different doses of the commercial product "NeemAzal-T", a liquid formulation containing 1% of the active ingredient azadirachtin.

It was hypothesized that the root uptake of azadirachtin may depend on the organic matter content in the soil. Thus half of the plants were grown in commercial substrate and the other half in a 1:1 mixture of substrate and sand. Within each of these portions of 30 plants, they received three different treatment dosages in sets of ten:

- 1 ml drench (amounting to 10 mg Azadirachtin) per kg of soil,
- 1.5 ml drench (amounting to 15 mg Azadirachtin) per kg of soil,
- 2 ml drench (amounting to 20 mg Azadirachtin) per kg of soil.

The dose recommended by the manufacturer is 1 ml per kg of soil.

Four values were recorded for each single pot:

- the total number of larvae that hatched from the eggs,
- the number of larvae found dead,
- the number of larvae that pupated,
- the number of pupae found dead.

The data are displayed in Figure 6 as larval and pupal mortalities. An important characteristic is that the numbers of whiteflies exposed to the neem treatments are highly variable between single plants, ranging from 27 to 275 (larvae) and from 1 to 185 (pupae).

Larval and pupal mortality proportions are likely to be correlated since they were observed from the same experimental units. We will evaluate them in 5.4.3 using the methods proposed in this thesis, focusing on two research questions. First, what is the impact of raising the dosage beyond the manufacturer's recommendation on larval and pupal mortality? And second, do we see a difference when sand is added to the substrate?

The data were taken from a series of experiments carried out by Josephine Karanja at the Institute of Horticultural Production Systems, Department Phytomedicine, Leibniz University Hannover in 2014, and published in Karanja et al. (2015). The raw values of the subset considered in this thesis are shown in Table 23 in Appendix C.

3 Methods

This chapter introduces some statistical concepts and methods that are to be used as building blocks for the longitudinal MCTs in Chapters 4 and 5. We review linear mixed-effects models (LMMs) and consider different structures for the random effects and residual covariances in 3.1. Selection of a model from a set of candidates using the AICc criterion is described in 3.2. Generalized linear models (GLMs) and generalized estimating equation (GEEs) for discrete data are introduced in 3.3 and 3.4. The idea of combining multiple marginal models is presented in 3.5. We describe general simultaneous inference in 3.6, and multiple contrast tests (MCTs) are in the spotlight of 3.7, with a critical assessment of existing MCT procedures for correlated data settings. 3.8 lists various small-sample approximations for the degrees of freedom (DF) to be used in longitudinal MCTs. Different definitions of power in multiple testing are summarized in 3.9. We conclude the chapter with an overview of missing data issues in 3.10.

3.1 Linear Mixed-Effects Models

3.1.1 General Concepts

A standard way of processing longitudinal data involves fitting a linear mixed-effects model (LMM) that contains fixed-effect parameters governing the model's mean structure, and random effects that usually reflect the randomization or hierarchical structure of the experiment and imply a certain covariance matrix. Motivated by Cnaan et al. (1997), we adopt here a very general notion of LMMs to emphasize that their great flexibility shall be exploited in many ways: random effects may be included or not; error variances may be modeled as homo- or heteroscedastic; and error covariances may or may not be taken as zero. Thus our *general*[1] LMM can be one that contains, apart from the error term, only fixed effects. Using notation similar to that of Laird and Ware (1982), we write the model as

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i \ , \quad i = 1, \ldots, n$$

where $\mathbf{y}_i$ is a vector of responses from subject i, $\mathbf{X}_i$ and $\mathbf{Z}_i$ are known design matrices for the fixed and random effects, and $\boldsymbol{\beta}$ are the fixed-effects coefficients shared by all subjects. The random effects parameters $\mathbf{b}_i$ and residual errors $\boldsymbol{\epsilon}_i$ are conventionally assumed to follow Gaussian distributions with mean zero and covariance matrices $\mathbf{D}$ (same for all $\mathbf{b}_i$) and $\mathbf{R}_i$:

$$\mathbf{b}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{D}) \quad \text{and} \quad \boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_i).$$

In addition, we assume that random effects and errors are independent of one another:

$$Cov(\mathbf{b}_i, \boldsymbol{\epsilon}_i) = \begin{pmatrix} \mathbf{D} & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_i \end{pmatrix}.$$

The marginal expectation of $\mathbf{y}_i$ is

$$E(\mathbf{y}_i) = \mathbf{X}_i\boldsymbol{\beta}$$

[1] This notion of *general* is not to be confused with the class of *generalized* linear models (see 3.3), which allow for error distributions other than Gaussian but are inherently fixed-effects models.

with marginal variance

$$Var(\mathbf{y}_i) = \mathbf{\Sigma}_i = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^T + \mathbf{R}_i,$$

that is composed of a between- and a within-individual portion, so in summary

$$\mathbf{y}_i \sim \mathcal{N}(\mathbf{X}_i\boldsymbol{\beta},\ \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^T + \mathbf{R}_i).$$

The generalized least squares (GLS) estimator for the fixed-effect parameters is

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n} \mathbf{X}_i^T \hat{\mathbf{\Sigma}}_i^{-1} \mathbf{X}_i\right)^{-1} \sum_{i=1}^{n} \mathbf{X}_i^T \hat{\mathbf{\Sigma}}_i^{-1} \mathbf{y}^{-1}$$

with estimated variance

$$Var(\hat{\boldsymbol{\beta}}) = \left(\sum_{i=1}^{n} \mathbf{X}_i^T \hat{\mathbf{\Sigma}}_i^{-1} \mathbf{X}_i\right)^{-1}$$

where $\hat{\mathbf{\Sigma}}_i$ is the restricted maximum likelihood (REML) solution to $\mathbf{\Sigma}_i$ (Patterson and Thompson 1971), which is commonly preferred to maximum likelihood (ML) as the latter is biased because it ignores the degrees of freedom used for estimating the fixed effects. We refer to the literature for minutiae on LMM theory and computation: excellent reference texts are Searle et al. (1992), Verbeke and Molenberghs (2000), Pinheiro and Bates (2000), McCulloch and Searle (2001), and Fitzmaurice et al. (2011).

We would like to spotlight two important special cases of our *general* LMM concept:

Conditional independence model (CIM): One common simplification relies on the so-called conditional independence assumption

$$\mathbf{R}_i = \sigma^2\mathbf{I}$$

i.e., the residual errors are assumed independent, conditional on the random effects. This implies that all within-subject association must be absorbed by the random-effects covariance matrix $\mathbf{D}$. In addition, the errors are assumed to be homoscedastic.

Extended linear model (ELM): The "other extreme" (so-termed by Cnaan et al. 1997) is to leave out any random effects other than the errors i.e., to set $\mathbf{Z}_i = \mathbf{0}$. This yields an extended linear (fixed-effects) model

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i$$

that may account for heteroscedasticity and correlation through its error covariance matrices $\mathbf{R}_i$ on which we may impose a pattern (e.g., compound symmetric or autoregressive) or leave them completely unstructured (see 3.1.2 for an overview of options). This model still comes within our notion of *general* LMMs even though it is downgraded—but note that the errors are also random.

In practice, the choice whether to fit a CIM, an ELM, or an LMM including both random effects (other than $\boldsymbol{\epsilon}$) and some structure on the error covariances will be subject-matter-driven, and we do not feel obliged to give any general "statistical" recommendation here. We discuss consequences of over- and underfitting in 3.1.3. In case of uncertainty, model selection tools may be consulted, on which we will elaborate in 3.2.

3.1.2 Residual Covariance Patterns

An open problem when fitting an LMM is to decide upon a model for the error covariances $\mathbf{R}_i = \sigma^2 \mathbf{\Lambda}_i$. The simplest variant is to set $\mathbf{\Lambda}_i = \mathbf{I}$ for all $i = 1, \ldots, n$, leading to the CIM. In practice, however, allowing for heteroscedasticity and/or correlatedness of within-subject errors can improve the model fit substantially. This is for sure indispensable in ELMs with no random effects (besides the errors) but may also be advantageous in any LMM where the random effects specification captures the within-subject associations insufficiently.

In any case the goal is not to devise the *correct* covariance structure (whatever that means) but rather to find a viable and economical approximation. Indeed estimating a full unstructured correlation matrix and heterogeneous variances is often self-defeating:

1. Parameters may be unestimable, especially for small datasets.
2. With increasing dimensionality (i.e., number of occasions) estimation quickly becomes a computational burden.
3. Lots of parameters are wasted even when a much more parsimonious structural model would serve the same purpose.

This undesirable situation can be resolved by

a) checking thoroughly to what degree heteroscedasticity needs to be modeled, and

b) imposing a sparse parametric pattern on the matrix of within-subject correlations.

General considerations on the choice of variance and correlation structure are given in the following. A good strategy is probably to work out a few plausible models for the $\mathbf{R}_i$ and "let the data choose" be means of a selection criterion (see 3.2).

To describe structural covariance modeling, we decompose the covariance matrix of the within-subject errors into a variance part $\mathbf{V}_i$ and a correlation part $\mathbf{\Lambda}_i$ which are independent of one another:

$$\mathbf{R}_i = \mathbf{V}_i \mathbf{\Lambda}_i \mathbf{V}_i$$

where $\mathbf{V}_i$ is a diagonal matrix with strictly positive entries on the main diagonal, and $\mathbf{\Lambda}_i$ is symmetric and positive definite with all diagonal elements equal to one.

Variance Structures: For any subject i belonging to treatment group k, the variance part is

$$\mathbf{V}_i = \begin{pmatrix} \sigma_{i1} & & \\ & \ddots & \\ & & \sigma_{im} \end{pmatrix}$$

with σ_{ij} the square root of the variance at measurement occasion j. Now we obviously do not want to model subject-specific variances; patterns of practical relevance allow the variance to vary across measurement occasions $j = 1, \ldots, m$ and/or treatment groups $k = 1, \ldots, q$ i.e., we work with σ_{jk}. We consider four variance schemes of increasing complexity:

- Fully homoscedastic: $\sigma_{jk} = \sigma$

Variances are assumed constant across both treatment groups and measurement occasions. This is the simplest model but highly unrealistic in any actual longitudinal dataset.

- Heteroscedastic over time: $\sigma_{jk} = \sigma_j$

  Variances changing in the course of time are a common occurrence in longitudinal studies, and it may be reasonable to assume that they are not considerably different in the treatment groups.

- Heteroscedastic over treatments: $\sigma_{jk} = \sigma_k$

  Variances being constant over time but different from treatment to treatment are unlikely to occur in longitudinal data but possibly in other repeated measurement settings e.g., with multiple endpoints.

- Fully heteroscedastic: σ_{jk}

  Variances are allowed to vary between measurement occasions *and* treatment groups. Such a detailed model will be difficult to fit and to motivate with small sample sizes but may be justified for larger datasets.

A rather parsimonious strategy to model heteroscedasticity could involve a variance function of some sort e.g., an exponential or power variance function.

Correlation Structures: The second component of the residual covariance matrix $\mathbf{R}_i$ is the correlation part $\mathbf{\Lambda}_i$. We list here some frequently used correlation patterns and discuss their applicability to longitudinal and repeated measures settings. The matrices are exemplified for $m = 4$, and since they are symmetric, only their upper triangles are displayed. The restriction $|\rho| \leq 1$ applies to all correlation parameters with or without subscripts.

- Independence (IND):

$$\mathbf{\Lambda}_i = \mathbf{I}_m = \begin{pmatrix} 1 & 0 & 0 & 0 \\ & 1 & 0 & 0 \\ & & 1 & 0 \\ & & & 1 \end{pmatrix}$$

  The most naive way to deal with a repeated measures situation is to flatly ignore any correlation among the time points. The assumption of independent errors (which is implicit in the standard linear model) is highly unrealistic for longitudinal or any other correlated data and will lead to grossly invalid standard errors (SEs).

- Compound symmetry (CS):

$$\mathbf{\Lambda}_i = \begin{pmatrix} 1 & \rho & \rho & \rho \\ & 1 & \rho & \rho \\ & & 1 & \rho \\ & & & 1 \end{pmatrix}$$

  Compound symmetry requires just one parameter ρ to be estimated but on the other hand implies that all measurements are equally correlated. This is a questionable assumption for longitudinal data, where the strength of association is likely to decrease with increasing separation in time.

- First-order autoregressive (AR(1)):

  $$\mathbf{\Lambda}_i = \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 \\ & 1 & \rho & \rho^2 \\ & & 1 & \rho \\ & & & 1 \end{pmatrix}$$

  AR(1) is equally parsimonious in parameters as CS but able to reflect that correlation decreases (exponentially) with increasing time gaps between occasions. This makes it a favored pattern for longitudinal data. As a limitation, it requires that measurements are obtained at equally spaced points in time. This restriction can be overcome with the generalization to CAR(1). Higher-order autoregressive structures are conceivable but rarely realized in practice.

- Continuous first-order autoregressive (CAR(1)):

  $$\mathbf{\Lambda}_i = \begin{pmatrix} 1 & \rho^{|t_2-t_1|} & \rho^{|t_3-t_1|} & \rho^{|t_4-t_1|} \\ & 1 & \rho^{|t_3-t_2|} & \rho^{|t_4-t_2|} \\ & & 1 & \rho^{|t_4-t_3|} \\ & & & 1 \end{pmatrix}$$

  The continuous generalizaton of AR(1) is appropriate if the measurements are not equally spaced in time as they take into account the lags $|t_{j'} - t_j|$ between time points t_j and $t_{j'}$, with $j, j' = 1, \ldots, m$. For data with constant lags, CAR(1) is the same as AR(1).

- Toeplitz (TOEP):

  $$\mathbf{\Lambda}_i = \begin{pmatrix} 1 & \rho & \rho_2 & \rho_3 \\ & 1 & \rho & \rho_2 \\ & & 1 & \rho \\ & & & 1 \end{pmatrix}$$

  Toeplitz structures assume that correlation of (equally spaced) occasions varies with their separation in time. Unlike with AR(1), however, there is no restriction to exponential decay. This flexibility comes at the cost of having to estimate $m-1$ correlation parameters instead of just one.

- Unstructured (UN):

  $$\mathbf{\Lambda}_i = \begin{pmatrix} 1 & \rho_{12} & \rho_{13} & \rho_{14} \\ & 1 & \rho_{23} & \rho_{24} \\ & & 1 & \rho_{34} \\ & & & 1 \end{pmatrix}$$

  A completely unstructured pattern will reflect the data's correlation structure most accurately, thus minimizing the risk of misspecification. However, the absence of constraints for the matrix elements inflates the number of parameters to $\frac{m(m+1)}{2}$.

Further correlation patterns include higher order autoregressive (e.g., AR(2)), moving average (MA), autoregressive moving average (ARMA), antedependence (ANTE), factor analytic (FA), spherical, Huynh-Feldt (HF), and various spatial structures; in addition, banding can be introduced where all entries in higher off-diagonals are set to zero. For an overview of covariance patterns for longitudinal and repeated measures designs consult e.g., Jennrich and Schluchter (1986), Diggle et al. (1994), Wolfinger (1993), Wolfinger (1996), and Littell et al. (2006).

3.1.3 Modeling Random Effects and Error Covariances

When facing a practical data problem, one has to decide on a model for $\mathbf{V}_i$ and $\boldsymbol{\Lambda}_i$ and/or the random effects. The dilemma here is that both under- and overfitting can be detrimental. In real-data problems, we need to distinguish between appropriate simplifications and problematic misspecifications that ignore vital features of the data. A realistic goal is to find "a good enough" model (Cheng et al. 2010) rather than "the best" or even "the correct" one.

Imposing a parametric structure always bears the risk of getting it wrong. Various researchers have studied misspecifications of random effects or covariance structures and their impact on fixed-effects inference. Jacqmin-Gadda et al. (2007) and Schielzeth and Forstmeier (2009) warned against overoptimistic inferences as a consequence of insufficient random-intercept-only models. Likewise, Gurka et al. (2011) showed that underspecification (e.g., only CS due to only a random intercept) inflates the type I error rate for fixed-effect tests.

To avoid such underfitting and "be on the safe side", one could be tempted to generally estimate UN correlation and separate variances per time point (in an ELM), or the most complex random-effects structure (in a CIM). This seems appealing at first sight and may be fine when the sample size is large relative to the number of time points. However, Littell et al. (2000) pointed out that UN always fits but the SE estimates may be unstable. Park et al. (2001) assessed type I error rates and power and stated that especially with large numbers of occasions, estimating UN covariances becomes inefficient. Lu and Mehrotra (2010) proposed to always use UN in order to avoid biased estimates of $\boldsymbol{\beta}$ under missingness at random (MAR) but also admitted this may lead to convergence problems. So for small to medium sample sizes (as they are common practice in the life sciences) one runs the risk of overfitting the data and getting very instable covariance estimates, or even of being unable to actually fit the model in the first place.

A simple illustration clarifies this matter. Assume that we have four treatment groups in our experiment and the endpoint is evaluated repeatedly at five successive time points. In the simplest case (pooled variance estimate across time points and parsimonious correlation structure such as AR(1)) only two parameters are spent to model the covariance structure. By contrast, fitting unstructured heteroscedastic model inflates the number of parameters for the covariance matrix to 15 (five variances and ten pairwise correlations). If we wanted to generalize the model even further and include separate covariance matrices for different groups, the number of parameters to be estimated would literally rocket upwards (20 variances and 40 pairwise correlations).

Thus in many cases a reasonable approximation with few parameters is more useful than an overparameterized correlation matrix or maximal random effects. And instead of choosing a covariance pattern oneself, one could set up a collection of candidate model and "let the data choose". This idea of selecting among various covariance structures appears in Jennrich and Schluchter (1986) and Wolfinger (1993, 1996) and is supposed to minimize the risk of gross misspecification. Keselman et al. (1998) compared AIC and the Bayesian information criterion (BIC) of Schwarz (1978) for this purpose, however with the rather irrelevant focus on their ability to pick the "correct" model. We will pursue the model selection approach in 3.2.

3.2 AICc Model Selection

In real-world life sciences, the "true" process that generated a given dataset is complex and *per se* unknown and unknowable. Hence often a model of some kind is employed to provide a passably simple approximation to the truth. Such a model is useful for obtaining estimates of parameters with a practical interpretation, for making inference about such parameters, and for predictions. At the end of the day, it is a tool that helps to understand what is going on. However, a model is not, and cannot be, truth itself. In other words, the intent of modeling is not to rebuild the true underlying mechanism but to grasp its most important features.

The predominant *modus operandi* of arriving at a model deemed good is to pick one (e.g., because it has been used before or others have used it before or just because it "feels" right) and then settle for it. Such an approach involves a good deal of arbitrariness. A more objective strategy is to assemble an entire set of plausible candidate models and then select the best approximating model by means of a sensible criterion. For this purpose we utilize a concept that is well-founded in information theory: the Kullback-Leibler distance (Kullback and Leibler 1951) measures the loss of information when the unknown true mechanism f is approximated with model g, indexed with parameters θ:

$$I(f,g) = \int f(x) \log \left(\frac{f(x)}{g(x|\theta)} \right) \, dx.$$

From this perspective it becomes obvious that a good model captures as much information about the true process as possible (i.e., does not underfit the data) without overinterpreting data structures that are essentially just noise (i.e., not overfitting the data). This is the familiar bias-variance tradeoff in statistics.

A straightforward way of harnessing information theory for practical model selection is Akaike's information criterion (AIC) (Akaike 1974), which estimates the relative expected Kullback-Leibler distance:

$$AIC = -2\ell(\hat{\theta}) + 2K$$

where $\ell(\hat{\theta}) = \log(\mathcal{L}(\hat{\theta}|y))$ is the maximized log-likelihood given the data and K the number of model parameters. When presented with a set of R candidate models, AIC chooses the one that provides the best approximation to the unknown truth in terms of Kullback-Leibler distance. Contrary to a common misbelief, suchlike model selection makes no assumption that the true model is among the candidates (Burnham and Anderson 2002).

When the sample size is small relative to the number of estimated model parameters, the so-called second-order AIC or AICc (Sugiura 1978; Hurvich and Tsai 1989) should be preferred to AIC:

$$\begin{aligned} AICc &= AIC + \frac{2K(K+1)}{n-K-1} \\ &= -2\ell(\hat{\theta}) + 2K \left(\frac{n}{n-K-1} \right) \end{aligned}$$

where n is the sample size. The small-sample bias adjustment has a noticable impact for $\frac{n}{K} < 40$, as recommended by Burnham and Anderson (2002). AIC and AICc are asymptotically equal.

A single AICc value is not interpretable; it becomes only meaningful in the context of the candidate set that contains R different models. Therefore the transformation

$$\Delta_l = AICc_l - AICc_{\min}$$

is useful where $AICc_{\min}$ is the minimum AICc among the set of R candidate models. The "best" model is assigned $\Delta = 0$, and all other models get values greater than zero.

Another handy measure is the Akaike weight

$$w_l = \frac{e^{-\frac{\Delta_l}{2}}}{\sum_{r=1}^{R} e^{-\frac{\Delta_r}{2}}},$$

which can be interpreted as the probability that model l is the best model in a Kullback-Leibler sense i.e., conditional on the data and the candidate set.

3.3 Generalized Linear Models

Generalized linear models (GLMs) as introduced by Nelder and Wedderburn (1972) and detailed in McCullagh and Nelder (1989) extend the classical linear model (e.g., Searle 1971) to discrete outcomes such as counts, rates, and proportions. Suppose a setting with observational units $i = 1, \ldots, n$ in treatment groups $k = 1, \ldots, q$, then a GLM is built on three components:

1. a distributional assumption for the outcome that is a member of the exponential family (such as Gaussian, binomial, or Poisson) and implicates an expression for the expected value of the random variables $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})$

   $$E(\mathbf{Y}_i) = \boldsymbol{\mu}_i$$

   as well as the variance

   $$Var(\mathbf{Y}_i) = \phi V(\boldsymbol{\mu}_i) = \phi \mathbf{a}_i,$$

2. a linear predictor

   $$\boldsymbol{\eta}_i = \sum_{k=1}^{q} x_{ik}\beta_k = \mathbf{x}_i\boldsymbol{\beta},$$

3. a link function

   $$g(\boldsymbol{\mu}_i) = \boldsymbol{\eta}_i$$

   that connects the linear predictor to the expected value of the outcome variable.

In consequence, we assume linearity and additivity of effects on the link.

We consider three widespread special cases of GLMs:

Gaussian: The standard linear model for Gaussian data can be viewed as a GLM with identity link function

$$g(\boldsymbol{\mu}_i) = \boldsymbol{\mu}_i,$$

and the variance is independent of the mean so that

$$\phi = \sigma^2 \quad \text{and} \quad V(\boldsymbol{\mu}_i) = 1.$$

Poisson: The canonical link under Poisson assumption is

$$g(\boldsymbol{\mu}_i) = \log(\boldsymbol{\mu}_i),$$

and the variance is equal to the mean:

$$V(\boldsymbol{\mu}_i) = \boldsymbol{\mu}_i.$$

Binomial: The canonical link function for binomial data is

$$g(\boldsymbol{\mu}_i) = \text{logit}(\boldsymbol{\mu}_i) = \log\left(\frac{\boldsymbol{\mu}_i}{1-\boldsymbol{\mu}_i}\right)$$

but other links such as probit or complementary log-log are conceivable as well. The binomial variance is

$$V(\boldsymbol{\mu}_i) = \boldsymbol{\mu}_i(1-\boldsymbol{\mu}_i).$$

For both Poisson and binomial GLMs we have $\phi = 1$ unless there is overdispersion, meaning that the data exhibit variability that exceeds the sampling variance according to Poisson or binomial theory. Then a variance inflation factor $\phi > 1$ can estimated from the data in a quasi-likelihood framework (Wedderburn 1974). Alternative remedies for overdispersion include estimating robust “sandwich” variances (Zeileis 2006) or switching to another distribution (e.g., use negative binomial for counts, or beta-binomial for proportions).

3.4 Generalized Estimating Equations

A popular way to model correlated discrete outcomes are generalized estimating equations (GEEs) as introduced by Liang and Zeger (1986) and Zeger and Liang (1986). This technique can be viewed as an extension of GLMs to non-independent settings. No joint likelihood has to be specified with GEEs.

Using notation from 3.3, consistent estimates of $\boldsymbol{\beta}$ are found by solving

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \mathbf{V}_i^{-1}(\mathbf{Y}_i - \boldsymbol{\mu}_i) = \mathbf{0}$$

with working covariance matrix

$$\mathbf{V}_i = \phi \mathbf{A}_i^{\frac{1}{2}} \mathbf{R}_i(\alpha) \mathbf{A}_i^{\frac{1}{2}}$$

where $\mathbf{A}_i$ is the diagonal matrix of the a_{ik} as defined in 3.3, and $\mathbf{R}_i(\alpha)$ a working correlation matrix depending only on a parameter α. This gives

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\sqrt{n} \overset{asym.}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{K})$$

with

$$\mathbf{K} = \lim_{n\to\infty} n \left(\sum_{i=1}^{n} \frac{\partial \boldsymbol{\mu}_i^T}{\partial \boldsymbol{\beta}} \mathbf{V}_i \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}^T} \right)^{-1} \left(\sum_{i=1}^{n} \frac{\partial \boldsymbol{\mu}_i^T}{\partial \boldsymbol{\beta}} \mathbf{V}_i Cov(\mathbf{Y}_i) \mathbf{V}_i \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}^T} \right) \left(\sum_{i=1}^{n} \frac{\partial \boldsymbol{\mu}_i^T}{\partial \boldsymbol{\beta}} \mathbf{V}_i \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}^T} \right)^{-1}.$$

The robust sandwich covariance estimator of $\mathbf{K}$ is obtained by inserting consistent estimates for $\boldsymbol{\beta}$, ϕ, and α, and $(\mathbf{Y}_i - \boldsymbol{\mu}_i)(\mathbf{Y}_i - \boldsymbol{\mu}_i)^T$ for $Cov(\mathbf{Y}_i)$ (e.g., Halekoh et al. 2006). This yields consistent estimates of the covariances even under misspecification of $\mathbf{R}_i(\alpha)$ (Hardin and Hilbe 2013). The alternative would be a naive covariance estimation directly from the data.

Another strategy for modeling correlated discrete outcomes is offered by the family of generalized linear mixed-effects models (GLMMs) (e.g., Breslow and Clayton 1993). They require to specify a complete joint distribution of the data and account for within-subject correlation by introducing random effects to the linear predictor of a GLM. This can make them practically challenging (Bolker et al. 2009) and also computationally instable (Zhang et al. 2011). And if the distributional assumptions for the random effects and sampling variability are violated, GLMM estimates will be biased.

Another crucial difference between GEEs and GLMMs lies in the interpretation of model parameters: GEE estimates are to be viewed as population-averaged effects whereas fixed-effect estimates from a GLMM have a subject-specific meaning i.e., they describe changes *within* individuals (e.g., Zeger et al. 1988; Young et al. 2007; Zhang et al. 2012). Note that there are covariates that cannot change within a subject such as sex, race, species, etc.

One way to overcome some of the limitations of GLMMs is to "marginalize" them i.e., model the marginal mean rather than a mean that is conditional on random effects. These marginalized multilevel models (Heagerty and Zeger 2000; Wang and Louis 2004; Griswold et al. 2013) extend the interpretability of GLMM parameters and are less sensitive to misspecifications.

3.5 Multiple Marginal Models

3.5.1 General Methodology

Instead of basing multiple inferences on a single model that includes all time points, Pipper et al. (2012) propose a different strategy. Their approach combines multiple marginal models (MMM) i.e., one model per time point, from which a joint correlation matrix is determined. Adjusted p-values and SCIs are computed by incorporating the correlation between the respective score contributions of the time points from the different models. A big advantage of this method is that one does not have to bother about how to model the covariance structure. Moreover, it is easy to include covariates and generalize the method to endpoints of different types (e.g., discrete, time-to-event). Strong FWER control, however, is ensured only asymptotically; the small-sample behavior of this approach is largely unexplored to date.

The parameters of interest are the group effects $\boldsymbol{\beta}_j$; they are estimated as $\hat{\boldsymbol{\beta}}_j$ through m marginal models fitted separately for time points $j = 1, \ldots, m$, and their correlations are obtained via "stacking" the score contributions (derivatives of the log-likelihood) of these parameter estimates.

We can see that asymptotically

$$
\begin{aligned}
(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)\sqrt{n} &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} -\mathcal{I}_j^{-1} \tilde{\Psi}_{ij} + o_P(1) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Psi_{ij} + o_P(1)
\end{aligned}
$$

where $\mathcal{I}_j^{-1}$ 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 (e.g., van der Vaart 1998, theorem 5.21). Now the idea of Pipper et al. (2012) is to "stack" the $\boldsymbol{\beta}_j$, $\hat{\boldsymbol{\beta}}_j$, and Ψ_{ij} over all $j = 1, \ldots, m$ 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:

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

We can estimate $\boldsymbol{\Sigma}$ consistently as

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

where the $\widehat{\boldsymbol{\Psi}}_i$ are obtained by plugging the parameter estimates from the m marginal models into $\boldsymbol{\Psi}_i$.

This approach is sufficiently general to allow for various classes of marginal models (e.g., logistic regression, Cox proportional hazards model, ...), even in the same analysis. Here we consider only simple linear regression and define design vectors

$$
\mathbf{x}_i^{(j)} = (1, x_{i1}^{(j)}, \ldots, x_{iq}^{(j)})^T
$$

for treatment groups $k = 1, \ldots, q$. Now if the effect we are interested in is in the first coordinate of $\boldsymbol{\gamma}^{(j)}$ in a linear marginal model

$$
\mathbf{y}_i^{(j)} = \mathbf{x}_i^{(j)T} \boldsymbol{\gamma}^{(j)} + \boldsymbol{\epsilon}_i^{(j)}, \quad \boldsymbol{\epsilon}_i^{(j)} \sim \mathcal{N}(0, \sigma^2)
$$

for the jth time point, then Ψ_{ij} is the first coordinate of

$$
-E\left(\frac{1}{\sigma^2} \mathbf{x}_i^{(j)} \mathbf{x}_i^{(j)T}\right)^{-1} \mathbf{x}_i^{(j)} \left(\mathbf{y}_i^{(j)} - \mathbf{x}_i^{(j)T} \boldsymbol{\gamma}^{(j)}\right)
$$

which we can use to get the estimate $\widehat{\Psi}_{ij}$ and hence $\widehat{\boldsymbol{\Psi}}_i$ and also $\widehat{\boldsymbol{\Sigma}}$, which provides us with the prerequisites to do simultaneous inference for the $\hat{\boldsymbol{\beta}}_j$ (see 3.7).

3.5.2 Parameterization

An important observation with finite samples is that the parameterization of the marginal models can have an impact on the estimated covariances and thus on the inference. Consider an occasion-specific linear model for individuals i in treatment groups k:

$$y_i = \sum_{k=1}^{q} \beta_k x_{ik} + \epsilon_i \ , \ i = 1, \ldots, n.$$

Two straightforward ways to parameterize the model are:

1. β_1 is the mean of a reference group and β_2 through β_k are the other group means' differences to β_1 (this is R's standard parameterization for factors, also known as "treatment contrasts");

2. β_k is the kth group mean.

Either parameterization will yield correct effect estimates, but the associated covariance estimates may differ. The crux of the matter is that the method of Pipper et al. (2012) uses the *observed* Fisher information $\mathcal{J}$ from which it calculates the *empirical* covariances. $\mathcal{J}$ is asymptotically equal to $\mathcal{I} = E(\mathcal{J})$, the *expected* Fisher information. With small to moderate sample sizes, however, $\mathcal{I}$ and $\mathcal{J}$ may take quite dissimilar values, and the empirical covariances may grossly differ from the expected ones.

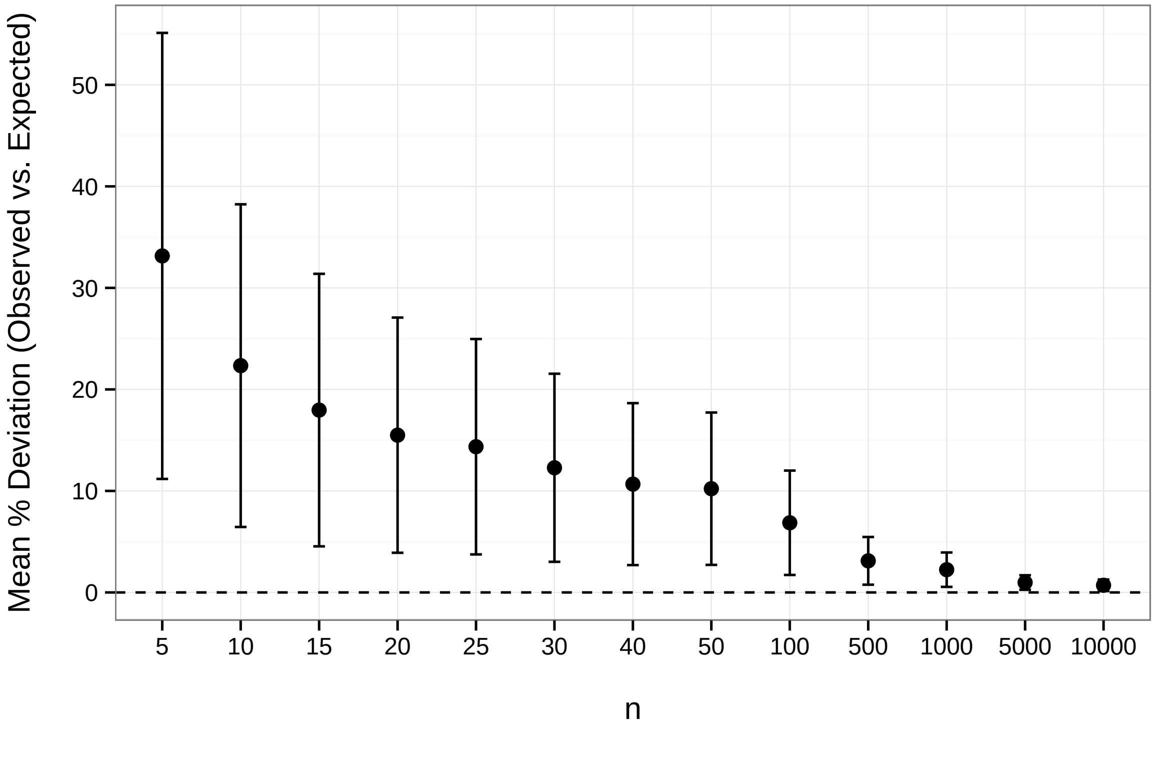


Figure 7: Mean percentage deviation (± standard deviation) of the observed and expected variance estimates for the comparison $\beta_2 - \beta_3$ in a setup involving three standard normal variates, each of them with sample size n.

We can illustrate this issue in a simple setup with three samples of n randomly drawn standard normal variates each, using the implementation of the method in the R package multcomp (Hothorn et al. 2015). Fitting a one-way ANOVA model to these data and estimating the variance of $\beta_2 - \beta_3$ from the model, we get the truly expected variance only with the second parameterization. Figure 7 shows the mean percentage deviation (plus/minus standard deviation) of the observed from the expected variance of $\beta_2 - \beta_3$ as estimated from 1000 simulated datasets. We recognize that the impact of parameterization is tremendous with small sample sizes: for $n = 5$, the observed variance (using the first parameterization) deviates from the expected one by more than 30% on average. Even with n as large as 50 independent units per sample, their mean deviation is still 10%. Several hundreds, or even thousands, of sampling units are required to force the mean percentage deviation down to a few percent.

What is happening here is that the first parameterization has trouble estimating covariances of zero with small to moderate sample sizes; it will yield non-zero values where there should be zeroes. Using the second parameterization, we are off the hook: it uses strictly different portions of the data to estimate different parameters, so it is preferable for our purpose. The whole problem vanishes with increasing sample size because the observed and expected Fisher information, and thus variance, are asymptotically equal.

3.6 Simultaneous Inference

3.6.1 General Concepts

Assessing more than one statistical hypothesis at one time is called simultaneous inference. This is like asking multiple specific questions and expecting one specific answer per question, and more often than not also an overall answer to the set of questions as a whole. Directing several enquiries to the same set of data, however, usually results in a multiple comparison problem. Multiplicity can arise from several group comparisons being made, several endpoints being investigated, several subgroups being analyzed, etc.

The common goal with simultaneous inference is to control a joint rate of type I errors occurring over the whole set (or "family") of hypotheses. Hochberg and Tamhane (1987, p. 5) defined a family as "any collection of inferences for which it is meaningful to take into account some combined measure of errors". In this context we appreciate the notion of a "claimwise" error rate as proposed by Phillips et al. (2013); this emphasizes that a claim can consist of diverse elementary hypotheses that are meaningful together.

Performing a series of level α tests for the elementary hypotheses without any further adjustment usually inflates the type I error rate of the entire claim. We want to focus here on methods that control the FWER in the strong sense i.e., the probability of incorrectly rejecting one or more true elementary nulls is to be bounded by α, no matter which and how many elementary nulls are true or false. The inferential procedures discussed in this work control the FWER (at least approximately) in the strong sense for a claim that may comprise comparisons among treatment groups as well as comparisons among time points.

When comparing multiple treatments separately and simultaneously at multiple occasions, it is sufficient for claiming an effect if at least one treatment difference is significant at least at one occasion. Likewise, when comparing multiple occasions separately and simultaneously for multiple treatment groups, an effect may be claimed if at least one occasion difference is significant for at least one treatment. Hence the claims we want to make are formulated as union-intersection tests (UITs), and adjustment is needed as the goal is to bound the FWER (approximately) by α. A UIT involves testing the intersection of elementary null hypotheses against the union of alternatives:

$$H_0 = \bigcap H_0^{(i)}$$

$$H_A = \bigcup H_A^{(i)}$$

The global H_0 is rejected if *at least one* elementary $H_0^{(i)}$ is rejected (Roy 1953).

The results of many simultaneous inference procedures can be expressed either as adjusted p-values or SCIs, but intervals are superior to p-values in multiple ways (Gardner and Altman 1986): both convey information about statistical significance, but SCIs in addition allow to assess the magnitude of an effect on the original scale, its direction (decrease or increase), and its subject-matter relevance. Therefore SCIs are much more useful for direct interpretation with respect to the research questions of interest.

3.6.2 Multiple Comparison Procedures

The simplest and best-known adjustment for multiplicity is based on Bonferroni's inequality (Bonferroni 1935, 1936)[2], leading to the corrected type I error bound

$$\tilde{\alpha} = \frac{\alpha}{z}$$

and to adjusted p-values (which are then to be compared with α) for hypotheses $h = 1, \ldots, z$ given by

$$\tilde{p}_h = \min(z p_h, 1).$$

This method is universally applicable to ensure strong FWER control but also notorious for being conservative unless test statistics are independent.

Just minimally more powerful is the correction of Šidák (1967) with its adjusted α bound of

$$\tilde{\alpha} = 1 - (1 - \alpha)^{\frac{1}{z}}.$$

Unlike the Bonferroni method, it controls α exactly (in a probabilistic sense), however only under independence of test statistics. In the presence of correlation among tests, its achieved type I error level can lie considerably below the nominal α.

This conservativity can be cushioned by incorporating dependence of test statistics by means of their *joint* parametric distribution. Tukey's pairwise comparisons using the studentized range distribution, Dunnett's comparisons with a control, or the analysis of means (ANOM) are examples how α can be better exploited under dependence. These are all single-step tests procedures (meaning that the same critical value applies to all test statistics), and all of them can be viewed as special cases of multiple contrast tests (MCTs), which we will consider in depth in 3.7.

Stepwise test procedures offer another strategy to lessen conservativity as they uniformly improve the power of corresponding single-step tests. The Bonferroni-Holm step-down

[2] In fact, Boole's inequality is applied here.

test (Holm 1979) is uniformly more powerful than Bonferroni; similarly, the single-step many-to-one test of Dunnett (1955) can be made uniformly more powerful in step-down (Naik 1975; Marcus et al. 1976; Dunnett and Tamhane 1991) or step-up (Dunnett and Tamhane 1992, 1995) variants[3].

The trouble with stepwise techniques is that compatible SCIs are, if available at all, cumbersome to derive and in most cases noninformative. Guilbaud (2008) and Strassburger and Bretz (2008) established SCI bounds corresponding to Holm-type step-down ("sequentially rejective") tests; these bounds stick to the margin δ_h (usually $\delta_h = 0 \; \forall \; h$) for all rejected hypotheses, and hence provide no additional information compared to the p-values unless *all* hypotheses are rejected. Compatible SCIs for step-down Dunnett tests, following suggestions by Bofinger (1987) and Stefansson et al. (1988), have the same unpleasant property. SCIs compatible with step-up Dunnett tests do not exist to date. We consider the unavailability of assuredly informative SCIs a major deficiency that outweighs the achievable gain in power, hence stepwise procedures will not play a role in the remainder of this work.

Elaborate treatise of simultaneous inference is provided in the textbooks by Hochberg and Tamhane (1987) and Hsu (1996) as well as a series of recent review articles (Dmitrienko and D'Agostino, Sr. 2013; Alosh et al. 2013; Dmitrienko et al. 2013; Huque et al. 2013). The books by Dmitrienko et al. (2010) and Dickhaus (2014) connect mathematical theory and biomedical applications. Software-specific overviews are delivered e.g., in Bretz et al. (2010, using `R`) or Westfall et al. (2011, using SAS).

3.7 Multiple Contrast Tests

3.7.1 General Methodology

Multiple contrast tests (MCTs) are a convenient and supremely flexible method for testing multiple hypotheses and obtaining compatible SCIs. They control the familywise α in the strong sense and take dependencies among tests into account. Basic MCT methodology was outlined in Mukerjee et al. (1987) and Bretz et al. (2001a) and extended to a wide class of (semi-)parametric models by Hothorn et al. (2008).

Many well-known and widely applied multiple comparison procedures can be formulated in terms of MCTs e.g., comparisons to a control (Dunnett 1955), all-pairwise comparisons (Tukey 1953), comparisons versus the grand mean ("analysis of means", see Pallmann and Hothorn 2016), changepoint alternatives (Hirotsu et al. 2011), and Williams' trend test (Williams 1971, 1972; Bretz 2006). Beyond that, arbitrary sets of contrasts can be tailored to a specific research problem.

MCTs provide both global and localized significance decisions, and the latter can be based upon either multiplicity-adjusted p-values or SCIs. So an MCT gives answers to three question:

- "Is there an effect?" $\longrightarrow$ global test,
- "Where is the effect?" $\longrightarrow$ local adjusted p-values,

[3] The step-down Dunnett test can be viewed as an extension of Holm (1979), in contrast to the step-up Dunnett test being an extension of Hochberg (1988).

- "Where is the effect, and how big is it?" ⟶ SCIs.

To begin with, we consider a one-way layout with randomized treatment groups $k = 1, \ldots, q$ and independent observations $i = 1, \ldots, n_k$ within each group, and we assume homoscedasticity across all treatments. We review testing a single contrast first before extending our considerations to multiple contrasts and in a second step.

A Single Contrast: The basic module of an MCT is a single contrast

$$\eta = \mathbf{c}^T\boldsymbol{\beta} = c_1\beta_1 + \cdots + c_q\beta_q$$

i.e., a linear combination of parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_q)^T$ defined by coefficients $\mathbf{c} = (c_1, \ldots, c_q)^T$ where more often than not $\sum_{k=1}^{q} c_k = 0$. Replacing $\boldsymbol{\beta}$ with an estimate $\hat{\boldsymbol{\beta}}$ gives the estimated contrast $\hat{\eta}$, and the standardized version of this is

$$Z = \frac{\hat{\eta} - \delta}{\mathrm{SE}(\eta)}$$

where δ is the expected value of η under H_0 (often chosen to be 0), and

$$\mathrm{SE}(\eta) = \sigma\sqrt{\sum_{k=1}^{q} \frac{c_k^2}{n_k}}$$

with σ being the square root of the common variance. Z constitutes a contrast test statistic about the null hypothesis

$$H_0: \ \eta = \delta.$$

$\mathrm{SE}(\eta)$ is a known quantity if σ is known, and then Z is standard normal under H_0:

$$Z \sim \mathcal{N}(0, 1).$$

In practice, σ is usually unknown and must be estimated from the data; replacing σ with its estimate s in $\mathrm{SE}(\eta)$ yields the test statistic

$$T = \frac{\eta - \delta}{\widehat{\mathrm{SE}}(\eta)},$$

which is t-distributed under H_0 with $\nu = n - q$ degrees of freedom:

$$T \sim \mathcal{T}(\nu).$$

Multiple Contrasts: If a set of several standardized contrasts $\mathbf{T} = (T_1, \ldots, T_z)$ is to be assessed simultaneously, the global H_0 is composed of elementary nulls

$$H_0^{(h)}: \ \eta_h = \delta_h,$$

and the joint null is the intersection

$$H_0 = \bigcap H_0^{(h)}.$$

In statistical practice, several contrasts are assembled in a family (or "claim") such that their joint investigation gives a detailed and meaningful answer to the research questions. This was described by Mukerjee et al. (1987) as vectors being "strategically located within the alternative region". The comparisons of interest are specified via a coefficient matrix

$$\mathbf{C} = (\mathbf{c}_1, \dots, \mathbf{c}_z) = (c_{hk})$$

that is composed of multiple coefficient vectors. The quantities of interest are the contrasts

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

and more often than not these contrasts are interdependent i.e., non-orthogonal, and then correlations can be exploited to sharpen the test procedure. Some of the most commonly used contrasts with their coefficient matrices (examplified for comparisons of four parameters) are:

- many-to-one comparisons to a common control:

$$\mathbf{C}^{Dunnett} = \begin{pmatrix} -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \end{pmatrix},$$

- all-pairwise comparisons:

$$\mathbf{C}^{Tukey} = \begin{pmatrix} -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \\ 0 & -1 & 1 & 0 \\ 0 & -1 & 0 & 1 \\ 0 & 0 & -1 & 1 \end{pmatrix},$$

- comparisons to the grand mean (for a balanced setup):

$$\mathbf{C}^{GrandMean} = \begin{pmatrix} \frac{3}{4} & -\frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{4} & \frac{3}{4} & -\frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{4} & -\frac{1}{4} & \frac{3}{4} & -\frac{1}{4} \\ -\frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} & \frac{3}{4} \end{pmatrix},$$

- Williams trend test (for a balanced setup):

$$\mathbf{C}^{Williams} = \begin{pmatrix} -1 & 0 & 0 & 1 \\ -1 & 0 & \frac{1}{2} & \frac{1}{2} \\ -1 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix}.$$

The correlations of test statistics depend on these coefficient matrices.

The test statistic corresponding to the hth contrast is

$$T_h = \frac{\hat{\eta}_h - \delta_h}{\widehat{\mathrm{SE}}(\eta_h)}$$

where

$$\hat{\eta}_h = \sum_{k=1}^{q} c_{hk}\hat{\beta}_k,$$

and the SE is now estimated as

$$\mathrm{SE}(\eta_h) = s\sqrt{\sum_{k=1}^{q} \frac{c_{hk}^2}{n_k}}.$$

The joint distribution of $\mathbf{T} = (T_1, \dots, T_z)$ is z-dimensional t (Cornish 1954; Dunn and Massey, Jr. 1965) with ν DF, and $\mathbf{R}$ describing the correlation among the test statistics:

$$\mathbf{T} \sim \mathcal{T}_z(\nu, \mathbf{R}).$$

With homoscedastic and independent errors, the DF is quite simply

$$\nu = \sum_{k=1}^{q}(n_k - 1),$$

and the elements of $\mathbf{R}$ are uniquely defined by the sample sizes and contrast coefficients i.e., by known quantities:

$$Corr(T_h, T_{h'}) = \rho_{hh'} = \frac{\sum_{k=1}^{q} \frac{c_{hk}c_{h'k}}{n_k}}{\sqrt{\left(\sum_{k=1}^{q} \frac{c_{hk}^2}{n_k}\right)\left(\sum_{k=1}^{q} \frac{c_{h'k}^2}{n_k}\right)}}.$$

For instance, the many-to-one test as defined via $\mathbf{C}^{Dunnett}$ compares all groups to the control, so this piece of information is shared by all comparisons, and it can easily be seen that the pairwise correlation of contrasts is 0.5 for balanced sample sizes.

For the two-sided case, we reject $H_0^{(h)}$ if

$$|T_h| > t_{z,1-\alpha}^{two}(\nu, \hat{\mathbf{R}})$$

where $t_{z,1-\alpha}^{two}(\nu, \hat{\mathbf{R}})$ is the two-sided equicoordinate $1-\alpha$ quantile of $\mathcal{T}_z(\nu, \hat{\mathbf{R}})$. The corresponding adjusted p-value is calculated as

$$p_h = 1 - \int_{-|T_h^{obs}|}^{|T_h^{obs}|} \dots \int_{-|T_h^{obs}|}^{|T_h^{obs}|} t_z(\mathbf{x}; \nu, \mathbf{R})\ d\mathbf{x}$$

where T_h^{obs} is an observed value of T_h, and $t_z(\mathbf{x}; \nu, \mathbf{R})$ is the PDF that corresponds to $\mathcal{T}_z(\nu, \mathbf{R})$. One elementary hypothesis is declared significant if its corresponding adjusted p-value is below the pre-defined familywise significance level, which is traditionally 0.05. The global null is declared significant if the minimum p-value is below the significance level (maximum test).

SCIs are often preferred to p-values as they give a measure for the uncertainty of the estimate on the scale of the observations rather than on an abstract probability scale. SCI bounds with coverage probability $1-\alpha$ are given by

$$\hat{\eta}_h \mp t_{z,1-\alpha}^{two}(\nu, \mathbf{R})\ \widehat{\mathrm{SE}}(\eta_h).$$

The hth elementary hypothesis is rejected if δ_h is not included in the corresponding SCI. The global null is rejected if at least one SCI does not include the corresponding δ_h.

These SCIs have two convenient properties:

1. They are compatible with test decisions from adjusted p-values: a $(1 - \alpha)$ SCI excluding the point of no effect is always associated with an adjusted p-value smaller than α; hence SCIs can be readily used for hypothesis testing.
2. Their bounds are always informative; this stands in contrast to other methods (e.g., stepwise procedures) whose usefulness is thereby limited.

Inference for directional one-sided problems can be carried out in a very similar way (e.g., Bretz et al. 2010). Efficient computation of probabilities and quantiles from a multidimensional t-distribution is described in Genz and Bretz (2009); see also Appendix A.2 for details on multivariate t-distributions.

Numerous extensions of MCTs have been proposed in the past few years. Hothorn et al. (2008) discussed multiple comparisons for a wide range of parametric models (ANOVA and regression models, GLMs, linear and nonlinear mixed-effects models, Cox and Weibull models). Hasler and Hothorn (2008) and Herberich et al. (2010) made suggestions for mastering heteroscedasticity. Dilba et al. (2006) and Hare and Spurrier (2007), among others, transfered the principle of MCTs to hypotheses involving ratios instead of differences of means.

In this thesis we study MCTs for settings when neither independence nor homoscedasticity can be assumed, and comparisons are to be made among randomized treatment groups and/or among repeated occasions. We will be dealing with modeling stategies for estimating $\boldsymbol{\beta}$ and the associated covariance matrix $\boldsymbol{\Sigma}$ and how to incorporate them in the MCT procedure. Before that, however, we review two similar MCT procedures that already exist.

3.7.2 MCTs with Correlated Outcomes

In recent years, MCT variants were developed for two inferential problems related to ours that both involve correlated outcomes. We review these multivariate MCTs and repeated measures MCTs briefly and expound why they are of limited avail for our specific matter.

MCTs for Multiple Endpoints: Hasler and Hothorn (2011) developed multivariate MCTs and showed the benefit of acknowledging the correlation between multiple endpoints. Their method could be applied to our longitudinal setting directly by treating measurement occasions as endpoints. However, this is nonsatisfying from different points of view:

1. The procedure only allows for comparisons among treatments per time point but not among time points per treatment.
2. It is not straightforwardly adaptable to data with additional covariates.
3. Covariances of endpoints are always assumed as heteroscedastic (variances) and unstructured (correlation), either for all treatment groups or even separately for each treatment (Hasler 2014a); in a scenario involving a mere four groups and five time points, say, there are 60 covariance parameters to estimate! It is obvious that this is undesirable, and may indeed be unnecessary, especially with small sample

sizes. We will aim at sparsity in parameters, which can be achieved with patterned covariance models or by introducing random effects.

4. It is fairly unclear how to deal with incomplete measurements, which disqualifies the method from being applied to longitudinal data where missing values and dropout are a frequent occurrence. Admittedly, there are workarounds, but they are either assumption-laden (such as estimating the means from all available data but the covariances only from the complete cases) or technically complicated (e.g., multiple imputation techniques).

MCTs for Repeated Measures: Hasler (2013) discussed MCTs for comparing means between the levels of a repeated factor in a single-treatment setup. Being confronted with the challenge of finding a workable DF approximation, he fabricated as a purely empirical solution

$$\nu_{Hasler} = n - 1 - \frac{z-1}{m-1}$$

where n is the number of independent units, z the number of comparisons, and m the number of time points, and approved its small-sample performance under H_0 via simulations. This procedure is evidently unsuited for our purposes:

1. The method addresses comparisons among time points but no comparisons of treatments.
2. Most real-world problems involve more than one treatment group, so the limitation to a single group is unacceptable.
3. Beyond that, it raises similar difficulties as the multi-endpoint MCTs: most importantly, it is inflexible as regards coping with additional covariates, and missingness cannot be adequately acknowledged without making restrictive assumptions.

In conclusion, the existing MCTs for multiple endpoints and repeated measures are inflexible in several respects. The goal of this dissertation is to widen their scope of application to practically relevant scenarios under less restrictive premises. Direct extensions of these two variants of MCTs, however, are unpromising as they will retain existing limitations; thus our strategy shall be based on more flexible modeling of the data.

3.8 Approximations to the Degrees of Freedom

As the multivariate t is only an *approximation* to the joint distribution of test statistics, a proper approximation for its degrees of freedom (DF) is especially relevant with small sample sizes. However, there is no hope to find any "correct" DF for arbitrary settings, and it is rather an open question which of the many suggestions circulating in the literature works best under given circumstances.

Any useful DF approximation should reflect the amount of "information" contained in the data (Faes et al. 2009). In the context of longitudinal data this could mean the DF needs to capture the number of independent subjects, the number of total repeated measurements, and the strength of within-subject correlation.

We collect here a couple of possibly useful approaches for small sample sizes, some of which are well-known and some rather innovative, and assess their applicability to multiple comparisons in longitudinal designs. As n grows large, all approximations tend to infinity, and the asymptotic reference distribution becomes multivariate normal.

Naive: The naive lower boundary post for any useful DF approximation is

$$\begin{aligned}\nu_{naive} &= n - \mathrm{rank}(\mathbf{X}) \\ &= n - mq\end{aligned}$$

where n is the number of independent subjects, m the number of occasions, and q the number of treatment groups. It basically ignores that there are repeated measurements and only takes independent replications into account; hence it most likely understates the amount of information in the data (unless all observations are perfectly correlated). Using ν_{naive} in the presence of small and moderate sample sizes should lead to conservative tests.

Residual: Another simplistic approximation is the DF of the residual error variance of each fixed factor

$$\begin{aligned}\nu_{res} &= N - \mathrm{rank}(\mathbf{X}) \\ &= N - mq\end{aligned}$$

where N is the total number of measurements. It acts as if the repeated measurements were independent observations and will therefore lead to overoptimistic results in the presence of longitudinal correlation, at least with finite samples. So ν_{res} may be viewed as an upper boundary post, and it is fairly obvious that any sensible DF approximation should lie in the interval $]\nu_{naive}, \nu_{res}[$.

Between-within: An attempt to achieve a tradeoff between the naive and residual DFs is known (especially to SAS users) as between-within method (Schluchter and Elashoff 1990) and calculated as

$$\nu_{bw} = N - \mathrm{rank}(\mathbf{X}) - \mathrm{rank}(\mathbf{Z}).$$

It has been shown to yield unsatisfactory results in simulation studies (see the end of this section) and will therefore not be pursued any further.

Containment: The approximation popularly referred to as the SAS containment DF computes

$$\nu_{cont} = N - \mathrm{rank}(\mathbf{X}\ \mathbf{Z})$$

which is the rank contribution of random terms that *syntactically* contain the fixed effect under consideration to the covariate matrix ($\mathbf{X}\ \mathbf{Z}$) (e.g., Verbeke and Molenberghs 1997; SAS Institute 2009). The implications of the term “syntactically” in this context need to be clarified: assume a mixed model with fixed effect `A`, random effect `B`, and interaction `A*B` which contains `A` syntactically. However, `A*B` could be denoted by `C` as well, which does not alter the model but *syntactically* (and only syntactically!) removes the interaction term involving `A` and thus change the value of ν_{cont}.

So a big drawback of the containment method is its ambiguity: one and the same statistical model can be written in different ways, possibly leading to different numerical results for ν_{cont}. The fact that computation of ν_{cont} is intrinsically tied to SAS syntax makes it hard to implement the method in other software packages such as `R`.

Pinheiro-Bates: The suggestion described by Pinheiro and Bates (2000, p. 91) and used in their `nlme` software (Pinheiro et al. 2015) is similar in spirit to the containment DF but untainted by peculiarities of the SAS syntax. In its general form the approximation is computed as

$$\nu_{PB} = m_i - (m_{i-1} + p_i)$$

where m_i denotes the number of units at level $i = 1, \ldots, Q+1$, and p_i the DF sum associated with the terms estimated at the ith level. This gives $m_0 = 1$ in the presence of a fixed-effects intercept and $m_0 = 0$ otherwise, and obviously $m_{Q+1} = N$.

For the case of a cell-means model without an intercept (which is a pseudo one-way model) the formula simplifies to

$$\nu_{PB_1} = N - n - mq + 1.$$

This is the DF for the interaction effect of treatments and time points which, however, turns out to be too large for our longitudinal MCTs. On the other hand, the simplification

$$\nu_{PB_2} = n - m$$

gives the DF for the treatment effect, and this is the DF we will use in the longitudinal MCTs in Chapter 4.

Satterthwaite: The approximation suggested by Satterthwaite (1941 1946) is found by matching first and second moments and is probably best known for appearing in Welch's t-test (Welch 1938, 1947), but has also been applied to fixed-effect tests in LMMs (Giesbrecht and Burns 1985; Fai and Cornelius 1996). It does not only depend on the experimental design (through the cell sample sizes n_b) but also on the data itself through the (estimated) cell variances $\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_s^2)$:

$$\nu_{Satt} = \frac{\left(\sum_{b=1}^{s} \frac{\sigma_b^2}{n_b}\right)^2}{\sum_{b=1}^{s} \frac{\left(\sigma_b^2/n_b\right)^2}{n_b - 1}}.$$

For homoscedastic data (i.e., constant σ_b^2), ν_{Satt} is equal to ν_{res}; and with increasingly heterogeneous variances over time (which are the same across treatments), ν_{Satt} tends to $\frac{\nu_{res}}{m}$ (Figure 8). If heterogeneity patterns were not restricted to be equal across treatment groups, ν_{Satt} would tend to $\frac{\nu_{res}}{mq}$ for increasingly non-constant variances. The Satterthwaite approximation is the basis for the Kenward-Roger method, which we will investigate in more detail.

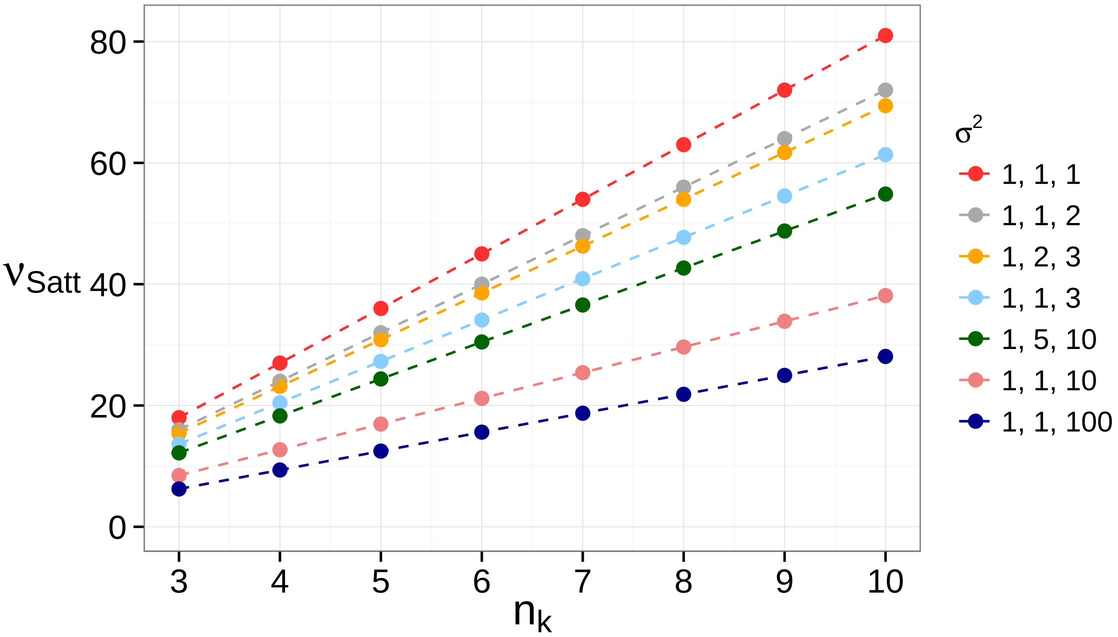


Figure 8: Satterthwaite approximation to the degrees of freedom in a setup with $q = 3$ treatments, $m = 3$ occasions, and $3 \leq n_k \leq 10$ subjects per treatment group under different configurations of the variances.

Kenward-Roger: Kenward and Roger (1997) proposed a small-sample approximation for testing (contrasts of) fixed effects from REML-based LMMs. They observed that fixed-effect tests motivated by asymptotic theory may perform poorly when sample sizes are small, and attributed this problem to two sources of bias:

1. The fact that the covariance matrix for the fixed effects is estimated from data and not known introduces extra variability, which is not properly acknowledged for, as shown by Kackar and Harville (1984).
2. The asymptotic estimator of the fixed-effect covariance matrix underestimates in the presence of small samples, as discussed by Harville and Jeske (1992).

Kenward and Roger explained how both issues—which are nicely reviewed in a broader context by Littell (2002)—can be straightened out using Taylor series expansions. Their method entails more than an adjustment to the DF; it is a corrective that comes in three parts:

1. a corrected (i.e., inflated) estimator of the fixed-effect covariance matrix that removes both sources of bias mentioned above, which is then used to build a Wald-type statistic F;
2. a scaling factor λ that is estimated from the data and multiplied by the Wald statistic F so that $F^* = \lambda F$ is approximately F-distributed with numerator DF z (the number of contrasts) and denominator DF ν;
3. an approximation to ν obtained by matching first and second moments of the Wald statistic F^* and its approximative distribution $\mathcal{F}_{z,\nu}$; this is essentially a Satterthwaite-type DF calculation on the corrected covariance estimate.

In addition, Kenward and Roger tweaked their solution a little to ensure that it gets λ and ν right (meaning "exact") for two special cases with true F distributions, namely the balanced one-way ANOVA and Hotelling's T^2 test.

Extended derivations and results for the Kenward-Roger method are given by Alnosaier (2007); among other things, he lists a number of conditions under which the DF are identical to the Satterthwaite method because the corrected covariance matrix is equal to the uncorrected one e.g., for all fixed-effects linear models but also many balanced LMMs.

Kenward and Roger (2009) themselves provide an improved bias correction for the covariance estimates under nonlinear covariance structures. Skene and Kenward (2010a) deal with inference for fixed effects with repeated measurements in mixed models in the presence of tiny sample sizes: they propose replacing the fixed-effect covariance matrix with a bias-adjusted sandwich estimator. However, this approach has seriously low power, which they attempt to tackle with a modified Box correction (Skene and Kenward 2010b).

Effective sample size: Taking the number of independent units as sample size (as in ν_{naive}) is likely to underestimate the information provided by the data whereas taking the total number of measurements (as in ν_{res}) probably overestimates it. The extent of under- or overstating depends on the correlation among the repeated measures. To obtain a well-balanced compromise for the amount of information available, a concept worth considering is effective sample size (ESS), which is the size of a (hypothetical) uncorrelated sample containing as much information as the correlated data at hand. Faes et al. (2009) propose it as an instrument to approximate the DF with small samples. In the general framework of an LMM, the ESS associated with the pth model parameter β_p is

$$\tilde{n}(\beta_p) = \frac{\widehat{\mathrm{Var}}(\hat{\beta}_p)}{\widetilde{\mathrm{Var}}(\hat{\beta}_p)} \sum_{i=1}^{N} n_i$$

where $\widehat{\mathrm{Var}}(\hat{\beta}_p)$ is the pth element of the variances of $\hat{\boldsymbol{\beta}}$

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}) = \left(\sum_{i=1}^{N} \mathbf{X}_i^T \boldsymbol{\Sigma}_i^{-1} \mathbf{X}_i \right)^{-1}$$

and $\widetilde{\mathrm{Var}}(\hat{\beta}_p)$ the pth element of

$$\widetilde{\mathrm{Var}}(\hat{\boldsymbol{\beta}}) = \left(\sum_{i=1}^{N} \mathbf{X}_i^T \mathbf{W}_i^{-1} \mathbf{X}_i \right)^{-1}$$

with $\mathbf{W}_i = \mathrm{diag}(\boldsymbol{\Sigma}_i)$.

They showed that calculation of the ESS simplifies e.g., when acting on the assumption of an AR(1) structure, which is reasonable with longitudinal data. Then we get

$$\tilde{n} = \frac{n[m - (m-2)\rho]}{1 + \rho}$$

where ρ is the AR(1) correlation coefficient; in practice ρ is estimated from the data and plugged in. We can construct the ESS-based DF by replacing the total number of experimental units n (as in ν_{naive}), or the total number of measurements N (as in ν_{res}), by the effective sample size $\tilde{n}$:

$$\nu_{ESS} = \tilde{n} - mq.$$

Notice that $\nu_{ESS} = \nu_{naive}$ with $\rho = 1$, and $\nu_{ESS} = \nu_{res}$ with $\rho = 0$ (Figure 9); these special cases are, of course, unlikely to arise with real-world data.

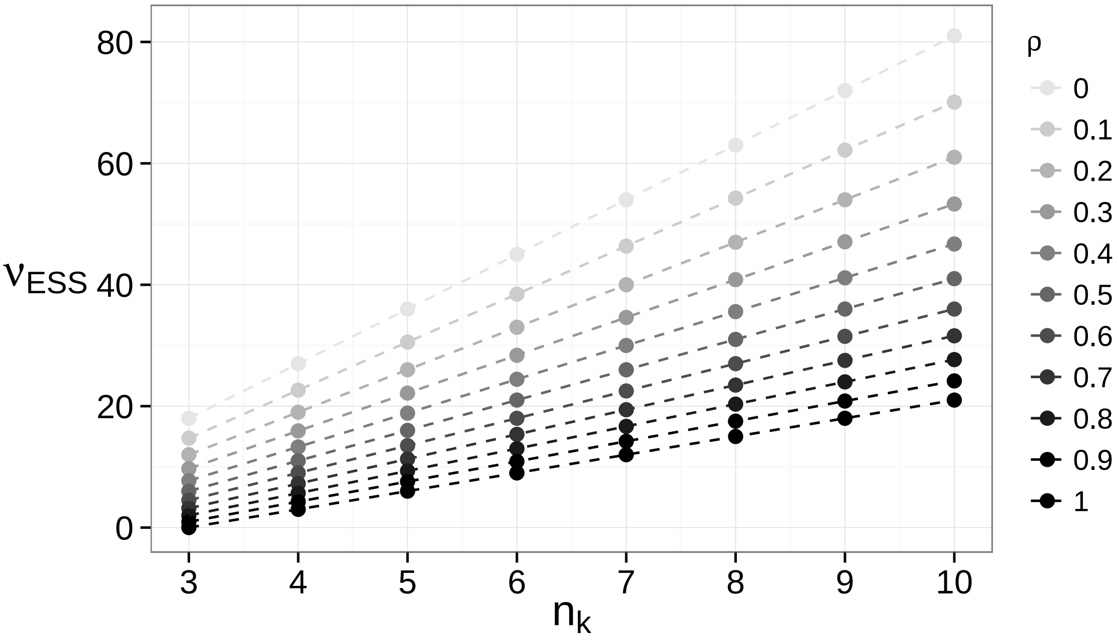


Figure 9: ESS approximation to the degrees of freedom in a setup with $q = 3$ treatments, $m = 3$ occasions, and $3 \leq n_k \leq 10$ subjects per treatment group under different values of the AR(1) correlation parameter ρ.

Adjusted: We have designed an approximation that "adjusts" ν_{naive} by adding a correction term that raises the DF with small sample sizes but vanishes as n grows large:

$$\nu_{adj} = n - mq + \frac{m^2q^2}{n}.$$

This behavior can be seen from Figure 10: ν_{adj} ascends less steeply than the other DF approximations with increasing n_k. Note that the "adjustment" leading to ν_{adj} does not have any profound theoretical background but rather arose empirically.

Whenever an approximation does not yield a whole-number DF (which can happen with Satterthwaite/Kenward-Roger, "adjusted", and ESS), we round it down to the nearest integer.

DFs for MMM: The residual DF associated with the jth marginal linear model is

$$\nu_{res}^{(j)} = n^{(j)} - q^{(j)}$$

where $n^{(j)}$ and $q^{(j)}$ are the numbers of independent units and treatment groups, respectively, that are present at occasion j. A practicable approximation to the DF for a set of

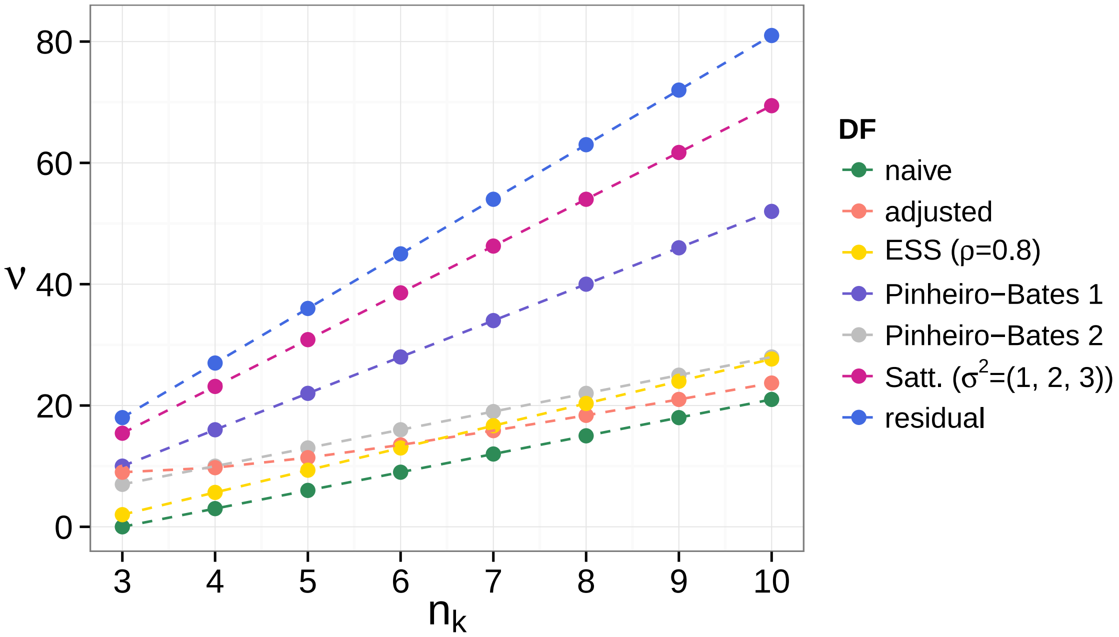


Figure 10: Comparison of different approximations to the degrees of freedom in a setup with $q = 3$ treatments, $m = 3$ occasions, and $3 \leq n_k \leq 10$ subjects per treatment group.

occasion-wise comparisons of treatments is the minimum of the marginal models' residual DFs:

$$\nu_{min} = \min_j \nu_{res}^{(j)}.$$

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

$$\bar{\nu} = \frac{1}{m} \sum_{j=1}^{m} \nu_{res}^{(j)}.$$

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

$$\nu_{res}^{(k)} = n_k - 1,$$

and the DF for the set of treatment-wise comparisons of occasions is again approximated as the minimum (or average) of the $\nu_{res}^{(k)}$.

Now we have assembled a variety of candidate DFs, and the question is which one(s) to use for our longitudinal MCTs. A number of simulation studies have been conducted to compare small-sample DFs under a multitude of conditions such as different covariance structures, imbalance of the data, and varying sample sizes. The optimal choice is influenced by the complexity of the covariance structure and the degree of imbalance, with small sample sizes making the impact of a poor choice on the type I error rate more severe; the results are usually very similar for moderate to large sample sizes.

Several studies showed the Kenward-Roger method to perform superior, or at least comparable, to other small-sample solutions under various data conditions e.g., by Guerin and Stroup (2000), Schaalje et al. (2002), Spilke et al. (2004, 2005), Alnosaier (2007), and Arnau et al. (2009). Nonetheless, inflation of the type I error rate has been reported for Kenward-Roger as well, particularly with very small samples and imbalance e.g., by Schaalje et al. (2002), Valderas Gomez et al. (2005), and Gregory (2011). We will evaluate the usefulness of various DF approximations for longitudinal MCTs with small sample sizes numerically in Chapter 4.

3.9 Power

The power $1-\beta$ of a hypothesis test is the probability of correctly rejecting H_0. This definition is straightforward for single tests; for a multiple comparison procedure (MCP), however, it is far from obvious what "the power" is. Indeed there exist various notions of power in the context of multiple testing; see e.g., Hochberg and Tamhane (1987, p. 129) or Horn and Vollandt (1998). We introduce here four widely adopted MCP power definitions in a nutshell.

We consider two-sided MCTs composed of elementary null hypotheses $H_0^h : \eta_h = \delta_h,\ h = 1,\dots,z$ stating that some linear combination of means, η_h, is equal to a pre-defined margin, δ_h. The global null hypothesis $H_0 = \bigcap_{h=1}^{b} H_0^h$ is the intersection of elementary nulls. We let $\mathcal{A} = \{h : H_A^h : \eta_h \neq \delta_h\}$ denote the subset of contrasts that are truly under the alternative. Further we let T_h denote the test statistic corresponding to the hth hypothesis and t an appropriately chosen critical point from a reference distribution such as multivariate t.

Per-pair power: $P(|T_h| > t \mid h \in \mathcal{A})$

The probability of rejecting a H_0^h for which the corresponding contrast η_h is truly under the alternative. This is a special case of the per-subset power (Einot and Gabriel 1975), and larger subsets than pairs are easily conceivable (triplets, quadruplets, ...).

Any-pair power: $P(\exists\ h \in \mathcal{A} : |T_h| > t)$

The probability of rejecting at least one H_0^h for which the corresponding contrast η_h is truly under the alternative (Ramsey 1978). Any-pair power is also called minimal power (e.g., Westfall et al. 2011) or disjunctive power (Senn and Bretz 2007).

All-pairs power: $P(|T_h| > t\ \forall\ h \in \mathcal{A})$

The probability of simultaneously rejecting all H_0^h for which the corresponding contrasts η_h are truly under the alternative (Ramsey 1978). Other names for all-pairs power are complete power (e.g., Westfall et al. 2011) and conjunctive power (Senn and Bretz 2007).

Global power: $P(\exists\ h : |T_h| > t)$

The probability of rejecting at least one H_0^h, no matter if the corresponding contrast η_h (or, in fact, any contrast) is truly under the alternative (e.g., Hayter and Liu 1992).

Power according to this definition is "contaminated" with type I errors from contrasts that are truly under the null.

For one-sided tests there is the additional problem of incorrect directional decisions, sometimes called type III errors (Harter 1957), but this is usually negligible unless the effects are very close to zero.

3.10 Missing Data

3.10.1 General Problem

Missing values are observations that were not recorded but should have been. Data points may be missing intermittently (e.g., due to failure of recording values because of technical errors) or there can be dropout towards the end (e.g., due to death or loss of follow-up in clinical studies). In longitudinal setups, incomplete measurements are the rule and not some curious exception. Thus any method that cannot handle missings in a proper way is practically very limited.

When pondering whether a method yields valid estimates and inferences in the presence of incomplete data, it is crucial to consider the underlying missingness process i.e., the reason why certain data points are unavailable. When ignoring this, a data analyst may easily run into trouble and obtain biased estimates, underestimate SEs, lose control over type I error rates and coverage probabilities, or impair the power of a test.

Key questions to be asked are: can the observed data be regarded a random subsample of the (hypothetical) complete data? And: how is a data point's probability to be missing related to observed and unobserved values? Unfortunately, the missingness process is usually unknown and cannot be made out from the data. Therefore it can be wise to avoid all too restrictive assumptions about the causes of missingness.

The standard classification of missingness mechanisms that dates back to Rubin (1976) is recapitulated subsequently, followed by a brief review of various techniques how to handle missing data in practice.

3.10.2 Missingness Mechanisms

We denote the (intended) complete set of measurements for the ith individual by

$$\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{im})^T$$

and assign response indicators

$$R_{ij} = \begin{cases} 1 & \text{if } Y_{ij} \text{ is observed,} \\ 0 & \text{if } Y_{ij} \text{ is missing} \end{cases}$$

that can be summarized in an indicator vector

$$\mathbf{R}_i = (R_{i1}, \ldots, R_{im})^T,$$

and divide Y_i into observed and missing values:

$$\mathbf{Y}_i = (Y_i^O, Y_i^M)^T.$$

To draw a distinction between three general mechanisms of missingness, we adopt the nomenclature in Little and Rubin (2002):

Missing completely at random (MCAR): The probability to be missing depends neither on the observed values nor on the unobserved ones:

$$P(R_i|Y_i^O, Y_i^M, X_i) = P(R_i|X_i).$$

Missing at random (MAR): The probability to be missing is independent of the unobserved values, conditional on the observed ones:

$$P(R_i|Y_i^O, Y_i^M, X_i) = P(R_i|Y_i^O, X_i).$$

Not missing at random (NMAR): The probability to be missing cannot be assumed independent of the observed and unobserved values:

$$P(R_i|Y_i^O, Y_i^M, X_i).$$

One should not be confused by the somewhat deceptive terminology: MAR involves a good deal of what could be perceived as "non-randomness" in everyday speech. Similarly, MCAR and MAR are sometimes called "ignorable" missingness, in contrast to the term "non-ignorable" for NMAR. These names are clearly misleading in the sense that they suggest one need not bother about the missingness process. In fact, "ignorability" means that when adopting a direct likelihood approach (see 3.10.3), one does not have to model the missingness process explicitly.

3.10.3 Remedies

It is obvious that a valid missingness method should yield consistent estimates, and resulting tests should have correct size unter H_0. Various solutions have been proposed to deal with the problem of missing values, among them many *ad hoc* approaches that are at best inefficient and at worst seriously flawed. On the other hand, better solutions can be computer-intensive—or, as a matter of fact, super-simple. We will review a few of the most popular strategies for coping with incomplete data and reveal their strengths, limitations, and pitfalls.

Many missing data techniques involve some form of imputation, which can be described as sort of a "repair method" for datasets with missing values. A dataset is "completed" by filling the gaps in the original data with "plausible" values. Then the suchlike imputed data set is analyzed as though there had been no missings, or with appropriate modifications. The usefulness of the results patently depends on the definition of what is considered a "plausible" value, and whether additional uncertainty is accounted for.

Last value carried forward (LVCF): For every subject, missing entries are filled by inserting the last observed value before the gap. This is probably the most detrimental strategy around since it is prone to be biased even under MCAR, and the bias may go either upwards or downwards (Molenberghs et al. 2004). The condition necessary for LVCF to yield consistent estimates under MCAR (i.e., that the value remains constant after dropout) is hardly ever met in practice. In addition, the approach suffers from unduly low SE estimates. Contrary to common belief, LVCF does not necessarily lead to conservative inferences. Related approaches such as carrying the baseline forward or carrying the "worst" value forward are equally bad.

Single mean imputation: This method comes in two variants. The simpler one is to fill gaps in the dataset with the *unconditional* (arithmetic) mean of the observed values of a variable. In a refined version, imputation values are estimated *conditionally* as predicted means from regression on the other variables using only complete cases (Buck 1960). Both versions require MCAR, and SEs are likely to be underestimated.

Single (stochastic) imputation (SI): For each missing data point one imputation value is randomly drawn from

$$f(Y_i^M|Y_i^O, X_i)$$

i.e., from the conditional distribution of missing values given the observed values and covariates. The process of imputation, however, gives rise to an additional source of uncertainty which must be accounted for in the analysis. Deriving an analytical expression for the SE can become inutterably complicated.

Multiple (stochastic) imputation (MI): This difficulty of SI can be overcome by repeatedly imputing values and evaluating each of these "completed" datasets. Several imputation values are drawn from

$$f(Y_i^M|Y_i^O, X_i)$$

for each missing entry, thus generating a series of "completed" datasets, each of which is evaluated and then the results are combined in a thoughtful manner. This makes it straightforward to incorporate the uncertainty associated with imputation. MI techniques are most valuable when both response values and covariates are missing. A widely applied MI technique is multivariate imputation by chained equations (Rubin 1987; van Buuren and Groothuis-Oudshoorn 2011; van Buuren 2012). Imputation is performed v times, thereby creating v "filled-up" datasets, each of which is evaluated separately to obtain estimates $\hat{\beta}^{(a)}$ for $a = 1, \ldots, v$. In a last step, the v results are combined adequately. The combined estimate is the unweighted average

$$\hat{\beta} = \bar{\beta} = \frac{1}{v}\sum_{a=1}^{v} \hat{\beta}^{(a)},$$

and the combined estimated covariance is

$$\widehat{Cov}(\hat{\beta}) = W + (1 + \frac{1}{v})B.$$

This captures within-imputation variability

$$W = \frac{1}{v}\sum_{a=1}^{v}\widehat{Cov}\left(\hat{\beta}^{(a)}\right)$$

as well as variability between imputations

$$B = \frac{1}{v-1}\sum_{a=1}^{v}(\hat{\beta}^{(a)} - \bar{\beta})(\hat{\beta}^{(a)} - \bar{\beta})^T.$$

MI yields valid results under the assumption of MAR, and thus as well when MCAR is assumed.

Apart from these imputation-based methods, there are also approaches that do not require the gappy data to be completed with "plausible" values.

Complete case analysis: All subjects with missing observations are removed i.e., only completers are included in the analysis. This requires MCAR for obtaining unbiased estimates, but even under that assumption it is obviously inefficient to throw away data, in particular if relatively many subjects have relatively few missing entries. As a consequence, SEs will be unnecessarily large, CIs too wide, and p-values too high. The bias under MAR can go in either direction (Molenberghs et al. 2004; Kenward and Roger 2009).

Available case analysis: The data are used as they are i.e., neither are parts of the data deleted (as with the complete case analysis) nor filled up artificially (as with imputation methods). Although stunningly simple, this is a valid approach under both MAR and MCAR if the joint distribution of responses is modeled with a full likelihood (e.g., using an LMM), and both means and within-subject association are correctly specified in the model. Laird (1988) and Kenward and Molenberghs (1998) pointed out that the covariance of $\hat{\boldsymbol{\beta}}$ has to be estimated from the observed and not the expected information matrix.

There are other techniques for coping with missingness that lead to utilizable results under M(C)AR but will not be discussed here in depth e.g., inverse probability weighting (Seaman and White 2013) or predictive mean matching (Rubin 1986; Little 1988). In case of MNAR, the missingness process has to be included, which is achieved by jointly modeling $\mathbf{Y}_i$ and $\mathbf{R}_i$ e.g., by using a selection model or pattern mixture model (Kenward and Molenberghs 1999).

In practice it is often helpful to resort to methods that deal with missingness in a natural way, such as joint modeling with LMMs. There is a wealth of simulation studies that advise to use an LMM (or MI) and warn against naive approaches like complete case analysis or LVCF (Mallinckrodt et al. 2001ab; Liu and Gould 2002; Mallinckrodt et al. 2004; Molenberghs et al. 2004; Beunckens et al. 2005; Barnes et al. 2008; Siddiqui et al. 2009). Siddiqui (2011) further claims that the full likelihood model is preferable to MI.

The MMM approach corresponds to an available case analysis per time point, but there is no joint likelihood involved for the entire set of data. Therefore it can lead to bias in the case of MAR. GEEs do not specify a full likelihood either and suffer from the same drawback under MAR.

4 Longitudinal MCTs with Gaussian Endpoints

Biological and medical research frequently generates continuous outcome variables for which it is reasonable to assume that the observed values are symmetrically distributed and tend to accumulate around their mean, and also that the variance is independent of the mean. These attributes in their entirety suggest the assumption of a Gaussian distribution for the data, which paves the way for a wealth of statistical analysis techniques to be applicable.

If the task is to compare multiple means of Gaussian data from *uncorrelated* samples (e.g., due to randomization in controlled experiments), standard MCTs using parameter and variance estimates from a classical linear model are straightforward (see 3.7). Matters become more intricate as soon as not all samples involved are stochastically independent e.g., because of experimental units being measured repeatedly at consecutive points in time.

We present and compare two competing modeling strategies for such data: the first one is to fit a joint model (an ELM, a CIM, or some more general LMM) that captures all observations at all occasions; the alternative is to fit one marginal model per occasion and thereupon combine them so as to derive the joint correlation across time points. Either modeling approach can be used to estimate mean parameters and covariances for use in a subsequent MCT procedure.

We develop a unified framework for multiple hypothesis tests and SCIs in setups with dependency structures. Specifically, we apply MCTs to arbitrary linear contrasts of *correlated* means under heteroscedasticity. In the presence of large samples we can rely on asymptotic theory; the crunchpoint is how to get (approximately) valid simultaneous tests and SCIs with small sample sizes. This is practically relevant in many fields of application but still a more or less open question to date. We aim to tweak MCTs for the sake of proper type I error rate control with small sample sizes, evaluate the power of longitudinal MCTs based on joint and multiple marginal models and quantify the superiority over Bonferroni-style adjustments.

This chapter is structured as follows: we develop and characterize simultaneous comparisons of multiple treatment means at several occasions in 4.1, and simultaneous comparisons of multiple occasion means within several treatment groups in 4.2. Then the next step is to hybridize them: in 4.3 we discuss strategies how comparisons of both types can be amalgamated in one and the same family under FWER control. Examples of real-data analyses showing the proposed methods in action are presented in 4.4. Approximate power calculations for longitudinal MCTs with Gaussian outcomes are presented in 4.5.

4.1 Comparing Multiple Treatments at Multiple Occasions

We start with simultaneous inference for several treatment groups at each of several points in time. The desire to protect an FWER over the entire set of elementary hypotheses requires us to account for the multiplicity of treatments as well as occasions. On the other hand, the longitudinal correlation of occasions can be exploited to cushion the conservatism of the procedure compared to a Bonferroni adjustment. We show in 4.1.1 how different modeling strategies may be employed to estimate parameters for subsequent multiple contrast testing and construction of SCIs so that we can pinpoint interesting effects with (approximate) control of α. A short numerical illustration of how the longitudinal correlation influences the correlation matrix of test statistics in the longitudinal MCT is given in 4.1.2. Results of α and power simulations are presented in 4.1.3.

4.1.1 Procedure

Suppose we have a random variable Y_{jki} that describes a Gaussian outcome measured from independent individuals $i = 1, \ldots, n_k$ in randomized treatment groups $k = 1, \ldots, q$ at repeated occasions $j = 1, \ldots, m$. The realizations of Y_{jki} are denoted by y_{jki}. 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). The combinations of occasions and treatment groups are indexed with $b = 1, \ldots, s$ where $s = mq$.

To reflect the data's mean structure, variability, and dependencies among repeated measurements in a model framework, we may take either of two basic modeling approaches: joint modeling with an LMM (as outlined in 3.1), or a combination of marginal linear models (as in 3.5). Both approaches provide us with estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ that we can use for simultaneous inference about the fixed-effects parameters in our longitudinal MCTs.

Joint modeling: We focus on two basic types of joint models that are special cases of our general LMM notion:

- an extended linear model (ELM) where the residuals may exhibit correlation and heteroskedasticity over time and possibly also across treatment groups, or
- a conditional independence model (CIM) that assumes constant residual variance and independent errors, conditional on the random effects.

Whichever LMM we fit, the fixed effects will be parameterized in cell means style i.e., one mean parameter for every combination of treatment group and occasion. This entails maximum flexibility for all shapes of means across time and treatments. The cell means design matrix (for complete and balanced data without additional covariates) is

$$\mathbf{X} = \mathbf{I}_{mq} \otimes \mathbf{1}_{n_q}$$

where $\mathbf{I}$ is an identity matrix and $\mathbf{1}$ is a column vector of ones. Handling incompleteness and including covariates is straightforward.

Considerably harder is the choice of a covariance structure for the joint model. The CIM requires to model random effects that reflect the dependencies among repeated measurements. With the simplest random subject effects, the random-effects design matrix (for complete and balanced data without additional random factors) is

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

If we allow for occasion-specific random subject effects, it becomes

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

where $\mathbf{e}_j$ is the jth unit vector of size m. Likewise, we could make the random subject effects both occasion- and treatment-specific, but this might already run us into computational trouble when actually trying to fit the model.

Many practical experiments are more complex and contain additional hierarchies e.g., the subjects might be hospitalized in different clinics (if they are patients), live in different cages (if they are lab animals), or belong to different spatial blocks (if they are plants in the field). The CIM can smoothly incorporate the hierarchical architecture of the experiment via random effects.

In the ELM framework we are required to make two choices as the residual covariance matrix is assembled from a variance portion and a correlation portion; we described a variety of possible choices in 3.1. In case we decide to fit an LMM that includes both random effects and some residual covariance structure, the agony of choice intensifies.

If unsure about which is the most appropriate model—and this should really be the standard situation—a reasonable option is to let an information criterion do the job. So having assembled several plausible models for the random effects and/or residual covariances, we entrust AICc (see 3.2) with picking the "best" of them. We must keep in mind, however, that AICc (or any other information-based criterion) judges the "goodness" of a model *relative* to its competitors in the candidate set. It says nothing about whether it is good in an *absolute* sense—it might just be the best in a set of terrible models.

Multiple marginal models: The other strategy is to fit one linear model separately for every measurement occasion $j = 1, \ldots, m$. The jth of these marginal models is

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

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, and possible covariates[4]. The dependency structure across time points is established from

$$-E\left(\frac{1}{\sigma^2}\mathbf{X}^{(j)}\mathbf{X}^{(j)T}\right)^{-1}\mathbf{X}^{(j)}\left(\mathbf{y}^{(j)} - \mathbf{X}^{(j)T}\boldsymbol{\beta}^{(j)}\right)$$

as described in 3.5 and Pipper et al. (2012).

Multiple Contrast Tests: Whatever modeling approach (ELM, CIM, some other LMM, or marginal models) has provided us with estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$, we continue with simultaneous inference under the assumption of the $\hat{\boldsymbol{\beta}}$ being multivariate normal and $\hat{\boldsymbol{\Sigma}}$ a consistent estimator of $\boldsymbol{\Sigma}$.

We define the comparisons of interest in a coefficient matrix $\mathbf{C}$. Assume the model parameterization is such that all treatment means at the first occasion come first, followed by the treatment means at the second occasion, and so on. Then the coefficient matrix is block-diagonal and can be constructed as the Kronecker product of an m-dimensional identity matrix and an "elementary" coefficient matrix $\mathbf{C}_0$ as

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

[4] The covariates may, at least in principle, even differ between models.

As an example, the full contrast coefficient matrix for many-to-one comparisons among $q = 3$ treatment groups separately and simultaneously at $m = 4$ occasions is

$$
\begin{aligned}
\mathbf{C}^{trt}_{Dun} &= \mathbf{I}_4 \otimes \mathbf{C}_{Dun} \\
&= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \otimes \begin{bmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix} \\
&= \begin{bmatrix}
-1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0
\end{bmatrix}.
\end{aligned}
$$

The subsequent computation of MCTs i.e., adjusted p-values and SCIs, is to a great extent identical to the standard MCT procedure described in 3.7, albeit with one crucial difference: the correlation matrix of the test statistics no longer contains just known quantities (sample sizes and contrast coefficients) but also covariances reflecting the longitudinal correlation and possible heteroscedasticity in the repeated measurements. These covariances are *a priori* unknown, and we get by with plugging in their estimates from MMM or a joint model, which makes our longitudinal MCT procedure approximate.

By means of the coefficients in $\mathbf{C}$ we formulate contrasts

$$\eta_h = \sum_{b=1}^{s} c_{hb}\beta_b$$

and use them to specify a set of elementary linear hypotheses, the hth pair of which (for two-sided inference) is

$$H_0^{(h)} : \eta_h = \delta_h \quad \text{versus} \quad H_A^{(h)} : \eta_h \neq \delta_h$$

where more often than not $\delta_h = 0 \ \forall \ h$. The hth test statistic is computed as

$$T_h = \frac{\hat{\eta}_h - \delta_h}{\sqrt{\mathbf{c}_h \widehat{\mathbf{\Sigma}} \mathbf{c}_h^T}}$$

with estimated contrast

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

The exact distribution of $\mathbf{T}$ under H_0 is unclear but may be approximated as z-variate t with ν DF and correlation $\tilde{\mathbf{\Gamma}}$:

$$\mathbf{T} \overset{appr.}{\sim} \mathcal{T}_z(\nu, \tilde{\mathbf{\Gamma}}).$$

It is not straightforward to see what the DF should be in general, and especially not in the presence of unbalanced data e.g., when numbers of repeated observations differ between experimental units. A whole range of possible DF approximations for small samples were listed in 3.8.

The asymptotic distribution of $\mathbf{T}$ is z-variate normal with correlation $\tilde{\boldsymbol{\Gamma}}$:

$$\mathbf{T} \overset{asym.}{\sim} \mathcal{N}_z(\mathbf{0}, \tilde{\boldsymbol{\Gamma}}).$$

Details on the multivariate normal and t-distribution are given in Appendix A. We will explore the performances of both the asymptotic and the approximate procedure with different DF approximations in a simulation study in 4.1.3.

The covariance matrix of test statistics $\mathbf{T} = (T1, \dots, T_z)$ under H_0 is

$$\tilde{\boldsymbol{\Sigma}} = \mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^T$$

so that we obtain $\tilde{\boldsymbol{\Gamma}}$ as

$$\tilde{\boldsymbol{\Gamma}} = \mathbf{V}\tilde{\boldsymbol{\Sigma}}\mathbf{V}$$

where $\mathbf{V} = \text{diag}(\tilde{\boldsymbol{\Sigma}})^{-\frac{1}{2}}$ is the inverse of a matrix with the square root of the diagonal elements from $\tilde{\boldsymbol{\Sigma}}$ on its diagonal and all off-diagonal elements zero. As $\boldsymbol{\Sigma}$ is unknown in practice, we plug in a consistent estimate $\widehat{\boldsymbol{\Sigma}}$ (from a joint model or the MMM approach) so as to get $\widehat{\tilde{\boldsymbol{\Sigma}}}$ and $\widehat{\tilde{\boldsymbol{\Gamma}}}$.

An elementary hypothesis $H_0^{(h)}$ gets rejected if

$$|T_h| > t_{z,1-\alpha}^{two}(\nu, \widehat{\tilde{\boldsymbol{\Gamma}}})$$

with $t_{z,1-\alpha}^{two}(\nu, \widehat{\tilde{\boldsymbol{\Gamma}}})$ the two-sided equicoordinate $(1-\alpha)$ quantile of $\mathcal{T}_z(\nu, \widehat{\tilde{\boldsymbol{\Gamma}}})$. Bounds of SCIs with coverage probability $1-\alpha$ are obtained as

$$\hat{\eta}_h \mp t_{z,1-\alpha}^{two}(\nu, \widehat{\tilde{\boldsymbol{\Gamma}}})\sqrt{\mathbf{c}_h \widehat{\boldsymbol{\Sigma}} \mathbf{c}_h^T}.$$

Adjusted p-values are given by

$$p_h = 1 - \int_{-|T_h^{obs}|}^{|T_h^{obs}|} \dots \int_{-|T_h^{obs}|}^{|T_h^{obs}|} t_z(\mathbf{x}; \nu, \widehat{\tilde{\boldsymbol{\Gamma}}}) \; d\mathbf{x}$$

where T_h^{obs} designates an observed value of the test statistic T_h, and $t_z(\mathbf{x}; \nu, \widehat{\tilde{\boldsymbol{\Gamma}}})$ is the PDF corresponding to $\mathcal{T}_z(\nu, \widehat{\tilde{\boldsymbol{\Gamma}}})$. The z-dimensional integral needs to be solved numerically e.g., using a software implementation of the Genz-Bretz algorithm (Genz and Bretz 2009).

Since the global null hypothesis is the intersection of elementary nulls

$$H_0 = \bigcap_{h=1}^{z} H_0^{(h)}$$

we reject the global H_0 if at least one of the $H_0^{(h)}$ is rejected. This implies a maximum-type test with test statistic

$$T^{max} = \max_h |T_h|$$

whose p-value is computed as

$$p = \min_h p_h.$$

One-sided testing problems are addressed in a similar fashion; see e.g., Bretz et al. (2010).

4.1.2 Numerical Illustration

Imagine the simple case of n experimental units randomized to $q = 2$ treatments called A and B, and some continuous outcome is measured twice from each of them ($m = 2$), at two different points in time. Assume that observations are independent across treatments (due to randomization) but correlated across time points (due to repeated measurements) with covariance matrix

$$\mathbf{\Sigma} = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}$$

for both treatments A and B. We are interested in comparisons of treatments A and B in a cell means model sseparately and simultaneously for occasions 1 and 2, requiring the coefficient matrix

$$\mathbf{C}_2 = \begin{bmatrix} -1 & 0 & 1 & 0 \\ 0 & -1 & 0 & 1 \end{bmatrix}.$$

The covariance matrix of the cell means model is

$$\mathbf{\Lambda}_2 = \mathbf{I}_2 \otimes \mathbf{\Sigma} = \begin{bmatrix} 1 & 0.9 & 0 & 0 \\ 0.9 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0.9 \\ 0 & 0 & 0.9 & 1 \end{bmatrix}.$$

Note that the means we want to compare (μ_{A1} vs. μ_{B1} and μ_{A2} vs. μ_{B2}) are indeed uncorrelated! The two test statistics, however, are not:

$$\tilde{\mathbf{\Sigma}}_2 = \mathbf{C}_2 \mathbf{\Lambda}_2 \mathbf{C}_2^T = \begin{bmatrix} 2 & 1.8 \\ 1.8 & 2 \end{bmatrix}.$$

Standardizing $\tilde{\mathbf{\Sigma}}_2$ by its diagonal elements gives the correlation matrix

$$\tilde{\mathbf{\Gamma}}_2 = \text{diag}(\tilde{\mathbf{\Sigma}}_2)^{-\frac{1}{2}}\ \tilde{\mathbf{\Sigma}}_2\ \text{diag}(\tilde{\mathbf{\Sigma}}_2)^{-\frac{1}{2}} = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}.$$

The identity $\tilde{\mathbf{\Gamma}} = \mathbf{\Sigma}$ is of course not true for general $\mathbf{\Sigma}$ and $\mathbf{C}$.

Extending the case to $q = 3$ treatment groups shows how the longitudinal MCT puts figures to the “off-diagonal blocks” of $\tilde{\mathbf{\Sigma}}$ that are occupied by zeroes in standard MCTs. Suppose we want to carry out many-to-one comparisons among treatments A, B, and C separately and simultaneously for occasions 1 and 2, using the coefficient matrix

$$\mathbf{C}_3 = \begin{bmatrix} -1 & 0 & 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 1 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

The model covariance matrix is

$$\mathbf{\Lambda}_3 = \mathbf{I}_3 \otimes \mathbf{\Sigma} = \begin{bmatrix} 1 & 0.9 & 0 & 0 & 0 & 0 \\ 0.9 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0.9 & 0 & 0 \\ 0 & 0 & 0.9 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0.9 \\ 0 & 0 & 0 & 0 & 0.9 & 1 \end{bmatrix},$$

so again the means to be compared are uncorrelated. The covariance matrix of test statistics

$$\tilde{\mathbf{\Sigma}}_3 = \mathbf{C}_3 \mathbf{\Lambda}_3 \mathbf{C}_3^T = \begin{bmatrix} 2 & 1 & 1.8 & 0.9 \\ 1 & 2 & 0.9 & 1.8 \\ 1.8 & 0.9 & 2 & 1 \\ 0.9 & 1.8 & 1 & 2 \end{bmatrix}$$

has got non-zero elements in the “off-diagonal blocks”, and so does the correlation matrix

$$\tilde{\mathbf{\Gamma}}_3 = \mathrm{diag}(\tilde{\mathbf{\Sigma}}_3)^{-\frac{1}{2}}\ \tilde{\mathbf{\Sigma}}_3\ \mathrm{diag}(\tilde{\mathbf{\Sigma}}_3)^{-\frac{1}{2}} = \begin{bmatrix} 1 & 0.5 & 0.9 & 0.45 \\ 0.5 & 1 & 0.45 & 0.9 \\ 0.9 & 0.45 & 1 & 0.5 \\ 0.45 & 0.9 & 0.5 & 1 \end{bmatrix}.$$

We see how the longitudinal correlation “propagates” to the correlation matrix of test statistics and at the end of the day makes the test less conservative than Bonferroni.

4.1.3 Simulation Study

Our methods have *asymptotically* correct size when using the multivariate normal reference distribution, and are only *approximate* with the reference distribution being multivariate t; therefore we are interested in their actual behavior especially with small to moderate sample sizes as they are common in biological and clinical practice. We investigate via simulation what sizes of samples are required for the asymptotic procedure to achieve acceptable performance. Then we study the small-sample variant using an approximation to the DF. The fact that there is no generally “correct” DF leads to uncertainty about which approximation to use in practice. We undertake simulations to assess the performance of a number of DFs numerically. Furthermore, we compare the powers of longitudinal MCTs based on joint models and MMM with that of simplistic Bonferroni-type procedures.

Type I error: To assess the behavior of the asymptotic procedures, we consider balanced settings with $q = \{3, 4, 5\}$ 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 group are drawn from an m-variate normal distribution $\mathcal{N}_m(\boldsymbol{\mu}, \mathbf{\Lambda})$ with mean vector $\boldsymbol{\mu} = (10, \ldots, 10)$ and joint covariance matrix

$$\mathbf{\Lambda} = \mathbf{ABA}$$

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 $\mathbf{A} = \mathrm{diag}(\sqrt{1}, \sqrt{2}, \ldots, \sqrt{m})$. With $m = 4$, for instance, this yields

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

We simulate 5000 datasets under H_0 and carry out many-to-one (Dunnett), all-pairwise (Tukey), or grand-mean (ANOM) comparisons among treatments per measurement occasion for two-sided hypotheses using parameter and covariance estimates obtained from

a) a joint ELM assuming AR(1) correlation and heterogeneous variances over 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.

The simulation results are displayed in Figure 11. The type I error rates of all methods are slightly inflated to around 7 to 8% with $n_k = 10$ but level off at the nominal 5% once the sample sizes exceed 30 or 40. This holds true for many-to-one, all-pairwise, and grand-mean 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.

Having seen that at least 30 to 40 subjects per group are necessary for the asymptotic procedure to keep the type I error rate, we study the small-sample properties of the approximate procedure with different DFs for the multivariate t-distribution. The simulation setup is identical to the asymptotic one, with the only difference that we reduce the sample size per treatment to $n_k = \{4, 6, 8, 10, 12, 14\}$. Then we fit models as described above and compute longitudinal MCTs using

- ν_{naive}, ν_{ESS}, ν_{adj}, ν_{PB}, and ν_{res} as DF approximations for the joint ELM with AR(1) correlation and heterogeneous variances over time,
- the same DF methods and additionally ν_{KR} for the CIM with occasion-related random subject effects, and
- the minimum ν_{res} of all marginal models for MMM.

So in total each simulation dataset is evaluated twelve times, and we record for each evaluation the number of rejections of the global H_0.

Results from 5000 simulation runs are displayed in Figure 12. Unsurprisingly, the residual DF, when applied to a CIM or an ELM, provokes liberal test decisions with type I error rates of 10% and above in the presence of small sample sizes; on the other hand, the naive DF makes the test severely conservative with almost no power for $n_k \leq 6$. So quite predictably, ν_{naive} and ν_{res} act as boundaries for the remaining DFs.

All CIM-based tests are more conservative (or less liberal) than their ELM counterparts using the same DF approximation, especially for $m > 3$. A similar behavior could be observed in the asymptotic simulations as well.

The Pinheiro-Bates DF performs excellent even when n_k is very small. The DF based on the ESS is conservative for $n_k \leq 8$ but comes very close to the nominal α of 0.05 as the sample size increases. The Kenward-Roger method for the CIM performs very well in most situations and only gets a little conservative with $m = 4$ and $m = 5$, but this is probably due to the CIM and not the DF itself. Our own “adjusted” DF does a fairly good job across most scenarios but not substantially better than other DFs. The MMM approach also comes very close to a size of 5% even with the smallest sample sizes and

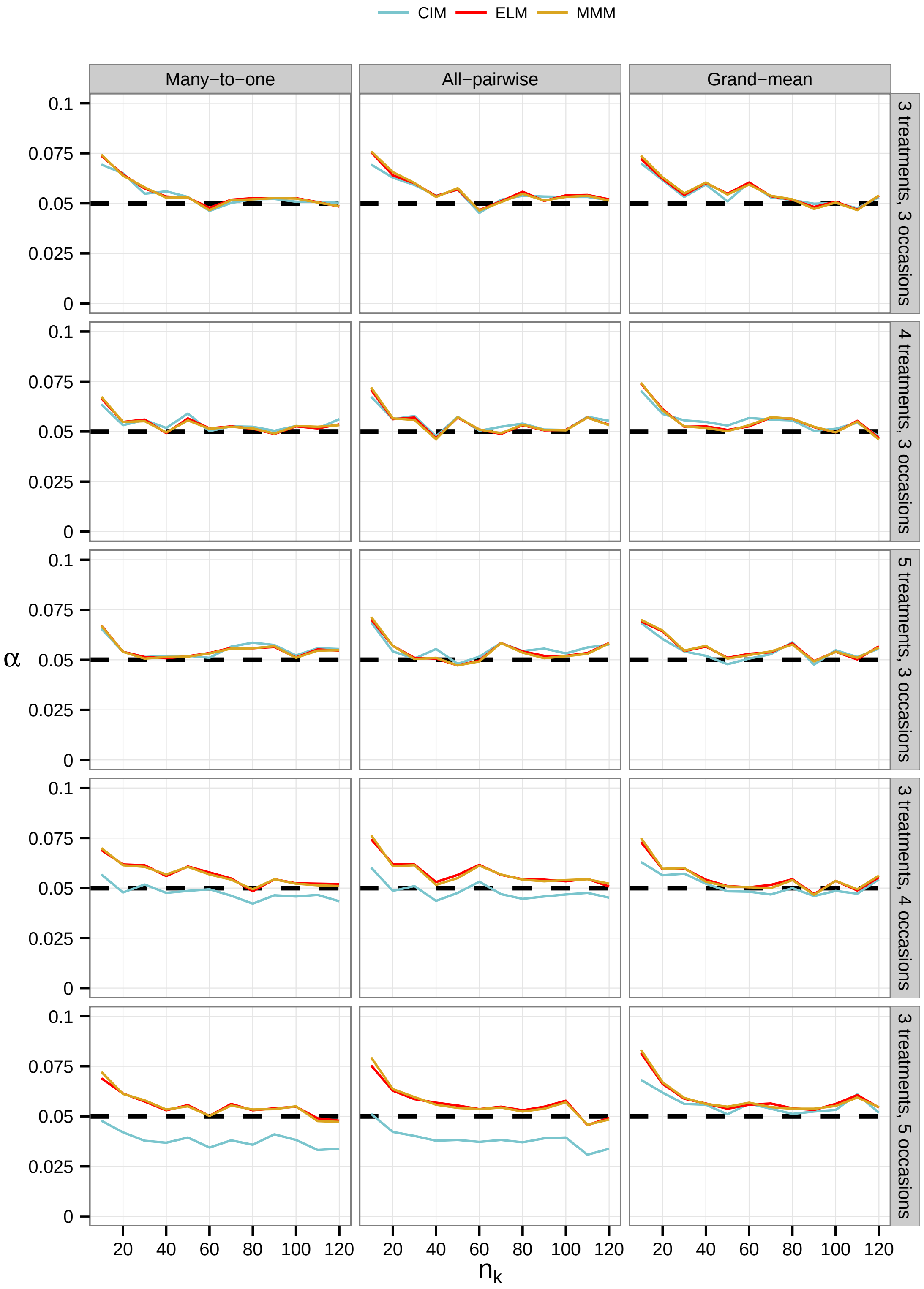


Figure 11: Simulated type I error rates for asymptotic many-to-one, all-pairwise, and grand-mean comparisons among $q = \{3, 4, 5\}$ Gaussian treatment means separately and simultaneously at $m = \{3, 4, 5\}$ occasions, with n_k independent subjects per treatment group (5000 simulation runs).

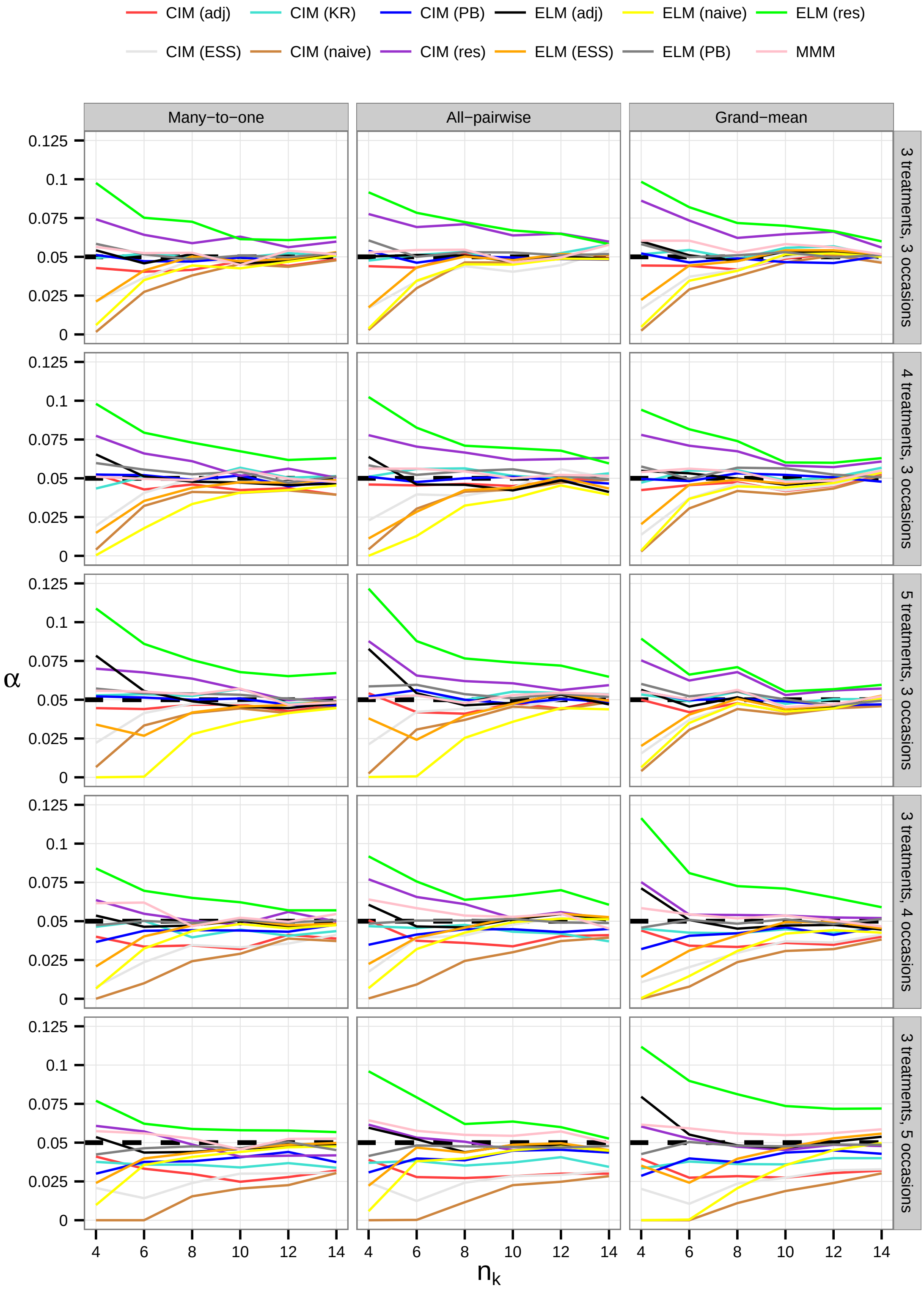


Figure 12: Simulated type I error rates for many-to-one, all-pairwise, and grand-mean comparisons among $q = \{3, 4, 5\}$ Gaussian treatment means separately and simultaneously at $m = \{3, 4, 5\}$ occasions, with n_k independent subjects per treatment group, and different small-sample degrees of freedom (5000 simulation runs).

shows consistently good control of α across all simulated setups. There are no sailent discrepancies between many-to-one, all-pairwise, and grand-mean contrasts.

Power: Now that we have figured out sample sizes n_k that ensure α is kept for the asymptotic procedures based on either joint or multiple marginal models, and also detected DF approximations that work well for small sample sizes, one may wonder if the methods differ in terms of statistical power. Another relevant question is 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 MCTs occasion-wise, or even pairwise z- or t-tests, followed by a Bonferroni adjustment?

We focus our asymptotic power investigations on many-to-one, all-pairwise, and grand-mean 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 for the α simulations, 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 $\boldsymbol{\mu} = (\mu, \dots, \mu, \mu + \Delta)$ with non-centralities $\Delta = \{0, 0.1, \dots, 1.5\}$. This leads to a scenario with exactly one many-to-one and one grand-mean comparison being under H_A; the number of all-pairwise tests under the alternative is $q - 1$.

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

1. Calculate standard z-tests for all single comparisons and adjust the resulting p-values with Bonferroni. This means turning a blind eye to any correlations among test statistics.
2. Perform an (asymptotic) MCT within each time point and adjust via Bonferroni for the multiplicity of time points. This approach incorporates the portion of correlation that originates from multiple test statistics being built with overlapping subsets of $\widehat{\boldsymbol{\beta}}$, but ignores the correlation among time points.
3. Base 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 with the MMM method 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 13 (in most panels the blue, brown, and green curves overlap, and so do the gray and pink ones). All three methods that acknowledge dependence among 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 curves indicate a power advantage of 7 to 8 percentage points at the steepest point near $\Delta = 0.7$. 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 advantage of Bonferroni-corrected single MCTs over Bonferroni-corrected single t-tests is marginal (around 1%) throughout the simulation settings.

All power considerations up to this point assumed a true underlying Toeplitz correlation with $\rho_1 = 0.90$ 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 investigate four different correlation patterns:

$$\mathbf{B}_1 = \begin{pmatrix} 1 & 0.95 & 0.90 \\ & 1 & 0.95 \\ & & 1 \end{pmatrix}, \quad \mathbf{B}_2 = \begin{pmatrix} 1 & 0.90 & 0.80 \\ & 1 & 0.90 \\ & & 1 \end{pmatrix},$$

$$\mathbf{B}_3 = \begin{pmatrix} 1 & 0.80 & 0.50 \\ & 1 & 0.80 \\ & & 1 \end{pmatrix}, \quad \mathbf{B}_4 = \begin{pmatrix} 1 & 0.50 & 0.20 \\ & 1 & 0.50 \\ & & 1 \end{pmatrix}.$$

The corresponding power curves are displayed in Figure 14 (again we see lots of overlap just like before). 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 is actually negligible unless $\rho_1 \geq 0.9$. Across all settings considered, the ELM assuming AR(1) and MMM perform nearly identically whereas the CIM can have slightly lower or higher power, depending on the correlation $\boldsymbol{\rho}$; this is again indication of the conditional independence assumption being not quite appropriate in these cases.

The simulation setting for the small-sample power is similar to the asymptotic one, but now exact rather than asymptotic tests are used for the Bonferroni-corrected methds (i.e., separate t-tests instead of z-tests, and separate MCTs that have a multivariate t reference distribution instead of multivariate normal), and we use small-sample adjustments for our longitudinal MCTs: ν_{PB} with the joint ELM, ν_{KR} with the joint CIM, and the minimum ν_{res} of the marginal models for MMM. We consider true treatment differences of $\Delta = \{0, 0.2, \ldots, 4\}$ in one group at one point in time in settings with $n_k = 10$.

The overall picture of the simulation results shown in Figures 60 and 61 in Appendix D is very similar to that of the asymptotic procedure. The power curves of all methods that incorporate longitudinal correlation are very alike, and the power advantage over Bonferroni-type adjustments grows with the number of occasions and with the strength of the longitudinal correlation.

4.2 Comparing Multiple Occasions within Multiple Treatments

Having developed a procedure for multiple comparisons of treatments at several occasions in 4.1, we now change direction and go into multiple comparisons of occasions within several treatment groups. We expound the method, which is indeed very similar, in 4.2.1, and illustrate it numerically in 4.2.2. Simulation results under H_0 and H_A are presented in 4.2.3 for both the asymptotic and the small-sample procedures.

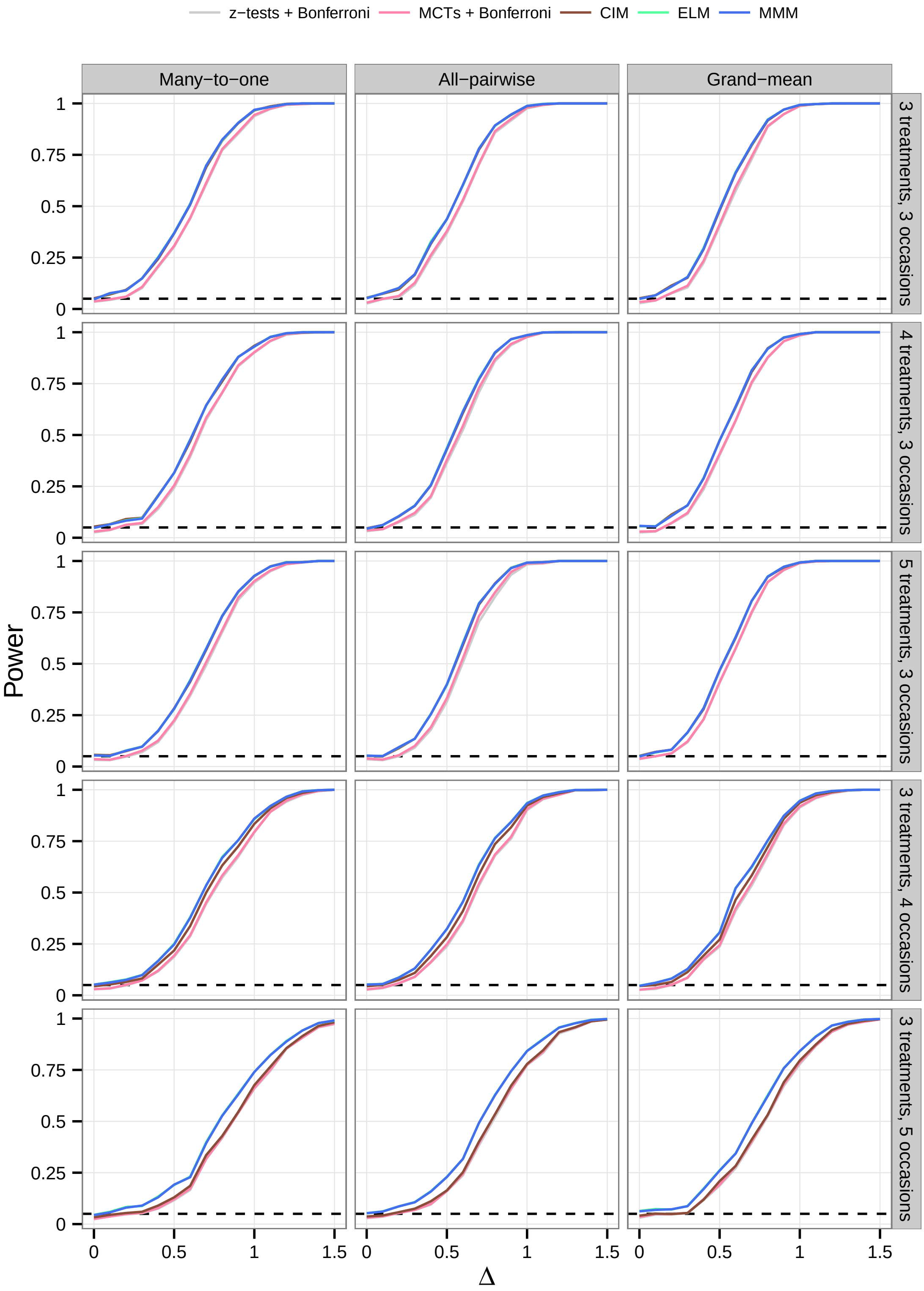


Figure 13: Simulated powers for asymptotic many-to-one, all-pairwise, and grand-mean comparisons among $q = \{3, 4, 5\}$ Gaussian treatment means separately and simultaneously at $m = \{3, 4, 5\}$ occasions, with $n_k = 100$ independent subjects per treatment group (1000 simulation runs).

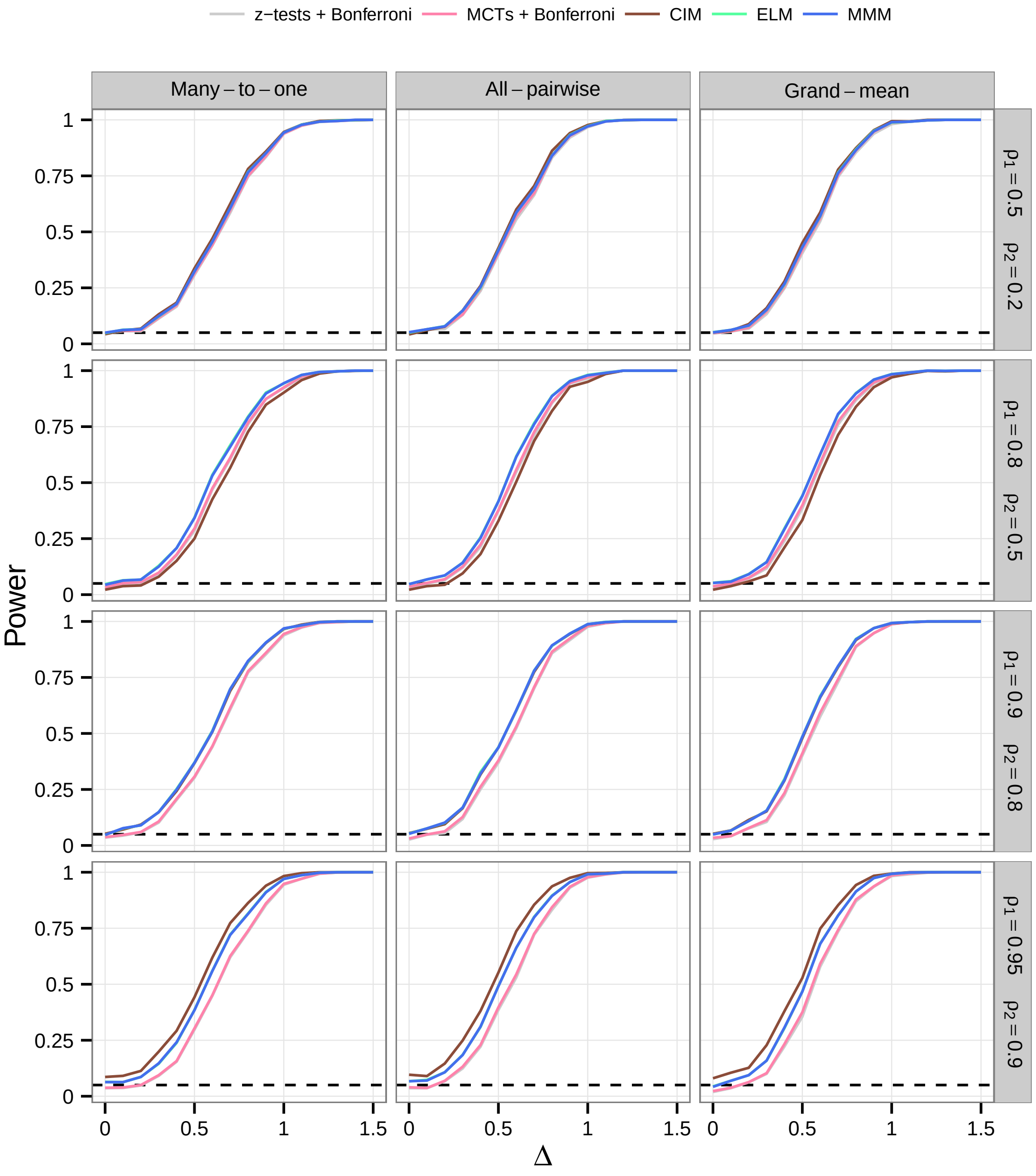


Figure 14: Simulated powers for asymptotic many-to-one, all-pairwise, and grand-mean comparisons among $q = 3$ Gaussian treatment means separately and simultaneously at $m = 3$ occasions, with $n_k = 100$ independent subjects per treatment group, and different longitudinal correlations (1000 simulation runs).

4.2.1 Procedure

The modeling step for comparisons of occasion means within several treatment groups is the same as described in 4.1.1: we may either fit a joint model embracing all time points (or fit several candidate joint models and hand them over to AICc), or m occasion-specific marginal models and then combine them with the MMM method to devise the joint correlation over time. Having obtained model estimates of both $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$, we can continue with simultaneous inference. The coefficient matrix $\mathbf{C}$ for the MCT is conveniently built as the Kronecker product of an "elementary" coefficient matrix and a q-dimensional identity matrix as

$$\mathbf{C} = \mathbf{C}_0 \otimes \mathbf{I}_q.$$

We exemplify this for Dunnett-type comparisons of $m = 4$ occasions separately and simultaneously within $m = 3$ treatment groups and get

$$\begin{aligned}
\mathbf{C}_{Dun}^{occ} &= \mathbf{C}_{Dun} \otimes \mathbf{I}_3 \\
&= \begin{bmatrix} -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \end{bmatrix} \otimes \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \\
&= \begin{bmatrix}
-1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
-1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & -1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
-1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{bmatrix}.
\end{aligned}$$

The remainder of the MCT procedure is along the lines of what we described for comparisons among treatment means at multiple occasions in 4.1.1, and we investigate the same DF approximations for small samples.

4.2.2 Numerical Illustration

Similar to the illustration in 4.1.2, suppose there are $q = 2$ treatments (A and B) to be compared separately and simultaneously at two points in time whose longitudinal correlation is

$$\mathbf{\Sigma} = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix},$$

and the comparisons are defined by the coefficient matrix

$$\mathbf{C}_2 = \begin{bmatrix} -1 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 \end{bmatrix}.$$

As opposed to 4.1, now we compare *correlated* mean parameters:

$$\mathbf{\Lambda}_2 = \mathbf{I}_2 \otimes \mathbf{\Sigma} = \begin{bmatrix} 1 & 0.9 & 0 & 0 \\ 0.9 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0.9 \\ 0 & 0 & 0.9 & 1 \end{bmatrix}.$$

The test statistics themselves, however, are *uncorrelated*:

$$\tilde{\mathbf{\Sigma}}_2 = \mathbf{C}_2 \mathbf{\Lambda}_2 \mathbf{C}_2^T = \begin{bmatrix} 0.2 & 0 \\ 0 & 0.2 \end{bmatrix},$$

and correspondingly

$$\tilde{\mathbf{\Gamma}}_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

4.2.3 Simulation Study

Just as with comparisons among treatments, it is unclear what sample sizes are necessary for the asymptotic procedures to preserve α, and what DF approximation to use with smallish sample sizes. Thus we hope to find answers via simulation. Beyond that, we compare the powers of the longitudinal MCTs with Bonferroni-style adjustments.

Type I error: The setup of our simulation study for type I error rates is identical to the one from 4.1.3, except that the comparisons are now between occasion means within multiple treatment groups. Simulation results of the asymptotic procedures for n_k between 10 and 120 are shown in Figure 15. They give a slightly different picture compared to comparisons among treatments per time point. In the MMM approach the realized α for all-pairwise comparisons skyrockets to 14 to 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. Tests relying on the CIM can again turn a little conservative as the number of occasions rises.

For the approximate small-sample procedure, simulated type I error rates based on 5000 simulation runs are presented in Figure 16. We see once more that the naive DF makes the procedure extremely conservative; on the contrary, MCTs using the residual DF are prone to rejecting too many true null hypotheses; so both these DFs are out of the question.

The ESS-based, Pinheiro-Bates and “adjusted” DFs keep the nominal α but tend to conservatism, especially for $n_k < 10$ in setups with $m > 3$ occasions. The Kenward-Roger method performs well across most scenarios but seems to turn slightly liberal with increasing number of occasions. MMM can become fairly liberal especially for $m > 3$.

Power: The simulation setup is again very similar to that in 4.1.3 but now with comparisons between time points instead of treatment groups. The joint models and MMM remain the same as before but their competitors change somewhat: as two-sample tests and MCTs based on simple linear modeling would be inappropriate for comparing the correlated occasion means, we rather use CIMs with random subject effects to base our tests on, and then adjust them via Bonferroni for the multiplicity of treatment groups (and also of the comparisons in case of pairwise tests). All power simulations are carried out with 1000 replications.

Figure 17 shows that for comparisons of occasions within each treatment group, the superiority of joint modeling or the MMM approach compared to procedures that adjust for multiple time points with Bonferroni is beyond any doubt. The ELM with AR(1) correlation, the CIM with random subject effects, and MMM performs equally well. Their power advantage is bigger for all-pairwise contrasts compared to many-to-one and grand-mean because the number of comparisons under the alternative is $m - 1$ for the former but only 1 for the latter. Increasing the number of time points diminishes the power of all procedures.

The impact of the strength of longitudinal correlation is illustrated in Figure 18. The gap between the asymptotic power curves of procedures with or without incorporation

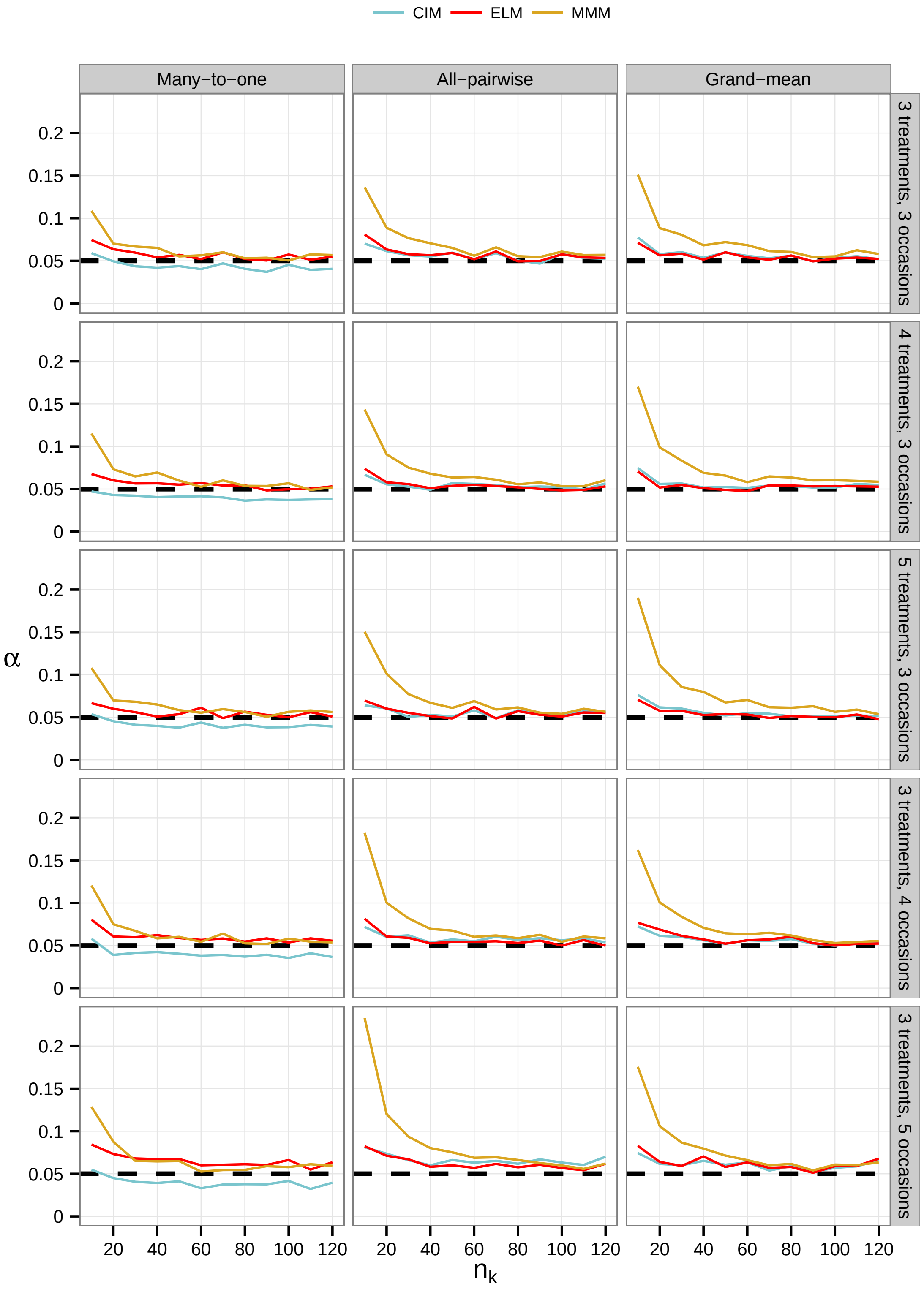


Figure 15: Simulated type I error rates for asymptotic many-to-one, all-pairwise, and grand-mean comparisons among $m = \{3, 4, 5\}$ Gaussian occasion means separately and simultaneously for $q = \{3, 4, 5\}$ treatments, with n_k independent subjects per treatment group (5000 simulation runs).

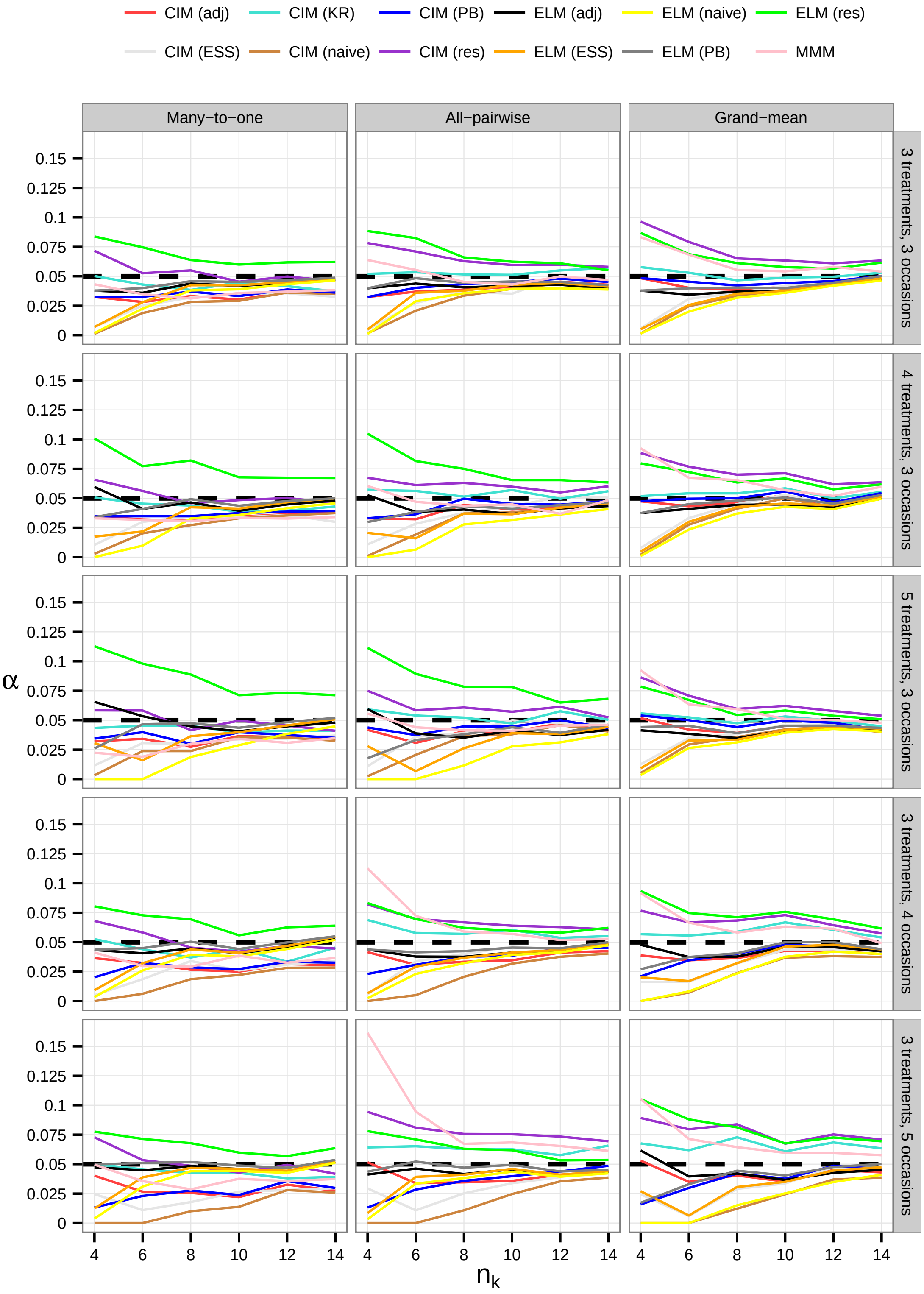


Figure 16: Simulated type I error rates for many-to-one, all-pairwise, and grand-mean comparisons among $m = \{3, 4, 5\}$ Gaussian occasion means separately and simultaneously for $q = \{3, 4, 5\}$ treatments, with n_k independent subjects per treatment group, and different small-sample degrees of freedom (5000 simulation runs).

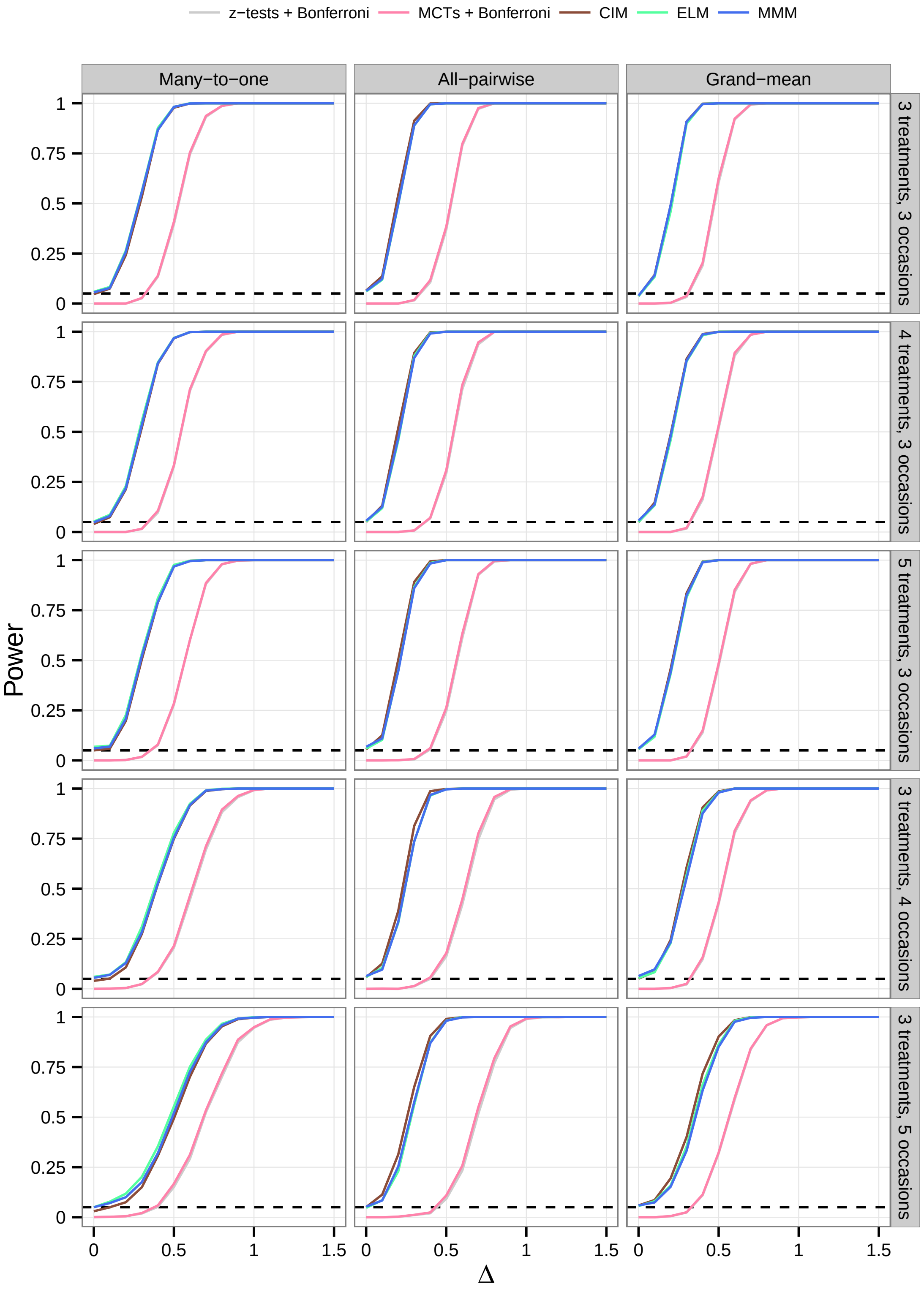


Figure 17: Simulated powers for asymptotic many-to-one, all-pairwise, and grand-mean comparisons among $m = \{3, 4, 5\}$ Gaussian occasion means separately and simultaneously for $q = \{3, 4, 5\}$ treatments, with $n_k = 100$ independent subjects per treatment group (1000 simulation runs).

of longitudinal correlation increases with the correlation. Even under relatively weak dependence ($\rho_1 = 0.50$, $\rho_2 = 0.20$), the gain in power is similar to the maximum achievable gain for comparisons of treatment groups within time points—but the latter requires very high correlation ($\rho_1 = 0.95$, $\rho_2 = 0.90$) for this.

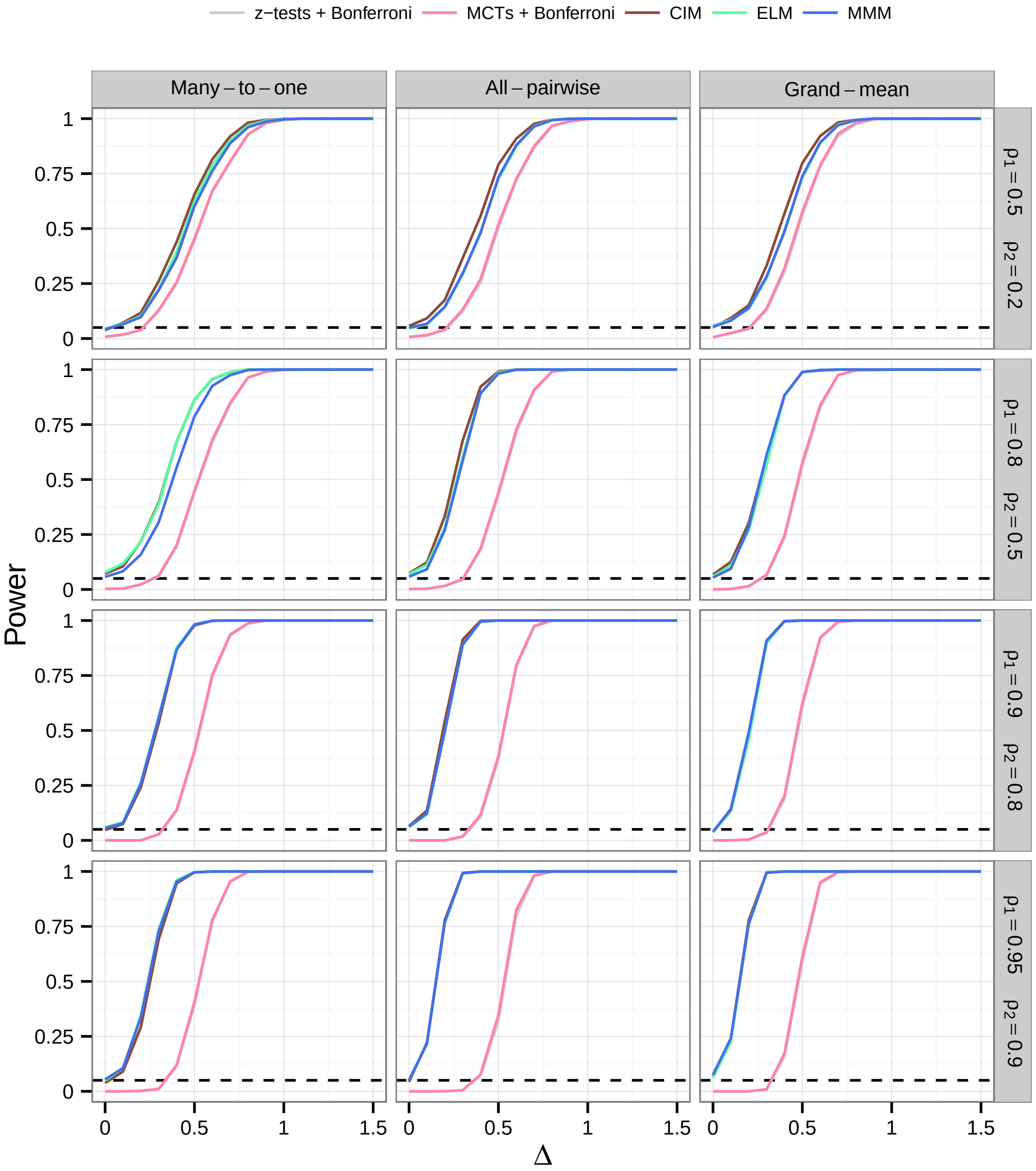


Figure 18: Simulated powers for asymptotic many-to-one, all-pairwise, and grand-mean comparisons among $m = 3$ Gaussian occasion means separately and simultaneously for $q = 3$ treatments, with $n_k = 100$ independent subjects per treatment group, and different longitudinal correlations (1000 simulation runs).

We use the setup, models, and DFs as in 4.1.3 for simulating the power with small sample sizes. Figure 62 in Appendix D shows results based on 1000 simulation runs. Again the procedures that exploit the longitudinal correlation are clearly superior to the simple

Bonferroni-type adjustments, but now MMM has somewhat lower power compared to the joint modeling approaches.

The small-sample power comparisons of different correlations with $m = q = 3$ (Figure 63 in Appendix D) also show this pattern: the power advantage of procedures that account for longitudinal correlation increases dramatically as the correlation increases, but MMM is not as quite as powerful as joint modeling.

In summary, comparisons among occasions per treatment group benefit much more from incorporating longitudinal correlation than comparisons among treatments per occasion, and substantial gain is achieved already with relatively weak longitudinal dependence. With finite samples, the MMM approach can have lower power compared to joint modeling.

4.3 A Duplex Procedure

Now that we have established procedures for multiple comparisons of treatments at each of several occasions as well as for multiple comparisons of occasions within each of several treatment groups, the next step is to combine them. Suppose the subject-matter issue is such that comparisons of both types are of interest, and we are willing to claim an effect if at least one comparison of treatments is significant at least at one occasion, or if at least one comparison of occasions is significant at least for one treatment group i.e., any significant effect in any direction is sufficient to reject the global H_0. This calls for a duplex procedure that embraces comparisons among treatments *and* among occasions and controls the FWER over the entire set of hypotheses, at least approximately with small sample sizes.

The strategy shall be to take the longitudinal MCTs from 4.1 and 4.2 as building blocks. We estimate $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ using either joint modeling or MMM. Elementary hypotheses are specified in a coefficient matrix $\mathbf{C}$ that includes contrasts of both types. One apparent challenge here is how to find a critical value (or values) if the corresponding DFs are widely different for the two types of comparisons. This is less of a problem when sample sizes are relatively large but should be properly addressed for smallish dimensions.

We describe the method and several approaches to approximate the small-sample DFs in 4.3.1. The impact of longitudinal dependence on the correlation of test statistics is illustrated in 4.3.2. We show simulation results for size and power in 4.3.3.

4.3.1 Procedure

The statistical procedure itself resembles the basic one described in 4.1.1, with the difference that we have now comparisons among treatments *and* among occasions at the same time. The coefficient matrix that reflects this set of comparisons (again exemplified for many-to-one comparisons of $q = 3$ treatments and $m = 4$ occasions) can be assembled from $\mathbf{C}_{Dun}^{trt}$ of 4.1.1 and $\mathbf{C}_{Dun}^{occ}$ of 4.2.1 as

$$\mathbf{C}_{Dun}^{both} = \begin{bmatrix} \mathbf{C}_{Dun}^{trt} \\ \mathbf{C}_{Dun}^{occ} \end{bmatrix} = \begin{bmatrix} -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 \\ -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

Inferences from this asymptotic procedure will be under control of the FWER with large sample sizes. Matters get more complicated, however, for small samples because the DFs to be used are even less obvious than with comparisons of treatments or occasions alone. The point is that, depending on the DF method used, the DFs assigned to each type of comparison (within treatment groups or within time points) can be substantially different.

Suppose we are using the Kenward-Roger method to determine DFs. This yields a vector $\boldsymbol{\nu} = (\nu_1, \dots, \nu_z)$ where ν_h is the DF associated with the hth comparison. The question is how to incorporate these multiple DFs into our inference framework with a multivariate t-distribution from which an equicoordinate critical point is to be calculated. We make three suggestions and investigate their performances later on:

1. Apply the minimum DF
$$\nu_{min} = \min_h(\nu_1, \dots, \nu_z)$$
to all comparisons i.e., use
$$\mathcal{T}_z(\nu_{min}, \tilde{\mathbf{\Gamma}})$$
as reference distribution for all test statistics. This approach is "exact" (to the extent the Kenward-Roger method is "exact") only in completely balanced and homoscedastic settings, and likely to be conservative otherwise.

2. Use the average DF
$$\bar{\nu} = \frac{1}{z}\sum_{h=1}^{z} \nu_h$$
for all comparisons and thus
$$\mathcal{T}_z(\bar{\nu}, \tilde{\mathbf{\Gamma}})$$
as joint reference distribution of all test statistics. Just like the ν_{min} approach, this is only "exact" (in the above sense) with complete balance and homoscedasticity.

In general, it should be “exact on average” (again in the above sense) meaning that nominal test size should be achieved; however, some adjusted p-values should be too small and others too large, and likewise, some SCIs should be too narrow and some too wide.

We may run into trouble with strategies 1 and 2 when there is heavy imbalance and/or heteroscedasticity. One possible corrective is to use different DFs for different comparisons. Games and Howell (1976) proposed such an approach for all-pairwise mean comparisons, and Hasler and Hothorn (2008) extended it to arbitrary linear contrasts using MCTs. They used Satterthwaite DFs, but without a Kenward-Roger adjustment of the covariances.

3. Apply the contrast-specific DFs as computed by the Kenward-Roger approximation so that
$$\mathcal{T}_z(\nu_h, \tilde{\mathbf{\Gamma}})$$
is used as reference distribution for the hth test statistic. The adjusted p-value corresponding to the hth elementary (two-sided) hypothesis is found by integrating from $-|T_h|$ to $|T_h|$ over $\mathcal{T}_z(\nu_h, \tilde{\mathbf{\Gamma}})$. This should be “exact” (again in the above sense) even with imbalance and heterogeneous variances. The method is computationally tedious as it involves multiple integration steps; an `R` program is given in Appendix E.

We also need to point out that a procedure assigning different DFs to elementary hypotheses leads to different critical values for single test statistics of the same family to be compared with; what follows is no longer a simultaneous test procedure in the sense of Gabriel (1969).

These three general ideas can be applied in a similar way to DFs that, unlike the Kenward-Roger method, do not produce contrast-specific DFs but “only” one DF for comparisons among treatments and one DF for comparisons among time points. In this case averaging will be weighted according to the number of comparisons of each type i.e.,

$$\bar{\nu} = \frac{z_{trt}\nu_{trt} + z_{occ}\nu_{occ}}{z}$$

where ν_{trt} and ν_{occ} are the DFs for comparisons of treatments per occasion and occasions per treatment, respectively, z_{trt} and z_{occ} are the numbers of comparisons of each type, and $z_{trt} + z_{occ} = z$.

We study the performance of ν_{min}, $\bar{\nu}$, and contrast-specific ν_h via simulation in 4.3.3.

4.3.2 Numerical Illustration

We draw on the heuristical illustrations of 4.1 and 4.2 and start again by assuming a longitudinal correlation of

$$\mathbf{\Sigma} = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}$$

in each of treatment groups A and B i.e., an overall correlation matrix

$$\mathbf{\Lambda}_2 = \mathbf{I}_2 \otimes \mathbf{\Sigma} \begin{bmatrix} 1 & 0.9 & 0 & 0 \\ 0.9 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0.9 \\ 0 & 0 & 0.9 & 1 \end{bmatrix}$$

for the cell means model. The coefficient matrix now comprises comparisons among treatments as well as among time points:

$$\mathbf{C}_2 = \begin{bmatrix} -1 & 0 & 1 & 0 \\ 0 & -1 & 0 & 1 \\ -1 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 \end{bmatrix}.$$

This leads to a covariance matrix of test statistics

$$\tilde{\mathbf{\Sigma}}_2 = \mathbf{C}_2\mathbf{\Lambda}_2\mathbf{C}_2^T = \begin{bmatrix} 2 & 1.8 & 0.1 & -0.1 \\ 1.8 & 2 & -0.1 & 0.1 \\ 0.1 & -0.1 & 0.2 & 0 \\ -0.1 & 0.1 & 0 & 0.2 \end{bmatrix}$$

and the corresponding correlation matrix

$$\tilde{\mathbf{\Gamma}}_2 = \begin{bmatrix} 1 & 0.9 & 0.16 & -0.16 \\ 0.9 & 1 & -0.16 & 0.16 \\ 0.16 & -0.16 & 1 & 0 \\ -0.16 & 0.16 & 0 & 1 \end{bmatrix}.$$

Again we see that the two test statistics for comparisons of treatments per occasion (T_1 and T_2) are highly correlated whereas those for comparisons of occasions per treatment (T_3 and T_4) are uncorrelated. In addition, there is a weak positive correlation between T_1 and T_3 as well as between T_2 and T_4; these are the comparisons that have the same figure (–1 or 1) in the same place of their respective contrast coeffient vectors. There is also a weak negative correlation between T_1 and T_4 as well as between T_2 and T_3, which are the comparisons that have figures with opposite signs in the same place of their respective contrast coeffient vectors.

4.3.3 Simulation Study

Similar to the simulation study in 4.1.3 and 4.2.3, we explore the asymptotic properties of our test procedure and seek to determine what sample sizes are required to keep the nominal type I error level. Then we address the problem which of the DF strategies (minimum, average, comparison-specific) works best with small sample sizes. Power analyses will focus on the question how much can be gained in comparison to performing separate sets of comparisons among treatments and occasions (like in 4.1 and 4.2) and adjusting with Bonferroni.

Type I error: Type I error simulations of the asymptotic procedures are carried out as described in 4.1.3, but with contrast coefficient matrices such that comparisons of treatment groups per time point and time points per treatment group are under joint control of $\alpha = 0.05$. Each setting is repeated 1000 times.

Figure 19 shows that the simulated type I error rates of the CIM-based asymptotic procedures range between 6 and 12% for $n_k = 10$ but stabilize around 5% as soon as n_k exceeds 30 or 40. The asymptotic MMM method can be substantially more liberal with type I error rates between 12 and 20% for $n_k = 10$, and requires around 50 subjects per group to achieve acceptable control of α.

For small samples of $n_k = \{4, 6, 8, 10, 12, 14\}$, we employ a joint CIM and the MMM approach, but the reference distribution is no longer multivariate normal. Instead we use the minimum, (weighted) average, or comparison-specific DFs as described in 4.3.1 for the multivariate t reference distribution.

The simulation results in Figures 20, and 21 show that, just as expected, the minimum DF usually yields the lowest and the average DF the highest type I error rates, and the separate DF approach lies inbetween. For the CIM with Kenward-Roger DFs (Figure 20), all three approaches are fairly similar; only when the number of time points is larger than the number of treatments, the minimum DF tends to make the tests a bit conservative. The discrepancy between the three DF approaches is larger for MMM (Figure 21). Here the average DF substantially inflates the type I error rates. The separate DF solution performs very well, and using the minimum DF leads to conservative decisions.

In summary, the approach with separate DFs proves to be the best choice for most cases. In situations with very small sample sizes where the type I error level is slightly exceeded, the minimum DF can actually improve α control, but otherwise it makes the procedure conservative. Averaging the DFs is obviously the poorest solution and likely to lead to type I error inflation. All these considerations become practically irrelevant with larger sample sizes.

Power: We simulate the asymptotic powers of longitudinal MCTs based on a CIM or MMM, either jointly or separately for comparisons of treatment groups and time points. Jointly means that all elementary hypotheses are part of one big family of comparisons that are performed under joint type I error control. Separately means that all elementary hypotheses of comparisons among treatment means are part of one family and all elementary hypotheses of comparisons among occasion means are part of another family, and each of them is tested separetely with control of α, followed by a Bonferroni adjustment of the resulting adjusted p-values (i.e., multiplication by two) to keep the overall type I error rate. This way we can assess the additional power benefit of incorporating the correlation between comparisons among treatments and comparisons among occasions. Settings with $\Delta = \{0, 0.1, \ldots, 1\}$ and $n_k = 100$ are simulated.

We see from Figure 22 that the large-sample power of the asymptotic duplex procedure does not depend on the underlying modeling strategy. The maximum power difference between joint and separate testing of contrast (the vertical distance between the dashed and solid line of the same color) is between 3 and 5 percentage points. The power of the asymptotic procedure increases with the longitudinal correlation of measurement occasions (Figure 23), but the power difference between the joint and separate procedure does not.

For the small-sample procedure simulated with $n_k = 10$ for $\Delta = \{0, 0.2, \ldots, 2\}$, the power of the CIM-based procedure with the Kenward-Roger method is discernibly greater than with MMM (Figures 24 and Figure 25). We also observe that the power advantage of using the joint test compared to separate tests for both types of comparisons is now larger than with the asymptotic procedure.

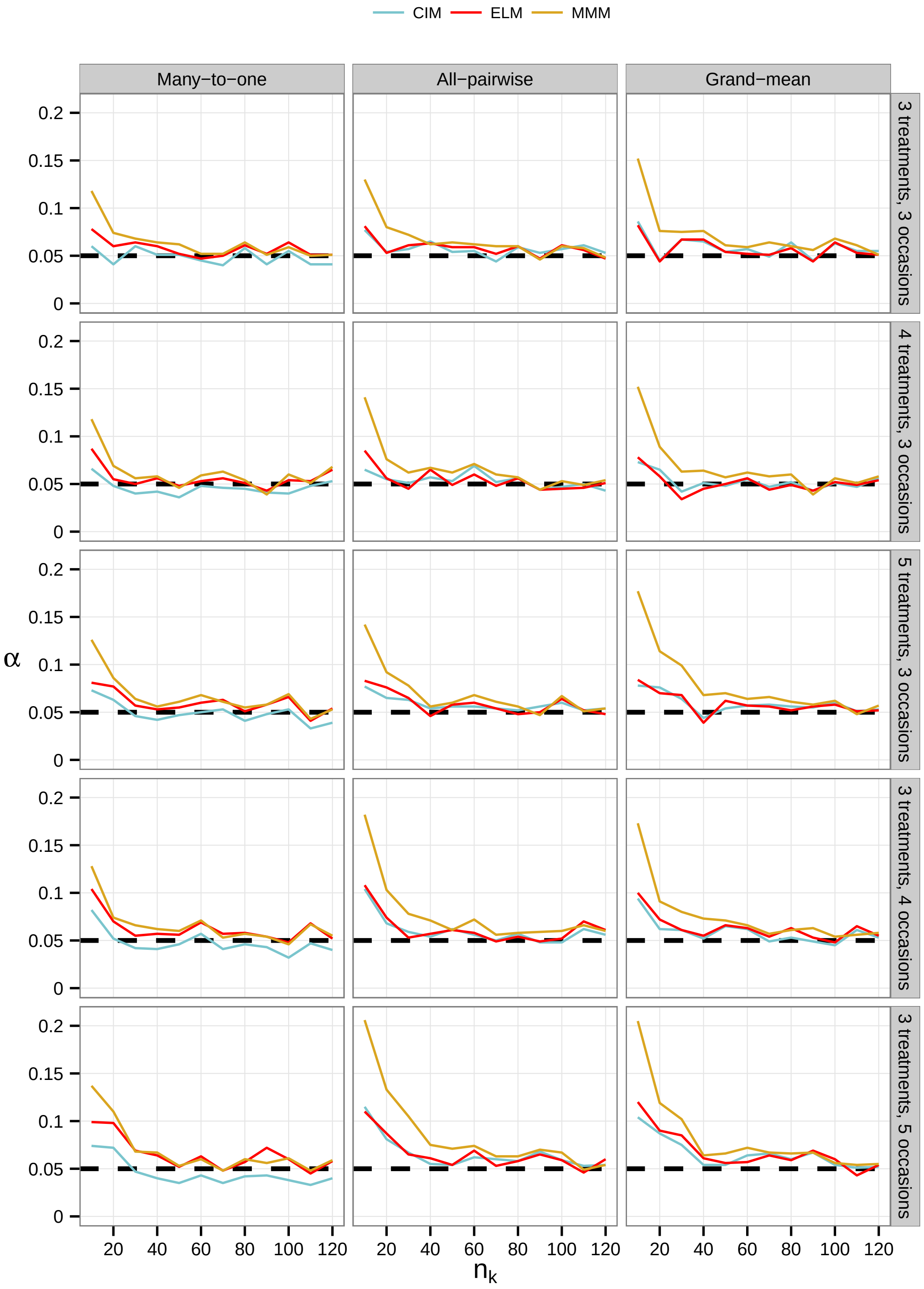


Figure 19: Simulated type I error rates for asymptotic many-to-one, all-pairwise, and grand-mean comparisons among $q = \{3, 4, 5\}$ Gaussian treatment means separately and simultaneously at $m = \{3, 4, 5\}$ occasions and among $m = \{3, 4, 5\}$ Gaussian occasion means separately and simultaneously for $q = \{3, 4, 5\}$ treatments, with n_k independent subjects per treatment group (1000 simulation runs).

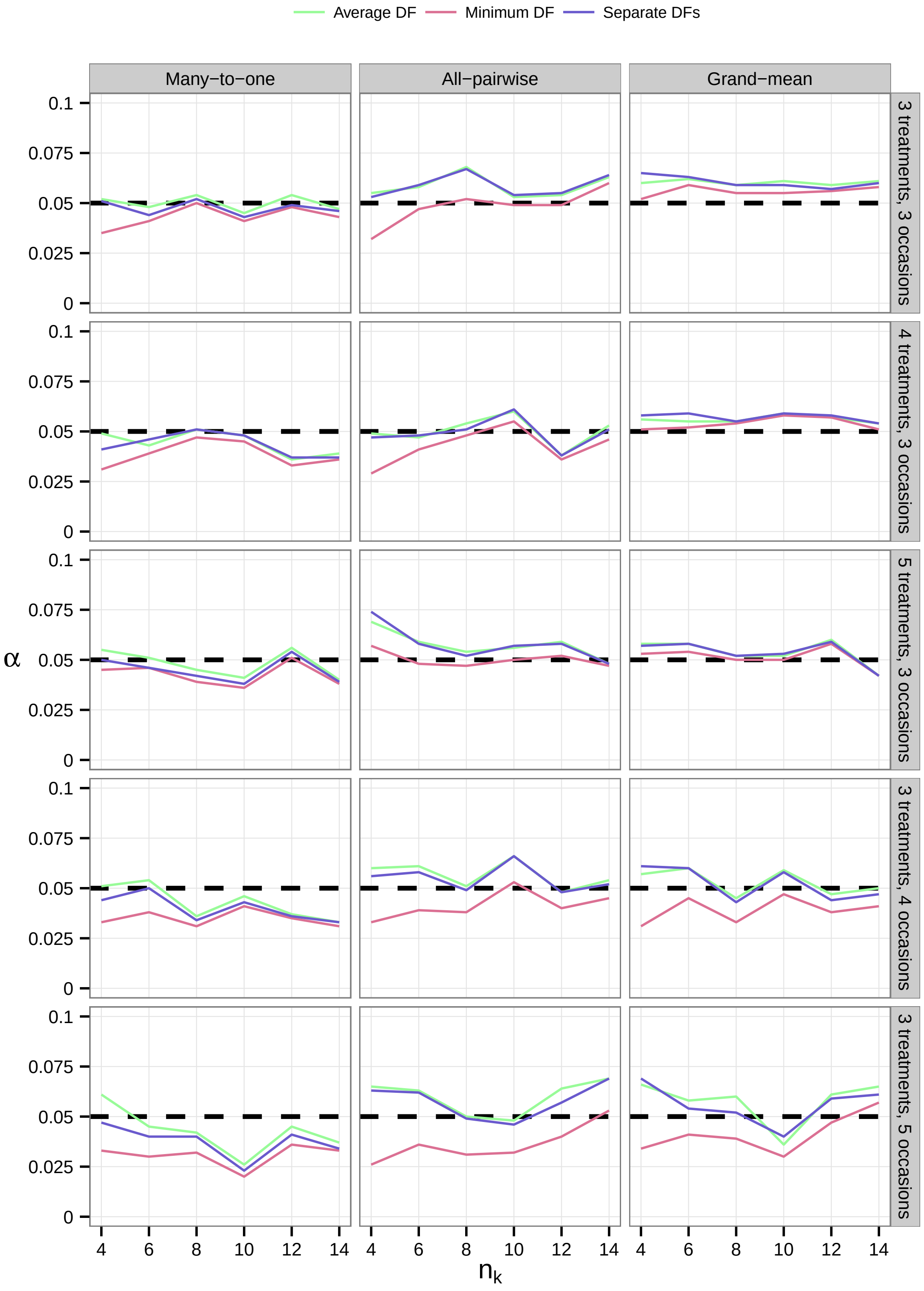


Figure 20: Simulated type I error rates for many-to-one, all-pairwise, and grand-mean comparisons among $q = \{3, 4, 5\}$ Gaussian treatment means separately and simultaneously at $m = \{3, 4, 5\}$ occasions and among $m = \{3, 4, 5\}$ Gaussian occasion means separately and simultaneously for $q = \{3, 4, 5\}$ treatments, with n_k independent subjects per treatment group, based on a conditional independence model with average, minimum, and separate Kenward-Roger DFs (1000 simulation runs).

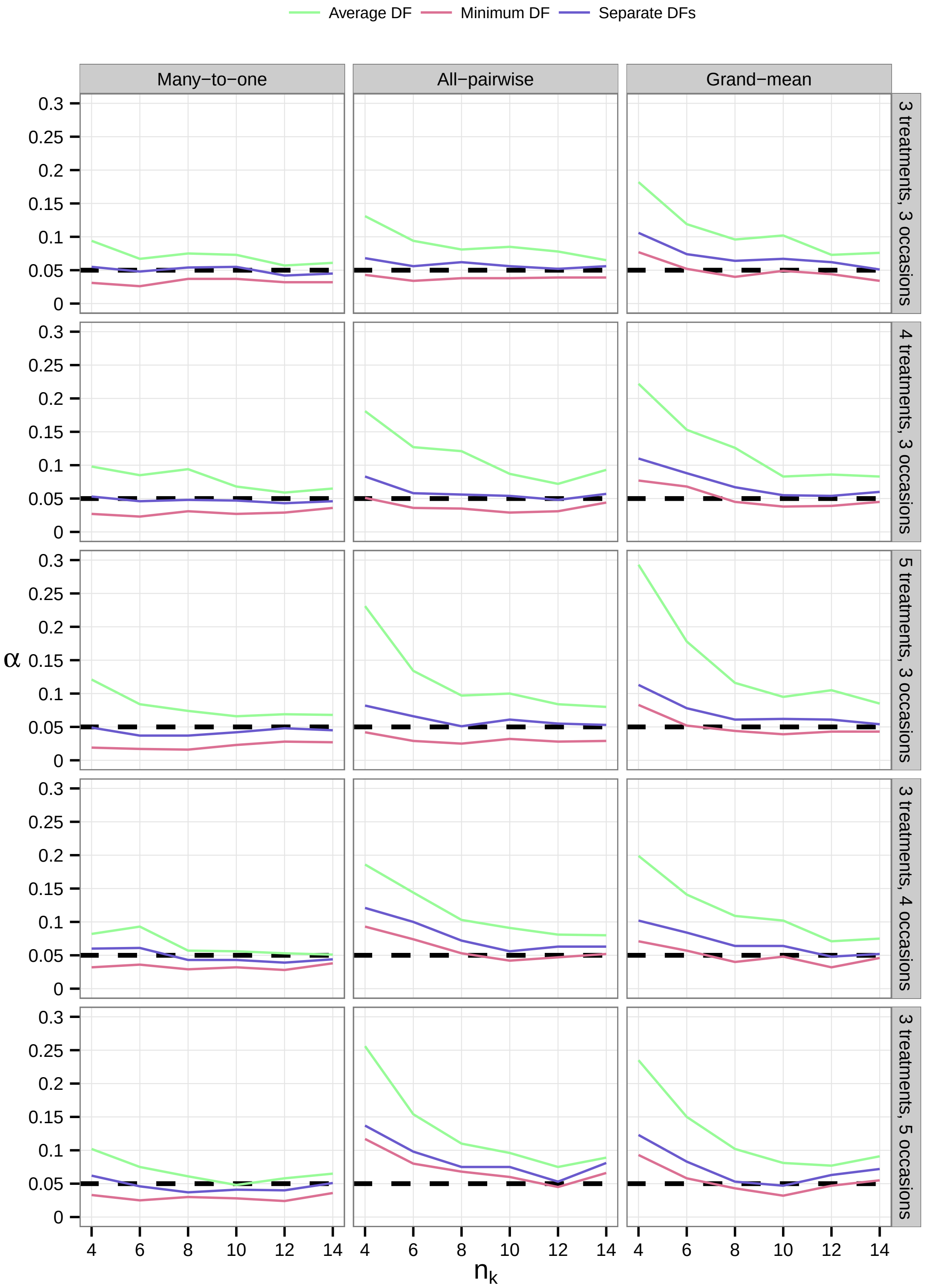


Figure 21: Simulated type I error rates for many-to-one, all-pairwise, and grand-mean comparisons among $q = \{3, 4, 5\}$ Gaussian treatment means separately and simultaneously at $m = \{3, 4, 5\}$ occasions and among $m = \{3, 4, 5\}$ Gaussian occasion means separately and simultaneously for $q = \{3, 4, 5\}$ treatments, with n_k independent subjects per treatment group, based on multiple marginal models with average, minimum, and separate DFs (1000 simulation runs).

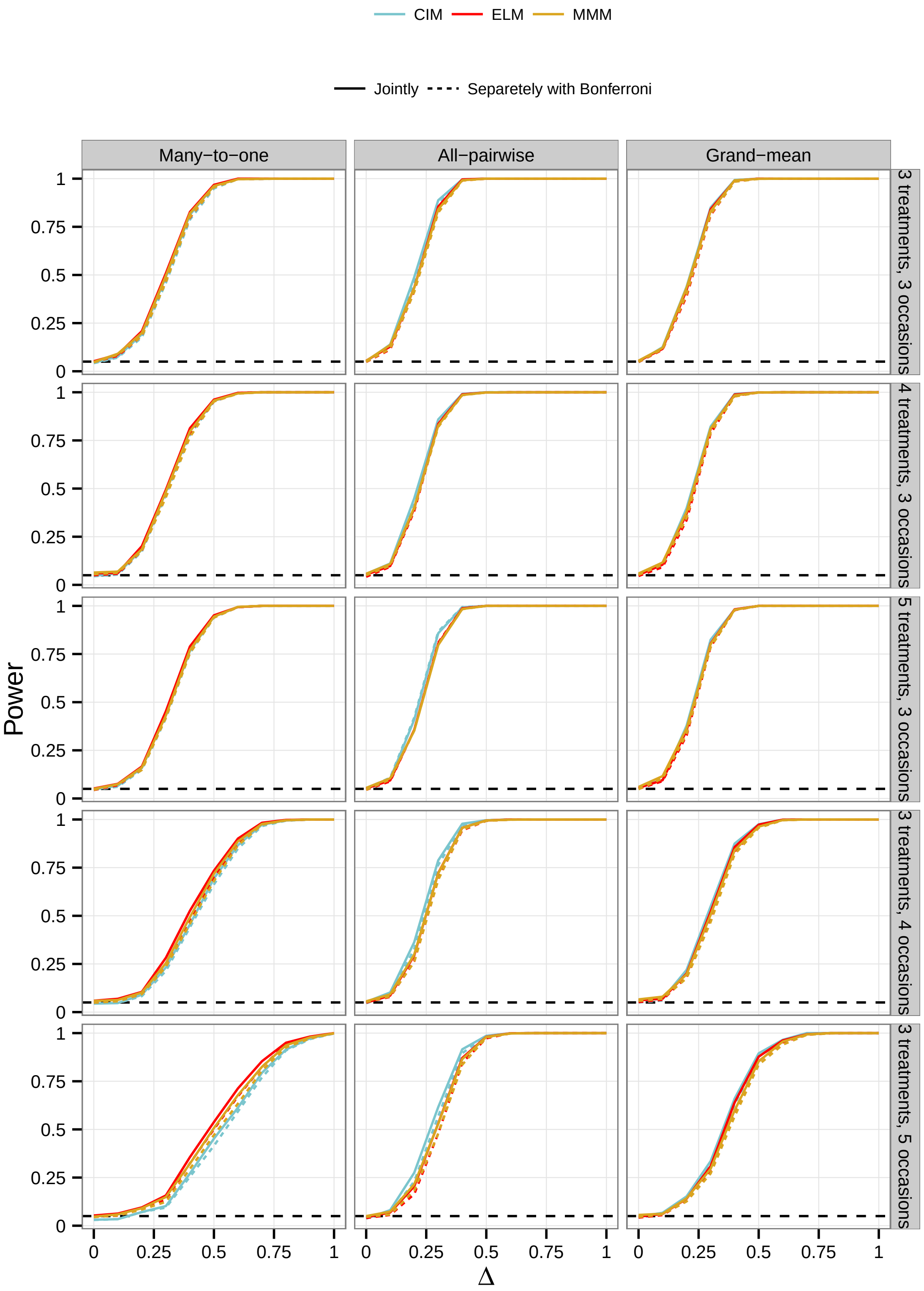


Figure 22: Simulated powers for asymptotic many-to-one, all-pairwise, and grand-mean comparisons among $q = \{3, 4, 5\}$ Gaussian treatment means separately and simultaneously at $m = \{3, 4, 5\}$ occasions and among $m = \{3, 4, 5\}$ Gaussian occasion means separately and simultaneously for $q = \{3, 4, 5\}$ treatments, with $n_k = 100$ independent subjects per treatment group (1000 simulation runs).

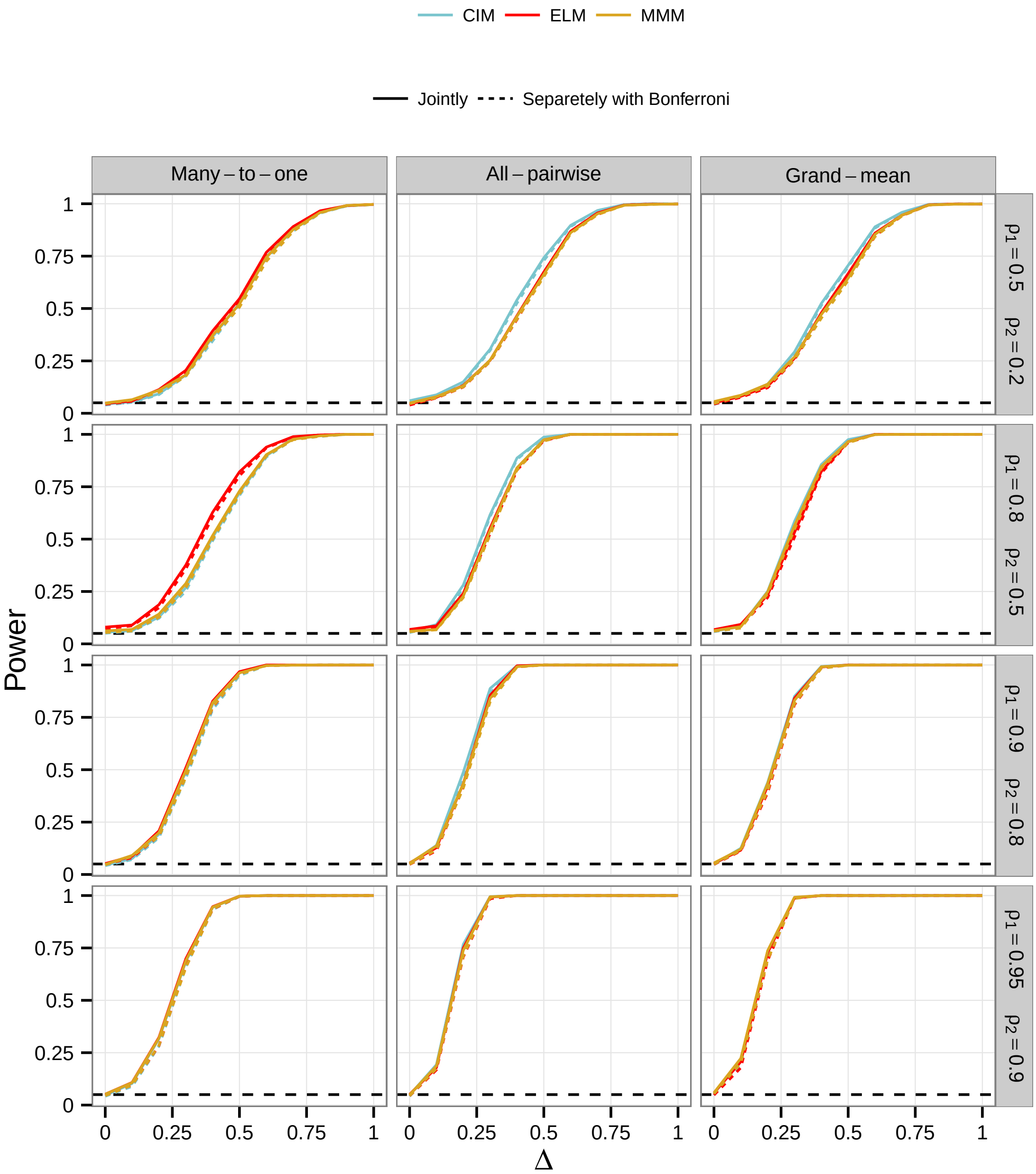


Figure 23: Simulated powers for asymptotic many-to-one, all-pairwise, and grand-mean comparisons among $q = 3$ Gaussian treatment means separately and simultaneously at $m = 3$ occasions and among $m = 3$ Gaussian occasion means separately and simultaneously for $q = 3$ treatments, with $n_k = 100$ independent subjects per treatment group and different longitudinal correlations (1000 simulation runs).

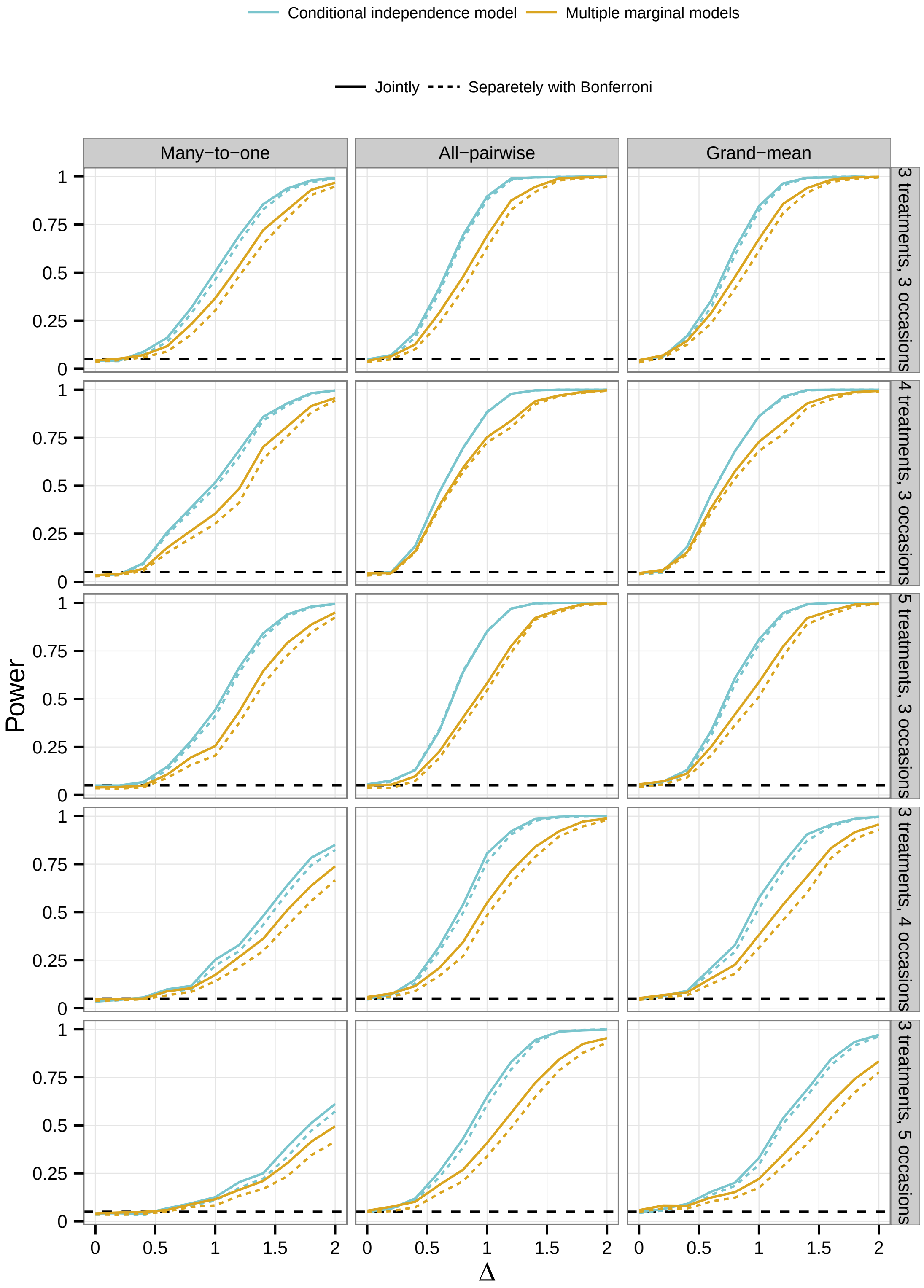


Figure 24: Simulated powers for many-to-one, all-pairwise, and grand-mean comparisons among $q = \{3, 4, 5\}$ Gaussian treatment means separately and simultaneously at $m = \{3, 4, 5\}$ occasions and among $m = \{3, 4, 5\}$ Gaussian occasion means separately and simultaneously for $q = \{3, 4, 5\}$ treatments, with $n_k = 10$ independent subjects per treatment group (1000 simulation runs).

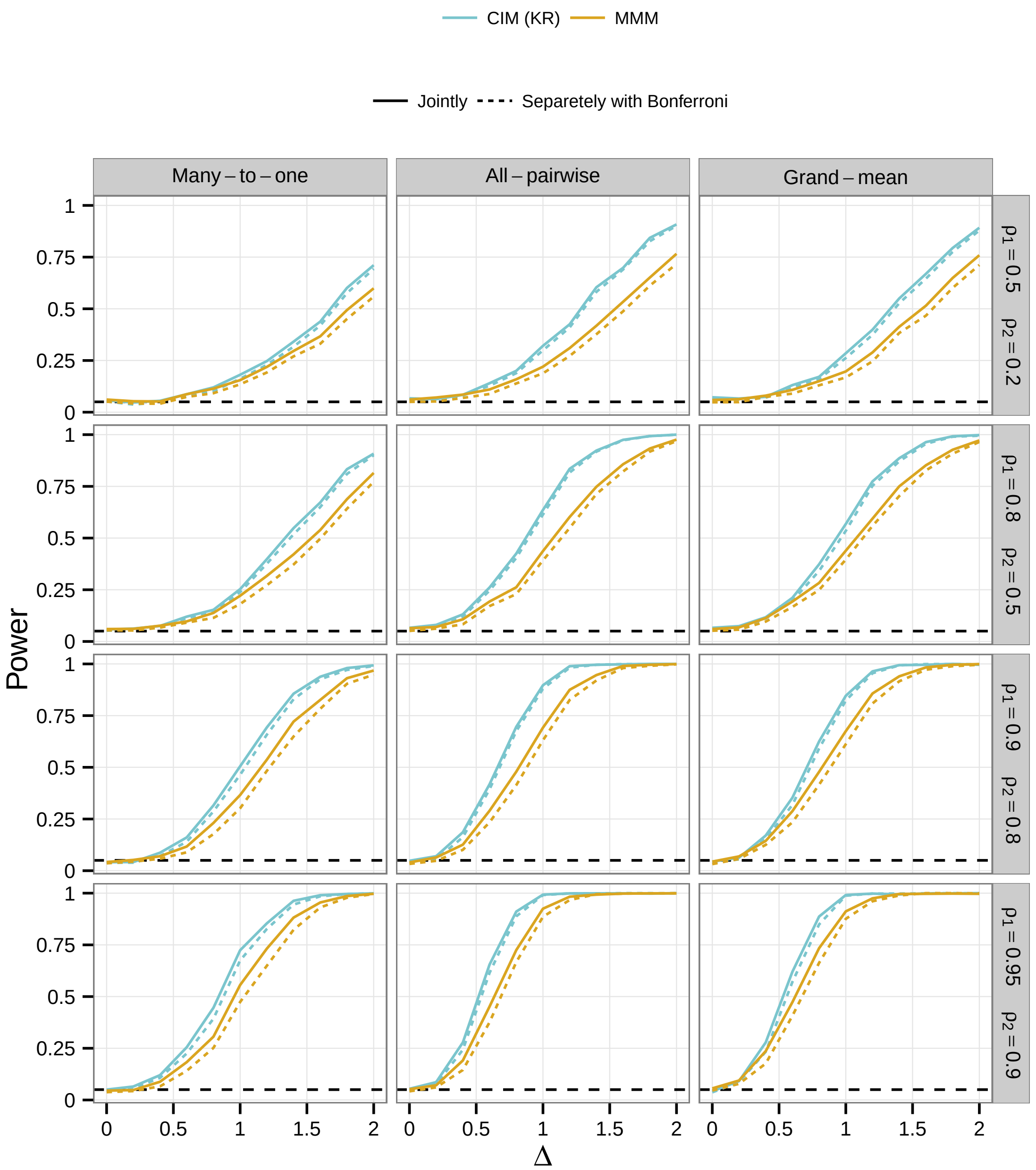


Figure 25: Simulated powers for many-to-one, all-pairwise, and grand-mean comparisons among $q = 3$ Gaussian treatment means separately and simultaneously at $m = 3$ occasions and among $m = 3$ Gaussian occasion means separately and simultaneously for $q = 3$ treatments, with $n_k = 10$ independent subjects per treatment group and different longitudinal correlations (1000 simulation runs).

4.4 Application to Example Data

We illustrate the use of our longitudinal MCTs for Gaussian outcomes with the dataset on bradykinin receptor antagonism, mercuric chloride, and heart rates that we introduced in 2.1, 2.2 and 2.3.

4.4.1 Bradykinin Receptor Antagonism

The data on bradykinin receptor antagonism were presented in 2.1 and will now be evaluated with longitudinal MCTs based on either joint or multiple marginal models. A similar analysis of this dataset appeared in Pallmann et al. (2015).

Comparing multiple drugs simultaneously at multiple occasions: One relevant research question 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 non-baseline occasions. This we can achieve with longitudinal MCTs using estimates from an AICc-selected ELM or 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., multiplication by four).

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 UN correlation and heteroscedasticity between occasions and treatments; all other candidate models are substantially inferior (ΔAICc > 25).

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 UN 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 for many-to-one comparisons of treatments and occasions (Figures 26 and 27). Notice that the most complex random-effects structure would correspond to UN 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 4. We have seen in the simulation studies in 4.1.3, 4.2.3, and 4.3.3 that sample sizes of $n_k > 30$ are sufficient for the asymptotic procedure to control α; therefore we begin with an asymptotic analysis of the bradykinin data.

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

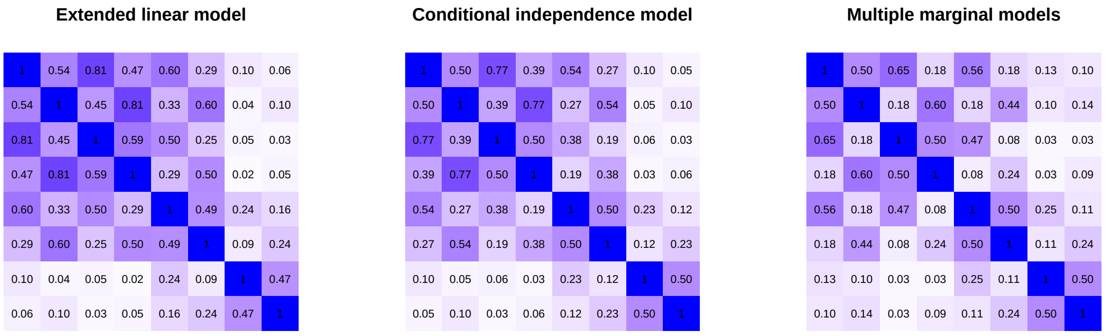


Figure 26: Bradykinin data: correlation matrices of test statistics for many-to-one comparisons of treatments per time point (except baseline).

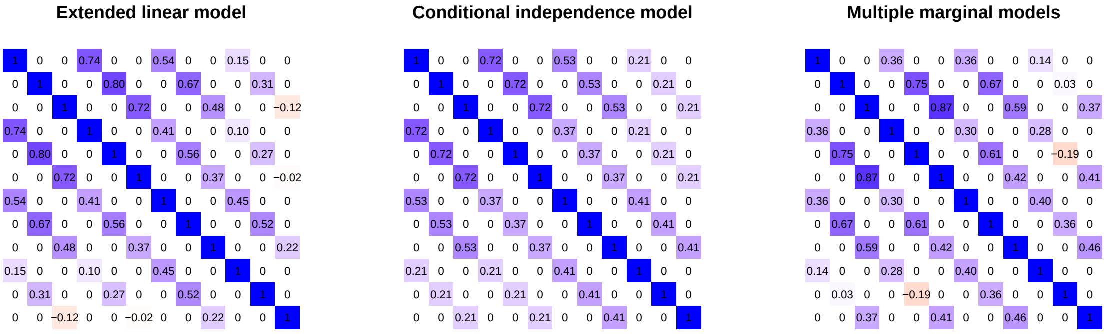


Figure 27: Bradykinin data: correlation matrices of test statistics for many-to-one comparisons of time points per treatment group.

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, although not statistically significant. 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 SEs 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 estimates SEs 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 increased p-values, indicating that there is 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 SCIs displayed in Figure 28. Intervals obtained from CIM or MMM analyses are

Table 1: Simultaneous inference for the bradykinin data: estimated differences of D-dimer log-concentrations, standard errors, and adjusted p-values for occasion-wise many-to-one comparisons of HOE 140 and EACA against placebo.

	Estimate	SE(ELM)	SE(CIM, MMM)	p(ELM)	p(CIM)	p(MMM)	p(Bonf)
30 min on-pump: EACA - Placebo	-0.246	0.170	0.176	0.607	0.656	0.672	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	0.010
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	0.439	0.369	0.382	0.511
Postoperative day 1: HOE140 - Placebo	-0.290	0.127	0.143	0.133	0.240	0.250	0.310

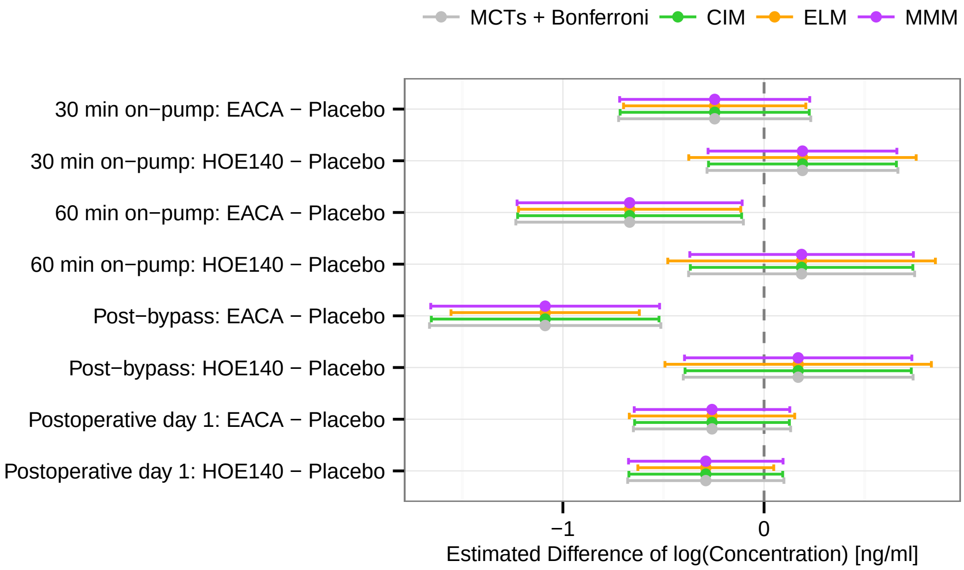


Figure 28: Bradykinin data: 95% simultaneous confidence intervals for asymptotic many-to-one comparisons of treatments per time point (except baseline) based on AICc-selected joint models (CIM and ELM), combined marginal models (MMM), and Bonferroni-adjusted MCTs per time point.

about the same width whereas those from ELM analysis can be a little wider or narrower due to their using comparison-specific SE estimates. The Bonferoni-basd SCIs are always a little wider than those of the CIM- and MMM-based longitudinal MCTs that both use the same SE estimates.

Table 2: Simultaneous inference for the bradykinin data: DFs for the many-to-one comparisons of treatments at all occasions except baseline.

	Naive	Residual	ESS	“Adjusted”	Pinheiro-Bates	Kenward-Roger	MMM
DF	97	545	171	99	107	109	109

So far we have been relying on asymptotics because we felt the sample sizes per group (37 for placebo and EACA, 38 for HOE 140) were large enough to do so. To assess whether this was justified, we compare the results of the asymptotic evaluation with some “small-sample” analyses. Table 2 shows different DF approximations for the bradykinin data: we see that most of them are not much larger than $\nu_{naive} = 97$, only $\nu_{ESS} = 171$ (based on an estimated $\hat{\rho} = 0.71$ in the AR(1) pattern) is considerably different, and of course

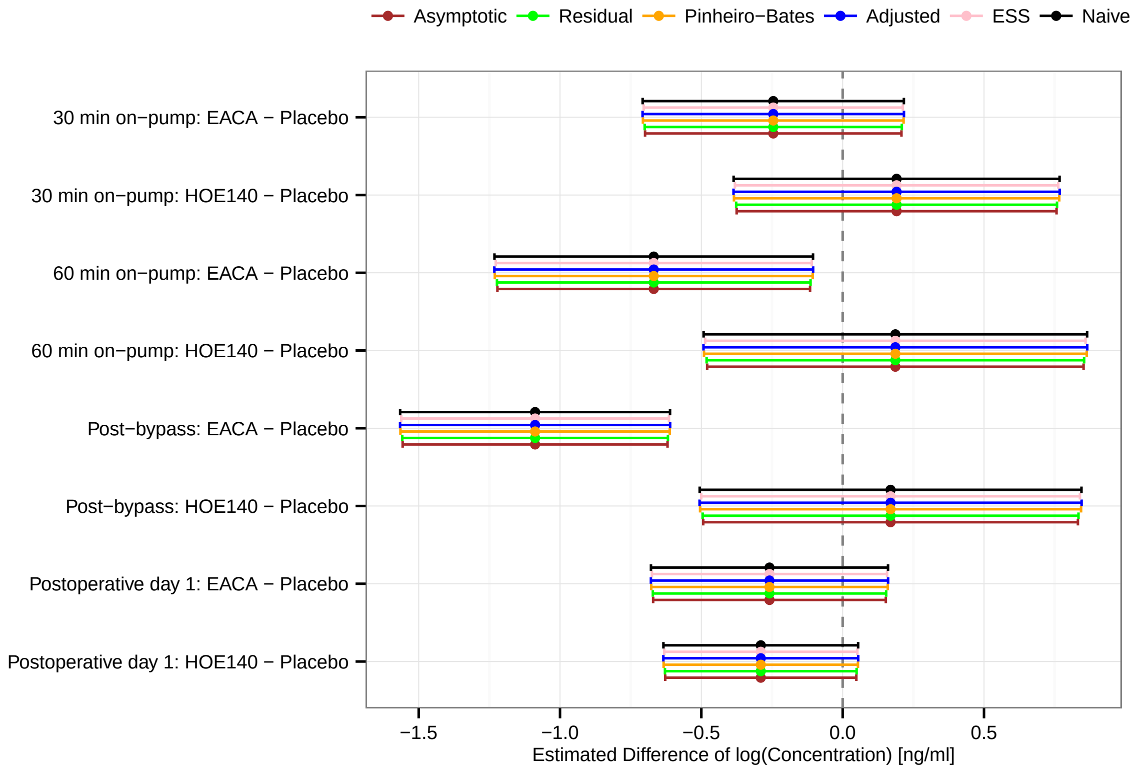


Figure 29: Bradykinin data: 95% simultaneous confidence intervals for many-to-one comparisons of treatments per time point (except baseline) based on the AICc-selected joint ELM, with different approximations to the degrees of freedom, and asymptotic results (for comparison).

$\nu_{res} = 545$. MMM refers to the smallest residual DF of the occasion-specific marginal linear models, which happens to be identical to the Kenward-Roger DF here.

These differences between DFs have a negligible impact on the results of the analysis. For example, the 95% equicoordinate quantile of the multivariate t-distribution with $\widehat{\mathbf{\Sigma}}$ estimated from the AICc-selected ELM that is used for constructing SCIs is 7.092 with 97 DF and 7.041 with 109 DF. We can view the practical impact of different DF approximation on the width of the ELM-based SCIs in Figure 29. Using the asymptotic reference distribution (multivariate normal) or a multivariate t-distribution with ν_{res} makes the intervals slightly narrower, but this small difference is most likely irrelevant in practice. Therefore the asymptotic analysis seems justified for this dataset.

Comparing multiple occasions simultaneously for multiple drugs: 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.

Again we use the AICc-selected joint models as well as the MMM approach and perform an asymptotic analysis, and again there is considerable discrepancy between pooled-treatment and treatment-specific SE estimates, and they bring about p-values that differ quite a bit between the modeling strategies (Table 3). The widths of the corresponding

95% SCIs shown in Figure 30 also vary according to the SE 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.

Table 3: Simultaneous inference for the bradykinin data: estimated differences of D-dimer log-concentrations, standard errors, and adjusted p-values for treatment-wise many-to-one comparisons of occasions against baseline.

	Estimate	SE(ELM)	SE(CIM)	SE(MMM)	SE(Bonf)	p(ELM)	p(CIM)	p(MMM)	p(Bonf)
Placebo: 30 min on-pump - Baseline	0.361	0.100	0.098	0.094	0.134	0.004	0.003	0.001	0.075
Placebo: 60 min on-pump - Baseline	0.886	0.149	0.138	0.161	0.134	<0.001	<0.001	<0.001	<0.001
Placebo: Post-bypass - Baseline	1.511	0.149	0.155	0.164	0.134	<0.001	<0.001	<0.001	<0.001
Placebo: Postoperative day 1 - Baseline	1.354	0.143	0.139	0.124	0.134	<0.001	<0.001	<0.001	<0.001
EACA: 30 min on-pump - Baseline	0.210	0.081	0.098	0.059	0.126	0.098	0.289	0.004	0.838
EACA: 60 min on-pump - Baseline	0.312	0.107	0.138	0.081	0.126	0.039	0.220	0.001	0.140
EACA: Post-bypass - Baseline	0.517	0.117	0.155	0.128	0.126	<0.001	0.010	0.001	0.001
EACA: Postoperative day 1 - Baseline	1.189	0.151	0.139	0.146	0.126	<0.001	<0.001	<0.001	<0.001
HOE140: 30 min on-pump - Baseline	0.615	0.113	0.097	0.129	0.163	<0.001	<0.001	<0.001	0.002
HOE140: 60 min on-pump - Baseline	1.135	0.150	0.136	0.161	0.163	<0.001	<0.001	<0.001	<0.001
HOE140: Post-bypass - Baseline	1.744	0.186	0.153	0.159	0.163	<0.001	<0.001	<0.001	<0.001
HOE140: Postoperative day 1 - Baseline	1.127	0.118	0.137	0.141	0.163	<0.001	<0.001	<0.001	<0.001

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.

Comparing multiple drugs and multiple occasions simultaneously: We can also combine all hypotheses into one family so that the set of elementary comparisons now contains both comparisons among treatments and among occasions, and we strive after protecting a joint FWER of 5%. We see from Table 4 that the p-values increase but not by much, which means the correlation between contrasts is exploited. All of the hypotheses that could be rejected in the previous analysis can still be rejected now, as we see also from the SCIs in Figure 31.

4.4.2 Mercuric Chloride

Data from a study investigating toxic effects of mercuric chloride were presented in 2.2 and will now be analyzed using longitudinal MCTs. Here we will see the effect of missing values on the effect sizes estimated with our different modeling approaches.

Comparing multiple doses simultaneously at multiple occasions: The research question with the mercuric chloride data is: by how much do the treatment doses (2.5 and 5 mg/kg $HgCl_2$) differ from control at our three measurement time points of interest? Hence the main objective of our analysis shall be to construct one-sided 95% upper SCI bounds for the differences of low and high dose versus control separately and simultaneously for all three measurement occasions.

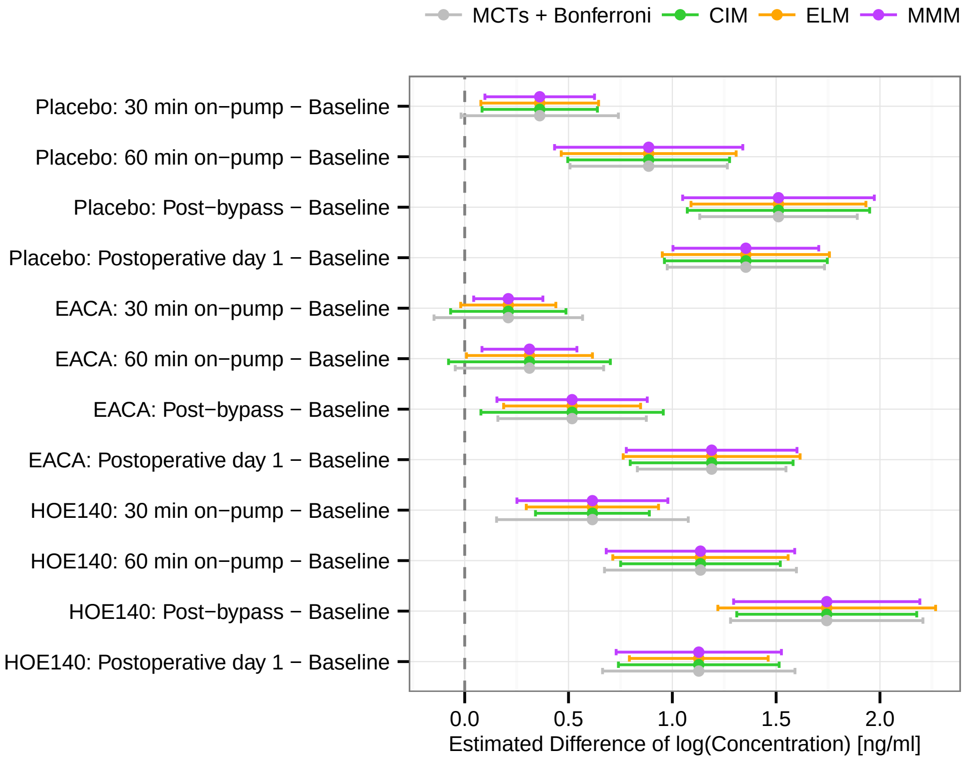


Figure 30: Bradykinin data: 95% simultaneous confidence intervals for asymptotic many-to-one comparisons of time points per treatment arm based on AICc-selected joint models (CIM and ELM), combined marginal models (MMM), and Bonferroni-adjusted MCTs per treatment arm.

This is again a fairly large dataset with sample sizes of 60 for control and 56 for each of the active dose groups at the first measurement occasion. However, the sample sizes are substantially smaller at the end of the study due to animals being sacrificed for interim measurements of body and organ weights.

We model the data using either a joint CIM or ELM or occasion-specific marginal models. Various different CIMs are conceivable, and it is not clear from the start which random-effects structure suits best. Animal-specific intercepts seem indispensable, but do animal-specific slopes bring about an improvement? Moreover, several animals were housed together in cages, so we could add cage-specific intercepts, and possibly also slopes? And can we assume that random intercepts and slopes are uncorrelated?

Our indecision calls for the assistance of a selection criterion. We build several LMMs with body weight as dependent variable, the interactions of dose levels and occasions as independent factor, and different random-effects structures as described above. AICc selection indicates that the best LMM fit is achieved with animal-specific intercepts and slopes but without random cage effects. Other reasonable models include, in addition to the random animal effects, cage-specific random intercepts and slopes (ΔAICc $= 0.36$), cage-specific random intercepts only (ΔAICc $= 1.95$), or cage-specific random intercepts and slopes that are uncorrelated (ΔAICc $= 2.12$). On the contrary, models without animal-specific slopes are out of the question with ΔAICc > 100. So we choose as our final model to be used for simultaneous inference the one including random animal effects

Table 4: Simultaneous inference for the bradykinin data: estimated differences of D-dimer log-concentrations, standard errors, and adjusted p-values for occasion-wise many-to-one comparisons of HOE 140 and EACA against placebo and treatment-wise many-to-one comparisons of occasions against baseline.

	Estimate	SE(ELM)	SE(CIM)	SE(MMM)	p(ELM)	p(CIM)	p(MMM)
30 min on-pump: EACA - Placebo	-0.246	0.170	0.176	0.176	0.878	0.908	0.905
30 min on-pump: HOE140 - Placebo	0.191	0.213	0.174	0.174	0.997	0.985	0.984
60 min on-pump: EACA - Placebo	-0.668	0.208	0.208	0.208	0.022	0.023	0.023
60 min on-pump: HOE140 - Placebo	0.186	0.250	0.207	0.207	1.000	0.997	0.997
Post-bypass: EACA - Placebo	-1.088	0.176	0.211	0.211	<0.001	<0.001	<0.001
Post-bypass: HOE140 - Placebo	0.169	0.249	0.210	0.210	1.000	0.999	0.999
Postoperative day 1: EACA - Placebo	-0.259	0.155	0.144	0.144	0.734	0.654	0.651
Postoperative day 1: HOE140 - Placebo	-0.290	0.127	0.143	0.143	0.289	0.475	0.473
Placebo: 30 min on-pump - Baseline	0.361	0.100	0.098	0.094	0.006	0.004	0.002
Placebo: 60 min on-pump - Baseline	0.210	0.081	0.098	0.059	0.143	0.395	0.007
Placebo: Post-bypass - Baseline	0.615	0.113	0.097	0.129	<0.001	<0.001	<0.001
Placebo: Postoperative day 1 - Baseline	0.886	0.149	0.138	0.161	<0.001	<0.001	<0.001
EACA: 30 min on-pump - Baseline	0.312	0.107	0.138	0.081	0.061	0.310	0.002
EACA: 60 min on-pump - Baseline	1.135	0.150	0.136	0.161	<0.001	<0.001	<0.001
EACA: Post-bypass - Baseline	1.511	0.149	0.155	0.164	<0.001	<0.001	<0.001
EACA: Postoperative day 1 - Baseline	0.517	0.117	0.155	0.128	<0.001	0.015	0.001
HOE140: 30 min on-pump - Baseline	1.744	0.186	0.153	0.159	<0.001	<0.001	<0.001
HOE140: 60 min on-pump - Baseline	1.354	0.143	0.139	0.124	<0.001	<0.001	<0.001
HOE140: Post-bypass - Baseline	1.189	0.151	0.139	0.146	<0.001	<0.001	<0.001
HOE140: Postoperative day 1 - Baseline	1.127	0.118	0.137	0.141	<0.001	<0.001	<0.001

on both intercepts and slopes.

The choice of a joint ELM is not obvious either, therefore we select one from a candidate set as in the bradykinin example (twelve models) using AICc again. The best-fitting ELM includes occasion-specific variances and an UN residual correlation matrix. Modeling variance separately for occasions and doses is considerably inferior (ΔAICc $= 1.75$), and all other models are very poor with ΔAICc > 60.

Compared to these considerations on LMM specification, defining the marginal linear models for each of the three occasions is child's play: the dependent variable is body weight, and dose is the independent factor. If we wanted to include random cage effects, combining marginal LMMs as in Jensen et al. (2015) would be the method of choice.

The correlation matrices of test statistics are very similar for ELM-, CIM-, and MMM-based comparisons (Figure 32). However, unlike in the bradykinin example, now the point estimates differ between CIM-, ELM-, and MMM-based analyses (Table 5); this is the case for all comparisons that involve unequal numbers of recorded observations due to dropout. We recall that the MMM estimates the effects of interest as "raw" averages whereas the LMMs borrow information from adjoining time points via the intra-animal correlation. The estimates of this correlation differ between the ELM and CIM, and in consequence the point estimates differ, too.

One crucial issue in this context is the assumed missing data mechanism, which can be MCAR or MAR for the LMM but must be MCAR for MMM. The animals to be sacrificed for interim analyses were picked at random, so their missingness should be unrelated to both measured and unmeasured outcomes. The ones found moribund or dead, however, might easily have had lower body weights i.e., their dropping out could be causally determined by previously measured values, so MCAR is certainly debatable for these animals, not least because all non-completers within the first year come from the $HgCl_2$ groups and none from the control group.

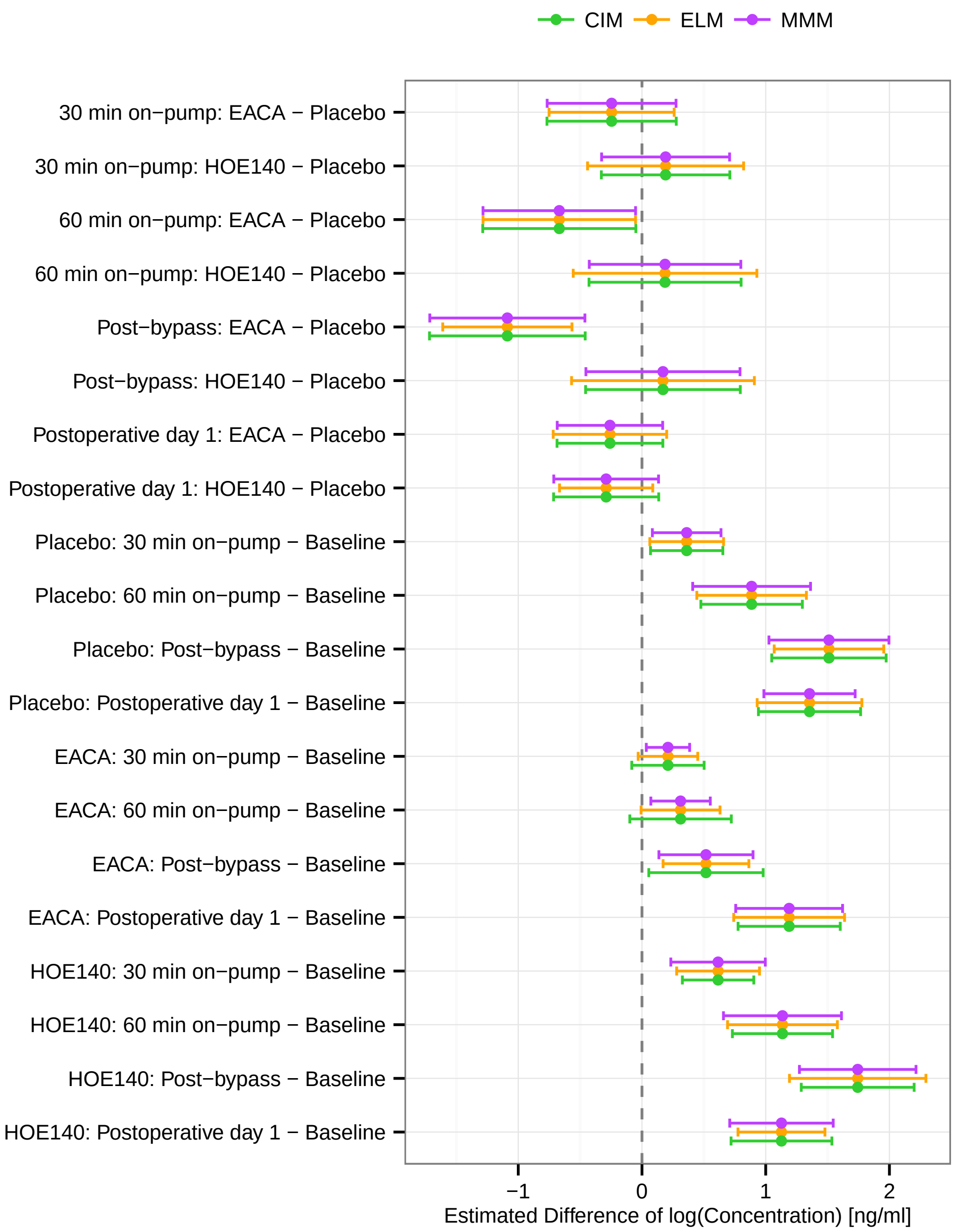
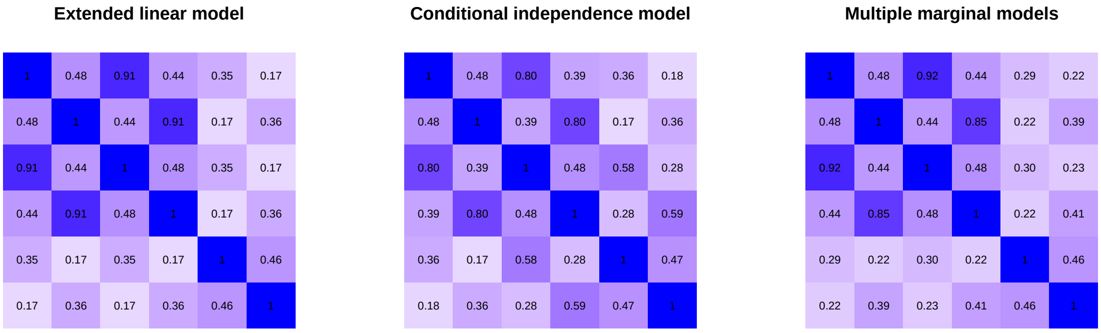


Figure 31: Bradykinin data: 95% simultaneous confidence intervals for asymptotic many-to-one comparisons of treatments per time point (except baseline) and time points per treatment arm simultaneously based on AICc-selected joint models (CIM and ELM) and combined marginal models (MMM).

We see from the 95% SCIs in Figure 33 that the body weight reduction in animals treated with $HgCl_2$ in comparison to control animals is significant for both dose groups at all measurement time points. Whereas the upper SCI bounds are fairly close to the point estimates for week 53 and for the interim analysis (week 65), the uncertainty is sizeably greater at the end of the study, and the bounds for the low-dose group (2.5 mg/kg) are fairly close to zero so that the effect may, although statistically significant, no longer be biologically relevant. The SEs are now much larger than before because of the reduced sample sizes.



Extended linear model

1	0.48	0.91	0.44	0.35	0.17
0.48	1	0.44	0.91	0.17	0.36
0.91	0.44	1	0.48	0.35	0.17
0.44	0.91	0.48	1	0.17	0.36
0.35	0.17	0.35	0.17	1	0.46
0.17	0.36	0.17	0.36	0.46	1

Conditional independence model

1	0.48	0.80	0.39	0.36	0.18
0.48	1	0.39	0.80	0.17	0.36
0.80	0.39	1	0.48	0.58	0.28
0.39	0.80	0.48	1	0.28	0.59
0.36	0.17	0.58	0.28	1	0.47
0.18	0.36	0.28	0.59	0.47	1

Multiple marginal models

1	0.48	0.92	0.44	0.29	0.22
0.48	1	0.44	0.85	0.22	0.39
0.92	0.44	1	0.48	0.30	0.23
0.44	0.85	0.48	1	0.22	0.41
0.29	0.22	0.30	0.22	1	0.46
0.22	0.39	0.23	0.41	0.46	1

Figure 32: Mercuric chloride data: correlation matrices of test statistics for many-to-one comparisons of treatment groups per time point.

Table 5: Simultaneous inference for the mercuric chloride data: estimated differences of body weights, standard errors, and adjusted p-values for occasion-wise many-to-one comparisons of 2.5 and 5 mg/kg mercuric chloride against control.

	Est(ELM)	Est(CIM)	Est (MMM)	SE(ELM)	SE(CIM)	SE(MMM)	p(ELM)	p(CIM)	p(MMM)	p(Bonf)
Week 53: 2.5 mg/kg - control	-22.338	-22.338	-22.338	3.225	3.466	3.225	<0.001	<0.001	<0.001	<0.001
Week 53: 5 mg/kg - control	-29.251	-29.251	-29.251	3.225	3.466	3.225	<0.001	<0.001	<0.001	<0.001
Week 65: 2.5 mg/kg - control	-30.491	-30.432	-30.206	3.774	3.646	3.705	<0.001	<0.001	<0.001	<0.001
Week 65: 5 mg/kg - control	-35.881	-36.098	-36.938	3.774	3.646	3.705	<0.001	<0.001	<0.001	<0.001
Week 105: 2.5 mg/kg - control	-24.734	-23.913	-26.367	8.685	7.675	9.141	0.011	0.005	0.009	0.023
Week 105: 5 mg/kg - control	-35.920	-32.687	-38.600	8.535	7.580	8.959	<0.001	<0.001	<0.001	<0.001

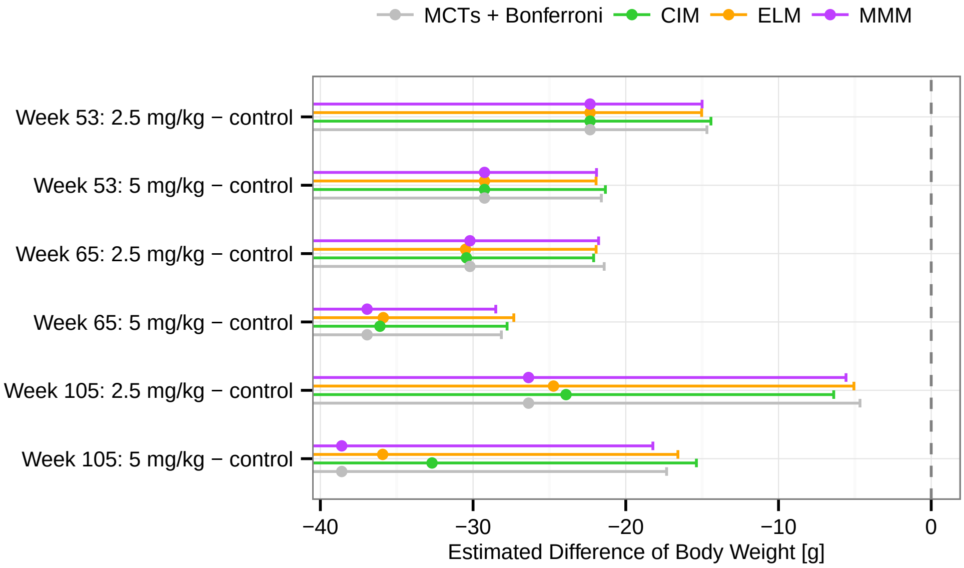


Figure 33: Mercuric chloride data: 95% simultaneous confidence intervals for asymptotic many-to-one comparisons of treatments per time point based on AICc-selected joint models (CIM and ELM), combined marginal models (MMM), and Bonferroni-adjusted MCTs per time point.

So far we have relied on asymptotics because we felt the sample sizes were sufficiently large. We now take the asymptotic analysis based on the joint CIM and compare it to one using the Kenward-Roger adjustment. As is to be expected in the presence of substantial dropout and heteroscedasticity, the Kenward-Roger DFs differ considerably between comparisons (Table 6). This raises the question which DF to use for the longitudinal MCTs: straightforward choices are the mean DF (171.1, rounded down to 171) or the minimum DF (141.4, rounded down to 141).

Table 6: Simultaneous inference for the mercuric chloride data: Kenward-Roger DFs for the many-to-one comparisons of treatments at all three occasions.

	Week 53	Week 65	Week 105
2.5 mg/kg vs. control	181.8	189	143.7
5 mg/kg vs. control	181.8	189	141.4

Table 7 shows the impact of the Kenward-Roger adjustment on the variance estimates. They are identical to the unadjusted ones for all groups in week 53 and the control group in week 65 (those without missing values). The inflation for the active dose groups in week 65 is minimal because there is only one data point missing in each group. The substantial amount of missingness at the last time points is reflected by variance inflation that is still small but noticeable.

Table 7: Simultaneous inference for the mercuric chloride data: unadjusted and Kenward-Roger-adjusted variance estimates.

		Unadjusted	Kenward-Roger
Week 53	Control	5.7983	5.7983
	2.5 mg/kg	6.2125	6.2125
	5 mg/kg	6.2125	6.2125
Week 65	Control	6.3947	6.3947
	2.5 mg/kg	6.8955	6.8959
	5 mg/kg	6.8955	6.8959
Week 105	Control	27.1003	27.3544
	2.5 mg/kg	31.8122	32.1706
	5 mg/kg	30.3539	30.6645

From the resulting SCIs in Figure 34 we see that there is practically no difference between the average and minimum DF here, and both make the SCIs just marginally wider than the asymptotic ones. The difference in width between asymptotic and finite-sample approaches is slightly bigger for week 105, but this is mainly due to the inflated covariance matrix.

4.4.3 Heart Rates

We evaluate the heart rate data introduced in 2.3 using longitudinal MCTs. Here we have truly small sample sizes (eight per treatment arm) and certainly cannot rely on asymptotics as with the previous examples in 4.4.1 and 4.4.2.

Comparing multiple drugs simultaneously at multiple occasions: The first question we want to answer is: when do which active treatments differ significantly from control, and by how much? To this end we perform Dunnett comparisons separately and simultaneously for each of the four time points, based on either a joint ELM, a joint CIM, or marginal models.

Once again we employ AICc model selection and use the same structures for heteroscedasticity, residual correlation, and random effects for the candidate models as in the previous

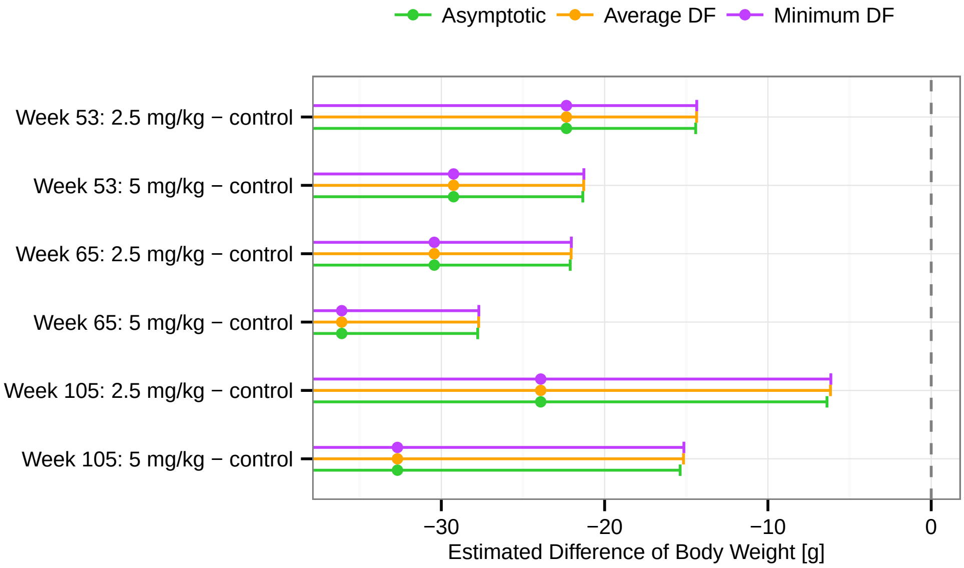
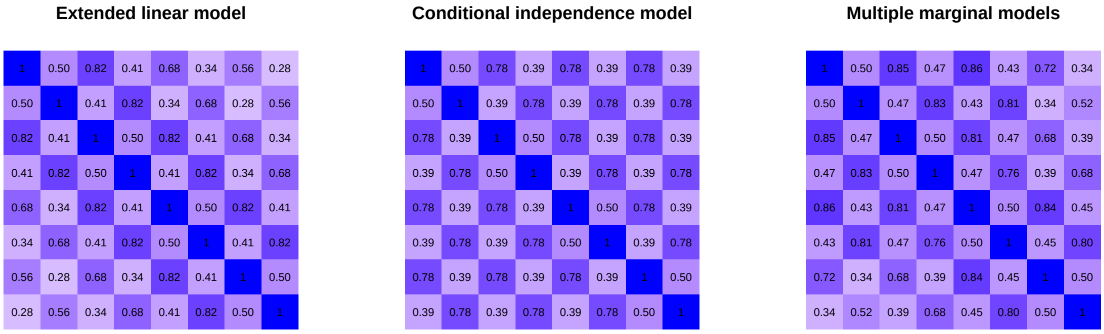


Figure 34: Mercuric chloride data: 95% simultaneous confidence intervals for many-to-one comparisons of treatments per time point based on the AICc-selected joint CIM with average and minimum Kenward-Roger degrees of freedom, and asymptotic results (for comparison).



Extended linear model

1	0.50	0.82	0.41	0.68	0.34	0.56	0.28
0.50	1	0.41	0.82	0.34	0.68	0.28	0.56
0.82	0.41	1	0.50	0.82	0.41	0.68	0.34
0.41	0.82	0.50	1	0.41	0.82	0.34	0.68
0.68	0.34	0.82	0.41	1	0.50	0.82	0.41
0.34	0.68	0.41	0.82	0.50	1	0.41	0.82
0.56	0.28	0.68	0.34	0.82	0.41	1	0.50
0.28	0.56	0.34	0.68	0.41	0.82	0.50	1

Conditional independence model

1	0.50	0.78	0.39	0.78	0.39	0.78	0.39
0.50	1	0.39	0.78	0.39	0.78	0.39	0.78
0.78	0.39	1	0.50	0.78	0.39	0.78	0.39
0.39	0.78	0.50	1	0.39	0.78	0.39	0.78
0.78	0.39	0.78	0.39	1	0.50	0.78	0.39
0.39	0.78	0.39	0.78	0.50	1	0.39	0.78
0.78	0.39	0.78	0.39	0.78	0.39	1	0.50
0.39	0.78	0.39	0.78	0.39	0.78	0.50	1

Multiple marginal models

1	0.50	0.85	0.47	0.86	0.43	0.72	0.34
0.50	1	0.47	0.83	0.43	0.81	0.34	0.52
0.85	0.47	1	0.50	0.81	0.47	0.68	0.39
0.47	0.83	0.50	1	0.47	0.76	0.39	0.68
0.86	0.43	0.81	0.47	1	0.50	0.84	0.45
0.43	0.81	0.47	0.76	0.50	1	0.45	0.80
0.72	0.34	0.68	0.39	0.84	0.45	1	0.50
0.34	0.52	0.39	0.68	0.45	0.80	0.50	1

Figure 35: Heart rate data: correlation matrices of test statistics for many-to-one comparisons of treatments per time point.

examples. The selected ELM has an AR(1) residual correlation pattern and equal variance across treatment groups and occasions. The second-best model specifies separate variance parameters for the treatment groups, but it has already a relatively large ΔAICc of 3.05. The selected CIM is also a simple one with only random patient effects; the runner-up (treatment-specific random effects) has a ΔAICc of 9.40. Given the low sample size of the data, it is not surprising that very simple models with few parameters are selected.

We see the consequences of these model choices in the estimated correlation matrices of test statistics (Figure 35): the selected CIM assumes equicorrelation whereas the selected ELM reveals an AR(1) pattern with decreasing correlation as the distance between time points increases. MMM estimates a matrix with lower correlation between later measurements than between earlier ones.

Table 8: Simultaneous inference for the heart rate data: estimated differences of heart rates, standard errors, and adjusted p-values for occasion-wise many-to-one comparisons of AX23 and BWW9 against control.

	Estimate	SE(ELM)	SE(CIM)	SE(MMM)	p(ELM)	p(CIM)	p(MMM)	p(Bonf)
T1: AX23 - control	-2.250	2.850	2.890	2.762	0.918	0.918	0.897	1.000
T1: BWW9 - control	9.000	2.850	2.890	2.762	0.026	0.021	0.018	0.028
T2: AX23 - control	8.125	2.850	2.890	3.132	0.048	0.042	0.074	0.125
T2: BWW9 - control	11.625	2.850	2.890	3.132	0.004	0.002	0.007	0.010
T3: AX23 - control	9.500	2.850	2.890	2.794	0.018	0.014	0.013	0.021
T3: BWW9 - control	7.125	2.850	2.890	2.794	0.097	0.089	0.081	0.137
T4: AX23 - control	2.125	2.850	2.890	2.859	0.935	0.935	0.928	1.000
T4: BWW9 - control	8.750	2.850	2.890	2.859	0.031	0.025	0.029	0.045

We use ν_{KR} as a DF approximation for the longitudinal MCTs based on the CIM, ν_{ESS} for the ELM-based procedure, and the ν_{res} of the marginal models for MMM. The results in Table 8 show that AX23 leads to a significantly increased heart rate only at T2 and T3 whereas the increase in the BWW9 group is significant at all time points except T3—this is exactly where the mean profiles of AX23 and BWW9 (Figure 3) cross paths. The adjusted p-values are similar across modeling approaches, as are the SCIs (Figure 36); discrepancies are due to different SE and correlation estimates and small differences in DFs. Ignoring the correlation and adjusting with Bonferroni results in larger p-values and wider SCIs.

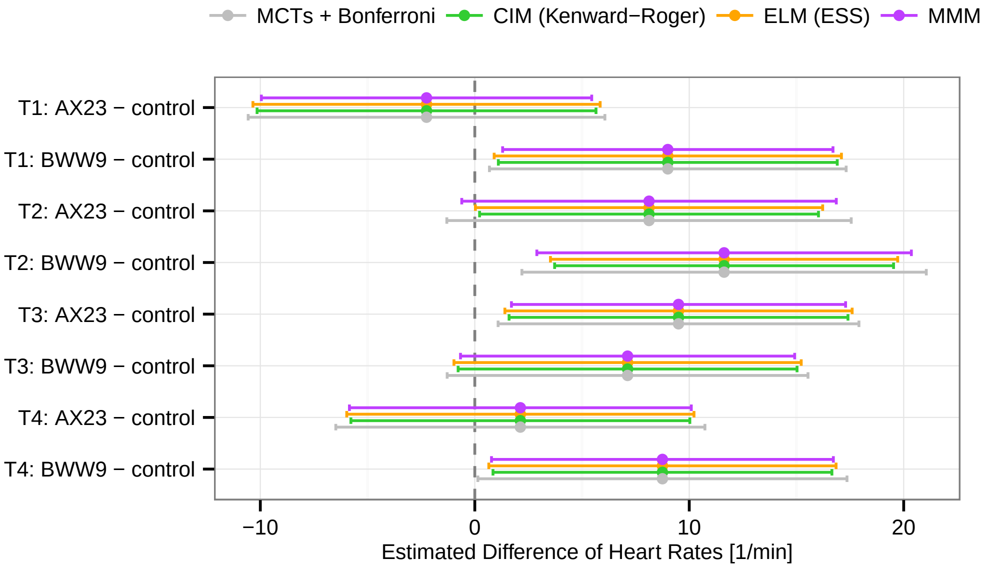


Figure 36: Heart rate data: 95% simultaneous confidence intervals for many-to-one comparisons of treatments per time point based on AICc-selected joint models (CIM and ELM), combined marginal models (MMM), and Bonferroni-adjusted MCTs per time point, with small-sample approximations to the degrees of freedom.

Comparing different DF approximations for the AICc-selected ELM, we find that ν_{ESS}, ν_{adj}, and ν_{PB2} yield very similar SCIs (Figure 37). Not surprisingly, the intervals constructed with the naive DF are unnecessarily wide, and using the residual DF makes the SCIs almost as narrow as for the asymptotic procedure, which is clearly inadequate given the small sample sizes.

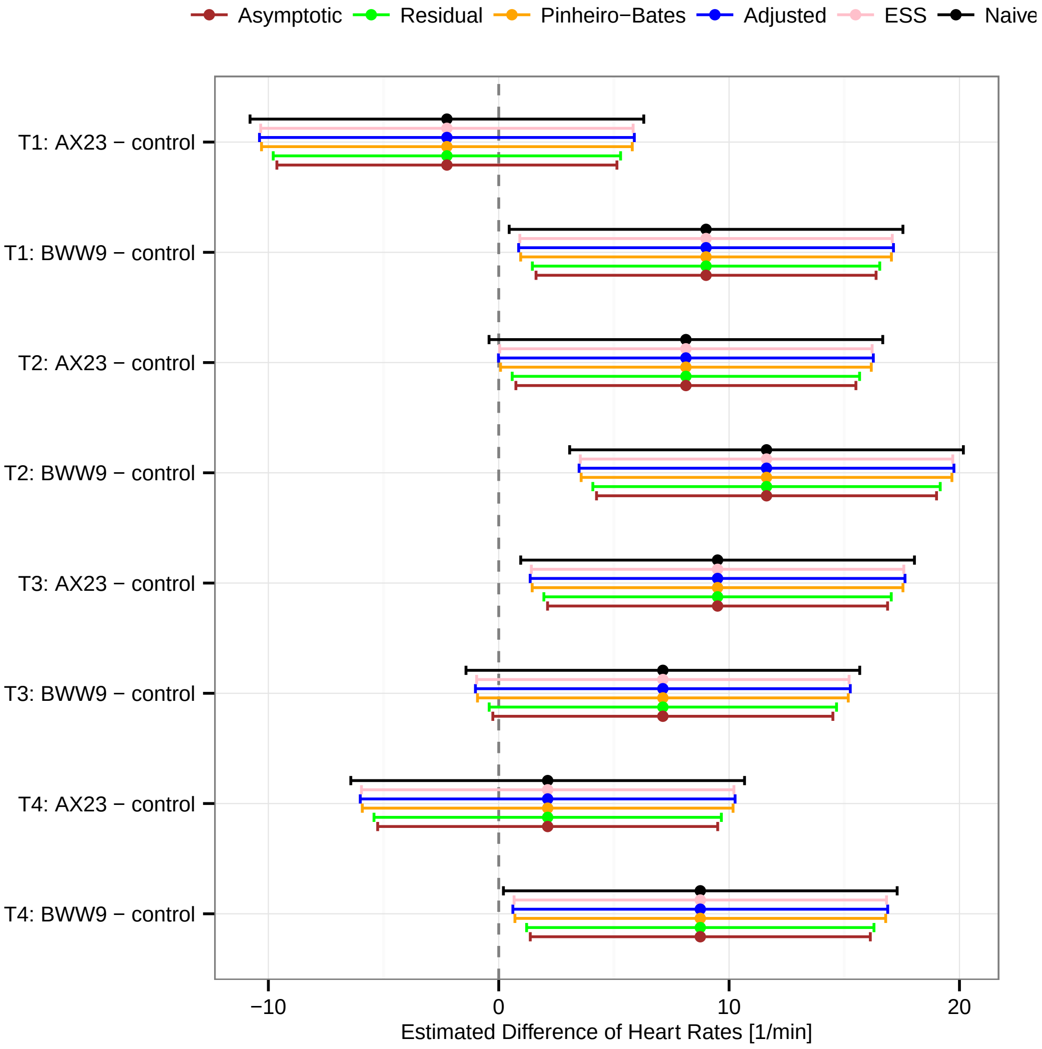


Figure 37: Heart rate data: 95% simultaneous confidence intervals for many-to-one comparisons of treatments per time point based on the AICc-selected ELM with different approximations to the degrees of freedom, and asymptotic results (for comparison).

Comparing multiple occasions simultaneously for multiple drugs: Another interesting problem is to assess by how much the heart rates differ across time points in the three treatment arms. This can be solved with Tukey comparisons separately and simultaneously for all treatments. Again we compare the AICc-selected joint models and MMM. The correlation matrices of test statistics are displayed in Figure 38.

In the AX23 arm, the heart rate is significantly higher at T2 and T3 compared to T1 and T4 (Table 9) and SCIs (Figure 39). Such clear effects cannot be found for BWW9 or control, whose mean profiles are relatively constant over time. Only T2 and T3 differ significantly for BWW9. The Bonferroni-adjusted separate MCTs do not always yield larger p-values and wider SCIs as would perhaps be expected, simply because each of them uses its own SE estimate whereas the AICc-selected joint models only estimate one common SE.

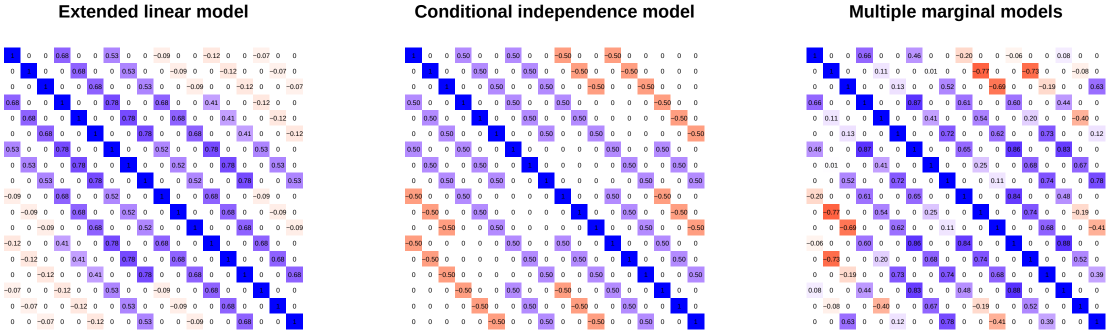


Figure 38: Heart rate data: correlation matrices of test statistics for all-pairwise comparisons of time points per treatment group.

Table 9: Simultaneous inference for the heart rate data: estimated differences of heart rates, standard errors, and adjusted p-values for treatment-wise all-pairwise comparisons of occasions.

	Estimate	SE(ELM)	SE(CIM)	SE(MMM)	p(ELM)	p(CIM)	p(MMM)	p(Bonf)
AX23: T2 - T1	10.000	1.200	1.365	1.417	<0.001	<0.001	0.002	<0.001
AX23: T3 - T1	10.500	1.621	1.365	1.066	<0.001	<0.001	<0.001	<0.001
AX23: T4 - T1	2.625	1.898	1.365	1.324	0.850	0.529	0.504	0.468
AX23: T3 - T2	0.500	1.200	1.365	1.673	1.000	1.000	1.000	1.000
AX23: T4 - T2	-7.375	1.621	1.365	1.924	0.003	<0.001	0.056	<0.001
AX23: T4 - T3	-7.875	1.200	1.365	1.319	<0.001	<0.001	0.005	<0.001
BWW9: T2 - T1	2.250	1.200	1.365	1.532	0.551	0.725	0.779	0.855
BWW9: T3 - T1	-3.125	1.621	1.365	1.411	0.516	0.291	0.395	0.242
BWW9: T4 - T1	-2.000	1.898	1.365	2.217	0.964	0.836	0.976	1.000
BWW9: T3 - T2	-5.375	1.200	1.365	1.943	0.004	0.003	0.207	0.004
BWW9: T4 - T2	-4.250	1.621	1.365	1.925	0.175	0.042	0.398	0.033
BWW9: T4 - T3	1.125	1.200	1.365	1.544	0.983	0.996	0.993	1.000
Control: T2 - T1	-0.375	1.200	1.365	0.816	1.000	1.000	1.000	1.000
Control: T3 - T1	-1.250	1.621	1.365	1.009	0.995	0.990	0.884	1.000
Control: T4 - T1	-1.750	1.898	1.365	1.613	0.984	0.918	0.937	1.000
Control: T3 - T2	-0.875	1.200	1.365	0.778	0.997	0.999	0.925	1.000
Control: T4 - T2	-1.375	1.621	1.365	1.438	0.991	0.981	0.967	1.000
Control: T4 - T3	-0.500	1.200	1.365	0.891	1.000	1.000	0.999	1.000

Comparing multiple occasions and multiple drugs simultaneously: In a final step, we combine the Dunnett comparisons of treatment groups and the Tukey comparisons of time points into one claim for which we want to protect a joint FWER of 5%. We apply the duplex procedure introduced in 4.3. Based on our AICc-selected joint CIM, the Kenward-Roger DF for comparisons of treatments is 29 and for comparisons of occasions it is 63. Since the model involves constant variance and there are no missing values, the Kenward-Roger DF is the same for all comparisons of each type.

The resulting SCIs are shown in Figure 40. For comparisons of time points, the method using separate DFs yields slightly shorter SCIs because it can use 63 DF here whereas the minimum DF is 29 and the (weighted) average is 52. For comparisons of treatment groups, the SCIs using minimum and separate DF are identical because they use both 29 DF whereas the intervals using the average DF of 52 are narrower (in fact narrower than they should be).

Compared to the previous analyses, the “bigger” claim including more elementary hypotheses makes the SCIs wider than those in Figures 36 and 39. As a consequence of this, we can no longer claim that BWW9 raises the heart rate significantly at the first and last

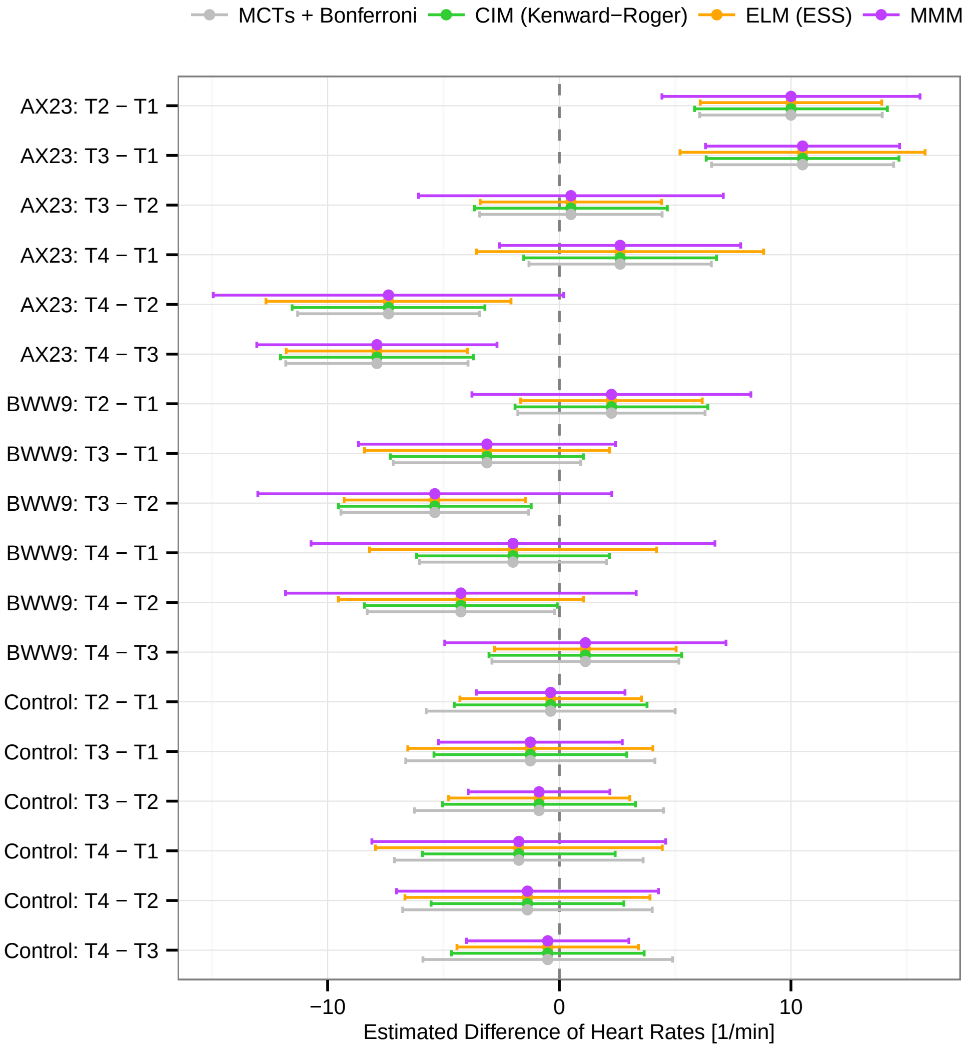


Figure 39: Heart rate data: 95% simultaneous confidence intervals for all-pairwise comparisons of time points per treatment group based on AICc-selected joint models (CIM and ELM), combined marginal models (MMM), and Bonferroni-adjusted MCTs per treatment arm, with small-sample approximations to the degrees of freedom.

time point. This is the price we pay for controlling the FWER at 5% over comparisons of treatment groups and time points simultaneously.

4.5 Power

All power investigations in this chapter were done via simulation, but it is also possible to work out the power of longitudinal MCTs—at least approximately. Exact any-pair power calculations for standard MCTs are outlined in Genz and Bretz (1999) and Bretz et al. (2001b). We extend their work in that we show how to calculate various types of powers (see 3.9) for our longitudinal MCTs. Our subsequent considerations are, however, only approximate because correlations of time points have to be estimated from sample values and hence the reference quantiles are random variables rather than fixed quantities.

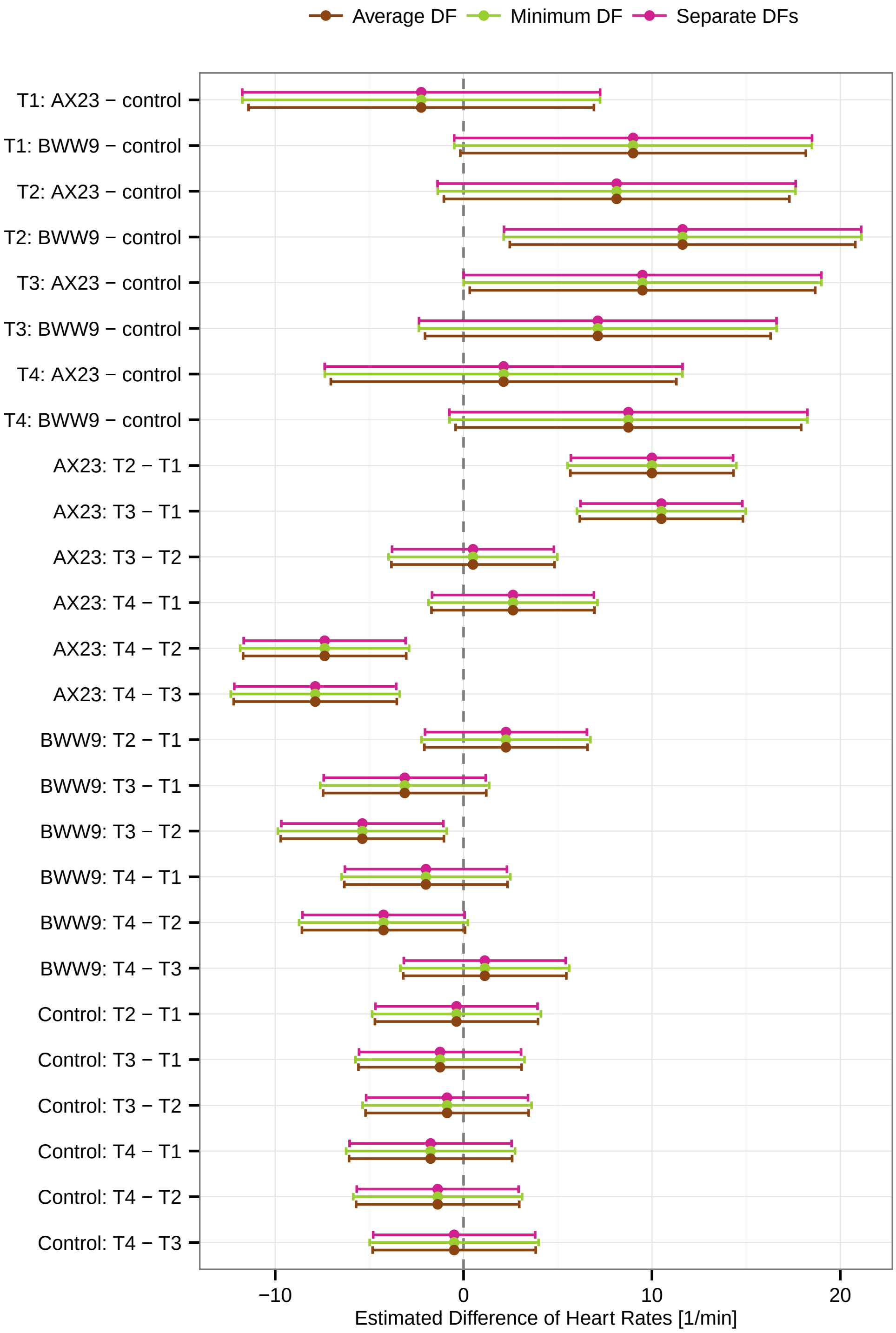


Figure 40: Heart rate data: 95% simultaneous confidence intervals for many-to-one comparisons of treatments per time point and all-pairwise comparisons of time points per treatment group simultaneously based on the AICc-selected joint CIM with average, minimum, and separate (comparison-specific) Kenward-Roger degrees of freedom.

Global Power: The global power for two-sided problems can be approximated as

$$P\{\exists\ h : |T_h| > t^{two}_{z,1-\alpha}(\nu, \mathbf{\Gamma})\} = 1 - T_z(-t^{two}_{z,1-\alpha}(\nu, \mathbf{\Gamma}), t^{two}_{z,1-\alpha}(\nu, \mathbf{\Gamma}), \mathbf{\Gamma}, \nu, \boldsymbol{\kappa})$$

where $T_z(a, b, \mathbf{\Gamma}, \nu, \boldsymbol{\kappa})$ is the z-variate t-probability with integration bounds a and b (same in all z dimensions), correlation matrix $\mathbf{\Gamma}$, degrees of freedom ν, and noncentrality vector $\boldsymbol{\kappa}$. The one-sided analogues are

$$P\{\exists\ h : T_h > t^{one}_{z,1-\alpha}(\nu, \mathbf{\Gamma})\} = 1 - T_z(-\infty, t^{one}_{z,1-\alpha}(\nu, \mathbf{\Gamma}), \mathbf{\Gamma}, \nu, \boldsymbol{\kappa})$$

and for the opposite direction

$$P\{\exists\ h : T_h < -t^{one}_{z,1-\alpha}(\nu, \mathbf{\Gamma})\} = 1 - T_z(-t^{one}_{z,1-\alpha}(\nu, \mathbf{\Gamma}), \infty, \mathbf{\Gamma}, \nu, \boldsymbol{\kappa}).$$

Any-Pair Power: The approximate any-pair power for testing two-sided hypotheses is

$$P\{\exists\ h \in \mathcal{A} : |T_h| > t^{two}_{z,1-\alpha}(\nu, \mathbf{\Gamma})\} = 1 - T_{z^*}(-t^{two}_{z,1-\alpha}(\nu, \mathbf{\Gamma}), t^{two}_{z,1-\alpha}(\nu, \mathbf{\Gamma}), \mathbf{\Gamma}^*, \nu, \boldsymbol{\kappa}^*)$$

with T_{z^*} a z^*-variate t-probability, and z^* is the number of elementary hypotheses that is under the alternative. Now $\mathbf{\Gamma}^*$ is the submatrix of $\mathbf{\Gamma}$ containing the elements of $\mathbf{\Gamma}$ that correspond to the elementary hypotheses under the alternative, and likewise $\boldsymbol{\kappa}^*$ is the subvector of $\boldsymbol{\kappa}$ that contains only the z^* values referring to the set $\mathcal{A}$ (see 3.9). Note that $1 \leq z^* \leq z$; in case all contrasts are under H_0, the any-pair power is set to α. The expressions for upper- and lower-tailed tests, respectively, are

$$P\{\exists\ h \in \mathcal{A} : T_h > t^{one}_{z,1-\alpha}(\nu, \mathbf{\Gamma})\} = 1 - T_{z^*}(-\infty, t^{one}_{z,1-\alpha}(\nu, \mathbf{\Gamma}), \mathbf{\Gamma}^*, \nu, \boldsymbol{\kappa}^*)$$

and

$$P\{\exists\ h \in \mathcal{A} : T_h < -t^{one}_{z,1-\alpha}(\nu, \mathbf{\Gamma})\} = 1 - T_{z^*}(-t^{one}_{z,1-\alpha}(\nu, \mathbf{\Gamma}), \infty, \mathbf{\Gamma}^*, \nu, \boldsymbol{\kappa}^*).$$

All-Pairs Power: An approximation for the all-pairs power with two-sided hypotheses is

$$P\{|T_h| > t^{two}_{z,1-\alpha}(\nu, \mathbf{\Gamma})\ \forall\ h \in \mathcal{A}\}$$

and is best found via simulation. If all contrasts are under the null, the all-pairs power is α by definition. The corresponding expressions for one-sided alternatives are

$$P\{T_h > t^{one}_{z,1-\alpha}(\nu, \mathbf{\Gamma})\ \forall\ h \in \mathcal{A}\} = T_{z^*}(-t^{one}_{z,1-\alpha}(\nu, \mathbf{\Gamma}), \infty, \mathbf{\Gamma}^*, \nu, \boldsymbol{\kappa}^*)$$

and

$$P\{T_h < t^{one}_{z,1-\alpha}(\nu, \mathbf{\Gamma})\ \forall\ h \in \mathcal{A}\} = T_{z^*}(-\infty, t^{one}_{z,1-\alpha}(\nu, \mathbf{\Gamma}), \mathbf{\Gamma}^*, \nu, \boldsymbol{\kappa}^*)$$

with the symbols defined as for the any-pair power.

It is arguable which type of power to consider for a given problem. The choice should depend on the *win criterion* (Dmitrienko et al. 2013): if the research question is such that an effect can be successfully claimed if *at least* one difference (and no matter which one) is significant, then the any-pair power is appropriate. For practical purposes, however, the global power is probably more relevant; it is “contaminated” with elementary type I errors, but in practice they will be indistinguishable from true effects. The all-pairs power is in demand if we want to show that *all* differences are significant, a situation which is rather hard to think of in the context of longitudinal MCTs.

An R program for computing the approximate global, any-pair, and all-pairs power is given in Appendix E.

5 Longitudinal MCTs with Discrete Endpoints

Numerous biological and medical outcomes are not measured on a continuous scale but rather as counts or proportions. Such discrete data have signature characteristics that are incompatible with the assumption of a Gaussian distribution:

- Proportions are always between 0 and 1.
- Counts are integers and cannot become negative.
- The variance is not independent of the mean.

Therefore the methods proposed in the previous Chapter 4 are not directly applicable to discrete data, but they can be modified. A joint model for discrete data may come from the class of generalized linear mixed-effects models (GLMMs) or be fitted with generalized estimating equations (GEEs). Likewise, the MMM method can be built on occasion-specific generalized linear models (GLMs).
We describe asymptotic approaches to comparisons of treatments at multiple occasions and comparisons of occasions within multiple treatment groups for discrete endpoints, focusing on binomial proportions in 5.1 and Poisson rates in 5.2, and we outline how to create a duplex procedure (analogous to that for Gaussian data in 4.3) in the presence of discrete outcomes in 5.3. We assess via simulation the finite-sample properties of these asymptotic procedures, devise their achieved size under H_0, and compare their powers. Moreover, we show how to evaluate three real datasets in 5.4.

5.1 Binomial Data

Imagine an experiment where binomial proportions are measured repeatedly at a number of occasions within multiple treatment groups i.e., a number of successes and failures is recorded for each treatment at each time point. We are interested in multiple comparisons (adjusted p-values and SCIs) of binomial proportions π among treatment groups at several occasions or among occasions within several treatment groups under control of α. Such comparisons can be useful for repeatable or reversible events; on the other hand, irreversible events like death are better analyzed with time-to-event methods.

We propose two basic approaches, one based on a joint model for the entire timespan of interest, and the other using marginal occasion-specific models, in 5.1.1. As the procedures are asymptotic in nature, satisfactory control of type I error rates cannot be guaranteed with small sample sizes. Thus we employ numerical simulations in 5.1.2 to assess minimum sample sizes that ensure α control within a sufficient range. In addition, we compare the powers of the joint modeling and MMM approaches with each other and also with simple Bonferroni adjustments that ignore any correlation over time. Application to real data from a study in pest control is showcased in 5.4.3.

5.1.1 Procedure

Suppose the outcome of an experiment with independent units $i = 1, \ldots, n$ randomized to treatment groups $k = 1, \ldots, q$ is binomial, and it is measured repeatedly at occasions $j = 1, \ldots, m$. The random variable of outcomes $\mathbf{Y}$ has realizations $\mathbf{y} = (y_{111}, \ldots, y_{nmq})^T$

where y_{ijk} refers to the observation recorded from individual i belonging to treatment group k at time point j. Further n_k denotes the number of individuals randomized to the kth treatment group, and $n^{(j)}$ is the number of individuals measured at occasion j, which may vary (because of dropout etc.) from occasion to occasion. Superscript index (j) pertains to the marginal model for the jth occasion.

Within every single occasion j the observations are (assumed to be) distributed as

$$Y_i^{(j)} \sim Bin(\pi_i^{(j)}, w_i^{(j)})$$

where $w_i^{(j)}$ denotes the total number of success and failures for unit i at occasion j. According to binomial theory,

$$E(Y_i^{(j)}) = w_i^{(j)} \pi_i^{(j)}$$

and

$$Var(Y_i^{(j)}) = w_i^{(j)} \pi_i^{(j)} (1 - \pi_i^{(j)}).$$

Our aim is now to compare treatment means at multiple occasions or occasion means at multiple treatments using a procedure that is a variation of the longitudinal MCTs for Gaussian endpoints described in 4.1.1 and 4.2.1. Instead of fitting a joint LMM, however, we now use binomial GEEs (as in 3.4) with some working correlation $\mathbf{R}_i(\alpha)$ as a joint model. Solving the GEE yields consistent estimates of $\boldsymbol{\beta}$, and the sandwich estimate $\widehat{\boldsymbol{\Sigma}}$ is a consistent estimate of $\boldsymbol{\Sigma}$ even under misspecification of the working correlation; naive covariance estimates may be calculated as well. These estimates can readily be used for tests and SCIs as detailed in 4.1.1. Multiple testing of GEE parameters was also described in Orelien et al. (2002). We do not consider GLMMs as joint models here because we want to focus on the interpretation of population rather than within-subject effects.

The alternative to joint modeling is once again to combine multiple occasion-specific models. Instead of marginal linear models, we now fit one marginal binomial GLM (as in 3.3)

$$\text{logit}(\boldsymbol{\pi}^{(j)}) = \log\left(\frac{\boldsymbol{\pi}^{(j)}}{1 - \boldsymbol{\pi}^{(j)}}\right) = \boldsymbol{\eta}^{(j)} = \mathbf{X}^{(j)} \boldsymbol{\beta}^{(j)}$$

per time point j with $\boldsymbol{\pi} = (\pi_1^{(j)}, \ldots, \pi_{n^{(j)}}^{(j)})^T$ and $\boldsymbol{\eta} = (\eta_1^{(j)}, \ldots, \eta_{n^{(j)}}^{(j)})^T$. In a simple one-way model without covariates, $\mathbf{X}^{(j)}$ will be an $n^{(j)} \times q$ matrix, and $\boldsymbol{\beta}^{(j)} = (\beta_1^{(j)}, \ldots, \beta_q^{(j)})^T$. The correlation among the parameters $\boldsymbol{\beta} = (\boldsymbol{\beta}^{(1)}, \ldots, \boldsymbol{\beta}^{(m)})^T$ can be estimated as described in 3.5 by stacking their respective score contributions from the m marginal binomial GLMs. The Ψ_{ij} are now obtained from the first coordinate of

$$-E\left(\boldsymbol{\pi}^{(j)}(1 - \boldsymbol{\pi}^{(j)}) \mathbf{x}_i^{(j)} \mathbf{x}_i^{(j)T}\right)^{-1} \mathbf{x}_i^{(j)} \left(\mathbf{y}_i^{(j)} - \boldsymbol{\pi}^{(j)}\right).$$

Following the further steps in 3.5 provides us with estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ that can be used for testing hypotheses and building SCIs.

Comparisons of treatments per occasion or occasions per treatment are defined in coefficient matrices $\mathbf{C}$ as outlined in 4.1.1 and 4.2.1. The testing procedure controls the FWER asymptotically at level α, and the SCIs have an asymptotic coverage probability of $(1 - \alpha)$. Finite-sample properties of this asymptotic method have not been investigated so far, so we will dedicate the next subsection to an exploration of which sample sizes are necessary to achieve a reasonable test size.

5.1.2 Simulation Study

As pointed out by Pipper et al. (2012) the procedure combining marginal GLMs is asymptotic i.e., it may fail to control the prespecified rate of type I errors with small sample sizes. Thus we set up a simulation study to gauge which sample sizes are necessary to come close to nominal α control. Moreover, we are interested in power comparisons of the joint GEE and multiple marginal GLM approaches.

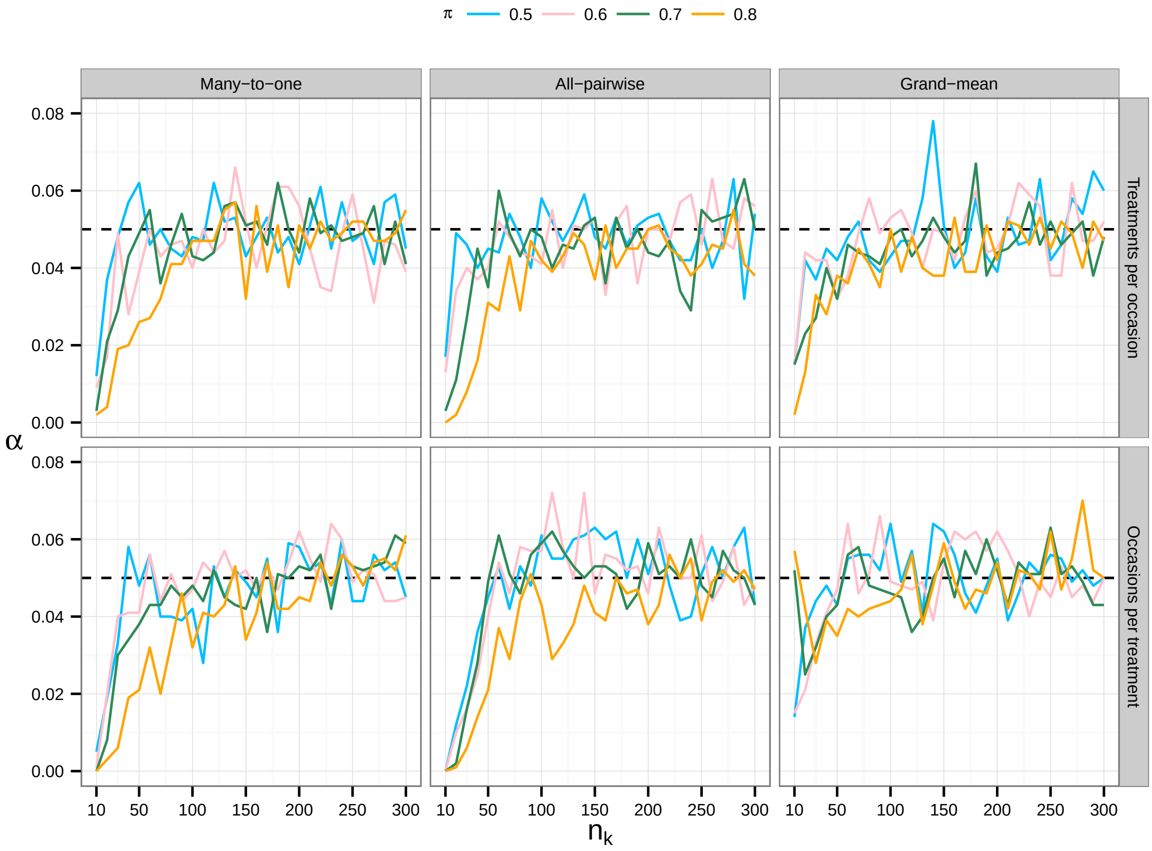


Figure 41: Simulated type I error rates for asymptotic many-to-one, all-pairwise, and grand-mean comparisons of binomial data ($\pi = \{0.5, 0.6, 0.7, 0.8\}$) involving $q = 3$ treatments and $m = 3$ occasions, with n_k independent subjects per treatment group, based on multiple marginal GLMs (1000 simulation runs). Top: comparisons of treatment means separately and simultaneously within occasions; bottom: comparisons of occasion means separately and simultaneously within treatment groups.

Type I error: To investigate the actual size of the tests under H_0, we generate correlated multivariate binary data using Gaussian copulae (see Appendices B and E) in balanced longitudinal settings with $n_k = \{10, 20, \dots, 300\}$ independent units in each of $q = \{3, 4, 5\}$ treatment groups, measured at $m = \{3, 4, 5\}$ points in time. For layouts with $q = 3$ and $m = 3$ we investigate four different correlations:

$$\mathbf{B}_1 = \begin{pmatrix} 1 & 0.95 & 0.90 \\ & 1 & 0.95 \\ & & 1 \end{pmatrix}, \quad \mathbf{B}_2 = \begin{pmatrix} 1 & 0.90 & 0.80 \\ & 1 & 0.90 \\ & & 1 \end{pmatrix},$$

$$\mathbf{B}_3 = \begin{pmatrix} 1 & 0.80 & 0.50 \\ & 1 & 0.80 \\ & & 1 \end{pmatrix}, \quad \mathbf{B}_4 = \begin{pmatrix} 1 & 0.50 & 0.20 \\ & 1 & 0.50 \\ & & 1 \end{pmatrix}.$$

For settings with $q = 4$ or 5 treatments and $m = 3$, we use $\mathbf{B}_2$ as correlation matrix. For $m = 4$ with $q = 3$ we use

$$\mathbf{B}_5 = \begin{pmatrix} 1 & 0.90 & 0.80 & 0.70 \\ & 1 & 0.90 & 0.80 \\ & & 1 & 0.90 \\ & & & 1 \end{pmatrix},$$

and for $m = 5$ with $q = 3$

$$\mathbf{B}_6 = \begin{pmatrix} 1 & 0.90 & 0.80 & 0.70 & 0.60 \\ & 1 & 0.90 & 0.80 & 0.70 \\ & & 1 & 0.90 & 0.80 \\ & & & 1 & 0.90 \\ & & & & 1 \end{pmatrix}.$$

All these settings are simulated with binomial probabilities $\pi = \{0.5, 0.6, 0.7, 0.8\}$ for the simplest case where $w_i^{(j)} = 1$ for all independent units at all occasions, so the outcomes y_{ijk} can only take values 0 or 1. . We apply two different modeling strategies: fitting a binomial GEE with AR(1) working correlation, or fitting one binomial GLM per time point and proceed with MMM. Based on either of these, we simultaneously test many-to-one, all-pairwise, or grand-mean contrasts of treatment means per occasion, or occasion means per treatment, at a nominal $\alpha = 0.05$ level. Then we record for each of 1000 simulation runs whether the minimum adjusted p-value falls below 0.05.

Figure 41 displays simulated type I error rates with $m = q = 3$ and correlation $\mathbf{B}_2$ for the multiple marginal GLMs approach. We see that this asymptotic procedure is conservative in the presence of small samples. The extent of conservatism tends to be more pronounced as π diverges from 0.5. For samples as small as $n_k = 10$, many-to-one and all-pairwise comparisons of occasions within several treatment groups have practically no power whereas the realized α level for comparisons of treatments at multiple time points is between 0 and 0.02 when $n_k = 10$. The situation appears somewhat more relaxed with grand-mean contrasts. Broadly speaking, a minimum sample size of at least 50 (in the setting where all $w_i^{(j)} = 1$) seems to be required for type I error rates reliably close to the nominal α.

The situation is slightly different for the GEE-based procedure (Figure 42). When comparing multiple treatments per occasion, using the naive SE makes the tests conservative with small sample sizes (where small refers to $n_k < 50$) whereas the robust SE appears to be very unstable if the sample size is pathologically small ($n_k < 10$), especially for π equal to 0.7 and 0.8. For comparisons of occasions per treatment, using the naive SE leads to a slightly liberal procedure with type I error rates around 0.07 to 0.08 even for large samples.

Simulation results with m and q values other than 3 are shown in Appendix D: Figures 64, 65, and 66 for multiple marginal GLMs, and Figures 67, 68, and 69 for joint GEEs.

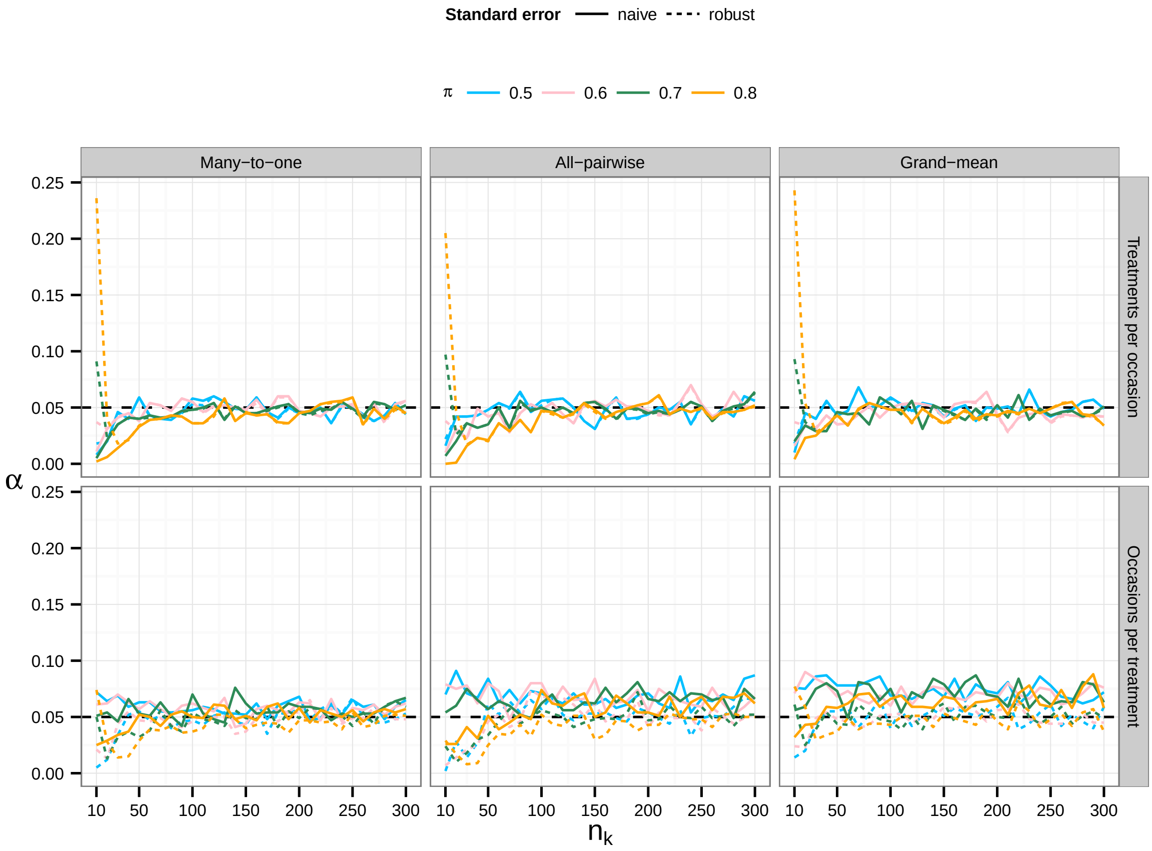


Figure 42: Simulated type I error rates for asymptotic many-to-one, all-pairwise, and grand-mean comparisons of binomial data ($\pi = \{0.5,\ 0.6,\ 0.7,\ 0.8\}$) involving $q = 3$ treatments and $m = 3$ occasions, with n_k independent subjects per treatment group, based on GEEs (1000 simulation runs). Top: comparisons of treatment means separately and simultaneously within occasions; bottom: comparisons of occasion means separately and simultaneously within treatment groups.

Power: We study power with the sample size fixed at $n_k = 100$ for which we have observed proper control of α in the preceding simulations. We investigate the same scenarios as above i.e., $m = \{3, 4, 5\}$ and $q = \{3, 4, 5\}$ with correlations $\mathbf{B}_1$ through $\mathbf{B}_6$. The true proportion of success is set to $\pi + \Delta$ with $\Delta = \{-0.30, -0.28, \ldots, 0.28, 0.30\}$ for the qth treatment at the mth occasion, and π for all other combinations of k and j. We apply the following asymptotic procedures to compare treatments per occasion or occasions per treatment:

1. pairwise two-sample tests (z-tests) of GLM parameters, adjusted with Bonferroni for the multiplicity of treatments and occasions (this ignores any correlation among test statistics),
2. one standard (asymptotic) MCT of GLM parameters per time point, adjusted with Bonferroni for the multiplicity of occasions (i.e., ignoring correlation over time),
3. longitudinal MCTs based on multiple marginal GLMs with MMM,
4. longitudinal MCTs based on a joint GEE with AR(1) working correlation.

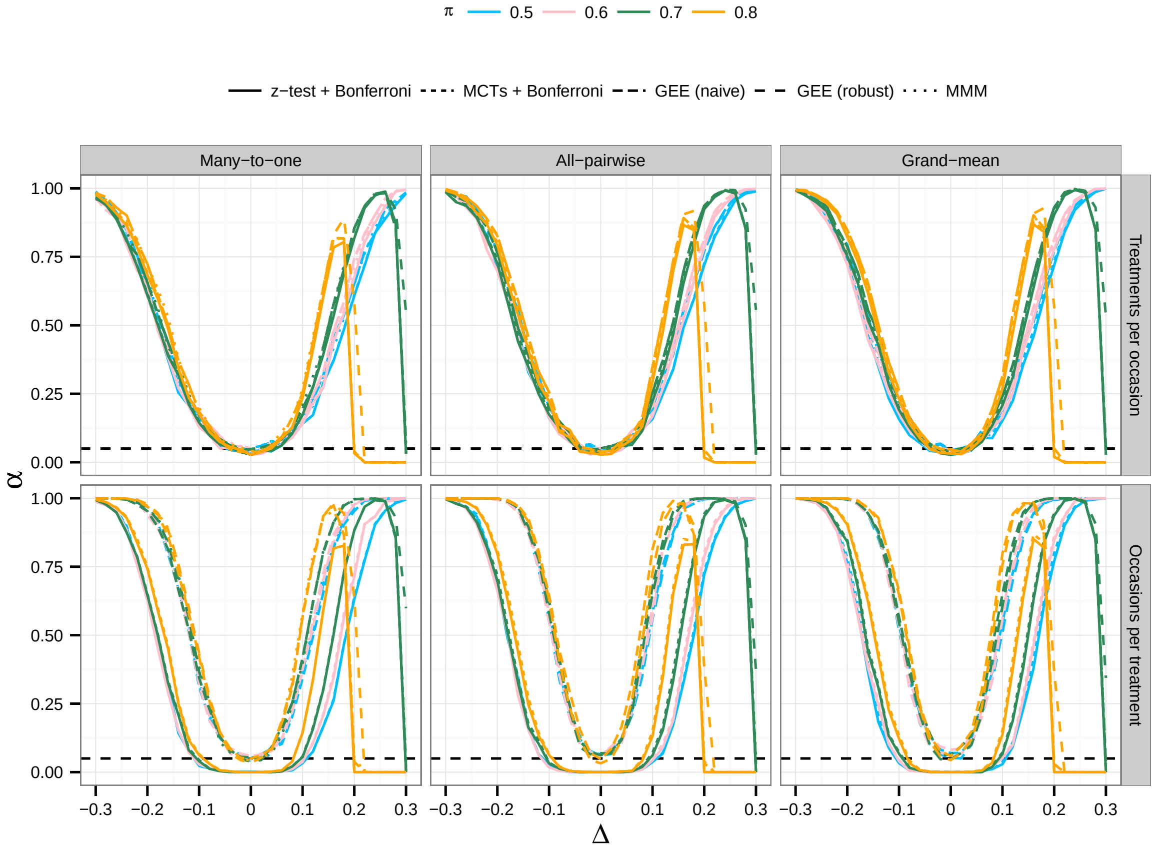


Figure 43: Simulated powers for many-to-one, all-pairwise, and grand-mean comparisons of binomial data ($\pi = \{0.5, 0.6, 0.7, 0.8\}$) involving $q = 3$ treatments and $m = 3$ occasions, with $n_k = 100$ independent subjects per treatment group (1000 simulation runs). Top: comparisons of treatment means separately and simultaneously within occasions; bottom: comparisons of occasion means separately and simultaneously within treatment groups.

Power curves from 1000 simulation runs are presented in Figure 43. We see that the power advantage of longitudinal MCTs over Bonferroni-based procedures is only a few percent when treatments are compared per occasion. There is almost no difference in power between joint GEEs and multiple marginal GLMs. However, when comparing occasions per treatment, the Bonferroni-based procedures that ignore any longitudinal correlation are severely inferior and have practically no power for $-0.1 < \Delta < 0.1$. All this is in line with the findings for Gaussian data in 4.1.3 and 4.2.3. The curves are asymmetric around $\Delta = 0$ because the theoretical upper limit of the π or $\pi + \Delta$ is 1, and the power goes down to zero as soon as the bound of 1 is reached.

Further simulation results for $m, q > 3$ are depicted in Figure 71, and for $m = q = 3$ with correlation matrices $\mathbf{B}_1$, $\mathbf{B}_3$, and $\mathbf{B}_4$ in Figure 72, both in Appendix D.

5.2 Poisson Data

Now suppose the outcome of an experiment are Poisson rates (e.g., counts of events within a defined period of time) measured repeatedly from individuals that were randomized to treatment groups. The goal is to perform multiple comparisons of Poisson rates λ between treatment groups at several occasions and/or between occasions within several treatment groups, so as to obtain adjusted p-values as well as SCIs.

This section is structured similar to 5.1: the basic methodology for longitudinal MCTs based on joint modeling as well as a set of marginal GLMs is to be presented in 5.2.1. We devise minimum sample size requirements for proper α control via simulation in 5.2.2 and compare the respective powers of procedures incorporating and ignoring longitudinal correlation. Application to real data from an experiment in entomology and a clinical trial is illustrated in 5.4.1 and 5.4.2.

5.2.1 Procedure

Assume an experiment where independent units $i = 1, \ldots, n$ are randomized to treatment groups $k = 1, \ldots, q$, and some Poisson outcome is measured repeatedly at occasions $j = 1, \ldots, m$. The random variable of outcomes $\mathbf{Y}$ has realizations $\mathbf{y} = (y_{111}, \ldots, y_{nmq})^T$ where y_{ijk} refers to the observation recorded from individual i belonging to treatment group k at time point j. Further n_k denotes the number of individuals randomized to the kth treatment group, and $n^{(j)}$ is the number of individuals measured at occasion j, which may vary (because of dropout etc.) from occasion to occasion. Superscript index (j) pertains to the marginal model for the jth occasion.

Within every single occasion j the observations are (assumed to be) distributed as

$$Y_i^{(j)} \sim Pois(\lambda_i^{(j)}),$$

and we know from Poisson theory that

$$E(Y_i^{(j)}) = Var(Y_i^{(j)}) = \lambda_i^{(j)}.$$

We want to carry out comparisons among treatment means at multiple occasions or occasion means for multiple treatment groups using an asymptotic longitudinal MCT procedure similar to the one introduced for binomial data in 5.1.1 but now for Poisson outcomes. The joint modeling approach is now a Poisson GEE with some working correlation $\mathbf{R}_i(\alpha)$, and the alternative is the MMM-type combination of occasion-specific Poisson GLMs

$$\log(\boldsymbol{\lambda}^{(j)}) = \boldsymbol{\eta}^{(j)} = \mathbf{X}^{(j)}\boldsymbol{\beta}^{(j)}$$

with $\boldsymbol{\lambda} = (\lambda_1^{(j)}, \ldots, \lambda_{n^{(j)}}^{(j)})^T$, and $\boldsymbol{\eta}$, $\mathbf{X}$, and $\boldsymbol{\beta}$ as defined in 5.1.1.

We estimate the correlation among the parameter estimates $\boldsymbol{\beta} = (\boldsymbol{\beta}^{(1)}, \ldots, \boldsymbol{\beta}^{(m)})^T$ as outlined in 3.5 by stacking their respective score contributions from the m marginal Poisson GLMs and obtain the Ψ_{ij} from the first coordinate of

$$-E\left(\boldsymbol{\lambda}^{(j)}\mathbf{x}_i^{(j)}\mathbf{x}_i^{(j)T}\right)^{-1}\mathbf{x}_i^{(j)}\left(\mathbf{y}_i^{(j)} - \boldsymbol{\lambda}^{(j)}\right).$$

Completing the steps in 3.5 yields estimates of both $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ that can be used for tests with asymptotic control of α and SCIs with asymptotic coverage probability of $(1-\alpha)$. We investigate the finite-sample properties of this method in the following.

5.2.2 Simulation Study

We have noted that the method combining multiple binomial GLMs controls α only asymptotically, and that small sample sizes may lead to undesirable behavior. Similar concerns pertain to combining marginal Poisson GLMs, which we will investigate numerically. We seek to assess minimum sample size requirements for the asymptotic method to control α in an acceptable range.

Type I error: Correlated multivariate Poisson data under H_0 are drawn by means of Gaussian copulae (see Appendices B and E) as in 5.1.2, with the difference that the marginal distributions are now Poisson with rates $\lambda = \{3, 5, 10, 20\}$. The simulation setups resemble those for the binomial investigation but with $n_k = \{4, 6, \ldots, 50\}$ independent units within each of $q = \{3, 4, 5\}$ treatment groups measured at $m = \{3, 4, 5\}$ occasions. We use as correlation matrices $\mathbf{B}_1$ through $\mathbf{B}_6$ from 5.1.2, and the variances are obviously determined by the choice of λ.

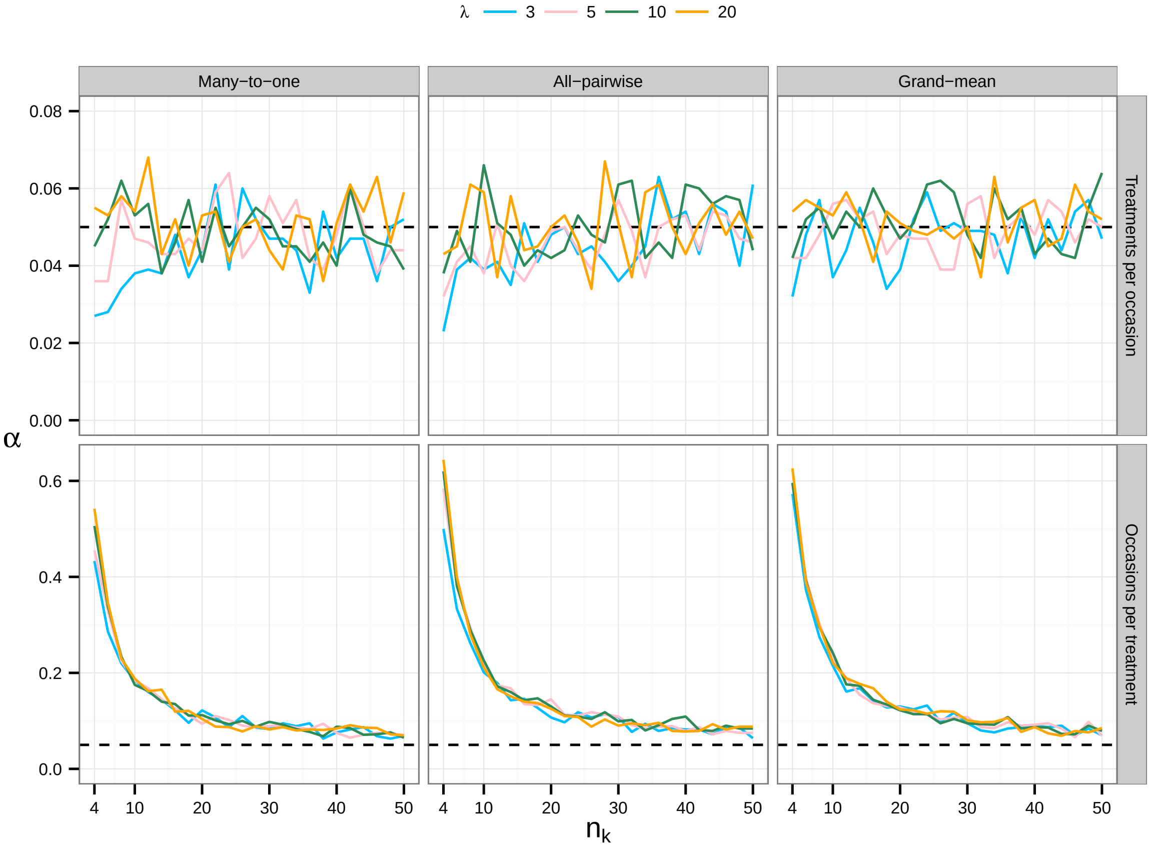


Figure 44: Simulated type I error rates for asymptotic many-to-one, all-pairwise, and grand-mean comparisons of Poisson data ($\lambda = \{3,\ 5,\ 10,\ 20\}$) involving $q = 3$ treatments and $m = 3$ occasions, with n_k independent subjects per treatment group, based on multiple marginal GLMs (1000 simulation runs). Top: comparisons of treatment means separately and simultaneously within occasions; bottom: comparisons of occasion means separately and simultaneously within treatment groups. Note that the vertical axes differ widely in their scaling.

Estimates of the test size based on multiple marginal GLMs with 1000 simulation runs are shown in Figure 44. We find comparisons of treatment means at multiple occasions to be unproblematic even with small samples: they may get a little conservative (around 3 to 4 %) for $n_k < 10$, but apart from that they stick to the nominal 5% level very closely. Comparisons of occasions within multiple treatment groups, however, are exceedingly liberal with type I error rates of up to 60% under H_0, and approach their nominal size rather slowly. Only with $n_k = 50$ acceptable α levels around 6 to 7% are achieved.

When using GEE estimates as a basis for longitudinal MCTs, the tests can as well become very liberal if the sample size is too small, for both comparisons of treatments and occassions (Figure 45). Interestingly, the naive SE estimate seems to cushion the extent of conservatism somewhat. With the naive SE, sample sizes n_k between 20 and 30 are sufficient to control the level α whereas at least 40 to 50 are needed with the robust SE estimates.

Appendix D holds more simulation results: type I error estimates for scenarios with m or $q > 3$ are shown in Figures 73, 74, and 75 for multiple marginal GLMs, and in Figures 76, 77, and 78 for joint GEEs.

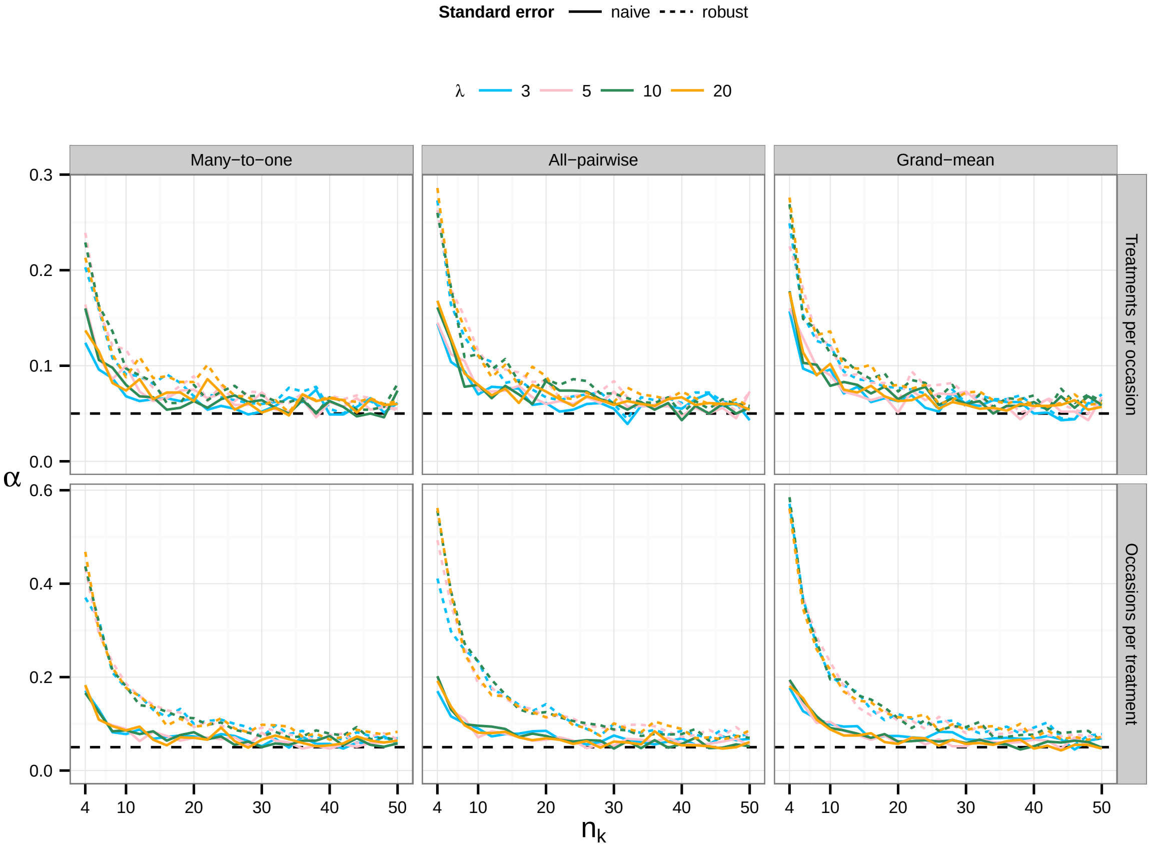


Figure 45: Simulated type I error rates for asymptotic many-to-one, all-pairwise, and grand-mean comparisons of Poisson data ($\lambda = \{3, 5, 10, 20\}$) involving $q = 3$ treatments and $m = 3$ occasions, with n_k independent subjects per treatment group, based on GEEs (1000 simulation runs). Top: comparisons of treatment means separately and simultaneously within occasions; bottom: comparisons of occasion means separately and simultaneously within treatment groups.

Power: To simulate the finite sample power, we fix the sample size at $n_k = 50$ and study scenarios involving $m = \{3, 4, 5\}$ and $q = \{3, 4, 5\}$ with correlation matrices $\mathbf{B}_1$ through $\mathbf{B}_6$. The true Poisson rate for the qth treatment at the mth occasion is set to $\lambda + \Delta$ with $\Delta = \{0, 0.02, \ldots, 0.28, 0.30\}$, and λ for all other combinations of k and j.

The power curves based on 1000 simulation runs (Figure 46) look somewhat similar to those for Gaussian data in 4.1.3 and 4.2.3. Longitudinal MCTs based on GEEs or MMM have almost identical power, and they are marginally superior to the Bonferroni-based procedures for comparing treatments per occasion, and clearly more powerful when occasions are compared per treatment group. These power advantages increase with λ, but on the other hand, increasing λ reduces the power of all procedures because in a Poisson framework the mean equals the variance.

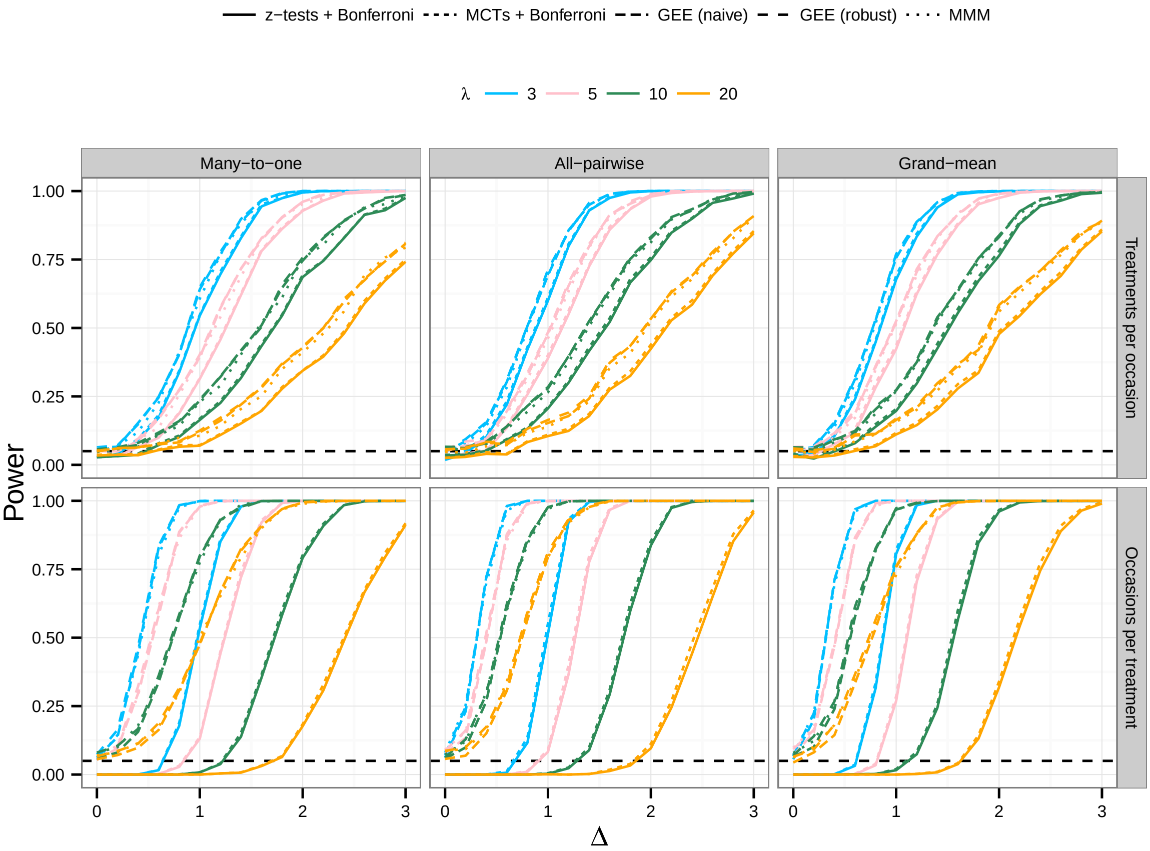


Figure 46: Simulated powers for many-to-one, all-pairwise, and grand-mean comparisons of Poisson data ($\lambda = \{3,\ 5,\ 10,\ 20\}$) involving $q = 3$ treatments and $m = 3$ occasions, with $n_k = 50$ independent subjects per treatment group (1000 simulation runs). Top: comparisons of treatment means separately and simultaneously within occasions; bottom: comparisons of occasion means separately and simultaneously within treatment groups.

Additional power curves are presented in Appendix D with Figure 80 for $m, q > 3$, and Figure 81 for correlation matrices $\mathbf{B}_1$, $\mathbf{B}_3$, and $\mathbf{B}_4$ with $m = q = 3$.

5.3 A Duplex Procedure

We want to extend our asymptotic longitudinal MCTs for discrete data along the lines of the duplex procedure in 4.3 i.e., we wish to enable ourselves to obtain adjusted p-values and SCIs for sets of elementary hypotheses that contain both simultaneous comparisons of treatment means and simultaneous comparisons of occasion means, with asymptotic FWER control over all elementary hypotheses.

5.3.1 Procedure

We have learned from the simulations in 5.1.2 and 5.2.2 that the sample size requirements for sufficient α control are usually higher with comparisons of occasions within multiple treatment groups than with comparisons of treatments at multiple occasions. Now that we want to blend the two types of tests, the data samples should obviously be large enough for both components (comparisons of treatments per occasion, and comparisons of occasions per treatment) to achieve proper control of α.

Longitudinal MCTs for discrete data are inherently asymptotic, and thus the duplex procedure is considerably easier to contrive than the one for Gaussian outcomes: one does not have to worry about how to deal with widely different DFs because there are none involved.

5.3.2 Simulation Study

We study the finite-sample type I error rate and power of the asymptotic duplex procedure based on either a GEE or multiple marginal GLMs via simulation, for both binomial and Poisson data. The simulation setups are similar to those in 5.1.2 and 5.2.2, but the contrast matrix now embraces comparisons among treatments as well as among occasions as shown in 4.3.

Type I error: For binomial outcomes, the tests based on GEEs using robust SEs or multiple marginal GLMs are severely conservative for $n_k < 50$ whereas the GEE with naive SEs comes closer to the nominal α (Figure 47). As π moves away from 0.5, the degree of conservatism tends to increase. With sufficiently large sample sizes all methods keep the desired type I error rate.

For Poisson counts, all asymptotic methods are wildly liberal when the sample size is too small (Figure 48). With $n_k = 4$, the GEE with naive SEs leads to type I error rates around 20 to 25%, whereas they can even be inflated up to 40 to 60% for the GEE with robust SEs and for multiple marginal GLMs. As a consequence, n_k should be at least around 30 for the former and 40 to 50 for the latter methods to ensure reasonable control of α. The choice of λ does not seem to have any influence here.

Detailed simulation results with different numbers of treatment groups and time points are shown in Figure 70 for binomial outcomes and in Figure 79 for Poisson counts.

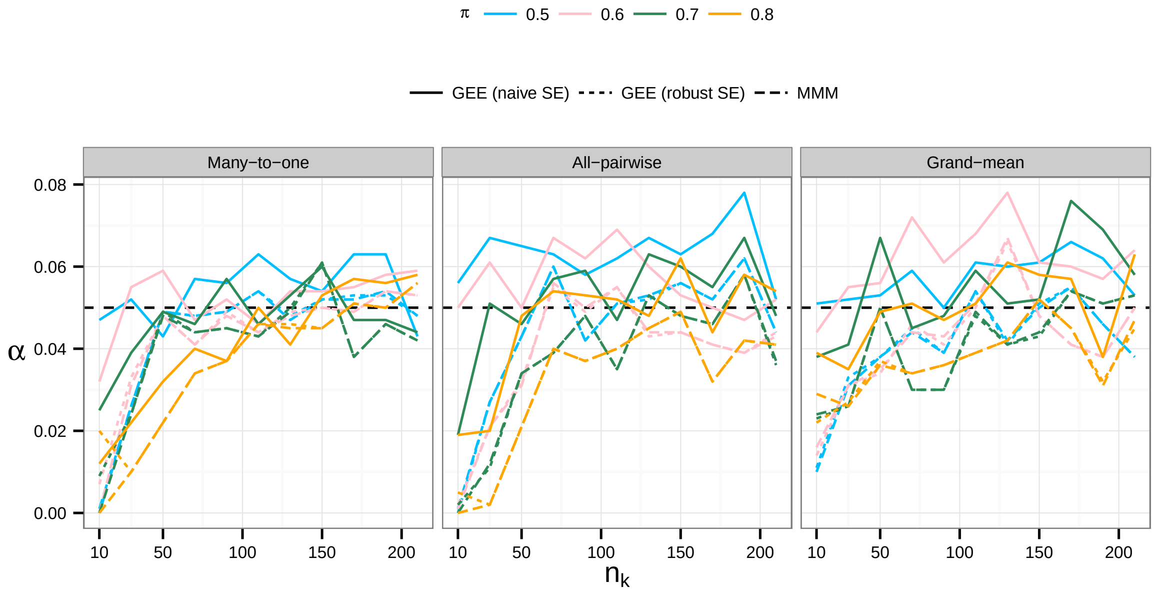


Figure 47: Simulated type I error rates for asymptotic many-to-one, all-pairwise, and grand-mean comparisons of binomial data ($\pi = \{0.5, 0.6, 0.7, 0.8\}$) among $q = 3$ treatment means separately and simultaneously at $m = 3$ occasions and among $m = 3$ occasion means separately and simultaneously for $q = 3$ treatments, with n_k independent subjects per treatment group, based on GEEs or multiple marginal GLMs (1000 simulation runs).

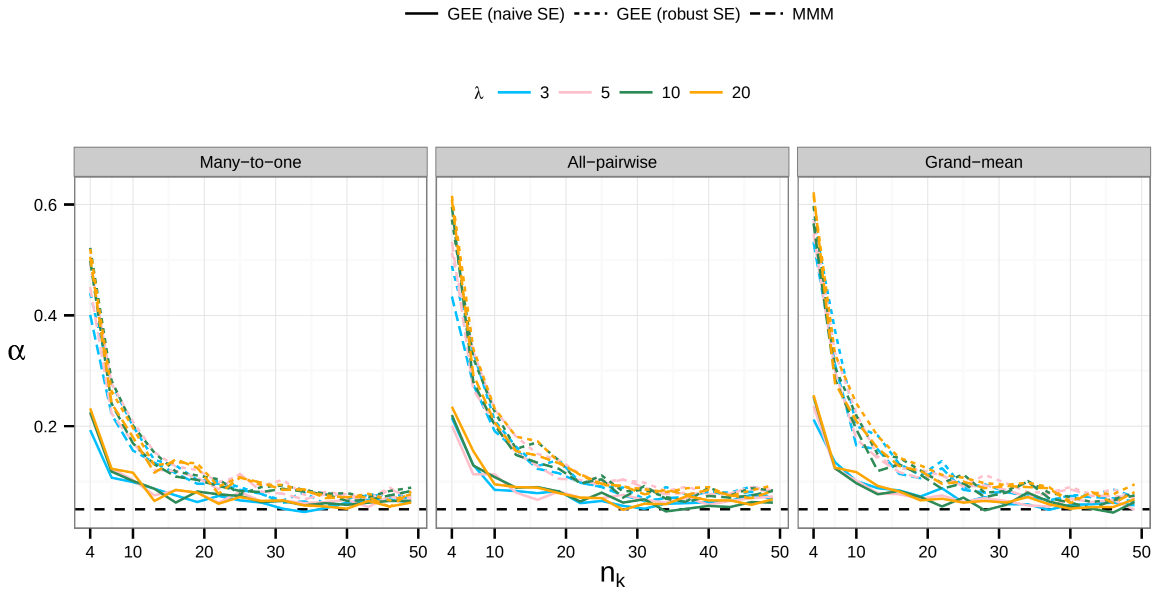


Figure 48: Simulated type I error rates for asymptotic many-to-one, all-pairwise, and grand-mean comparisons of Poisson data ($\lambda = \{3, 5, 10, 20\}$) among $q = 3$ treatment means separately and simultaneously at $m = 3$ occasions and among $m = 3$ occasion means separately and simultaneously for $q = 3$ treatments, with n_k independent subjects per treatment group, based on GEEs or multiple marginal GLMs (1000 simulation runs).

Power: We want to gauge the gain in power when uniting all comparisons in one family (or "claim") compared to performing separate longitudinal MCTs for comparisons among treatments and comparisons among occasions and then adjusting with Bonferroni (i.e., multiplying the p-values by two). The resulting power curves are shown in Figure 49 for binomial and in Figure 50 for Poisson data. The power advantage of the joint analysis (all comparisons in the same test family) is never more than 2 to 3 percentage points, and the powers of the GEE- and MMM-based analyses are practically the same.

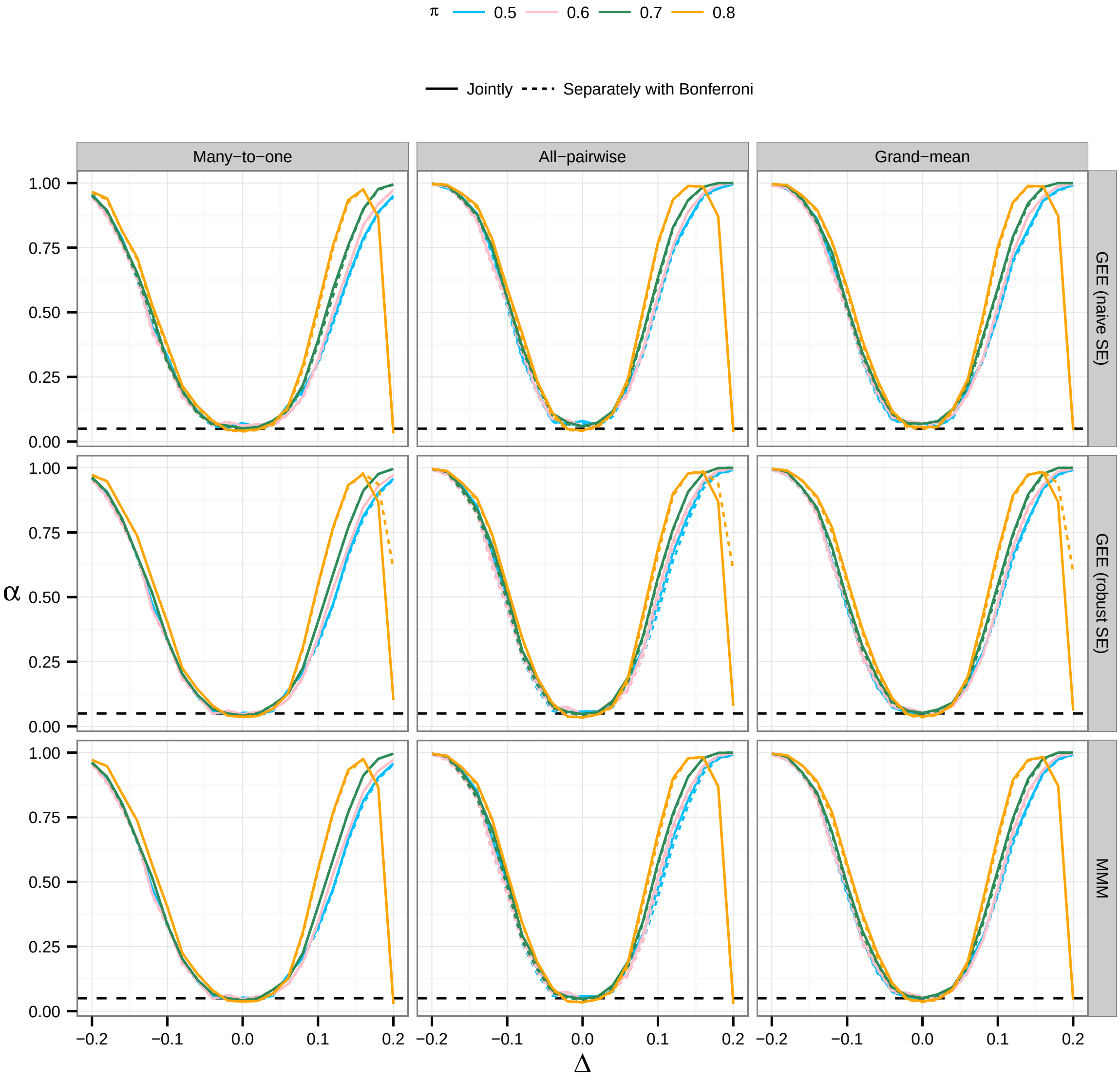


Figure 49: Simulated powers for asymptotic many-to-one, all-pairwise, and grand-mean comparisons of binomial data ($\pi = \{0.5, 0.6, 0.7, 0.8\}$) among $q = 3$ treatment means separately and simultaneously at $m = 3$ occasions and among $m = 3$ occasion means separately and simultaneously for $q = 3$ treatments, with $n_k = 100$ independent subjects per treatment group, based on GEEs or multiple marginal GLMs (1000 simulation runs).

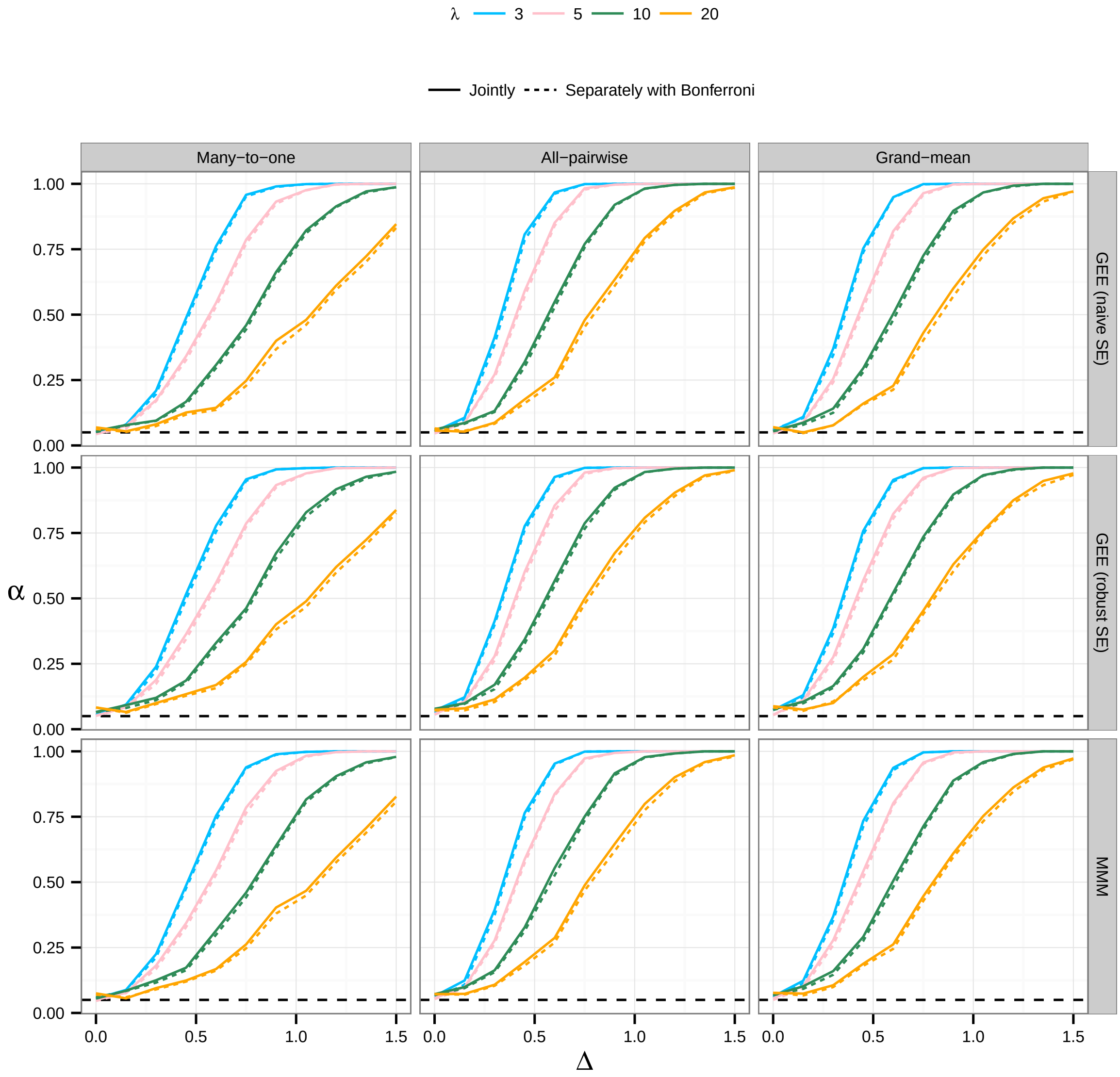


Figure 50: Simulated powers for asymptotic many-to-one, all-pairwise, and grand-mean comparisons of Poisson data ($\lambda = \{3, 5, 10, 20\}$) among $q = 3$ treatment means separately and simultaneously at $m = 3$ occasions and among $m = 3$ occasion means separately and simultaneously for $q = 3$ treatments, with $n_k = 50$ independent subjects per treatment group, based on GEEs or multiple marginal GLMs (1000 simulation runs).

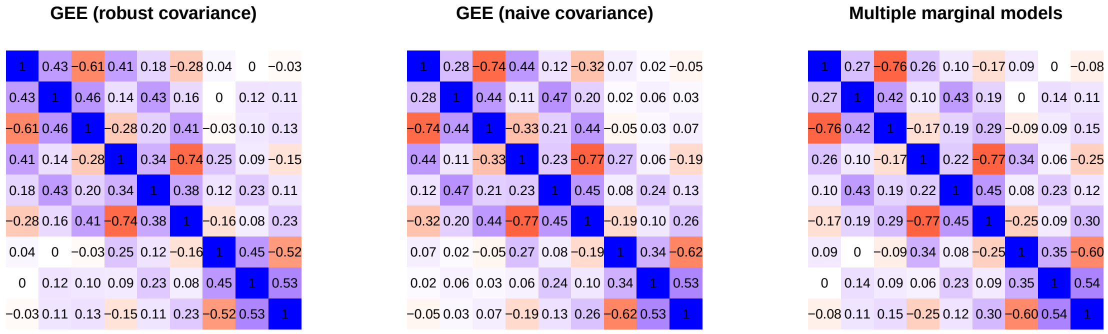
5.4 Application to Example Data

We illustrate longitudinal MCTs for discrete outcomes with the count data of greenhouse whiteflies and epileptic seizures that we introduced in 2.4 and 2.5 and with the proportions of larval and pupal mortalities under azadirachtin treatment as exposed in 2.6.

5.4.1 Greenhouse Whiteflies

We evaluate the greenhouse whitefly data of 2.4 with longitudinal MCTs for discrete endpoints. The sample sizes are large enough to justify the asymptotic method. A reasonable distributional assumption is Poisson: the outcome is measured as counts, and the data are skewed to the right.

Comparing multiple environments simultaneously for multiple plant parts: We want to investigate whether the type of microclimatic environment (glasshouse or one of two types of foil tunnels) makes an impact on how many whiteflies are found at the bottom, middle, and top of the plants after introduction of the predator *Macrolophus pygmaeus*. To this end we perform all-pairwise comparisons of the environments separately and simultaneously for each plant part.



GEE (robust covariance)

1	0.43	−0.61	0.41	0.18	−0.28	0.04	0	−0.03
0.43	1	0.46	0.14	0.43	0.16	0	0.12	0.11
−0.61	0.46	1	−0.28	0.20	0.41	−0.03	0.10	0.13
0.41	0.14	−0.28	1	0.34	−0.74	0.25	0.09	−0.15
0.18	0.43	0.20	0.34	1	0.38	0.12	0.23	0.11
−0.28	0.16	0.41	−0.74	0.38	1	−0.16	0.08	0.23
0.04	0	−0.03	0.25	0.12	−0.16	1	0.45	−0.52
0	0.12	0.10	0.09	0.23	0.08	0.45	1	0.53
−0.03	0.11	0.13	−0.15	0.11	0.23	−0.52	0.53	1

GEE (naive covariance)

1	0.28	−0.74	0.44	0.12	−0.32	0.07	0.02	−0.05
0.28	1	0.44	0.11	0.47	0.20	0.02	0.06	0.03
−0.74	0.44	1	−0.33	0.21	0.44	−0.05	0.03	0.07
0.44	0.11	−0.33	1	0.23	−0.77	0.27	0.06	−0.19
0.12	0.47	0.21	0.23	1	0.45	0.08	0.24	0.13
−0.32	0.20	0.44	−0.77	0.45	1	−0.19	0.10	0.26
0.07	0.02	−0.05	0.27	0.08	−0.19	1	0.34	−0.62
0.02	0.06	0.03	0.06	0.24	0.10	0.34	1	0.53
−0.05	0.03	0.07	−0.19	0.13	0.26	−0.62	0.53	1

Multiple marginal models

1	0.27	−0.76	0.26	0.10	−0.17	0.09	0	−0.08
0.27	1	0.42	0.10	0.43	0.19	0	0.14	0.11
−0.76	0.42	1	−0.17	0.19	0.29	−0.09	0.09	0.15
0.26	0.10	−0.17	1	0.22	−0.77	0.34	0.06	−0.25
0.10	0.43	0.19	0.22	1	0.45	0.08	0.23	0.12
−0.17	0.19	0.29	−0.77	0.45	1	−0.25	0.09	0.30
0.09	0	−0.09	0.34	0.08	−0.25	1	0.35	−0.60
0	0.14	0.09	0.06	0.23	0.09	0.35	1	0.54
−0.08	0.11	0.15	−0.25	0.12	0.30	−0.60	0.54	1

Figure 51: Greenhouse whitefly data: correlation matrices of test statistics for all-pairwise comparisons of environments per part of the plant.

One strategy is to fit a joint GEE with Poisson assumption, logarithmic link function, and an unstructured working correlation matrix; the other one is to fit separate quasi-Poisson GLMs for each of the plant parts. The estimated overdispersion parameters are 8.32, 9.21, and 4.71, so the quasi-likelihood analysis is obviously justified. Figure 51 shows the estimated correlations based on the joint model and MMM, which are all very similar.

Table 10: Simultaneous inference for the greenhouse whitefly data: estimated differences of log-numbers of whiteflies, standard errors, and adjusted p-values for part-wise all-pairwise comparisons of environments.

	Est(GEE)	Est(MMM)	SE(robust)	SE (naive)	SE(MMM)	p(robust)	p(naive)	p(MMM)	p(Bonf)
Bottom: Tunnel 1 - Glasshouse	-0.398	-0.453	0.207	0.236	0.262	0.331	0.478	0.455	0.574
Bottom: Tunnel 2 - Glasshouse	0.161	0.161	0.185	0.176	0.188	0.953	0.938	0.956	1.000
Bottom: Tunnel 2 - Tunnel 1	0.559	0.614	0.210	0.252	0.279	0.061	0.180	0.189	0.208
Middle: Tunnel 1 - Glasshouse	-0.841	-0.834	0.277	0.269	0.304	0.020	0.014	0.047	0.048
Middle: Tunnel 2 - Glasshouse	-0.084	-0.084	0.200	0.193	0.217	0.999	0.999	0.999	1.000
Middle: Tunnel 2 - Tunnel 1	0.757	0.750	0.282	0.293	0.331	0.057	0.073	0.162	0.175
Top: Tunnel 1 - Glasshouse	-0.030	0.043	0.619	0.839	0.661	1.000	1.000	1.000	1.000
Top: Tunnel 2 - Glasshouse	0.141	0.141	0.624	0.776	0.625	1.000	1.000	1.000	1.000
Top: Tunnel 2 - Tunnel 1	0.171	0.098	0.654	0.931	0.736	1.000	1.000	1.000	1.000

Results of asymptotic all-pairwise comparisons of environments are presented in Table 10. We know there are a few values missing for foil tunnel 1; this is why the estimates differ between GEE and MMM for all comparisons that involve tunnel 1. The robust and naive SEs from the GEE and those from the MMM approach are mostly similar, but a bit higher with MMM for the bottom and middle parts, and with the naive GEE approach for the top part. Overall the SEs associated with measurements from the top part are around three times as large as those from the bottom and middle of the plant, which is due to the low counts with many zeroes.

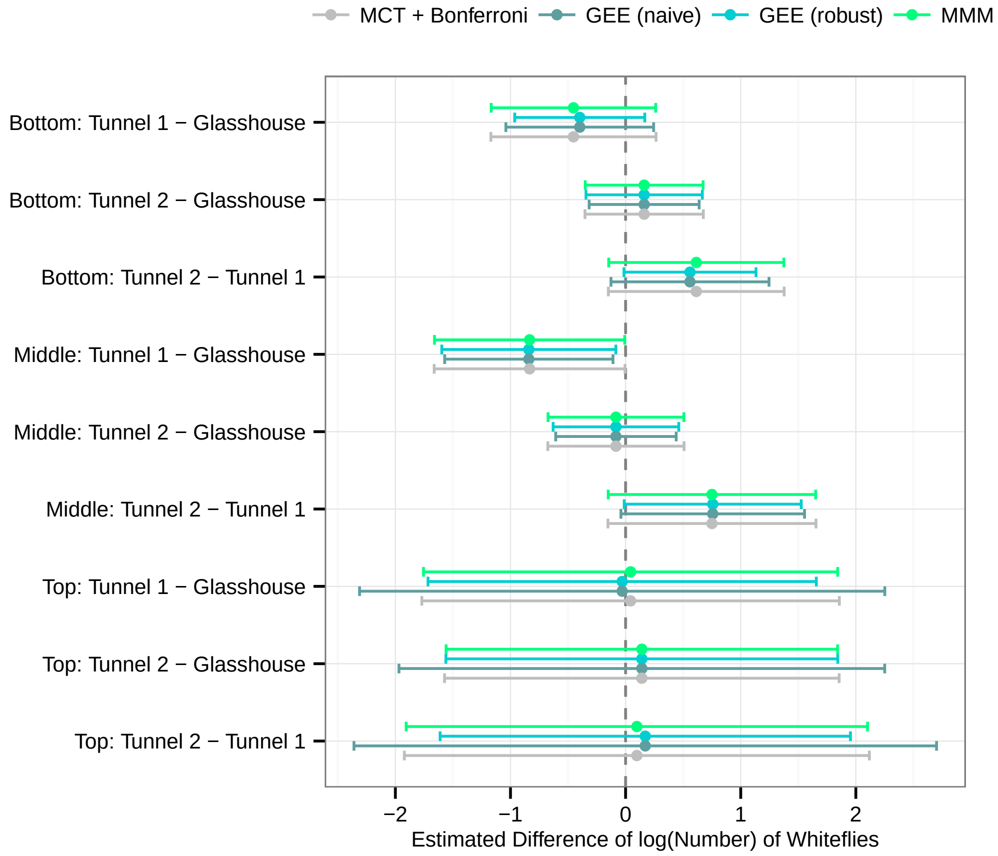
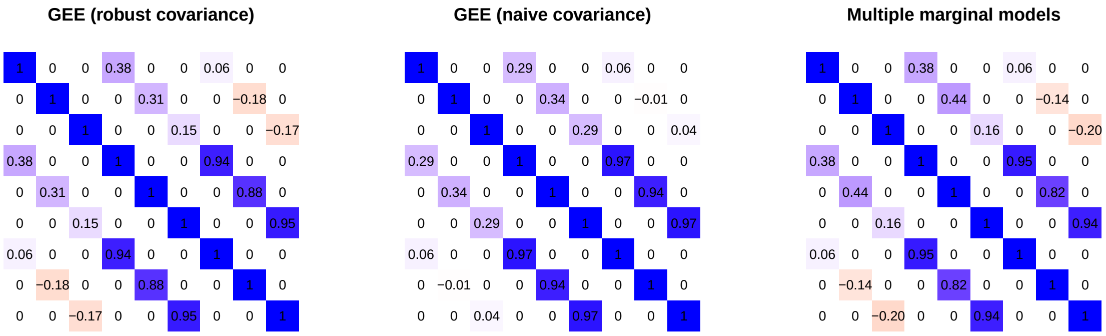


Figure 52: Greenhouse whitefly data: 95% simultaneous confidence intervals for asymptotic all-pairwise comparisons of environments per plant part based on a joint GEE (with robust or naive covariance estimation), combined marginal GLMs (MMM), and Bonferroni-adjusted MCTs per plant part.

In the environment of tunnel 1 infestation with whiteflies at the middle of the plant is significantly reduced in comparison to the glasshouse; the upper boundaries of the SCIs are, however, very close to the point no effect (Figure 52), implying that the difference might not be seen as biologically important. There is also some indication that tunnel 1 is better than tunnel 2 in reducing infestation at the bottom and middle, although these comparisons are not significant at the familywise 5% level. For the top of the plant all effect estimates are close to 0, with very wide SCIs around them. The p-values obtained from separate MCTs per plant part that were adjusted with Bonferroni for the multiplicity of measurements are always larger than those from the GEE with robust SE estimates and MMM, and correspondingly the SCIs are always wider.

Comparing multiple plant parts simultaneously in multiple environments: Next we want compare the mean numbers of whiteflies at the different plant parts separately and simultaneously for each type of environment. The relevant correlation matrices of test statistics are depicted in Figure 53, and we see clearly that the bottom and middle parts are highly correlated with each other but not with the top.

Table 11 summarizes the results of the analysis using all-pairwise comparisons. Again,



GEE (robust covariance)

1	0	0	0.38	0	0	0.06	0	0
0	1	0	0	0.31	0	0	−0.18	0
0	0	1	0	0	0.15	0	0	−0.17
0.38	0	0	1	0	0	0.94	0	0
0	0.31	0	0	1	0	0	0.88	0
0	0	0.15	0	0	1	0	0	0.95
0.06	0	0	0.94	0	0	1	0	0
0	−0.18	0	0	0.88	0	0	1	0
0	0	−0.17	0	0	0.95	0	0	1

GEE (naive covariance)

1	0	0	0.29	0	0	0.06	0	0
0	1	0	0	0.34	0	0	−0.01	0
0	0	1	0	0	0.29	0	0	0.04
0.29	0	0	1	0	0	0.97	0	0
0	0.34	0	0	1	0	0	0.94	0
0	0	0.29	0	0	1	0	0	0.97
0.06	0	0	0.97	0	0	1	0	0
0	−0.01	0	0	0.94	0	0	1	0
0	0	0.04	0	0	0.97	0	0	1

Multiple marginal models

1	0	0	0.38	0	0	0.06	0	0
0	1	0	0	0.44	0	0	−0.14	0
0	0	1	0	0	0.16	0	0	−0.20
0.38	0	0	1	0	0	0.95	0	0
0	0.44	0	0	1	0	0	0.82	0
0	0	0.16	0	0	1	0	0	0.94
0.06	0	0	0.95	0	0	1	0	0
0	−0.14	0	0	0.82	0	0	1	0
0	0	−0.20	0	0	0.94	0	0	1

Figure 53: Greenhouse whitefly data: correlation matrices of test statistics for all-pairwise comparisons of plant parts per environment.

the comparisons involving tunnel 1 have slightly different estimates with GEE and MMM. From the SEs in Table 11 and the SCIs in Figure 54 we can tell that the differences between middle and bottom parts are estimated much more precisely than their differences to the top parts. Numbers of whiteflies do not differ significantly between the middle and the bottom in any microclimatic environment.

Table 11: Simultaneous inference for the greenhouse whitefly data: estimated differences of log-numbers of whiteflies, standard errors, and adjusted p-values for environment-wise all-pairwise comparisons of plant parts.

	Est(GEE)	Est(MMM)	SE(robust)	SE (naive)	SE(MMM)	p(robust)	p(naive)	p(MMM)	p(Bonf)
Glasshouse: Middle - Bottom	-0.021	-0.021	0.143	0.111	0.127	1.000	1.000	1.000	1.000
Tunnel 1: Middle - Bottom	-0.463	-0.402	0.229	0.245	0.322	0.262	0.328	0.790	0.900
Tunnel 2: Middle - Bottom	-0.266	-0.266	0.149	0.155	0.177	0.409	0.445	0.613	0.822
Glasshouse: Top - Bottom	-2.957	-2.957	0.434	0.475	0.394	<0.001	<0.001	<0.001	<0.001
Tunnel 1: Top - Bottom	-2.589	-2.461	0.477	0.712	0.564	<0.001	0.002	<0.001	<0.001
Tunnel 2: Top - Bottom	-2.977	-2.977	0.457	0.626	0.489	<0.001	<0.001	<0.001	<0.001
Glasshouse: Top - Middle	-2.936	-2.936	0.402	0.455	0.365	<0.001	<0.001	<0.001	<0.001
Tunnel 1: Top - Middle	-2.126	-2.059	0.461	0.671	0.511	<0.001	0.011	<0.001	<0.001
Tunnel 2: Top - Middle	-2.712	-2.712	0.459	0.600	0.493	<0.001	<0.001	<0.001	<0.001

Comparing multiple environments and multiple plant parts simultaneously: We can combine the comparisons among environments and among plant parts into one claim and analyze it under joint control of α at 5%. The results of this analysis are summarized in Table 12, and the corresponding SCIs are plotted in Figure 55. Unsurprisingly, we find that the adjusted p-values are a bit higher and the SCIs a bit wider than before, but this does not change the conclusions fundamentally.

Given the large number of zero counts, in particular for observations from the top part of the plant, an alternative for this dataset could be a zero-inflated Poisson model.

5.4.2 Epileptic Seizures

We analyze the counts of epileptic seizures introduced in 2.5 with longitudinal MCTs for Poisson data.

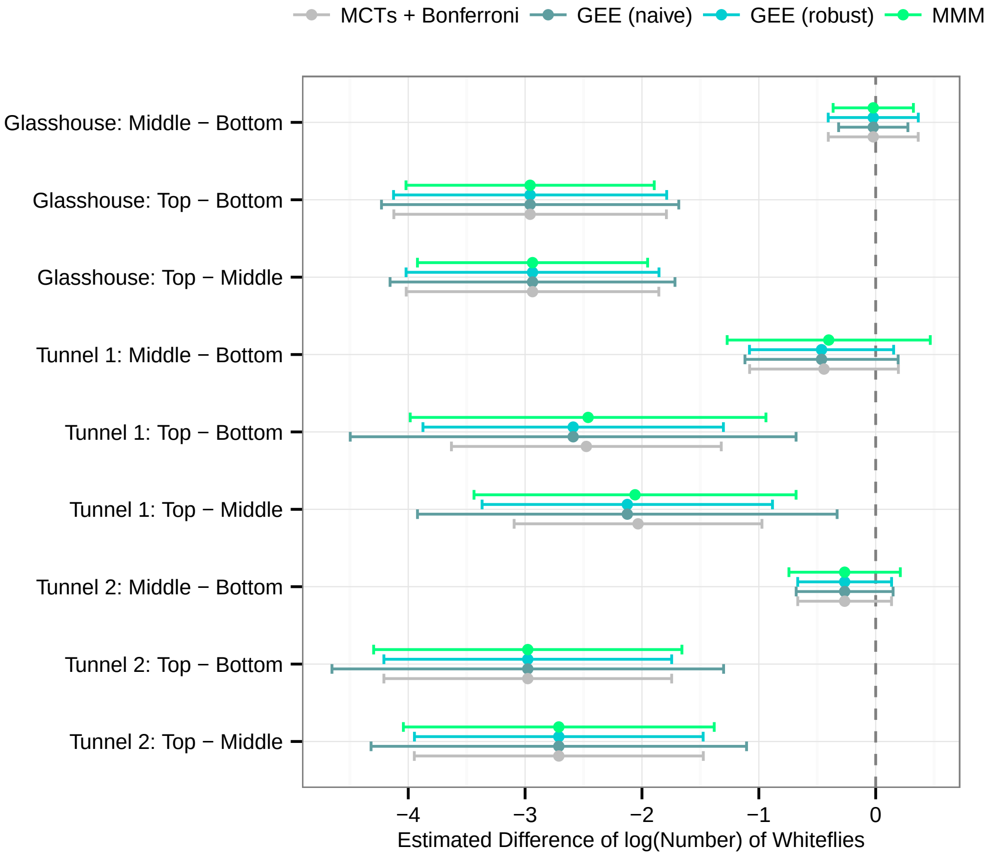


Figure 54: Greenhouse whitefly data: 95% simultaneous confidence intervals for asymptotic all-pairwise comparisons of plant parts per environment based on a joint GEE (with robust or naive covariance estimation), combined marginal GLMs (MMM), and Bonferroni-adjusted MCTs per environment.

Comparing active drug and placebo at multiple occasions: The question of interest is whether progabide accomplishes an improvement over placebo (i.e., fewer seizures) at any of the time points except baseline, and if so, at which one(s). We apply the longitudinal MCT procedures for count data introduced in 5.2 and rely on asymptotics, which should be acceptable given the sample sizes of 31 in the progabide group and 28 with placebo. This results in four two-sample comparisons of progabide versus placebo, one at each of the four measurement times. Since only a reduction compared to placebo would be an interesting effect, we carry out one-sided comparisons.

One option is GEE modeling with a Poisson assumption, a logarithmic link function for the number of seizures, AR(1) working correlation, and log(age) as a covariate. The alternative is marginal quasi-Poisson GLMs, also with a logarithmic link and log(age) as a covariate, separately for each time point. The GLMs reveal a substantial amount of overdispersion at all time points. Figure 56 shows that MMM estimates slightly lower correlation between time points than the GEE.

The test results are summarized in Table 13. The effect estimates from GEE and MMM are similar, but the SEs differ quite a bit: especially the naive SEs from the GEE are much more homogeneous than the robust ones and those from MMM. The conclusions

Table 12: Simultaneous inference for the greenhouse whitefly data: estimated differences of log-numbers of whiteflies, standard errors, and adjusted p-values for part-wise all-pairwise comparisons of environments and environment-wise all-pairwise comparisons of plant parts.

	Est(GEE)	Est(MMM)	SE(robust)	SE (naive)	SE(MMM)	p(robust)	p(naive)	p(MMM)
Bottom: Tunnel 1 - Glasshouse	-0.398	-0.453	0.207	0.236	0.262	0.473	0.639	0.617
Bottom: Tunnel 2 - Glasshouse	0.161	0.161	0.185	0.176	0.188	0.989	0.983	0.990
Bottom: Tunnel 2 - Tunnel 1	0.559	0.614	0.210	0.252	0.279	0.100	0.275	0.291
Middle: Tunnel 1 - Glasshouse	-0.841	-0.834	0.277	0.269	0.304	0.034	0.024	0.079
Middle: Tunnel 2 - Glasshouse	-0.084	-0.084	0.200	0.193	0.217	1.000	1.000	1.000
Middle: Tunnel 2 - Tunnel 1	0.757	0.750	0.282	0.293	0.331	0.094	0.119	0.253
Top: Tunnel 1 - Glasshouse	-0.030	0.043	0.619	0.839	0.661	1.000	1.000	1.000
Top: Tunnel 2 - Glasshouse	0.141	0.141	0.624	0.776	0.625	1.000	1.000	1.000
Top: Tunnel 2 - Tunnel 1	0.171	0.098	0.654	0.931	0.736	1.000	1.000	1.000
Glasshouse: Middle - Bottom	-0.021	-0.021	0.143	0.111	0.127	1.000	1.000	1.000
Tunnel 1: Middle - Bottom	-0.463	-0.402	0.229	0.245	0.322	0.402	0.486	0.905
Tunnel 2: Middle - Bottom	-0.266	-0.266	0.149	0.155	0.177	0.576	0.619	0.771
Glasshouse: Top - Bottom	-2.957	-2.957	0.434	0.475	0.394	<0.001	<0.001	<0.001
Tunnel 1: Top - Bottom	-2.589	-2.461	0.477	0.712	0.564	<0.001	0.004	<0.001
Tunnel 2: Top - Bottom	-2.977	-2.977	0.457	0.626	0.489	<0.001	<0.001	<0.001
Glasshouse: Top - Middle	-2.936	-2.936	0.402	0.455	0.365	<0.001	<0.001	<0.001
Tunnel 1: Top - Middle	-2.126	-2.059	0.461	0.671	0.511	<0.001	0.021	0.001
Tunnel 2: Top - Middle	-2.712	-2.712	0.459	0.600	0.493	<0.001	<0.001	<0.001

about the treatment effect, however, are the same with all methods: progabide leads to no relevant and no significant improvement over placebo at any of the four time points. This becomes also clear from the upper SCI bounds in Figure 57, which are all on the “wrong” side, with point estimates close to 0 (on the logarithmic scale). Even though we could not detect any “significant” effects, at least our longitudinal MCTs enable us to quantify the uncertainty around the estimates.

Table 13: Simultaneous inference for the epileptic seizures data: estimated differences of log-rates of seizures, standard errors, and adjusted p-values for occasion-wise comparisons of progabide and placebo (except baseline).

	Est(GEE)	Est(MMM)	SE(robust)	SE (naive)	SE(MMM)	p(robust)	p(naive)	p(MMM)	p(Bonf)
2 weeks	-0.090	-0.089	0.406	0.371	0.266	0.588	0.566	0.592	1.000
4 weeks	0.013	-0.006	0.293	0.384	0.319	0.692	0.676	0.718	1.000
6 weeks	-0.081	-0.087	0.419	0.382	0.486	0.599	0.579	0.656	1.000
8 weeks	-0.175	-0.184	0.327	0.410	0.313	0.458	0.490	0.483	1.000

5.4.3 Azadirachtin

We analyze the azadirachtin data presented in 2.6 with a view to comparing dose effects on both larval and pupal mortality of whiteflies. In particular, we seek to compare the mortalities associated with the different doses separately and simultaneously for both developmental stages. Once more we pursue either of two modeling strategies: a combination of two stage-specific binomial GLMs for larval and pupal mortalities, or alternatively a joint GEE approach. In addition we illustrate the parameterization issue with MMM that was described in 3.5.2.

Comparing multiple doses at multiple stages: We fit a joint binomial model for numbers of dead and alive whitefly larvae and pupae with GEE using the logit link and an AR(1) working correlation. Our independent variables are the combinations of developmental stages and dose levels, the culture substrates, and their interaction. The single plants are taken as clusters. A Wald-type test for sequentially added model terms

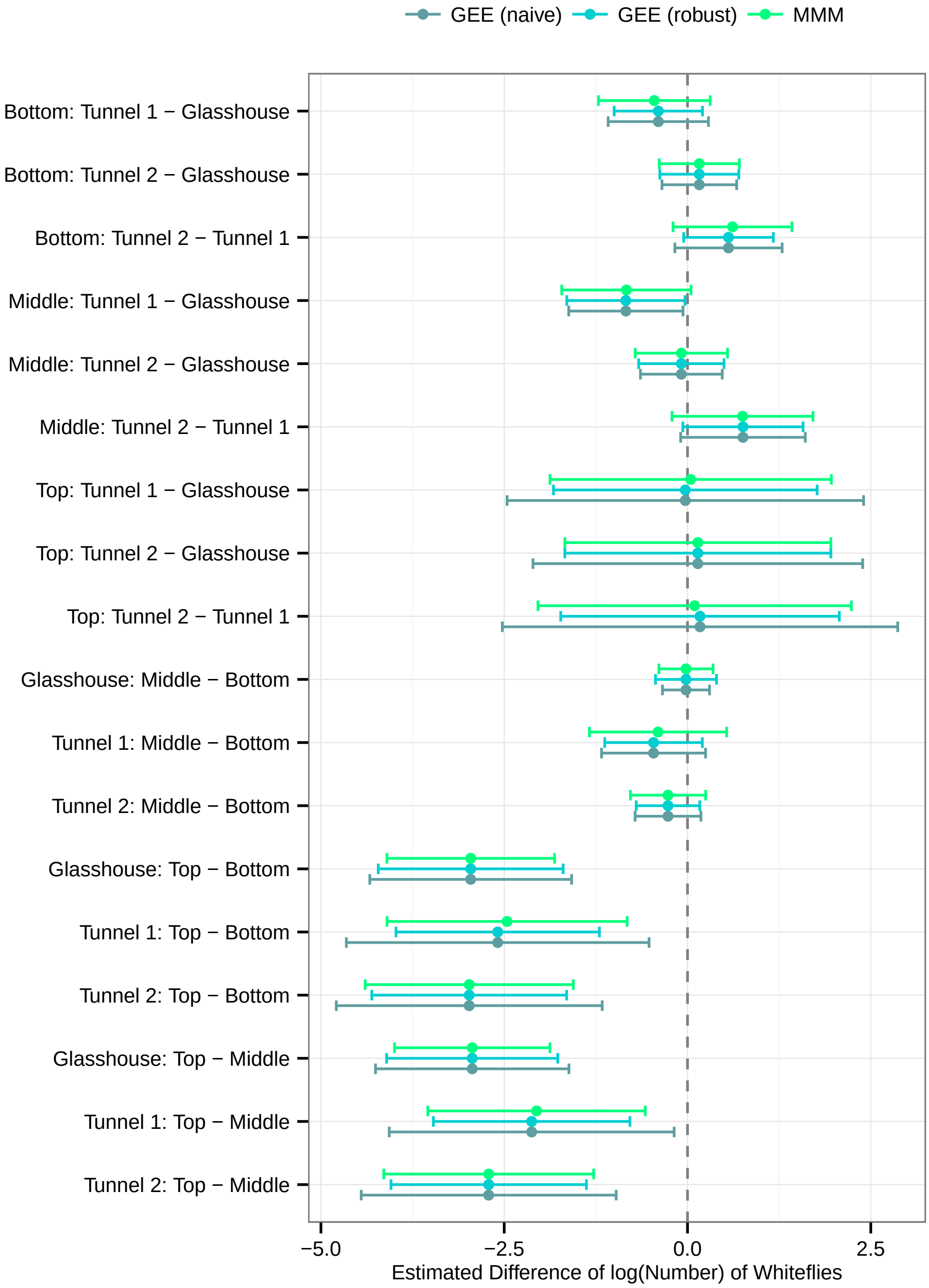


Figure 55: Greenhouse whitefly data: 95% simultaneous confidence intervals for asymptotic all-pairwise comparisons of environments per plant part and all-pairwise comparisons of plant parts per environment simultaneously based on a joint GEE (with robust or naive covariance estimation) and combined marginal GLMs (MMM).

yields a p-value of 0.560 for the interaction term, thus we can leave it out with confidence, and the model is simplified to main effects only.

In an alternative approach, we fit two separate quasibinomial GLMs with logit link function and overdispersion parameter ϕ to the larval and pupal data, respectively. The models include main effects for dose and substrate as well as an interaction term of the

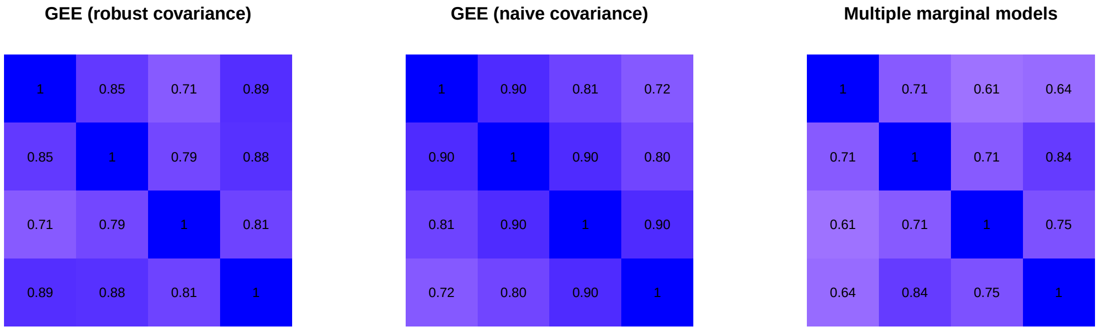


Figure 56: Epileptic seizures data: correlation matrices of test statistics for many-to-one comparisons of treatments per time point (except baseline).

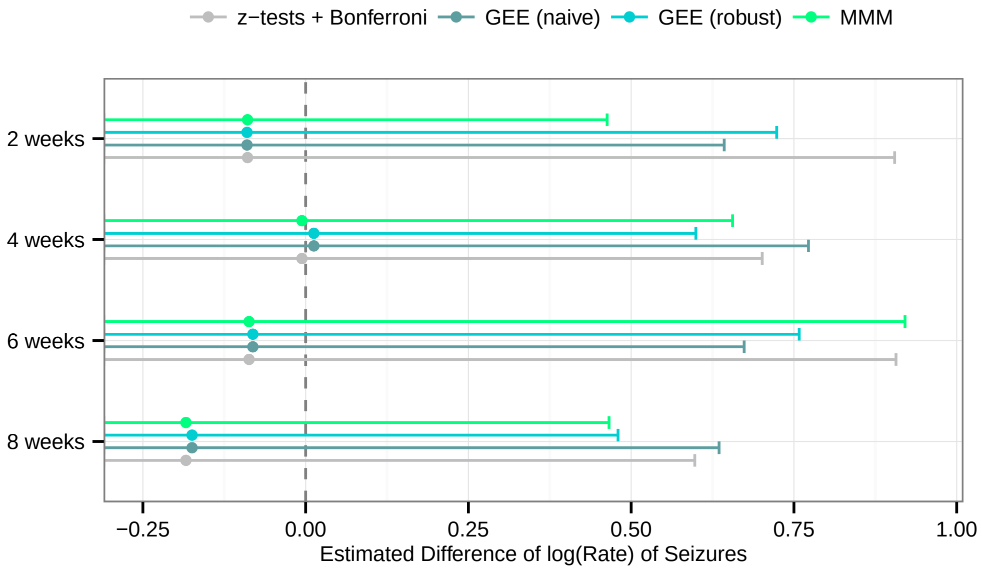


Figure 57: Epileptic seizures data: 95% simultaneous confidence intervals for asymptotic comparisons of treatments per time point (except baseline) based on a joint GEE (with robust or naive covariance estimation), combined marginal GLMs (MMM), and Bonferroni-adjusted z-tests per time point.

two. The interactions turn out to be nonsignificant in the analysis-of-deviance F-tests (p-values of 0.683 and 0.364) and are thus removed to simplify the models. The main effect of substrate is significant with both modeling strategies, but having said that, we focus our interpretation on the comparisons of dose levels separately and simultaneously for larvae and pupae.

The estimated mortality differences on the logit link (i.e., log odds ratios) for pairwise comparisons of neem doses separately for larvae and pupae are displayed in Table 14 along with SE estimates and adjusted p-values. When comparing GEE and MMM, the estimated differences, SEs, and p-values are just slightly different for larval mortality but there is substantial disagreement for pupae.

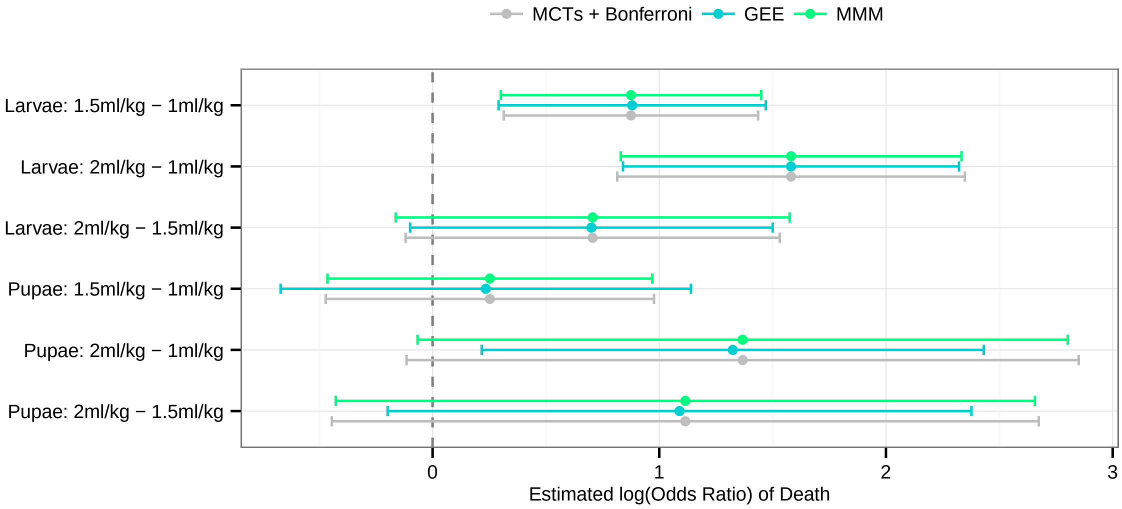


Figure 58: Azadirachtin data: 95% simultaneous confidence intervals for asymptotic all-pairwise comparisons of doses per developmental stage based on a joint GEE (with robust covariance estimation), combined marginal GLMs (MMM), and Bonferroni-adjusted MCTs per stage.

We observe that uncertainty of estimation is considerably larger for pupae than larvae, and it also tends to increase with dosage. As concerns larval mortality, both 1.5 and 2 ml/kg are found to be significantly superior to the manufacturer-recommended dose of 1 ml/kg ($p < 0.001$), and the SCI boundaries are far away from the point of no effect. By contrast, such clear effects cannot be detected for pupal mortality: only the difference between 1 and 2 ml/kg is significant—and only with the GEE-based procedure ($p = 0.011$) but not with MMM ($p = 0.069$). We see this also from the corresponding 95% SCIs in Figure 58: the discrepancy between methods is minor for comparisons of larval mortality but noticeable for comparisons of pupal mortality. One explanation for this may be that sample sizes are too small for the longitudinal MCTs. Nonetheless, the correlation matrices of test statistics look overall very similar (Figure 59).

Table 14: Simultaneous inference for the azadirachtin data: estimated log odds ratios of death, standard errors, and adjusted p-values for all-pairwise stage-wise comparisons of dose levels.

	Est(GEE)	Est(MMM)	SE(GEE)	SE(MMM)	p(GEE)	p(MMM)	p(Bonf)
Larvae: 1.5 vs. 1 ml/kg	0.880	0.875	0.228	0.223	<0.001	<0.001	<0.001
Larvae: 2 vs. 1 ml/kg	1.581	1.582	0.287	0.292	<0.001	<0.001	<0.001
Larvae: 2 vs. 1.5 ml/kg	0.701	0.706	0.310	0.338	0.113	0.163	0.132
Pupae: 1.5 vs. 1 ml/kg	0.235	0.253	0.350	0.279	0.947	0.853	1.000
Pupae: 2 vs. 1 ml/kg	1.324	1.368	0.429	0.558	0.011	0.068	0.085
Pupae: 2 vs. 1.5 ml/kg	1.089	1.115	0.499	0.600	0.135	0.260	0.293

Parameterization with multiple marginal GLMs: We have broached in 3.5.2 that parameterization of the marginal models is an issue with finite sample sizes and can have substantial impact on the covariance and SE estimates because the observed Fisher information may be widely different from the expected one due to estimation problems.

We can parameterize the quasi-binomial GLMs for larval and pupal mortalities in various ways. Table 15 lists four parameterizations of the stage-specific GLMs along with

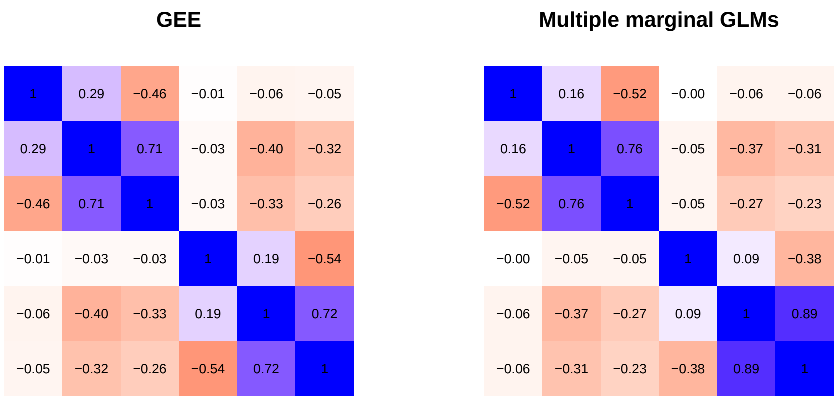


Figure 59: Azadirachtin data: correlation matrices of test statistics for all-pairwise comparisons of dose levels per developmental stage.

Table 15: Simultaneous inference for the azadirachtin data: four different options of parameterizing the quasi-binomial GLMs for larval and pupal mortalities. The effects are log odds of death, and differences are on the logit scale.

	β_1	β_2	β_3	β_4
a)	effect with 1 ml/kg	difference of 1.5 vs. 1 ml/kg	difference of 2 vs. 1 ml/kg	effect of sand
b)	effect with 1.5 ml/kg	difference of 1.5 vs. 1 ml/kg	difference of 2 vs. 1.5 ml/kg	effect of sand
c)	effect with 2 ml/kg	difference of 2 vs. 1 ml/kg	difference of 2 vs. 1.5 ml/kg	effect of sand
d)	effect with 1 ml/kg	effect with 1.5 ml/kg	effect with 2 ml/kg	effect of sand

the interpretations of the single parameters. Our longitudinal MCT procedure can be conducted with any of these parameterizations, but they lead to different results for the azadirachtin data.

Table 16: Simultaneous inference for the azadirachtin data: standard error estimates and adjusted p-values with different parameterizations of the quasi-binomial GLMs in the MMM approach for larval and pupal mortalities.

		a)		b), d)		c)	
Stage	Comparison	SE	p	SE	p	SE	p
Larvae	1.5 vs. 1 ml/kg	0.223	< 0.001	0.216	< 0.001	0.237	0.001
	2 vs. 1 ml/kg	0.292	< 0.001	0.290	< 0.001	0.296	< 0.001
	2 vs. 1.5 ml/kg	0.338	0.163	0.318	0.121	0.318	0.127
Pupae	1.5 vs. 1 ml/kg	0.279	0.853	0.282	0.856	0.442	0.970
	2 vs. 1 ml/kg	0.558	0.068	0.510	0.037	0.576	0.087
	2 vs. 1.5 ml/kg	0.600	0.261	0.606	0.268	0.606	0.279

The SE estimates and adjusted p-values associated with the treatment differences of interest are printed in Table 16. Note that they are identical for parameterizations b) and d), which we also used for Table 14, but considerably different from a) and c). This should make clear that SE estimates and p-values from multiple marginal GLMs are to be treated with caution as the experiment's sample size is probably insufficient.

6 Extensions and Alternatives

The methods studied in this thesis can be expanded in various directions, and we spotlight some ideas here: using ratios rather than differences of means (6.1); including further sources of "repeatedness" e.g., multiple endpoints in addition to repeated measurements over time (6.2); and employing profile likelihood test statistics for discrete data to better cope with small sample sizes (6.3). As possible alternatives to our methods, we discuss in brief multiple contrast rotation tests to speed up computation with larger dimensions (6.4), and nonparametric rank tests for longitudinal setups (6.5).

6.1 Ratios

It may be desirable for various applications to estimate effects and SCIs on a percentage scale. To this end our longitudinal MCTs could be modified so that SCIs for *ratios* instead of *differences* of treatments means per occasion, or occasion means per treatment, are computed.

A generic method for constructing a CI around a ratio of normal variates is suggested by the theorem of Fieller (1954)[5]. Based on this result, Zerbe et al. (1982) and Young et al. (1997) developed SCIs for fixed- and mixed-effects model parameters, respectively. Dilba et al. (2004 2006) extended these works to MCTs for ratios so that simultaneous tests and SCIs are available for arbitrary sets of linear contrasts involving ratios.

Ratio-type MCTs for multivariate Gaussian data were considered in Hasler and Hothorn (2012) and Hasler and Böhlendorf (2013). In similar fashion it should be a simple task to put the puzzle together and build a longitudinal MCT for Gaussian endpoints that yields SCIs for ratios of interest, in the framework of a joint LMM or multiple occasion-specific linear models.

6.2 Multivariate Longitudinal Data

As already hinted at in the introduction, there may be "more repeatedness" in the data than just longitudinal replication. For instance, several endpoints could be measured repeatedly over time from the same experimental units. This calls for a model that captures correlation both across endpoints and repeated measurements to be the basis for simultaneous inferences of endpoints and/or time points. Bandyopadhyay et al. (2011) and Verbeke et al. (2014) review joint modeling options for such multivariate longitudinal data. As an alternative, a recently published extension of the combination of multiple models (Jensen et al. 2015) allows to extend the method to marginal models with random effects i.e., one LMM for every endpoint.

[5] This theorem appeared in the literature way earlier than 1954 e.g., in Fieller (1940) and Fieller (1944).

6.3 Signed Likelihood Root Tests

We have observed rather poor performance of our longitudinal MCTs for discrete data in the presence of small sample sizes. This is not surprising due to the method being inherently asymptotic, but still accurate multiple tests and SCIs for linear combinations of GLM parameters might be required in practice from time to time.

Recently Gerhard (2014b) proposed replacing the common Wald-type test statistics in MCTs with profile likelihood statistics in order to achieve nicer small-sample properties, especially close to the borders of the parameter space. The basic idea is to use a signed root deviance statistic (Chen and Jennrich 1996) or alternatively a modified likelihood root relying on a higher order density approximation (Barndorff-Nielsen 1983), details on both of which are given in Brazzale and Davison (2008). It could be a worthwhile expansion of our longitudinal MCTs for discrete outcomes to use profile statistics.

6.4 Rotation Tests

Rotation tests are a flexible alternative for simultaneous inference based on a multivariate model or on multiple marginal models. Similar to resampling methods such as bootstrapping (Westfall and Young 1993), adjusted p-values and SCIs are found via Monte Carlo simulation, with the orthogonalized residuals being randomly *rotated* rather than *permuted* a large number of times (Langsrud 2005). Like with the MMM approach of 3.5, no specification of the covariance is required.

Rotation tests have recently gained attention for applications in high-dimensional gene expression analysis where the $n \ll p$ problem occurs (Dørum et al. 2009; Wu et al. 2010; Dørum et al. 2014; Solari et al. 2014) but are also useful in smaller-scale scenarios (Langsrud et al. 2007; Gerhard and Schaarschmidt 2015). In our specific setting of outcomes measured repeatedly over time, each occasion would be modeled separately and then the residuals could be combined for the rotation procedure. Extensions that use GLMs or LMMs for the marginal outcomes are available as well (Gerhard and Schaarschmidt 2015).

6.5 Nonparametric Rank Tests

Testing differences of means in correlated longitudinal data settings has been a research topic in the nonparametric statistics community for at least two decades. Most of their rank-based methods take account of the nonparametric Behrens-Fisher problem; this is sensible not only because repeated measures over time are often heteroscedastic themselves, but even if the original data have equal variances across time points, the ranks do not (Akritas 1990).

Konietschke et al. (2010) proposed MCTs involving relative effects p_j for comparing means of different time points $j = 1, \ldots, m$ in one treatment group. These p_j can be estimated using the ranks of the original observations, and they have interpretation as the probability that a randomly chosen measurement taken at time point j is larger than a randomly chosen measurement from the mean distribution. The null hypothesis

$$H_0 : \bigcap_{j=1}^{m} \{p_j = 0.5\}$$

can be specified in MCT representation as

$$H_0 : \mathbf{Cp} = \mathbf{0}$$

where $\mathbf{C}$ is an adequate contrast matrix, and $\mathbf{p} = (p_1, \ldots, p_m)^T$. This method yields threefold information just like classical MCTs: a global decision, adjusted p-values for multiple comparisons, and compatible SCIs. Unfortunately, it is slightly liberal as long as sample sizes are not large (type I error rates of 0.07 with $n = 10$ and 0.06 with $n = 20$).

A similar nonparametric strategy for one or two treatment groups was proposed before by Munzel and Tamhane (2002); however, as the authors pointed out, their method applies to large samples only and does not allow for comparisons among groups. The hypotheses of most other nonparametric procedures for (heteroscedastic) longitudinal data (e.g., Brunner and Langer 2000; Brunner and Puri 2001) contain marginal distribution functions as introduced by Akritas and Arnold (1994) instead of the relative effects $\mathbf{p}$ and hence do not offer SCIs for the p_i which limits their practical applicability. More on nonparametric inference in factorial longitudinal designs can be found in Brunner et al. (2002).

7 Discussion

In this thesis we proposed different approaches for simultaneous inference in longitudinal scenarios, evaluated their small- and finite-sample performances via numerical simulation, and exemplified their use with data from various areas of research in the life sciences. The methods are an extension and application of MCTs to settings where randomized units are measured repeatedly over time. This is useful whenever the research question of interest can be answered with *separate and simultaneous* comparisons of treatment groups and/or time points and FWER control over the entire set of comparisons is desirable. In a broader sense, the methods of this thesis can be applied whenever multiple comparisons are demanded in a setting with correlation such as longitudinal or spatial data, multiple endpoints, etc. MCT methods for correlated and repeated measurements were available before (Hasler and Hothorn 2011; Hasler 2013), but their lack of flexibility makes them inapplicable to many realistic longitudinal scenarios.

None of the test procedures and SCI methods we considered in this work is “exact” because we have to plug in the estimated covariance matrix $\widehat{\mathbf{\Sigma}}$ when calculating the MCTs. The MMM approach has an additional asymptotic element: the observed information matrix is used instead of the expected one. We gauged via simulation what sample sizes are required for the asymptotic procedures to achieve good control of the FWER, and they are usually in the order of a few dozens (see also Pallmann et al. 2015). Such sample sizes are not always feasible in real-world experiments or trials, therefore approximate small-sample methods are practically relevant. However, they come with the additional complication of having to approximate the DF of the multivariate t-distribution.

We do not want to create the impression that any approximation to the DF is “correct” in general. Our simulations show that various DF methods work well over a range of different situations. Some of them are perhaps more appealing because they are built on a solid theoretical footing (Kenward-Roger, ESS, Pinheiro-Bates) which gives hope that they maintain their decent performance also in scenarios not covered by simulations. On the other hand, approximations that behave nicely in simulation studies but lack any deeper justification, such as our own “adjusted” DF or the one devised by Hasler (2013), might fail in cases that go beyond those considered in simulations. So from this point of view, Kenward-Roger, ESS, and Pinheiro-Bates are preferable, but we need to keep in mind that they are only approximations to a problem that cannot be solved exactly to date. This being said, the question which DF approximation is applied only matters with small samples of less than 15 or even less than ten subjects per group. And as soon as the sample sizes exceed 30 or 40 per group, the discrepancy between results from asymptotic and small-sample longitudinal MCTs is negligible in practice.

Pipper et al. (2012) pointed out explicitly that the MMM method may break down with small sample sizes. We showed that this problem can be cushioned with appropriate DFs, and then the longitudinal MCT procedure based on MMM behaves well even with sample sizes of less than ten per group. While this is certainly true for comparisons of treatment means per time point, the procedure is much harder to tame when time points are compared per treatment group. This is consistent with the observation that the asymptotic procedures based on a joint LMM or MMM achieve very similar type I error rates under H_0 with small to moderate sample sizes when treatments are compared per occasion whereas MMM does considerably worse than joint modeling for comparisons of occasions per treatment group.

When it comes to deciding whether a joint model or MMM should be used to estimate the parameters for the longitudinal MCTs, there is no hard and fast answer as both have their strengths and weaknesses. MMM saves the trouble of devising a suitable structure for the random effects and/or error covariance matrix, which can be tricky with joint modeling. In practice, the decision which type of joint model to fit (ELM, CIM, or some more sophisticated LMM with both random effects and non-zero correlation of the residuals) will be guided by characteristics of the data-generating process. Failure to capture serial dependencies adequately may affect the joint model-based tests and SCIs; in addition, numerical convergence problems might prevent fitting a more complex (and more appropriate) model. Both these problems can be evaded to a certain degree with AICc model selection—and entirely by fitting and combining simple univariate occasion-specific models with MMM.

If model selection is applied, we think that AICc should be the criterion of choice because it is deeply rooted in information theory, and the small-sample version (AICc rather than AIC) is clearly preferable unless sample sizes are very large, as suggested by Burnham and Anderson (2002). The use of AIC(c) is not undisputed though: Littell et al. (2000), for instance, picked the BIC because they felt they needed a criterion that tends to select sparser models than AIC does.

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 MMM approach requires the harsher MCAR condition, and no information can be "borrowed" across time points. Missing values are indeed an important aspect with longitudinal data analysis; especially dropout is a common problem when subjects are measured repeatedly over time. We have pointed out that the validity of any statistical analysis in the presence of missing data will depend on the underlying mechanism of missingness. MCAR is the easiest one to deal with in practice but also involves the most stringent assumptions. Fitzmaurice et al. (2011, p. 498) recommend MAR as default "unless there is a strong and compelling rationale to support the MCAR assumption". From this point of view, methods that yield valid results under MAR are superior. Laborious techniques like multiple imputation fulfill this criterion but it is clearly preferable to avoid them if a joint LMM serves the same purpose—but only if its mean and covariance structures are correctly specified!

Our power simulations demonstrate that taking dependencies of time points into account pays off. In particular with high correlation, the longitudinal MCT procedures outperform simple Bonferroni adjustments. This effect is much more pronounced for comparisons of treatments than for comparisons of time points. On the other hand, the powers of the longitudinal MCTs based on joint modeling and the MMM approach are very similar across the board; only for small-sample comparisons among occasions MMM falls a little short of power.

Performing comparisons among treatment groups *and* among time points simultaneously within the same analysis is possible as well and extends our flexibility when formulating "claims". However, such a duplex procedure creates extra challenges, especially with small sample sizes and when the DFs for the two types of comparisons are widely different e.g., because there are relatively few comparisons of treatments and relatively many comparisons of time points, or *vice versa*. We argued that using comparison-specific DFs is the most accurate solution, but applying the minimum DF works well, too, and will not be very conservative in most settings. Despite its practical feasibility though, we think that the duplex procedure should be carried out with some care because it involves a lot of single comparisons even for setups with few treatment groups and time points, and one should always question whether all these comparisons are sensible and meaningful.

The extension to discrete endpoints such as Poisson counts and binomial proportions is important because many outcomes in biological and medical research are measured as rates, counts, proportions, etc. We choose GEEs over GLMMs to fit a joint model because only GEE estimates have the desired population-average interpretation. The alternative is again the MMM approach, now built on occasion-specific GLMs rather than linear models. When data points are missing, both MMM and the joint GEE require MCAR because the latter does not involve specification of a full likelihood for all time points. On the other hand, devising a joint model is less cumbersome in a GEE framework than with LMMs because consistent estimates are ensured even under misspecification of the working correlation.

The longitudinal MCTs for discrete data using either GEEs or MMM are inherently asymptotic, and we explored their finite-sample properties via simulation: we found that when sample sizes are insufficient, the tests can be either conservative or liberal, depending on whether the endpoint is binomial or Poisson, which modeling strategy is applied, whether treatments are compared per occasion or occasions per treatment group, and how the SEs are estimated with the GEE (naive or robust). As a rule of thumb, the sample sizes should be at least as large as for the asymptotic Gaussian procedure i.e., a minimum of 30 to 40 subjects per group are necessary for acceptable FWER control. Note that we have only studied asymptotic procedures for discrete outcomes, and better small-sample properties might be achieved with profile likelihood rather than Wald-type statistics as in Gerhard (2014a), or with DF approximations as suggested by Li and Redden (2015) in the context of GLMMs. Other methods for inference about repeatedly measured binary outcomes are described in Klingenberg and Satopää (2013).

The power curves of the longitudinal MCT procedures for discrete data are very similar for GEEs and the MMM approach, and both have a slight advantage over Bonferroni-adjusted pairwise tests when treatments are compared at multiple time points, and a substantial advantage when time points are compared within multiple treatment groups. We did not investigate overdispersed Poisson or binomial data explicitly, but we presume this might add complexity to the problem, especially if the amount of overdispersion varies over time.

We believe that gauging the magnitude and uncertainty of interesting effects is much more relevant than testing their “significance” in almost any real-world application. Therefore we suggest that results of our longitudinal MCTs be principally presented as SCIs and not—or at least not only—as adjusted p-values. The SCIs are compatible with the test decisions based on the p-values and always more informative than those.

We parameterized the fixed effects in our joint LMMs such that treatment groups and time points are combined in a single factor i.e., a pseudo-one-way layout or cell means model. This obviously gives us great flexibility to do all sorts of comparisons across treatment groups and time points but on the other hand requires to estimate lots of parameters and associated SEs. Littell et al. (2000) proposed that fitting a sparser mean model than cell means could be considered to get smaller SEs; a formal treatment of this issue is given by Altham (1984).

People raising fundamental objections to (multiple) hypothesis testing will most certainly argue against our methods. We shall not pursue this line of argumentation here but rather emphasize two important points related to such criticism.

First, we urge that time points should be carefully chosen so that each elementary comparison has its right to exist as part of the overall "claim". It makes no sense to analyze day 1, day 2, day 3, day 4, etc. all separately in a longitudinal study—simply because the data are available. This would be like going on a fishing trip for significances. Indeed, we want to stress that the time points in our data examples were never an arbitrary selection. The measurement occasions in the bradykinin trial, for instance, are all closely linked with important medical steps such as initiation of anesthesia or separation from the heart-lung machine.

And second, we do not believe that loads of p-values lead to any scientific insight. Our methods are inappropriate as soon as there are more than five or six time points, because then it can hardly be meaningful to direct one's attention to all of them separately. If the primary interest lies in the temporal evolution rather than specific points on the time line, summary measures like the slope or AUC are clearly favorable. Another possibility are simultaneous confidence bands (Sudhagoni and Djira 2012; Mun and Chun 2014), but they require a linear time trend. A more flexible option could be based on generalized additive mixed-effects models (GAMMs) as proposed by Herberich et al. (2014).

All simulations, data analyses, and graphics in this thesis were programmed in R (R Core Team 2015). Example code as well as some useful functions are provided in Appendix E.

In conclusion, we recommend using the MMM technique in many repeated measures settings where one would conventionally fit an ELM or CIM or any other joint LMM to obtain estimates for multiple testing or constructing SCIs. For the standard case where Gaussian treatment means are compared at multiple correlated occasions, one can simply model each time point separately without racking one's brain about random effects and residual correlation structures. We have shown that small sample sizes are not an impediment because a fairly straightforward DF approximation helps to achieve good FWER control, and the power is very similar to the joint modeling approach.

Caution should be exercised when comparing time points within multiple treatment groups: here the MMM approach can be substantially liberal with too small sample sizes, and in the approximate small-sample procedure it may also be less powerful than joint modeling. If there are missing values and MCAR is called into question, a joint LMM that allows to assume MAR—at least if its mean and covariance are adequately specified—is probably favorable.

MMM is also an attractive option for binomial and Poisson endpoints, although the asset of not having to think about covariance structures is not quite as relevant here because GEEs are capable of estimating the covariance consistently even if the working correlation is misspecified. On the other hand, the biggest advantage of joint modeling in the Gaussian setting also ceases to apply: GEEs require the stricter MCAR assumption for missing values because they are not based on a joint likelihood.

8 Conclusion

The key conclusions from this work are:

1. When comparisons of treatments per time point and/or comparisons of time points per treatment group are desired, classical methods for analyzing longitudinal Gaussian data, such as repeated measures ANOVA, are of little avail as they only produce p-values for the global hypothesis. By contrast, MCTs provide adjusted p-values and informative SCIs for single comparisons, plus a global decision.

2. Existing MCT methodology for correlated measurements is inflexible. Multivariate MCTs (Hasler and Hothorn 2011; Hasler 2014a) allow for comparisons among treatments only and cannot cope with covariates or missing values properly. Repeated measures MCTs (Hasler 2013) that allow for comparisons among occasions are limited to single-group designs and come with similar restrictions as the multivariate MCTs.

3. These limitations can be overcome by jointly modeling the data with an LMM where including covariates is straightforward and the assumption of MAR is sufficient for missing values. When faced with the choice between competing random-effects and/or residual covariance structures, AICc may serve as a decision criterion.

4. An elegant alternative strategy is to combine occasion-specific linear models (Pipper et al. 2012), which saves the trouble of specifying random effects and error covariances. One limitation of this MMM approach is that inference is only valid under MCAR, and no information can be "borrowed" across time points.

5. Longitudinal MCTs based on either of these methods control the FWER asymptotically. For small-sample inference, approximate control can be achieved using a DF method such as Kenward-Roger, Pinheiro-Bates, or a novel approach based on the ESS (Faes et al. 2009).

6. The powers of MCTs based on a joint LMM or MMM are very similar. There is more power to gain (in comparison to Bonferroni) when serial correlation is high and there are more occasions. Ignoring serial correlation is much more detrimental with comparisons of time points than with comparisons of treatment groups. It is also possible to calculate the power of the longitudinal MCTs, at least approximately.

7. Performing comparisons of treatments and time points simultaneously in a duplex procedure is possible, too. When the DFs for the two types of comparisons are widely different, including multiple comparison-specific DFs (as in Hasler and Hothorn 2008) can pay off. Using the minimum DF is always a safe and simple—but perhaps conservative—alternative.

8. The methods for Gaussian data are extendable to the case of binomial and Poisson endpoints. GLMMs provide a way to fit a joint model, but their estimates have an inherent subject-specific interpretation. GEEs on the other hand yield estimates with the desired population-average interpretation. The occasion-specific models for the MMM approach are now GLMs. The longitudinal MCTs based on either modeling strategy are asymptotic and may break down with small sample sizes. For Poisson GEEs, naive covariance estimation seems to be preferable to the robust

sandwich estimates. GEE- and MMM-based MCTs have practically the same large-sample power. Joint modeling with GEEs no longer has the advantage of allowing MAR.

Bibliography

Hirotugu Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974. doi:10.1109/TAC.1974.1100705.

Michael G. Akritas. The rank transform method in some two-factor designs. *Journal of the American Statistical Association*, 85(409):73–78, 1990. doi:10.2307/2289527.

Michael G. Akritas and Steven F. Arnold. Fully nonparametric hypotheses for factorial designs. I: Multivariate repeates measures designs. *Journal of the American Statistical Association*, 89(425):336–343, 1994. doi:10.2307/2291230.

Waseem S. Alnosaier. *Kenward-Roger approximate F test for fixed effects in mixed linear models*. PhD thesis, Oregon State University, Corvallis, OR, 2007. URL `http://ir.library.oregonstate.edu/xmlui/bitstream/handle/1957/5262/mydissertation.pdf`.

Mohamed Alosh, Frank Bretz, and Mohammad Huque. Advanced multiplicity adjustment methods in clinical trials. *Statistics in Medicine*, 33(4):693–713, 2013. doi:10.1002/sim.5974.

Patricia M. E. Altham. Improving the precision of estimation by fitting a model. *Journal of the Royal Statistical Society, Series B: Methodological*, 46(1):118–119, 1984.

Jaime Arnau, Roser Bono, and Guillermo Vallejo. Analyzing small samples of repeated measures data with the mixed-model adjusted F test. *Communications in Statistics—Simulation and Computation*, 38(5):1083–1103, 2009. doi:10.1080/03610910902785746.

J. M. Balaguer, C. Yu, J. G. Byrne, S. K. Ball, M. R. Petracek, N. J. Brown, and M. Pretorius. Contribution of endogeneous bradykinin to fibrinolysis, inflammation, and blood product transfusion following cardiac surgery: a randomized clinical trial. *Clinical Pharmacology and Therapeutics*, 93(4):326–334, 2013. doi:10.1038/clpt.2012.249.

S. Bandyopadhyay, B. Ganguli, and A. Chatterjee. A review of multivariate longitudinal data analysis. *Statistical Methods in Medical Research*, 20(4):299–330, 2011. doi:10.1177/0962280209340191.

Ole Barndorff-Nielsen. On a formula for the distribution of the maximum likelihood estimator. *Biometrika*, 70(2):343–365, 1983. doi:10.1093/biomet/70.2.343.

Sunni A. Barnes, Craig H. Mallinckrodt, Stacy R. Lindborg, and M. Kallin Carter. The impact of missing data and how it is handled on the rate of false-positive results in drug development. *Pharmaceutical Statistics*, 7(3):215–225, 2008. doi:10.1002/pst.310.

Kamil Barton. *MuMIn: Multi-model inference*, 2015. URL `http://CRAN.R-project.org/package=MuMIn`. R package version 1.15.1.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015a. doi:10.18637/jss.v067.i01.

Douglas Bates, Martin Mächler, Ben Bolker, Steven Walker, Rune Haubo Bojesen Christensen, Henrik Singmann, Bin Dai, and Gabor Grothendieck. *lme4: Linear mixed-effects models using 'Eigen' and S4*, 2015b. URL `http://CRAN.R-project.org/package=lme4`. R package version 1.1-10.

Caroline Beunckens, Geert Molenberghs, and Michael G. Kenward. Direct likelihood analysis versus simple forms of imputation for missing data in randomized clinical trials. *Clinical Trials*, 2(5):379–386, 2005. doi:10.1191/1740774505cn119oa.

Eve Bofinger. Step down procedures for comparison with a control. *Australian Journal of Statistics*, 29(3):348–364, 1987. doi:10.1111/j.1467-842X.1987.tb00751.x.

Benjamin M. Bolker, Mollie E. Brooks, Connie J. Clark, Shane W. Geange, John R. Poulsen, M. Henry H. Stevens, and Jada-Simone S. White. Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology and Evolution*, 24(3):127–135, 2009. doi:10.1016/j.tree.2008.10.008.

Carlo E. Bonferroni. Il calcolo delle assicurazioni su gruppi di teste. In *Studi in Onore del Professore Salvatore Ortu Carboni*, pages 13–60. Rome, Italy, 1935. In Italian.

Carlo E. Bonferroni. Teoria statistica delle classi e calcolo delle probabilità. In *Pubblicazioni del R. Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, volume 8, pages 3–62. Libreria Internazionale Seeber, Florence, Italy, 1936. In Italian.

Alessandra R. Brazzale and Anthony C. Davison. Accurate parametric inference for small samples. *Statistical Science*, 23(4):465–484, 2008. doi:10.1214/08-STS273.

Norman E. Breslow and David G. Clayton. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421):9–25, 1993. doi:10.2307/2290687.

Frank Bretz. An extension of the Williams trend test to general unbalanced linear models. *Computational Statistics and Data Analysis*, 50(7):1735–1748, 2006. doi:10.1016/j.csda.2005.02.005.

Frank Bretz, Alan Genz, and Ludwig A. Hothorn. On the numerical availability of multiple comparison procedures. *Biometrical Journal*, 43(5):645–656, 2001a. doi:10.1002/1521-4036(200109)43:5<645::AID-BIMJ645>3.0.CO;2-F.

Frank Bretz, Anthony J. Hayter, and Alan Genz. Critical point and power calculation for the studentized range test for general correlated means. *Journal of Statistical Computation and Simulation*, 71(2):85–97, 2001b. doi:10.1080/00949650108812136.

Frank Bretz, Torsten Hothorn, and Peter Westfall. *Multiple Comparisons Using R*. Chapman & Hall/CRC, Boca Raton, FL, 2010. ISBN 978-1-58488-574-0.

Edgar Brunner and Frank Langer. Nonparametric analysis of ordered categorical data in designs with longitudinal observations and small sample sizes. *Biometrical Journal*, 42 (6):663–675, 2000. doi:10.1002/1521-4036(200010)42:6<663::AID-BIMJ663>3.0.CO;2-7.

Edgar Brunner and Madan L. Puri. Nonparametric methods in factorial designs. *Statistical Papers*, 42(1):1–52, 2001. doi:10.1007/s003620000039.

Edgar Brunner, Sebastian Domhof, and Frank Langer. *Nonparametric Analysis of Longitudinal Data in Factorial Experiments.* John Wiley & Sons, New York, NY, 2002. ISBN 978-0-47144-166-3.

S. F. Buck. A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *Journal of the Royal Statistical Society, Series B: Methodological*, 22(2):302–306, 1960.

Kenneth P. Burnham and David R. Anderson. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. Second Edition.* Springer, New York, NY, 2002. ISBN 0-387-95364-7.

Jian-Shen Chen and Robert I. Jennrich. The signed root deviance profile and confidence intervals in maximum likelihood analysis. *Journal of the American Statistical Association*, 91(435):993–998, 1996. doi:10.2307/2291718.

Jing Cheng, Lloyd J. Edwards, Mildred M. Maldonado-Molina, Kelli A. Komro, and Keith E. Muller. Real longitudinal data analysis for real people: building a good enough mixed model. *Statistics in Medicine*, 29(504-520):4, 2010. doi:10.1002/sim.3775.

Avital Cnaan, Nan M. Laird, and Peter Slasor. Using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data. *Statistics in Medicine*, 16(20):2349–2380, 1997. doi:10.1002/(SICI)1097-0258(19971030)16:20<2349::AID-SIM667>3.0.CO;2-E.

William G. Cochran and Gertrude M. Cox. *Experimental Designs. Second Edition.* John Wiley & Sons, New York, NY, 1957. ISBN 978-0-471-54567-5.

Edmund A. Cornish. The multivariate t-distribution associated with a set of normal sample deviates. *Australian Journal of Physics*, 7(4):531–542, 1954.

Thorsten Dickhaus. *Simultaneous Statistical Inference with Application in the Life Sciences.* Springer, Heidelberg, Germany, 2014. ISBN 978-3-642-45181-2.

Peter J. Diggle, Kung-Yee Liang, and Scott L. Zeger. *Analysis of Longitudinal Data.* Oxford University Press, Oxford, UK, 1994. ISBN 0-19-852284-3.

Gemechis Dilba, Frank Bretz, Volker Guiard, and Ludwig A. Hothorn. Simultaneous confidence intervals for ratios with applications to the comparison of several treatments with a control. *Methods of Information in Medicine*, 43(5):465–469, 2004.

Gemechis Dilba, Frank Bretz, and Volker Guiard. Simultaneous confidence sets and confidence intervals for multiple ratios. *Journal of Statistical Planning and Inference*, 136(8):2640–2658, 2006. doi:10.1016/j.jspi.2004.11.009.

Alex Dmitrienko and Ralph D'Agostino, Sr. Traditional multiplicity adjustment methods in clinical trials. *Statistics in Medicine*, 32(29):5172–5218, 2013. doi:10.1002/sim.5990.

Alex Dmitrienko, Ajit C. Tamhane, and Frank Bretz, editors. *Multiple Testing Problems in Pharmaceutical Statistics.* Chapman & Hall/CRC, Boca Raton, FL, 2010. ISBN 978-1-58488-984-7.

Alex Dmitrienko, Ralph B. D'Agostino, Sr., and Mohammad F. Huque. Key multiplicity issues in clinical drug development. *Statistics in Medicine*, 32(7):1079–1111, 2013. doi:10.1002/sim.5642.

Guro Dørum, Lars Snipen, Margrete Solheim, and Solve Sæbø. Rotation testing in gene set enrichment analysis for small direct comparison experiments. *Statistical Applications in Genetics and Molecular Biology*, 8(1):1–24, 2009. doi:10.2202/1544-6115.1418.

Guro Dørum, Lars Snipen, Margrete Solheim, and Solve Sæbø. Rotation gene set testing for longitudinal expression data. *Biometrical Journal*, 56(5):1055–1075, 2014. doi:10.1002/bimj.201100178.

Olive J. Dunn and Frank J. Massey, Jr. Estimation of multiple contrasts using t-distributions. *Journal of the American Statistical Association*, 60(310):573–583, 1965. doi:10.2307/2282692.

Charles W. Dunnett. A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association*, 50(272):1096–1121, 1955. doi:10.2307/2281208.

Charles W. Dunnett and Ajit C. Tamhane. Step-down multiple tests for comparing treatments with a control in unbalanced one-way layouts. *Statistics in Medicine*, 10 (6):939–947, 1991. doi:10.1002/sim.4780100614.

Charles W. Dunnett and Ajit C. Tamhane. A step-up multiple test procedure. *Journal of the American Statistical Association*, 87(417):162–170, 1992. doi:10.1080/01621459.1992.10475188.

Charles W. Dunnett and Ajit C. Tamhane. Step-up multiple testing of parameters with unequal correlated estimates. *Biometrics*, 51(1):217–227, 1995. doi:10.2307/2533327.

Israel Einot and K. Ruben Gabriel. A study of the power of several methods of multiple comparisons. *Journal of the American Statistical Association*, 70(351a):574–583, 1975. doi:10.1080/01621459.1975.10482474.

Christel Faes, Geert Molenberghs, Marc Aerts, Geert Verbeke, and Michael G. Kenward. The effective sample size and an alternative small-sample degrees-of-freedom method. *The American Statistician*, 63(4):389–399, 2009. doi:10.1198/tast.2009.08196.

Alex Hrong-Tai Fai and Paul L. Cornelius. Approximate F-tests of multiple degree of freedom hypotheses in generalized least squares analyses of unbalanced split-plot experiments. *Journal of Statistical Computation and Simulation*, 54(4):363–378, 1996. doi:10.1080/00949659608811740.

Edgar C. Fieller. The biological standardization of insulin. *Journal of the Royal Statistical Society*, 7(1):1–64 (Supplement), 1940. doi:10.2307/2983630.

Edgar C. Fieller. A fundamental formula in the statistics of biological assay, and some applications. *Quarterly Journal of Pharmacy and Pharmacology*, 17, 1944.

Edgar C. Fieller. Some problems in interval estimation. *Journal of the Royal Statistical Society, Series B: Methodological*, 16(2):175–185, 1954.

Garrett M. Fitzmaurice, Nan M. Laird, and James H. Ware. *Applied Longitudinal Analysis. Second Edition.* John Wiley & Sons, Hoboken, NJ, 2011. ISBN 978-0-470-38027-7.

K. Ruben Gabriel. Simultaneous test procedures—some theory of multiple comparisons. *The Annals of Mathematical Statistics*, 40(1):224–250, 1969.

Paul A. Games and John F. Howell. Pairwise multiple comparison procedures with unequal N's and/or variances: a Monte Carlo study. *Journal of Educational Statistics*, 1:113–125, 1976. doi:10.2307/1164979.

Martin J. Gardner and Douglas G. Altman. Confidence intervals rather than P values: estimation rather than hypothesis testing. *British Medical Journal*, 292(6522):746–750, 1986.

Alan Genz and Frank Bretz. Numerical computation of multivariate t-probabilities with application to power calculation of multiple contrasts. *Journal of Statistical Computation and Simulation*, 63(4):103–117, 1999. doi:10.1080/00949659908811962.

Alan Genz and Frank Bretz. *Computation of Multivariate Normal and t Probabilities.* Springer, Heidelberg, Germany, 2009. ISBN 978-3-642-01688-2.

Alan Genz, Frank Bretz, Tetsuhisa Miwa, Xuefei Mi, Friedrich Leisch, Fabian Scheipl, and Torsten Hothorn. *mvtnorm: Multivariate normal and t distributions*, 2014. URL `http://CRAN.R-project.org/package=mvtnorm`. R package version 0.9-9997.

Daniel Gerhard. *mcroast: Multiple Contrast Rotation Tests*, 2014a. URL `https://github.com/daniel-gerhard/mcroast`. R package version 0.0-4.

Daniel Gerhard. Simultaneous small sample inference for linear combinations of generalized linear model parameters. *Communications in Statistics—Simulation and Computation*, 2014b. doi:10.1080/03610918.2014.895836.

Daniel Gerhard and Frank Schaarschmidt. Rotation tests for general linear hypotheses with family-wise error rate control. 2015. Submitted to *Biometrical Journal*.

Francis G. Giesbrecht and Joseph C. Burns. Two-stage analysis based on a mixed model: large-sample asymptotic theory and small-sample simulation results. *Biometrics*, 41 (2):477–486, 1985. doi:10.2307/2530872.

Karl B. Gregory. A comparison of denominator degrees of freedom approximation methods in the unbalanced two-way factorial mixed model. Master's thesis, Texas A&M University, College Station, TX, 2011. URL `http://www.learningace.com/doc/2798578/695791137a52ad51d325eaa95f220bd1`.

Michael E. Griswold, Bruce J. Swihart, Brian S. Caffo, and Scott L. Zeger. Practical marginal multilevel models. *Stat*, 2(1):129–142, 2013. doi:10.1002/sta4.22.

LeAnna Guerin and Walter W. Stroup. A simulation study to evaluate PROC MIXED analysis of repeated measures data. In *Proceedings of the 12th Kansas State University Conference on Applied Statistics in Agriculture*, pages 170–203, Manhattan, KS, 2000.

Olivier Guilbaud. Simultaneous confidence regions corresponding to Holm's step-down procedure and other closed-testing procedures. *Biometrical Journal*, 50(5):678–692, 2008. doi:10.1002/bimj.200710449.

Matthew J. Gurka, Lloyd J. Edwards, and Keith E. Muller. Avoiding bias in mixed model inference for fixed effects. *Statistics in Medicine*, 30(22):2696–2707, 2011. doi:10.1002/sim.4293.

Ulrich Halekoh and Søren Højsgaard. A Kenward-Roger approximation and parametric bootstrap methods for tests in linear mixed models—the R package pbkrtest. *Journal of Statistical Software*, 59(9):1–32, 2014a. doi:10.18637/jss.v059.i09.

Ulrich Halekoh and Søren Højsgaard. *pbkrtest: Parametric bootstrap and Kenward-Roger-based methods for mixed model comparison*, 2014b. URL `http://CRAN.R-project.org/package=pbkrtest`. R package version 0.4-2.

Ulrich Halekoh, Søren Højsgaard, and Jun Yan. The R package geepack for generalized estimating equations. *Journal of Statistical Software*, 15(2):1–11, 2006. doi:10.18637/jss.v015.i02.

James W. Hardin and Joseph M. Hilbe. *Generalized Estimating Equations. Second Edition*. Chapman & Hall/CRC, Boca Raton, FL, 2013. ISBN 978-1-4398-8813-2.

David R. Hare and John D. Spurrier. Simultaneous inference for ratios of linear combinations of general linear model parameters. *Biometrical Journal*, 49(6):854–862, 2007. doi:10.1002/bimj.200610333.

H. Leon Harter. Error rates and sample sizes for range tests in multiple comparisons. *Biometrics*, 13(4):511–536, 1957. doi:10.2307/2527975.

David A. Harville and Daniel R. Jeske. Mean squared error of estimation or prediction under a general linear model. *Journal of the American Statistical Association*, 87(419): 724–731, 1992. doi:10.2307/2290210.

Mario Hasler. Multiple contrasts for repeated measures. *The International Journal of Biostatistics*, 9(1):1–13, 2013. doi:10.1515/ijb-2012-0025.

Mario Hasler. Multiple contrast tests for multiple endpoints in the presence of heteroscedasticity. *The International Journal of Biostatistics*, 10(1):17–28, 2014a. doi:10.1515/ijb-2012-0015.

Mario Hasler. *SimComp: Simultaneous comparisons for multiple endpoints*, 2014b. URL `http://CRAN.R-project.org/package=SimComp`. R package version 2.2.

Mario Hasler and Kathrin Böhlendorf. Multiple comparisons for multiple endpoints in agricultural experiments. *Journal of Agricultural, Biological, and Environmental Statistics*, 18(4):578–593, 2013. doi:10.1007/s13253-013-0149-7.

Mario Hasler and Ludwig A. Hothorn. Multiple contrast tests in the presence of heteroscedasticity. *Biometrical Journal*, 50(5):793–800, 2008. doi:10.1002/bimj.200710466.

Mario Hasler and Ludwig A. Hothorn. A Dunnett-type procedure for multiple endpoints. *The International Journal of Biostatistics*, 7(1):article 3, 2011. doi:10.2202/1557-4679.1258.

Mario Hasler and Ludwig A. Hothorn. A multivariate Williams-type trend procedure. *Statistics in Biopharmaceutical Research*, 4(1):57–65, 2012. doi:10.1080/19466315.2011.633868.

Anthony J. Hayter and Wei Liu. A method of power assessment for tests comparing several treatments with a control. *Communications in Statistics—Theory and Methods*, 21(7):1871–1889, 1992. doi:10.1080/03610929208830885.

Patrick J. Heagerty and Scott L. Zeger. Marginalized multilevel models and likelihood inference. *Statistical Science*, 15(1):1–19, 2000.

Esther Herberich, Johannes Sikorski, and Torsten Hothorn. A robust procedure for comparing multiple means under heteroscedasticity in unbalanced designs. *PLoS One*, 5 (3):e9788, 2010. doi:10.1371/journal.pone.0009788.

Esther Herberich, Christine Hassler, and Torsten Hothorn. Multiple curve comparisons with an application to the formation of the dorsal funiculus of mutant mice. *The International Journal of Biostatistics*, 10(2):289–302, 2014. doi:10.1515/ijb-2013-0003.

Chihiro Hirotsu, Shoichi Yamamoto, and Ludwig A. Hothorn. Estimating the dose-response pattern by the maximal contrast type test approach. *Statistics in Biopharmaceutical Research*, 3(1):40–53, 2011. doi:10.1198/sbr.2010.08093.

Yosef Hochberg. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4):800–802, 1988. doi:10.2307/2336325.

Yosef Hochberg and Ajit C. Tamhane. *Multiple Comparison Procedures*. John Wiley & Sons, New York, NY, 1987. ISBN 0-471-82222-1.

Marius Hofert, Ivan Kojadinovic, Martin Maechler, and Jun Yan. *copula: Multivariate dependence with copulas*, 2014. URL `http://CRAN.R-project.org/package=copula`. R package version 0.999-9.

Wherly P. Hoffman, Justin Recknor, and Cindy Lee. Overall type I error rate and power of multiple Dunnett's tests on rodent body weights in toxicology studies. *Journal of Biopharmaceutical Statistics*, 18(5):883–900, 2008. doi:10.1080/10543400802287420.

Søren Højsgaard, Ulrich Halekoh, and Jun Yan. *geepack: Generalized estimating equation package*, 2014. URL `http://CRAN.R-project.org/package=geepack`. R package version 1.2-0.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Manfred Horn and Rüdiger Vollandt. Sample sizes for comparisons of k treatments with a control based on different definitions of the power. *Biometrical Journal*, 40(5):589–612, 1998. doi:10.1002/(SICI)1521-4036(199809)40:5<589::AID-BIMJ589>3.0.CO;2-8.

Harold Hotelling. The generalization of Student's ratio. *The Annals of Mathematical Statistics*, 2(3):360–378, 1931. doi:10.1214/aoms/1177732979.

Torsten Hothorn, Frank Bretz, and Peter Westfall. Simultaneous inference in general parametric models. *Biometrical Journal*, 50(3):346–363, 2008. doi:10.1002/bimj.200810425.

Torsten Hothorn, Frank Bretz, Peter Westfall, Richard M. Heiberger, André Schützenmeister, and Susan Scheibe. *multcomp: Simultaneous inference in general parametric models*, 2015. URL `http://CRAN.R-project.org/package=multcomp`. R package version 1.4-1.

Jason C. Hsu. *Multiple Comparisons. Theory and Methods.* Chapman & Hall, London, UK, 1996. ISBN 0-412-98281-1.

Mohammad F. Huque, Alex Dmitrienko, and Ralph D'Agostino. Multiplicity issues in clinical trials with multiple objectives. *Statistics in Biopharmaceutical Research*, 5(4): 321–337, 2013. doi:10.1080/19466315.2013.807749.

Clifford M. Hurvich and Chih-Ling Tsai. Regression and time series model selection in small samples. *Biometrika*, 76(2):297–307, 1989. doi:10.1093/biomet/76.2.297.

Huynh Huynh. Some approximate tests for repeated measurement designs. *Psychometrika*, 43(2):161–175, 1978. doi:10.1007/BF02293860.

ICH Expert Working Group. *ICH Harmonised Tripartite Guideline E9: Statistical Principles for Clinical Trials.* International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use, 1998. URL `http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E9/Step4/E9_Guideline.pdf`. Accessed on September 16th, 2014.

Hélène Jacqmin-Gadda, Solenne Sibillot, Cécile Proust, Jean-Michel Molina, and Rodolphe Thiébaut. Robustness of the linear mixed model to misspecified error distribution. *Computational Statistics and Data Analysis*, 51(10):5142–5154, 2007. doi:10.1016/j.csda.2006.05.021.

Robert I. Jennrich and Mark D. Schluchter. Unbalanced repeated-measures models with structured covariance matrices. *Biometrics*, 42(4):805–820, 1986. doi:10.2307/2530695.

Signe M. Jensen, Christian B. Pipper, and Christian Ritz. Evaluation of multi-outcome longitudinal studies. *Statistics in Medicine*, 34(12):1993–2003, 2015. doi:10.1002/sim.6461.

Raghu N. Kackar and David A. Harville. Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *Journal of the American Statistical Association*, 79(388):853–862, 1984. doi:10.2307/2288715.

Josephine Karanja, Hans-Michael Poehling, and Philip Pallmann. Efficacy and dose response of soil-applied neem formulations in substrates with different amounts of organic matter, in the control of whiteflies, *Aleyrodes proletella* and *Trialeurodes vaporariorum* (Hemiptera: Aleyrodidae). *Journal of Economic Entomology*, 2015. doi:10.1093/jee/tov047.

Michael G. Kenward and Geert Molenberghs. Likelihood based frequentist inference when data are missing at random. *Statistical Science*, 13(3):236–247, 1998. doi:10.1214/ss/1028905886.

Michael G. Kenward and Geert Molenberghs. Parametric models for incomplete continuous and categorical longitudinal data. *Statistical Methods in Medical Research*, 8(1): 51–83, 1999. doi:10.1177/096228029900800105.

Michael G. Kenward and James H. Roger. Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, 53(3):983–997, 1997. doi:10.2307/2533558.

Michael G. Kenward and James H. Roger. An improved approximation to the precision of fixed effects from restricted maximum likelihood. *Computational Statistics and Data Analysis*, 53(7):2583–2595, 2009. doi:10.1016/j.csda.2008.12.013.

Harvey J. Keselman, James Algina, Rhonda K. Kowalchuk, and Russell D. Wolfinger. A comparison of two approaches for selecting covariance structures in the analysis of repeated measurements. *Communications in Statistics—Simulation and Computation*, 27(3):591–604, 1998. doi:10.1080/03610919808813497.

Bernhard Klingenberg and Ville Satopää. Simultaneous confidence intervals for comparing margins of multivariate binary data. *Computational Statistics and Data Analysis*, 64:87–98, 2013. doi:10.1016/j.csda.2013.02.016.

Frank Konietschke, Arne C. Bathke, Ludwig A. Hothorn, and Edgar Brunner. Testing and estimation of purely nonparametric effects in repeated measures designs. *Computational Statistics and Data Analysis*, 54(8):1895–1905, 2010. doi:10.1016/j.csda.2010.02.019.

Samuel Kotz and Saralees Nadarajah. *Multivariate t Distributions and Their Applications.* Cambridge University Press, Cambridge, UK, 2004. ISBN 0-521-82654-3.

Samuel Kotz, Narayanaswamy Balakrishnan, and Norman L. Johnson. *Continuous Multivariate Distributions, Volume 1: Models and Applications. Second Edition.* John Wiley & Sons, New York, NY, 2000. ISBN 0-471-18387-3.

Solomon Kullback and Richard A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951. doi:10.1214/aoms/1177729694.

Alexandra Kuznetsova, Per Bruun Brockhoff, and Rune Haubo Bojesen Christensen. *lmerTest: Tests in linear mixed effects models*, 2015a. URL `http://CRAN.R-project.org/package=lmerTest`. R package version 2.0-29.

Alexandra Kuznetsova, Rune Haubo Bojesen Christensen, Cecile Bavay, and Per Bruun Brockhoff. Automated mixed ANOVA modeling of sensory and consumer data. *Food Quality and Preference*, 40(Part A):31–37, 2015b. doi:10.1016/j.foodqual.2014.08.004.

Nan M. Laird. Missing data in longitudinal studies. *Statistics in Medicine*, 7(1-2):305–315, 1988. doi:10.1002/sim.4780070131.

Nan M. Laird and James H. Ware. Random-effects models for longitudinal data. *Biometrics*, 38(4):963–974, 1982. doi:10.2307/2529876.

Øyvind Langsrud. Rotation tests. *Statistics and Computing*, 15(1):53–60, 2005. doi:10.1007/s11222-005-4789-5.

Øyvind Langsrud, Kjetil Jørgensen, Ragni Ofstad, and Tormod Næs. Analyzing designed experiments with multiple responses. *Journal of Applied Statistics*, 34(10):1275–1296, 2007. doi:10.1080/02664760701594246.

Russell Lenth. *lsmeans: Least-squares means*, 2015. URL `http://CRAN.R-project.org/package=lsmeans`. R package version 2.20-23.

Peng Li and David T. Redden. Comparing denominator degrees of freedom approximations for the generalized linear mixed model in analyzing binary outcome in small sample cluster-randomized trials. *BMC Medical Research Methodology*, 15:article 38, 2015. doi:10.1186/s12874-015-0026-x.

Kung-Yee Liang and Scott L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986. doi:10.2307/2336267.

Ramon C. Littell. Analysis of unbalanced mixed-model data: A case study comparison of ANOVA versus REML/GLS. *Journal of Agricultural, Biological, and Environmental Statistics*, 7(4):472–490, 2002. doi:10.1198/108571102816.

Ramon C. Littell, Jane Pendergast, and Ranjini Natarajan. Modelling covariance structure in the analysis of repeated measures data. *Statistics in Medicine*, 19(13):1793–1819, 2000. doi:10.1002/1097-0258(20000715)19:13<1793::AID-SIM482>3.3.CO;2-H.

Ramon C. Littell, George A. Milliken, Walter W. Stroup, Russell D. Wolfinger, and Oliver Schabenberger. *SAS for Mixed Models. Second Edition.* SAS Institute, Cary, NC, 2006. ISBN 978-1-59047-500-3.

Roderick J. A. Little. Missing-data adjustments in large surveys. *Journal of Business and Economic Statistics*, 6(3):287–296, 1988. doi:10.2307/1391878.

Roderick J. A. Little and Donald B. Rubin. *Statistical Analysis with Missing Data. Second Edition.* John Wiley & Sons, New York, NY, 2002. ISBN 978-0-471-18386-0.

Chunyan Liu, Timothy P. Cripe, and Mi-Ok Kim. Statistical issues in longitudinal data analysis for treatment efficacy studies in the biomedical sciences. *Molecular Therapy*, 18(9):1724–1730, 2010. doi:10.1038/mt.2010.127.

Guanghan Liu and A. Lawrence Gould. Comparison of alternative strategies for analysis of longitudinal data with dropouts. *Journal of Biopharmaceutical Statistics*, 12(2): 207–226, 2002. doi:10.1081/BIP-120015744.

Kaifeng Lu and Devan V. Mehrotra. Specification of covariance structure in longitudinal data analysis for randomized clinical trials. *Statistics in Medicine*, 29(4):474–488, 2010. doi:10.1002/sim.3820.

Craig H. Mallinckrodt, W. Scott Clark, and Stacy R. David. Accounting for dropout bias using mixed-effects models. *Journal of Biopharmaceutical Statistics*, 11(1-2):9–21, 2001a. doi:10.1081/BIP-100104194.

Craig H. Mallinckrodt, W. Scott Clark, and Stacy R. David. Type I error rates from mixed effects model repeated measures versus fixed effects ANOVA with missing values imputed via last observation carried forward. *Drug Information Journal*, 35(4):1215–1225, 2001b. doi:10.1177/009286150103500418.

Craig H. Mallinckrodt, Christopher J. Kaiser, John G. Watkin, Michael J. Detke, Geert Molenberghs, and Raymond J. Carroll. Type I error rates from likelihood-based repeated measures analyses of incomplete longitudinal data. *Pharmaceutical Statistics*, 3(3):171–186, 2004. doi:10.1002/pst.131.

Ruth Marcus, Eric Peritz, and K. Ruben Gabriel. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63(3):655–660, 1976. doi:10.1093/biomet/63.3.655.

Peter McCullagh and John A. Nelder. *Generalized Linear Models. Second Edition.* Chapman & Hall/CRC, Boca Raton, FL, 1989. ISBN 978-0-412-31760-5.

Charles E. McCulloch and Shayle R. Searle. *Generalized, Linear, and Mixed Models.* John Wiley & Sons, New York, NY, 2001. ISBN 0-471-19364-X.

George A. Milliken and Dallas E. Johnson. *Analysis of Messy Data. Volume I: Designed Experiments.* Chapman & Hall, London, UK, 1992. ISBN 0-412-99081-4.

Geert Molenberghs, Herbert Thijs, Ivy Jansen, Caroline Beunckens, Michael G. Kenward, Craig Mallinckrodt, and Raymond J. Carroll. Analyzing incomplete longitudinal clinical trial data. *Biostatistics*, 5(3):445–464, 2004. doi:10.1093/biostatistics/kxh001.

Hari Mukerjee, Tim Robertson, and Farrol T. Wright. Comparison of several treatments with a control using multiple contrasts. *Journal of the American Statistical Association*, 82(399):902–910, 1987. doi:10.2307/2288803.

Jungwon Mun and Hyonho Chun. Effective simultaneous confidence bands for repeated measurements in linear mixed-effect models. *Journal of Statistical Computation and Simulation*, 84(8):1748–1760, 2014. doi:10.1080/00949655.2013.764429.

Ullrich Munzel and Ajit C. Tamhane. Nonparametric multiple comparisons in repeated measures designs for data with ties. *Biometrical Journal*, 44(6):762–779, 2002. doi:10.1002/1521-4036(200209)44:6<762::AID-BIMJ762>3.0.CO;2-A.

Saralees Nadarajah and Dipak K. Dey. Multitude of multivariate t-distributions. *Statistics*, 39(2):149–181, 2005. doi:10.1080/02331880500031407.

Saralees Nadarajah and Samuel Kotz. Multitude of bivariate t distributions. *Statistics*, 38(6):527–539, 2004. doi:10.1080/02331880412331319305.

Umesh D. Naik. Some selection rules for comparing p processes with a standard. *Communications in Statistics*, 4(6):519–535, 1975. doi:10.1080/03610927508827267.

National Toxicology Program. Toxicology and Carcinogenesis Studies of Mercuric Chloride (CAS No. 7487-94-7) in F344 Rats and B6C3F$_1$ Mice (Gavage Studies). Technical Report 408, U.S. Department of Health and Human Services, Research Triangle Park, NC, February 1993. URL `http://ntp.niehs.nih.gov/ntp/htdocs/lt_rpts/tr408.pdf`. Accessed on July 22nd, 2015.

John A. Nelder and Robert W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society, Series A: General*, 135(3):370–384, 1972. doi:10.2307/2344614.

Roger B. Nelsen. *An Introduction to Copulas. Second Edition.* Springer, New York, NY, 2006. ISBN 978-0387-28659-4.

Peter C. O'Brien. Procedures for comparing samples with multiple endpoints. *Biometrics*, 40(4):1079–1087, 1984. doi:10.2307/2531158.

Jean G. Orelien, Jun Zhai, Richard Morris, and Rich Cohn. An approach to performing multiple comparisons with a control in GEE models. *Communications in Statistics—Theory and Methods*, 31(1):87–105, 2002. doi:10.1081/STA-120002436.

Philip Pallmann and Ludwig A. Hothorn. Analysis of means (ANOM): a generalized approach using R. *Journal of Applied Statistics*, 2016. doi:10.1080/02664763.2015.1117584.

Philip Pallmann, Mias Pretorius, and Christian Ritz. Simultaneous comparisons of treatments at multiple time points: combined marginal models versus joint modeling. *Statistical Methods in Medical Research*, 2015. doi:10.1177/0962280215603743.

Taesung Park, Jin-Kyung Park, and Charles S. Davis. Effects of covariance model assumptions on hypothesis tests for repeated measurements: analysis of ovarian hormone data and pituitary-pteryomaxillary distance data. *Statistics in Medicine*, 20(16):2441–2453, 2001. doi:10.1002/sim.859.

H. Desmond Patterson and Robin Thompson. Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58(3):545–554, 1971. doi:10.1093/biomet/58.3.545.

Alan Phillips, Chrissie Fletcher, Gary Atkinson, Eddie Channon, Abdel Douiri, Thomas Jaki, Jeff Maca, David Morgan, James H. Roger, and Paul Terrill. Multiplicity: discussion points from the Statisticians in the Pharmaceutical Industry multiplicity expert group. *Pharmaceutical Statistics*, 12(5):255–259, 2013. doi:10.1002/pst.1584.

José Pinheiro, Douglas Bates, Saikat DebRoy, Deepayan Sarkar, and R Core Team. *nlme: Linear and nonlinear mixed effects models*, 2015. URL `http://CRAN.R-project.org/package=nlme`. R package version 3.1-122.

José C. Pinheiro and Douglas M. Bates. *Mixed-Effects Models in S and S-PLUS*. Springer, New York, NY, 2000. ISBN 0-387-98957-9.

Christian B. Pipper, Christian Ritz, and Hans Bisgaard. A versatile method for confirmatory evaluation of the effects of a covariate in multiple models. *Journal of the Royal Statistical Society, Series C: Applied Statistics*, 61(2):315–326, 2012. doi:10.1111/j.1467-9876.2011.01005.x.

Hui Quan, Xiaohui Luo, and Tom Capizzi. Multiplicity adjustment for multiple endpoints in clinical trials with multiple doses of an active treatment. *Statistics in Medicine*, 24 (14):2151–2170, 2005. doi:10.1002/sim.2101.

R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2015. URL `http://www.R-project.org`.

Philip H. Ramsey. Power differences between pairwise multiple comparisons. *Journal of the American Statistical Association*, 73(363):479–485, 1978. doi:10.1080/01621459.1978.10480038.

Samarendra N. Roy. On a heuristic method of test construction and its use in multivariate analysis. *The Annals of Mathematical Statistics*, 24(2):220–238, 1953.

Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976. doi:10.2307/2335739.

Donald B. Rubin. Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business and Economic Statistics*, 4(1):87–94, 1986. doi:10.2307/1391390.

Donald B. Rubin. *Multiple Imputation for Nonresponse in Surveys.* John Wiley & Sons, New York, NY, 1987. ISBN 0-471-08705-X.

Abdul J. Sankoh, Mohammad F. Huque, and Satya D. Dubey. Some comments on frequently used multiple endpoint adjustment methods in clinical trials. *Statistics in Medicine*, 16(22):2529–2542, 1997. doi:10.1002/(SICI)1097-0258(19971130)16:223.0.CO;2-J.

Abdul J. Sankoh, Mohammad F. Huque, Heidy Kwan Russell, and Ralph B. D'Agostino, Sr. Global two-group multiple endpoint adjustment methods applied to clinical trials. *Drug Information Journal*, 33(1):119–140, 1999.

SAS Institute. *SAS/STAT 9.2 User's Guide. Second Edition.* SAS Institute, Cary, NC, 2009.

Franklin E. Satterthwaite. Synthesis of variances. *Psychometrika*, 6(5):309–316, 1941. doi:10.1007/BF02288586.

Franklin E. Satterthwaite. An approximate distribution of estimates of variance components. *Biometrics Bulletin*, 2(6):110–114, 1946. doi:10.2307/3002019.

G. Bruce Schaalje, Justin B. McBride, and Gilbert W. Fellingham. Adequacy of approximations to distributions of test statistics in complex mixed linear models. *Journal of Agricultural, Biological, and Environmental Statistics*, 7(4):512–524, 2002. doi:10.1198/108571102726.

Frank Schaarschmidt, Daniel Gerhard, and Martin Sill. *MCPAN: Multiple comparisons using normal approximation*, 2013. URL `http://CRAN.R-project.org/package=MCPAN`. R package version 1.1-15.

Holger Schielzeth and Wolfgang Forstmeier. Conclusions beyond support: overconfident estimates in mixed models. *Behavioral Ecology*, 20(2), 2009. doi:10.1093/beheco/arn145.

Mark D. Schluchter and Janet T. Elashoff. Small-sample adjustments to tests with unbalanced repeated measures assuming several covariance structures. *Journal of Statistical Computation and Simulation*, 37(1-2):69–87, 1990. doi:10.1080/00949659008811295.

Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2): 461–464, 1978. doi:10.1214/aos/1176344136.

Shaun R. Seaman and Ian R. White. Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research*, 22(3):278–295, 2013. doi:10.1177/0962280210395740.

Shayle R. Searle. *Linear Models*. John Wiley & Sons, New York, NY, 1971. ISBN 0-471-18499-3.

Shayle R. Searle, George Casella, and Charles E. McCulloch. *Variance Components*. John Wiley & Sons, Hoboken, NJ, 1992. ISBN 978-0-470-00959-8.

Stephen Senn and Frank Bretz. Power and sample size when multiple endpoints are considered. *Pharmaceutical Statistics*, 6(3):161–170, 2007. doi:10.1002/pst.301.

Qian Shi, Emily S. Pavey, and Rickey E. Carter. Bonferroni-based correction factor for multiple, correlated endpoints. *Pharmaceutical Statistics*, 11(4):300–309, 2012. doi:10.1002/pst.1514.

Mohammed M. Siddiqui. A bivariate t distribution. *The Annals of Mathematical Statistics*, 38(1):162–166, 1967. doi:10.1214/aoms/1177699066. Correction: 38(5):1594, 1967.

Ohidul Siddiqui. MMRM versus MI in dealing with missing data—a comparison based on 25 NDA data sets. *Journal of Biopharmaceutical Statistics*, 21(3):423–436, 2011. doi:10.1080/10543401003777995.

Ohidul Siddiqui, H. M. James Hung, and Robert O'Neill. MMRM vs. LOCF: a comprehensive comparison based on simulation study and 25 NDA datasets. *Journal of Biopharmaceutical Statistics*, 19(2):227–246, 2009. doi:10.1080/10543400802609797.

Simon S. Skene and Michael G. Kenward. The analysis of very small samples of repeated measurements I: an adjusted sandwich estimator. *Statistics in Medicine*, 29(27):2825–2837, 2010a. doi:10.1002/sim.4073.

Simon S. Skene and Michael G. Kenward. The analysis of very small samples of repeated measurements II: a modified Box correction. *Statistics in Medicine*, 29(27):2838–2856, 2010b. doi:10.1002/sim.4072.

Abe W. Sklar. Fonctions de répartition à n dimension et leurs marges. In *Publications de l'Institut de Statistique de l'Université de Paris*, volume 8, pages 229–231, 1959. In French.

Aldo Solari, Livio Finos, and Jelle J. Goeman. Rotation-based multiple testing in the multivariate linear model. *Biometrics*, 70(4):954–961, 2014. doi:10.1111/biom.12238.

Joachim Spilke, Hans-Peter Piepho, and Uwe Meyer. Approximating the degrees of freedom for ccontrast of genotypes laid out as subplots in an alpha-design in a split-plot experiment. *Plant Breeding*, 123(2):193–197, 2004. doi:10.1046/j.1439-0523.2003.00964.x.

Joachim Spilke, Hans-Peter Piepho, and Xiyuan Hu. A simulation study on tests of hypotheses and confidence intervals for fixed effects in mixed models for blocked experiments with missing data. *Journal of Agricultural, Biological, and Environmental Statistics*, 10(3):374–389, 2005. doi:10.1198/108571105X58199.

Gunnar Stefansson, Woo-Chul Kim, and Jason C. Hsu. On confidence sets in multiple comparisons. In Shanti S. Gupta and James O. Berger, editors, *Statistical Decision Theory and Related Topics IV*, pages 89–104. Springer, New York, NY, 1988. ISBN 978-1-4612-8365-2.

Andrew Stone and Christy Chuang-Stein. Strong control over multiple endpoints: are we adding value to the assessment of medicines? *Pharmaceutical Statistics*, 12(4):189–191, 2013. doi:10.1002/pst.1574.

Klaus Strassburger and Frank Bretz. Compatible simultaneous lower confidence bounds for the Holm procedure and other Bonferroni-based closed tests. *Statistics in Medicine*, 27(24):4914–4927, 2008. doi:10.1002/sim.3338.

Ramu G. Sudhagoni and Gemechis D. Djira. Multiple comparisons of parametric models in longitudinal studies. *Communications in Statistics—Theory and Methods*, 41(4), 2012. doi:10.1080/03610926.2010.522752.

Nariaki Sugiura. Further analysis of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics—Theory and Methods*, 7(1):13–26, 1978. doi:10.1080/03610927808827599.

Dei-In Tang, Clare Gnecco, and Nancy L. Geller. An approximate likelihood ratio test for a normal mean vector with nonnegative components with application to clinical trials. *Biometrika*, 76(3):577–583, 1989. doi:10.1093/biomet/76.3.577.

Peter F. Thall and Stephen C. Vail. Some covariance models for longitudinal count data with overdispersion. *Biometrics*, 46(3):657–671, 1990. doi:10.2307/2532086.

Yung L. Tong. *The Multivariate Normal Distribution*. Springer, New York, NY, 1990. ISBN 0-387-97062-2.

John W. Tukey. The problem of multiple comparisons. In Henry I. Braun, editor, *The Collected Works of John W. Tukey, Volume VIII*, pages 1–300. Chapman & Hall, New York, NY, 1994. ISBN 0-412-05121-4., 1953.

Elisa Valderas Gomez, G. Bruce Schaalje, and Gilbert W. Fellingham. Performance of the Kenward-Roger method when the covariance structure is selected using AIC and BIC. *Communications in Statistics—Simulation and Computation*, 34(2):377–392, 2005. doi:10.1081/SAC-200055719.

Stef van Buuren. *Flexible Imputation of Missing Data*. Chapman & Hall/CRC, Boca Raton, FL, 2012. ISBN 978-1-439-86824-9.

Stef van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3):1–67, 2011. doi:10.18637/jss.v045.i03.

Aad W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, Cambridge, UK, 1998. ISBN 0-521-49603-9.

Geert Verbeke and Geert Molenberghs, editors. *Linear Mixed Models in Practice: A SAS-Oriented Approach*. Springer, New York, NY, 1997. ISBN 0-387-98222-1.

Geert Verbeke and Geert Molenberghs. *Linear Mixed Models for Longitudinal Data*. Springer, New York, NY, 2000. ISBN 0-387-95027-3.

Geert Verbeke, Steffen Fieuws, Geert Molenberghs, and Marie Davidian. The analysis of multivariate longitudinal data: a review. *Statistical Methods in Medical Research*, 23 (1):42–59, 2014. doi:10.1177/0962280212445834.

Zbyněk Šidák. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62(318):626–633, 1967. doi:10.2307/2283989.

Zengri Wang and Thomas A. Louis. Marginalized binary mixed-effects models with covariate-dependent random effects and likelihood inference. *Biometrics*, 60(4):884–891, 2004. doi:10.1111/j.0006-341X.2004.00243.x.

James M. S. Wason, Lynne Stecher, and Adrian P. Mander. Correcting for multiple-testing in multi-arm trials: is it necessary and is it done? *Trials*, 15(1):article 364, 2014. doi:10.1186/1745-6215-15-364.

Gernot Wassmer, Peter Reitmeir, Meinhard Kieser, and Walter Lehmacher. Procedures for testing multiple endpoints in clinical trials: an overview. *Journal of Statistical Planning and Inference*, 82(1-2):69–81, 1999. doi:10.1016/S0378-3758(99)00032-4.

Robert W. M. Wedderburn. Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, 61(3):439–447, 1974. doi:10.1093/biomet/61.3.439.

Bernard L. Welch. The significance of the difference between two means when the population variances are unequal. *Biometrika*, 29(3-4):350–361, 1938. doi:10.2307/2332010.

Bernard L. Welch. The generalization of ‘Student’s’ problem when several different population variances are involved. *Biometrika*, 34(1-2):28–35, 1947. doi:10.1093/biomet/34.1-2.28.

Peter H. Westfall and S. Stanley Young. *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. John Wiley & Sons, New York, NY, 1993. ISBN 978-0-471-55761-6.

Peter H. Westfall, Randall D. Tobias, and Russell D. Wolfinger. *Multiple Comparisons and Multiple Tests Using SAS. Second Edition*. SAS Institute, Cary, NC, 2011.

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer, New York, NY, 2009. ISBN 978-0-387-98140-6.

Hadley Wickham and Winston Chang. *ggplot2: An implementation of the Grammar of Graphics*, 2015. URL `http://CRAN.R-project.org/package=ggplot2`. R package version 1.0.1.

D. A. Williams. A test for differences between treatment means when several dose levels are compared with a zero dose control. *Biometrics*, 27(1):103–117, 1971. doi:10.2307/2528930. Correction: 31(4):1019, 1975.

D. A. Williams. The comparison of several dose levels with a zero dose control. *Biometrics*, 28(2):519–531, 1972. doi:10.2307/2556164. Correction: 31(4):1019, 1975.

Russ Wolfinger. Covariance structure selection in general mixed models. *Communications in Statistics—Simulation and Computation*, 22(4):1079–1106, 1993. doi:10.1080/03610919308813143.

Russell D. Wolfinger. Heterogeneous variance: covariance structures for repeated measures. *Journal of Agricultural, Biological, and Environmental Statistics*, 1(2):205–230, 1996.

Di Wu, Elgene Lim, François Vaillant, Marie-Liesse Asselin-Labat, Jane E. Visvader, and Gordon K. Smyth. ROAST: Rotation gene set tests for complex microarray experiments. *Bioinformatics*, 26(17):2176–2182, 2010. doi:10.1093/bioinformatics/btq401.

Jun Yan. Enjoy the joy of copulas: with a package copula. *Journal of Statistical Software*, 21(4):1–21, 2007. doi:10.18637/jss.v021.i04.

David A. Young, Gary O. Zerbe, and William W. Hay, Jr. Fieller's theorem, Scheffé simultaneous confidence intervals, and ratios of parameters of linear and nonlinear mixed-effects models. *Biometrics*, 53(3), 1997. doi:10.2307/2533546.

Mary L. Young, John S. Preisser, Bahjat F. Qaqish, and Mark Wolfson. Comparison of subject-specific and population averaged models for count data from cluster-unit intervention trial. *Statistical Methods in Medical Research*, 16(2):176–184, 2007. doi:10.1177/0962280206071931.

Scott L. Zeger and Kung-Yee Liang. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, 42(1), 1986. doi:10.2307/2531248.

Scott L. Zeger, Kung-Yee Liang, and Paul S. Albert. Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, 44(4):1049–1060, 1988. doi:10.2307/2531734.

Achim Zeileis. Object-oriented computation of sandwich estimators. *Journal of Statistical Software*, 16(9):1–16, 2006. doi:10.18637/jss.v016.i09.

Gary O. Zerbe, Eugene Laska, Morris Meisner, and Howard B. Kushner. On multivariate confidence regions and simultaneous confidence limits for ratios. *Communications in Statistics—Theory and Methods*, 11(21):2401–2425, 1982. doi:10.1080/03610928208828398.

H. Zhang, Q. Yu, C. Feng, D. Gunzler, P. Wu, and X. M. Tu. A new look at the difference between the GEE and the GLMM when modeling longitudinal count responses. *Journal of Applied Statistics*, 39(9):2067–2079, 2012. doi:10.1080/02664763.2012.700452.

Hui Zhang, Naiji Lu, Changyong Feng, Sally W. Thurston, Yinglin Xia, Liang Zhu, and Xin M. Tu. On fitting generalized linear mixed-effects models for binary responses using different statistical packages. *Statistics in Medicine*, 30(20):2562–2572, 2011. doi:10.1002/sim.4265.

A Multivariate Normal and t-Distribution

A.1 Multivariate Normal Distribution

A random variable X is (univariate) normal with mean μ and variance $\sigma^2 > 0$

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

if it has density function (PDF)

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

The corresponding distribution function (CDF) is

$$F(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \int\limits_{-\infty}^{x} e^{-\frac{(t-\mu)^2}{2\sigma^2}} \, \mathrm{d}t.$$

The special case $\mu = 0$ and $\sigma^2 = 1$ is called standard normal.

The normal distribution be extended to the multivariate case as follows. A vector of random variables $\mathbf{X} = (X_1, \ldots, X_p)^T$ is nonsingular p-variate normal with mean vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^T$ and positive definite covariance matrix $\boldsymbol{\Sigma}^{p \times p}$

$$\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

if it has joint density function

$$f(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^p |\boldsymbol{\Sigma}|}} e^{-\frac{(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})}{2}}$$

where $|\boldsymbol{\Sigma}|$ denotes the determinant of $\boldsymbol{\Sigma}$. This is a central multivariate normal distribution if $\mu = \mathbf{0}$.

The distribution function is not available in closed form, except for a few special cases such as bivariate (Siddiqui 1967). Numerical methods for evaluating the integral

$$\Phi_p(\mathbf{x}; \mathbf{a}, \mathbf{b}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^p |\boldsymbol{\Sigma}|}} \int_{a_1}^{b_1} \ldots \int_{a_p}^{b_p} e^{-\frac{(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})}{2}} dx_p \ldots dx_1$$

are described in Genz and Bretz (2009). Extensive treatments of the properties of the multivariate normal distribution can be found in Tong (1990) and in chapter 2 of Kotz et al. (2000).

A.2 Multivariate t-Distribution

The distribution of a random variable T is (univariate) Student t with $\nu > 0$ degrees of freedom (DF)

$$T \sim \mathcal{T}(\nu)$$

if it has density function

$$\begin{aligned} g(t;\nu) &= \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}} \\ &= \frac{1}{\sqrt{\nu}B(\frac{1}{2},\frac{\nu}{2})}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}} \end{aligned}$$

where Γ denotes the gamma function

$$\Gamma(x) = \int_0^\infty a^{x-1}e^{-a}\,\mathrm{d}a$$

and B the beta function

$$B(x,y) = \int_0^1 a^{x-1}(1-a)^{y-1}\,\mathrm{d}a.$$

The corresponding distribution function is

$$G(t;\nu) = \frac{1}{2} + \frac{t}{2|t|}I_{\frac{t^2}{t^2+\nu}}\left(\frac{1}{2},\frac{\nu}{2}\right)$$

with I_z the regularized incomplete beta function

$$I_z(x,y) = \frac{B(z;x,y)}{B(x,y)}$$

where the numerator is the incomplete beta function

$$B(z;x,y) = \int_0^z a^{x-1}(1-a)^{y-1}\,\mathrm{d}a.$$

Student's t-distribution with ν DF can be derived as the ratio of a standard normal variable divided by an independent χ^2 variable with ν DF:

$$Z\sqrt{\frac{\nu}{\chi^2_\nu}} \sim \mathcal{T}(\nu)$$

where

$$Z \sim \mathcal{N}(0,1).$$

The special case of $\nu = 1$ is known as Cauchy, and for $\nu \to \infty$ the Student t converges in distribution to normality.

There exists no unique extension to the multivariate case. One typical form of the multivariate t-distribution (but not the only possible one) is achieved by dividing a multivariate Gaussian vector by a common χ^2 variable with ν DF. The "most natural" form (Kotz and Nadarajah 2004) defines the distribution of a vector of (univariate) t random variables $\mathbf{T} = (T_1,\ldots,T_n)^T$, and its density function is given by

$$g(\mathbf{t};\nu,\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{\Gamma(\frac{\nu+p}{2})}{\sqrt{(\pi\nu)^p}\Gamma(\frac{\nu}{2})\sqrt{|\boldsymbol{\Sigma}|}}\left(1+\frac{(\mathbf{t}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{t}-\boldsymbol{\mu})}{\nu}\right)^{-\frac{\nu+p}{2}}$$

where $\boldsymbol{\mu} = \mathbf{0}$ makes the distribution a central multivariate t. Setting $\nu = 1$ yields a multivariate Cauchy distribution, and for $\nu \to \infty$ the multivariate t converges in distribution to multivariate normality.

There is again no analytical expression for the distribution function, but Genz and Bretz (2009) provide numerical and computational methods for solving

$$T_p(\mathbf{t}; \mathbf{a}, \mathbf{b}, \nu, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{\Gamma(\frac{\nu+p}{2})}{\sqrt{(\pi\nu)^p}\Gamma(\frac{\nu}{2})\sqrt{|\boldsymbol{\Sigma}|}} \int_{a_1}^{b_1} \dots \int_{a_p}^{b_p} \left(1 + \frac{(\mathbf{t}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{t}-\boldsymbol{\mu})}{\nu}\right)^{-\frac{\nu+p}{2}} dx_p \dots dx_1.$$

A plethora of other possible multidimensional t-distributions (some of which are e.g., skewed or have variable DFs) are summarized in Nadarajah and Kotz (2004) and Nadarajah and Dey (2005).[6]

[6] These two journal articles are nearly identical to chapters 4 and 5 of Kotz and Nadarajah (2004).

B Copulas

A copula is a joint multivariate CDF with marginal univariate CDFs that are continuous uniform on the unit interval $[0, 1]$. Suppose the marginal CDFs of the random variables $X_1, \ldots, X_n$ are known to be $F_1, \ldots, F_n$ but their *joint* CDF F is unclear. We can define $U_j = F_j(X_j)$ for $j = 1, \ldots, n$ so that

$$U_j \sim \mathcal{U}(0, 1)$$

and the joint CDF of $U_1, \ldots, U_n$ is C.

By the theorem of Sklar (1959) any multivariate joint CDF F can be written as an n-dimensional copula of n univariate marginal CDFs $F_1, \ldots, F_n$:

$$F(x_1, \ldots, x_n) = C\{F_1(x_1), \ldots, F_n(x_n)\},$$

or likewise:

$$F\{F_1^{-1}(u_1), \ldots, F_n^{-1}(u_n)\} = C(u_1, \ldots, u_n).$$

Copulas can be used as a means for “gluing” marginal CDFs together so as to arrive at a joint multivariate CDF, given some dependence structure. We exploit this property in Chapter 5 to draw correlated binary and Poisson data. A more thorough introduction to copulas can be found in the textbook by Nelsen (2006).

C Example Datasets

The following tables show the raw data of all six example dataset described in Chapter 2 and evaluated in Chapters 4 and 5.

C.1 Bradykinin Receptor Antagonism

Table 17 displays the summary statistics of the original data on bradykinin receptor antagonism (Balaguer et al. 2013) that were used to create the simulated dataset that was described in 2.1 and analyzed in 4.4.1.

Table 17: Summary statistics of the log-concentrations of the original bradykinin dataset: number of independent units, sample means, sample covariances, and correlations for each treatment group.

	n	$\boldsymbol{\mu}$	$\boldsymbol{\Sigma}$	$\mathbf{R}$
Placebo	37	$\begin{pmatrix} 4.359 \\ 4.830 \\ 5.332 \\ 5.982 \\ 5.654 \end{pmatrix}$	$\begin{pmatrix} 0.470 & 0.314 & 0.216 & 0.217 & 0.172 \\ 0.314 & 0.435 & 0.368 & 0.278 & 0.072 \\ 0.216 & 0.368 & 0.690 & 0.287 & -0.007 \\ 0.217 & 0.278 & 0.287 & 0.805 & 0.287 \\ 0.172 & 0.072 & -0.007 & 0.287 & 0.594 \end{pmatrix}$	$\begin{pmatrix} 1 & 0.696 & 0.380 & 0.353 & 0.326 \\ 0.696 & 1 & 0.672 & 0.470 & 0.142 \\ 0.380 & 0.672 & 1 & 0.384 & -0.010 \\ 0.353 & 0.470 & 0.384 & 1 & 0.415 \\ 0.326 & 0.142 & -0.010 & 0.415 & 1 \end{pmatrix}$
EACA	37	$\begin{pmatrix} 4.275 \\ 4.586 \\ 4.695 \\ 4.882 \\ 5.224 \end{pmatrix}$	$\begin{pmatrix} 0.640 & 0.520 & 0.525 & 0.301 & 0.039 \\ 0.520 & 0.595 & 0.593 & 0.408 & 0.085 \\ 0.525 & 0.593 & 0.674 & 0.435 & 0.068 \\ 0.301 & 0.408 & 0.435 & 0.476 & 0.168 \\ 0.039 & 0.085 & 0.068 & 0.168 & 0.734 \end{pmatrix}$	$\begin{pmatrix} 1 & 0.857 & 0.796 & 0.548 & 0.057 \\ 0.857 & 1 & 0.937 & 0.766 & 0.128 \\ 0.796 & 0.937 & 1 & 0.767 & 0.097 \\ 0.548 & 0.766 & 0.767 & 1 & 0.285 \\ 0.057 & 0.128 & 0.097 & 0.285 & 1 \end{pmatrix}$
HOE 140	38	$\begin{pmatrix} 4.243 \\ 4.920 \\ 5.403 \\ 6.170 \\ 5.627 \end{pmatrix}$	$\begin{pmatrix} 0.532 & 0.371 & 0.278 & 0.261 & 0.074 \\ 0.371 & 0.686 & 0.654 & 0.563 & 0.141 \\ 0.278 & 0.654 & 0.898 & 0.470 & 0.184 \\ 0.261 & 0.563 & 0.470 & 1.092 & 0.236 \\ 0.074 & 0.141 & 0.184 & 0.236 & 0.407 \end{pmatrix}$	$\begin{pmatrix} 1 & 0.613 & 0.426 & 0.342 & 0.159 \\ 0.613 & 1 & 0.834 & 0.650 & 0.267 \\ 0.426 & 0.834 & 1 & 0.481 & 0.296 \\ 0.342 & 0.650 & 0.481 & 1 & 0.354 \\ 0.159 & 0.267 & 0.296 & 0.354 & 1 \end{pmatrix}$

ID	Drug	Baseline	Bypass (30 min)	Bypass (60 min)	Post-Bypass	Postoperative Day 1
1	Placebo	5.0328	5.3976	5.7022	6.9740	6.6857
2	Placebo	3.2558	4.6634	4.9376	4.7945	6.1281
3	Placebo	4.3291	4.9985	5.3642	4.6745	5.6563
4	Placebo	5.0445	5.6254	6.8620	7.1506	6.0146
5	Placebo	4.6330	5.3544	5.0056	5.7522	5.4053
6	Placebo	4.4899	4.7122	5.5447	5.9043	5.1270
7	Placebo	5.5365	5.4413	4.5876	6.6356	5.2526
8	Placebo	4.7983	5.0799	5.2646	6.6610	6.8176
9	Placebo	5.6800	6.3052	5.8627	7.5499	5.8322
10	Placebo	4.2362	4.6818	5.1980	5.4522	5.7017
11	Placebo	4.6783	5.3148	5.4520	6.3755	5.9469
12	Placebo	4.6946	4.8387	6.0578	5.2295	5.0096
13	Placebo	5.0146	4.6146	5.7012	4.5909	5.2785
14	Placebo	2.9159	4.1731	4.1696	6.6341	5.2684
15	Placebo	4.7121	3.9182	5.0801	5.7692	6.0655
16	Placebo	4.5448	4.9799	5.0404	6.5512	6.3418
17	Placebo	3.7075	4.4197	5.4676	5.6720	4.7158
18	Placebo	4.0451	4.2258	5.2906	5.4650	5.7607
19	Placebo	4.7075	5.4494	6.2947	6.3584	5.4620
20	Placebo	4.5409	4.6756	5.2056	5.4788	5.4775
21	Placebo	5.4544	5.6356	6.9267	5.7817	6.3965
22	Placebo	3.5622	4.6621	4.6821	5.5313	5.8226
23	Placebo	4.0237	4.1471	4.4944	4.8059	5.5482
24	Placebo	4.0242	4.6737	4.0138	4.9760	6.3968
25	Placebo	4.8555	5.2929	5.5312	6.8874	6.7966
26	Placebo	4.4740	4.9425	6.4587	5.6305	5.4560
27	Placebo	3.3500	3.7805	6.5075	6.6543	5.6083
28	Placebo	5.4953	5.7717	6.0502	5.9109	6.5771
29	Placebo	4.3722	4.1659	5.5897	6.1666	6.6083
30	Placebo	3.2259	4.4792	6.2273	6.3012	5.8045
31	Placebo	3.9048	4.0390	4.3449	5.9357	5.3479
32	Placebo	5.1379	5.1654	5.5610	6.3409	6.9182
33	Placebo	4.4870	4.9854	5.5398	5.6722	5.2836
34	Placebo	3.8467	4.0678	4.7957	7.1809	5.4543
35	Placebo	5.3440	5.2924	4.3547	7.2827	6.0409
36	Placebo	5.1262	4.0154	4.6562	5.2560	6.8244

37	Placebo	4.5716	5.2354	4.8145	5.7797	5.1237
38	EACA	3.6417	4.2443	4.3455	4.1988	5.9690
39	EACA	5.2347	5.4318	5.3544	5.0166	6.3620
40	EACA	2.6799	3.1637	2.7691	4.1240	5.7665
41	EACA	4.7959	4.7436	4.6865	4.6506	6.1241
42	EACA	5.2384	5.3986	5.3641	6.3473	6.9881
43	EACA	3.5313	3.3329	3.6316	4.2515	6.2711
44	EACA	5.4374	5.4060	5.3116	5.1482	5.7317
45	EACA	4.7497	5.2387	5.6755	5.9061	6.4901
46	EACA	3.9994	4.5213	4.4003	4.9266	4.7635
47	EACA	4.8602	5.1202	5.3856	4.6021	4.4033
48	EACA	3.9622	4.8755	5.2243	5.4294	6.4295
49	EACA	3.9181	4.7215	4.8748	4.8122	4.4851
50	EACA	3.3999	3.8630	4.6350	4.7032	4.8127
51	EACA	4.9221	5.1586	5.5928	5.3106	4.7746
52	EACA	5.2920	5.3436	5.4094	5.0067	4.4851
53	EACA	4.5566	5.4543	5.3121	5.4264	5.7615
54	EACA	4.7541	4.3777	4.9523	4.7395	5.4966
55	EACA	5.4079	5.1810	5.2174	5.3912	6.6474
56	EACA	5.1635	5.4551	5.5464	6.1374	6.1371
57	EACA	3.9581	4.3298	4.1104	5.0314	6.0722
58	EACA	3.4613	3.8509	4.4477	4.7059	4.6487
59	EACA	3.7405	4.2339	4.4179	5.1570	5.7681
60	EACA	5.6767	5.6736	5.7947	5.6614	6.8070
61	EACA	4.3312	4.7918	5.0311	5.2973	4.9566
62	EACA	4.3916	4.5870	4.1308	4.4881	5.0201
63	EACA	4.7908	4.2328	4.6065	4.7723	5.5932
64	EACA	4.0540	4.3076	4.3098	4.4834	4.6896
65	EACA	3.1456	2.8166	2.8126	2.9901	5.5030
66	EACA	3.7806	3.5349	3.3830	4.2426	5.1109
67	EACA	3.1163	3.5123	3.0519	3.7246	5.4771
68	EACA	5.3308	5.2923	5.6081	5.4353	4.3036
69	EACA	3.8724	3.3093	3.7313	4.0622	6.2560
70	EACA	4.2644	4.3283	3.7426	4.0981	6.1142
71	EACA	6.1016	5.7934	5.8101	5.3895	5.4321
72	EACA	4.0072	4.5687	4.8483	5.2821	5.3571
73	EACA	5.5426	5.6376	6.0201	4.6867	5.1257
74	EACA	3.2589	4.3041	4.3615	5.8699	6.2428
75	HOE140	4.5445	4.8873	5.4660	6.3073	5.5362
76	HOE140	3.9803	4.8262	5.4041	6.2802	5.1087
77	HOE140	3.9903	4.3823	5.4635	5.4369	4.8006
78	HOE140	3.7487	4.3809	4.4025	5.0761	5.4158
79	HOE140	4.5056	6.4541	7.2489	5.2038	5.2371
80	HOE140	4.7335	4.5281	4.9957	5.9588	5.5508
81	HOE140	3.8898	5.5183	6.1211	4.6945	4.5179
82	HOE140	5.0462	5.0754	4.7602	6.8555	6.1951
83	HOE140	4.2530	5.3592	5.8218	7.3369	5.5629
84	HOE140	5.2032	4.1962	4.6478	4.7834	5.6296
85	HOE140	4.8197	4.6761	4.6860	7.1969	6.2184
86	HOE140	3.8338	3.2804	4.4123	3.2687	5.3690
87	HOE140	4.3278	5.4680	7.8066	5.9094	6.4212
88	HOE140	4.7870	5.4737	6.2346	7.5477	5.5188
89	HOE140	5.3246	7.6445	9.1348	8.1383	5.4224
90	HOE140	4.4823	5.2752	5.6341	7.2637	6.0082
91	HOE140	4.8353	5.3661	5.7152	7.9531	4.9737
92	HOE140	4.0152	4.3153	4.0224	7.0902	5.7680
93	HOE140	5.2540	5.8040	5.2497	7.5609	6.5343
94	HOE140	4.5069	5.4298	5.0879	5.0988	5.7403
95	HOE140	5.2905	5.7389	5.3995	6.4169	4.7768
96	HOE140	3.7707	4.9903	5.8444	5.9301	6.1398
97	HOE140	4.5504	6.0234	7.2443	5.4204	5.8115
98	HOE140	3.4947	5.4485	6.5225	6.7342	6.5223
99	HOE140	4.7115	4.3043	5.4028	5.6985	5.5124
100	HOE140	4.2983	4.2313	4.4258	6.1352	5.1057
101	HOE140	3.9304	4.8874	4.9906	6.4953	4.9962
102	HOE140	4.7710	4.8861	5.6303	6.6411	5.9891
103	HOE140	4.6426	5.3947	5.6087	7.6143	5.7604
104	HOE140	3.2668	4.3548	4.4293	6.2312	5.1774
105	HOE140	4.0637	2.8750	3.9118	3.6200	5.1403
106	HOE140	5.1109	5.3725	5.5433	5.4071	4.6849
107	HOE140	4.0725	4.7567	5.0518	4.8867	5.2484
108	HOE140	4.6810	5.2553	5.8034	4.6827	5.9189
109	HOE140	3.9481	5.5836	5.8453	8.1567	5.2388

110	HOE140	4.3871	5.2834	6.1268	6.6257	5.6175
111	HOE140	4.5206	5.8390	6.3119	6.7712	5.6876
112	HOE140	4.3511	3.7497	4.6832	5.7740	5.9268

C.2 Mercuric Chloride

Animal	Cage	Dose	Week 53	Week 65	Week 105
30233	13	Vehicle Control	281.40	307.50	369.20
30234	13	Vehicle Control	263.00	284.90	297.90
30235	13	Vehicle Control	283.10	313.10	NA
30236	13	Vehicle Control	272.00	300.40	NA
30237	13	Vehicle Control	272.20	308.40	330.80
30238	14	Vehicle Control	277.80	333.60	NA
30239	14	Vehicle Control	246.20	281.20	NA
30240	14	Vehicle Control	291.60	317.40	360.60
30241	14	Vehicle Control	306.00	330.60	364.80
30242	14	Vehicle Control	284.50	315.00	NA
30243	15	Vehicle Control	257.30	286.00	339.10
30244	15	Vehicle Control	288.60	306.00	NA
30245	15	Vehicle Control	250.90	274.50	298.50
30246	15	Vehicle Control	300.30	330.50	334.80
30247	15	Vehicle Control	288.50	321.00	330.10
30248	16	Vehicle Control	274.10	310.60	NA
30249	16	Vehicle Control	297.70	327.00	361.50
30250	16	Vehicle Control	267.40	290.40	NA
30251	16	Vehicle Control	277.80	311.50	338.20
30252	16	Vehicle Control	286.20	324.50	349.30
30253	17	Vehicle Control	291.10	305.20	NA
30254	17	Vehicle Control	263.80	290.40	NA
30255	17	Vehicle Control	261.20	295.20	332.30
30256	17	Vehicle Control	244.70	285.90	NA
30257	17	Vehicle Control	286.40	317.30	271.70
30258	18	Vehicle Control	271.60	304.60	NA
30259	18	Vehicle Control	254.90	283.20	329.50
30260	18	Vehicle Control	297.00	334.30	NA
30261	18	Vehicle Control	289.10	322.30	352.50
30262	18	Vehicle Control	235.90	247.80	280.00
30263	19	Vehicle Control	244.80	266.20	299.10
30264	19	Vehicle Control	251.30	276.40	NA
30265	19	Vehicle Control	260.40	289.90	NA
30266	19	Vehicle Control	280.00	301.30	NA
30267	19	Vehicle Control	247.00	265.10	NA
30268	20	Vehicle Control	297.00	334.60	NA
30269	20	Vehicle Control	296.60	319.50	NA
30270	20	Vehicle Control	270.30	282.30	NA
30271	20	Vehicle Control	273.80	299.10	NA
30272	20	Vehicle Control	292.10	323.30	377.60
30273	21	Vehicle Control	273.60	309.40	343.90
30274	21	Vehicle Control	272.70	291.70	318.10
30275	21	Vehicle Control	261.50	291.10	324.70
30276	21	Vehicle Control	277.00	305.60	NA
30277	21	Vehicle Control	282.50	300.30	NA
30278	22	Vehicle Control	292.90	318.80	NA
30279	22	Vehicle Control	280.30	304.00	NA
30280	22	Vehicle Control	301.10	331.00	NA
30281	22	Vehicle Control	275.30	302.80	NA
30282	22	Vehicle Control	264.80	300.70	313.60
30283	23	Vehicle Control	280.80	320.50	345.70
30284	23	Vehicle Control	293.60	324.30	NA
30285	23	Vehicle Control	261.30	290.50	326.20
30286	23	Vehicle Control	279.50	307.40	334.50
30287	23	Vehicle Control	278.20	303.70	321.60
30288	24	Vehicle Control	258.70	288.20	254.90
30289	24	Vehicle Control	271.30	306.10	303.30
30290	24	Vehicle Control	266.20	296.70	327.20
30291	24	Vehicle Control	267.50	288.00	303.80
30292	24	Vehicle Control	268.30	302.60	315.40
30353	37	2.5 mg/kg	237.10	245.00	314.60
30354	37	2.5 mg/kg	253.30	287.20	326.90

30355	37	2.5 mg/kg	283.00	307.20	NA
30356	37	2.5 mg/kg	255.30	276.60	323.10
30357	37	2.5 mg/kg	287.40	314.50	NA
30358	38	2.5 mg/kg	283.10	305.70	NA
30359	38	2.5 mg/kg	232.40	255.40	NA
30360	38	2.5 mg/kg	262.50	273.00	NA
30362	38	2.5 mg/kg	247.70	278.40	309.60
30363	39	2.5 mg/kg	253.10	273.80	327.30
30364	39	2.5 mg/kg	247.90	275.40	NA
30365	39	2.5 mg/kg	258.00	274.70	319.00
30366	39	2.5 mg/kg	250.60	263.10	338.00
30367	39	2.5 mg/kg	257.70	277.60	NA
30368	40	2.5 mg/kg	231.80	253.50	NA
30369	40	2.5 mg/kg	255.70	279.90	324.90
30370	40	2.5 mg/kg	273.10	303.80	326.50
30371	40	2.5 mg/kg	257.80	289.30	278.60
30372	40	2.5 mg/kg	318.50	342.80	NA
30373	41	2.5 mg/kg	246.10	257.70	260.10
30374	41	2.5 mg/kg	271.70	286.00	NA
30375	41	2.5 mg/kg	245.30	258.80	335.30
30376	41	2.5 mg/kg	257.00	277.80	NA
30377	41	2.5 mg/kg	260.40	281.60	359.00
30378	42	2.5 mg/kg	252.80	256.00	NA
30379	42	2.5 mg/kg	230.50	250.80	257.70
30380	42	2.5 mg/kg	230.30	255.70	NA
30381	42	2.5 mg/kg	239.50	245.70	302.90
30382	42	2.5 mg/kg	243.50	263.10	308.00
30383	43	2.5 mg/kg	260.60	281.40	NA
30384	43	2.5 mg/kg	247.90	259.80	291.70
30385	43	2.5 mg/kg	258.70	291.10	NA
30386	43	2.5 mg/kg	276.50	303.30	NA
30387	43	2.5 mg/kg	259.30	301.40	345.60
30388	44	2.5 mg/kg	252.80	280.20	NA
30390	44	2.5 mg/kg	252.00	268.40	NA
30391	44	2.5 mg/kg	223.40	231.70	279.90
30392	44	2.5 mg/kg	253.30	279.50	303.50
30393	45	2.5 mg/kg	247.70	280.50	NA
30394	45	2.5 mg/kg	243.40	253.40	NA
30395	45	2.5 mg/kg	259.20	279.50	328.10
30396	45	2.5 mg/kg	238.00	251.00	NA
30397	45	2.5 mg/kg	237.60	NA	NA
30399	46	2.5 mg/kg	240.50	263.10	184.90
30400	46	2.5 mg/kg	238.70	260.00	NA
30401	46	2.5 mg/kg	287.70	301.50	227.60
30402	46	2.5 mg/kg	242.30	266.90	NA
30403	47	2.5 mg/kg	254.50	275.30	NA
30405	47	2.5 mg/kg	254.90	265.00	307.30
30406	47	2.5 mg/kg	241.20	255.60	NA
30407	47	2.5 mg/kg	228.20	244.60	NA
30408	48	2.5 mg/kg	211.70	242.00	267.00
30409	48	2.5 mg/kg	244.70	255.30	NA
30410	48	2.5 mg/kg	260.20	280.30	NA
30411	48	2.5 mg/kg	259.20	277.80	NA
30412	48	2.5 mg/kg	235.60	264.60	258.30
30473	61	5 mg/kg	227.60	248.50	258.40
30474	61	5 mg/kg	240.00	265.90	318.20
30476	61	5 mg/kg	237.10	264.20	298.10
30477	61	5 mg/kg	217.30	228.70	NA
30478	62	5 mg/kg	241.90	249.70	313.60
30479	62	5 mg/kg	247.80	274.20	NA
30480	62	5 mg/kg	230.40	269.40	NA
30481	62	5 mg/kg	268.20	295.80	327.90
30482	62	5 mg/kg	243.60	253.20	NA
30483	63	5 mg/kg	233.90	235.80	251.30
30485	63	5 mg/kg	265.10	306.30	332.60
30486	63	5 mg/kg	233.20	250.70	306.70
30487	63	5 mg/kg	208.30	236.50	260.10
30488	64	5 mg/kg	238.40	262.00	287.90
30489	64	5 mg/kg	275.40	291.80	304.40
30490	64	5 mg/kg	228.90	246.20	NA
30491	64	5 mg/kg	271.20	299.60	330.90
30492	64	5 mg/kg	232.70	238.40	NA
30493	65	5 mg/kg	264.90	280.90	273.60

30494	65	5 mg/kg	257.30	277.80	NA
30495	65	5 mg/kg	230.90	248.30	300.60
30496	65	5 mg/kg	224.10	233.50	279.50
30497	65	5 mg/kg	281.10	301.40	NA
30498	66	5 mg/kg	229.40	246.60	310.50
30499	66	5 mg/kg	248.10	272.10	NA
30500	66	5 mg/kg	246.90	259.20	290.50
30501	66	5 mg/kg	242.80	259.40	300.70
30502	66	5 mg/kg	268.80	280.40	NA
30503	67	5 mg/kg	244.00	246.60	293.70
30504	67	5 mg/kg	240.60	273.90	NA
30505	67	5 mg/kg	245.30	275.80	296.00
30507	67	5 mg/kg	256.50	290.20	NA
30508	68	5 mg/kg	249.70	256.60	290.10
30509	68	5 mg/kg	241.40	278.20	NA
30510	68	5 mg/kg	258.20	280.60	NA
30511	68	5 mg/kg	280.30	298.70	NA
30512	68	5 mg/kg	222.80	247.60	NA
30513	69	5 mg/kg	230.10	247.00	NA
30514	69	5 mg/kg	242.50	271.90	293.70
30515	69	5 mg/kg	251.10	269.10	302.40
30516	69	5 mg/kg	237.10	262.90	306.40
30517	69	5 mg/kg	259.60	298.10	NA
30518	70	5 mg/kg	200.60	226.40	177.10
30519	70	5 mg/kg	260.80	295.50	NA
30520	70	5 mg/kg	244.50	275.40	301.80
30521	70	5 mg/kg	235.80	271.50	NA
30522	70	5 mg/kg	249.70	269.10	NA
30523	71	5 mg/kg	247.70	284.70	NA
30524	71	5 mg/kg	237.30	262.70	NA
30525	71	5 mg/kg	251.50	288.70	NA
30526	71	5 mg/kg	300.20	NA	NA
30527	71	5 mg/kg	235.90	250.70	NA
30529	72	5 mg/kg	255.30	262.30	NA
30530	72	5 mg/kg	244.30	270.10	NA
30531	72	5 mg/kg	252.40	271.70	184.60
30532	72	5 mg/kg	235.30	250.50	272.00

C.3 Heart Rates

Person	Drug	Time 1	Time 2	Time 3	Time 4
1	AX23	72	86	81	77
2	AX23	78	83	88	81
3	AX23	71	82	81	75
4	AX23	72	83	83	69
5	AX23	66	79	77	66
6	AX23	74	83	84	77
7	AX23	62	73	78	70
8	AX23	69	75	76	70
9	BWW9	85	86	83	80
10	BWW9	82	86	80	84
11	BWW9	71	78	70	75
12	BWW9	83	88	79	81
13	BWW9	86	85	76	76
14	BWW9	85	82	83	80
15	BWW9	79	83	80	81
16	BWW9	83	84	78	81
17	Control	69	73	72	74
18	Control	66	62	67	73
19	Control	84	90	88	87
20	Control	80	81	77	72
21	Control	72	72	69	70
22	Control	65	62	65	61
23	Control	75	69	69	68
24	Control	71	70	65	63

C.4 Epileptic Seizures

Patient	Treatment	Baseline	2 Week	4 Weeks	6 Weeks	8 Weeks	Age
1	Placebo	11	5	3	3	3	31
2	Placebo	11	3	5	3	3	30
3	Placebo	6	2	4	0	5	25
4	Placebo	8	4	4	1	4	36
5	Placebo	66	7	18	9	21	22
6	Placebo	27	5	2	8	7	29
7	Placebo	12	6	4	0	2	31
8	Placebo	52	40	20	23	12	42
9	Placebo	23	5	6	6	5	37
10	Placebo	10	14	13	6	0	28
11	Placebo	52	26	12	6	22	36
12	Placebo	33	12	6	8	5	24
13	Placebo	18	4	4	6	2	23
14	Placebo	42	7	9	12	14	36
15	Placebo	87	16	24	10	9	26
16	Placebo	50	11	0	0	5	26
17	Placebo	18	0	0	3	3	28
18	Placebo	111	37	29	28	29	31
19	Placebo	18	3	5	2	5	32
20	Placebo	20	3	0	6	7	21
21	Placebo	12	3	4	3	4	29
22	Placebo	9	3	4	3	4	21
23	Placebo	17	2	3	3	5	32
24	Placebo	28	8	12	2	8	25
25	Placebo	55	18	24	76	25	30
26	Placebo	9	2	1	2	1	40
27	Placebo	10	3	1	4	2	19
28	Placebo	47	13	15	13	12	22
29	Progabide	76	11	14	9	8	18
30	Progabide	38	8	7	9	4	32
31	Progabide	19	0	4	3	0	20
32	Progabide	10	3	6	1	3	20
33	Progabide	19	2	6	7	4	18
34	Progabide	24	4	3	1	3	24
35	Progabide	31	22	17	19	16	30
36	Progabide	14	5	4	7	4	35
37	Progabide	11	2	4	0	4	57
38	Progabide	67	3	7	7	7	20
39	Progabide	41	4	18	2	5	22
40	Progabide	7	2	1	1	0	28
41	Progabide	22	0	2	4	0	23
42	Progabide	13	5	4	0	3	40
43	Progabide	46	11	14	25	15	43
44	Progabide	36	10	5	3	8	21
45	Progabide	38	19	7	6	7	35
46	Progabide	7	1	1	2	4	25
47	Progabide	36	6	10	8	8	26
48	Progabide	11	2	1	0	0	25
49	Progabide	151	102	65	72	63	22
50	Progabide	22	4	3	2	4	32
51	Progabide	42	8	6	5	7	25
52	Progabide	32	1	3	1	5	35
53	Progabide	56	18	11	28	13	21
54	Progabide	24	6	3	4	0	41
55	Progabide	16	3	5	4	3	32
56	Progabide	22	1	23	19	8	26
57	Progabide	25	2	3	0	1	21
58	Progabide	13	0	0	0	0	36
59	Progabide	12	1	4	3	2	37

C.5 Greenhouse Whiteflies

Sample	Environment	Bottom	Middle	Top
1	Greenhouse	0	0	0
2	Greenhouse	0	1	0

3	Greenhouse	0	3	0
4	Greenhouse	0	0	0
5	Greenhouse	0	0	0
6	Greenhouse	0	0	0
7	Greenhouse	8	0	0
8	Greenhouse	6	6	0
9	Greenhouse	10	4	2
10	Greenhouse	1	0	0
11	Greenhouse	0	1	0
12	Greenhouse	1	2	0
13	Greenhouse	0	0	0
14	Greenhouse	0	3	0
15	Greenhouse	0	9	0
16	Greenhouse	1	0	0
17	Greenhouse	0	6	0
18	Greenhouse	0	2	0
19	Greenhouse	2	0	0
20	Greenhouse	4	5	0
21	Greenhouse	6	2	0
22	Greenhouse	2	1	0
23	Greenhouse	3	4	0
24	Greenhouse	3	2	0
25	Greenhouse	3	4	0
26	Greenhouse	1	23	0
27	Greenhouse	9	13	7
28	Greenhouse	7	8	0
29	Greenhouse	8	13	0
30	Greenhouse	8	10	0
31	Greenhouse	13	41	0
32	Greenhouse	11	25	0
33	Greenhouse	10	26	4
34	Greenhouse	1	14	0
35	Greenhouse	2	24	0
36	Greenhouse	9	15	0
37	Greenhouse	10	15	0
38	Greenhouse	11	18	5
39	Greenhouse	12	40	0
40	Greenhouse	21	6	0
41	Greenhouse	7	7	0
42	Greenhouse	8	0	0
43	Greenhouse	36	34	0
44	Greenhouse	8	2	0
45	Greenhouse	7	10	0
46	Greenhouse	8	0	0
47	Greenhouse	5	7	0
48	Greenhouse	5	3	0
49	Greenhouse	7	5	0
50	Greenhouse	16	4	0
51	Greenhouse	10	4	0
52	Greenhouse	7	2	0
53	Greenhouse	23	0	0
54	Greenhouse	2	1	0
55	Greenhouse	23	11	0
56	Greenhouse	8	0	0
57	Greenhouse	14	3	0
58	Greenhouse	6	0	0
59	Greenhouse	12	0	0
60	Greenhouse	10	0	0
61	Greenhouse	1	4	0
62	Greenhouse	11	7	1
63	Greenhouse	33	6	0
64	Greenhouse	3	0	0
65	Greenhouse	4	6	0
66	Greenhouse	0	1	0
67	Greenhouse	14	20	0
68	Greenhouse	3	1	0
69	Greenhouse	9	2	0
70	Greenhouse	1	8	0
71	Greenhouse	2	1	0
72	Greenhouse	5	2	0
73	Greenhouse	3	6	0
74	Greenhouse	0	23	10
75	Greenhouse	30	5	2

76	Greenhouse	9	16	0
77	Greenhouse	6	10	1
78	Greenhouse	6	7	0
79	Greenhouse	55	32	0
80	Greenhouse	4	2	1
81	Greenhouse	8	9	0
82	Greenhouse	12	10	0
83	Greenhouse	3	4	0
84	Greenhouse	8	1	0
85	Tunnel 1	3	6	2
86	Tunnel 1	0	0	0
87	Tunnel 1	0	2	0
88	Tunnel 1	6	0	0
89	Tunnel 1	1	0	1
90	Tunnel 1	0	0	0
91	Tunnel 1	2	2	0
92	Tunnel 1	0	0	0
93	Tunnel 1	0	0	0
94	Tunnel 1	2	0	0
95	Tunnel 1	2	0	0
96	Tunnel 1	0	0	0
97	Tunnel 1	7	2	0
98	Tunnel 1	5	12	0
99	Tunnel 1	1	2	0
100	Tunnel 1	4	9	0
101	Tunnel 1	0	0	0
102	Tunnel 1	NA	NA	0
103	Tunnel 1	6	11	6
104	Tunnel 1	2	4	0
105	Tunnel 1	NA	0	0
106	Tunnel 1	9	0	0
107	Tunnel 1	10	0	0
108	Tunnel 1	NA	NA	0
109	Tunnel 1	10	0	0
110	Tunnel 1	5	6	0
111	Tunnel 1	NA	1	0
112	Tunnel 1	8	0	0
113	Tunnel 1	6	0	0
114	Tunnel 1	NA	NA	NA
115	Tunnel 1	10	1	0
116	Tunnel 1	NA	12	1
117	Tunnel 1	NA	1	2
118	Tunnel 1	5	4	0
119	Tunnel 1	5	1	0
120	Tunnel 1	NA	NA	NA
121	Tunnel 1	10	12	0
122	Tunnel 1	NA	18	1
123	Tunnel 1	NA	2	1
124	Tunnel 1	15	7	0
125	Tunnel 1	15	4	2
126	Tunnel 1	NA	NA	NA
127	Tunnel 2	3	6	0
128	Tunnel 2	0	0	0
129	Tunnel 2	0	0	0
130	Tunnel 2	0	0	0
131	Tunnel 2	0	4	0
132	Tunnel 2	0	0	0
133	Tunnel 2	1	6	0
134	Tunnel 2	0	0	0
135	Tunnel 2	0	2	0
136	Tunnel 2	0	0	0
137	Tunnel 2	9	5	0
138	Tunnel 2	0	2	0
139	Tunnel 2	10	12	0
140	Tunnel 2	4	7	0
141	Tunnel 2	18	8	0
142	Tunnel 2	3	15	0
143	Tunnel 2	9	7	0
144	Tunnel 2	3	6	0
145	Tunnel 2	15	3	0
146	Tunnel 2	4	4	0
147	Tunnel 2	7	9	0
148	Tunnel 2	20	14	0

149	Tunnel 2	6	20	0
150	Tunnel 2	7	7	0
151	Tunnel 2	28	20	0
152	Tunnel 2	8	9	0
153	Tunnel 2	8	1	0
154	Tunnel 2	15	11	0
155	Tunnel 2	25	0	0
156	Tunnel 2	12	0	0
157	Tunnel 2	6	11	2
158	Tunnel 2	11	4	3
159	Tunnel 2	10	5	0
160	Tunnel 2	6	0	0
161	Tunnel 2	18	16	0
162	Tunnel 2	25	0	0
163	Tunnel 2	9	1	1
164	Tunnel 2	9	13	1
165	Tunnel 2	13	18	2
166	Tunnel 2	13	6	0
167	Tunnel 2	22	25	2
168	Tunnel 2	16	9	8

C.6 Azadirachtin

Substrate	Dose	Larvae (Total)	Larvae (Dead)	Pupae (Total)	Pupae (Dead)
CS only	1ml/kg	50	48	2	2
CS only	1ml/kg	275	90	185	110
CS only	1ml/kg	266	112	154	75
CS only	1ml/kg	83	57	26	4
CS only	1ml/kg	81	75	6	6
CS only	1ml/kg	256	131	125	51
CS only	1ml/kg	106	79	27	5
CS only	1ml/kg	76	39	37	33
CS only	1ml/kg	133	62	71	66
CS only	1ml/kg	159	88	71	49
CS only	1.5ml/kg	87	73	14	10
CS only	1.5ml/kg	69	63	6	6
CS only	1.5ml/kg	151	99	52	44
CS only	1.5ml/kg	108	79	29	4
CS only	1.5ml/kg	149	102	47	31
CS only	1.5ml/kg	167	97	70	62
CS only	1.5ml/kg	50	45	5	5
CS only	1.5ml/kg	61	52	9	7
CS only	1.5ml/kg	187	137	50	25
CS only	1.5ml/kg	28	19	9	7
CS only	2ml/kg	54	52	2	2
CS only	2ml/kg	86	85	1	1
CS only	2ml/kg	28	26	2	2
CS only	2ml/kg	102	98	4	4
CS only	2ml/kg	74	62	12	6
CS only	2ml/kg	33	22	11	7
CS only	2ml/kg	55	41	14	11
CS only	2ml/kg	41	29	12	10
CS only	2ml/kg	56	54	2	2
CS only	2ml/kg	62	49	13	13
CS+sand	1ml/kg	99	71	28	23
CS+sand	1ml/kg	236	114	122	87
CS+sand	1ml/kg	86	78	8	7
CS+sand	1ml/kg	63	57	6	5
CS+sand	1ml/kg	156	108	48	29
CS+sand	1ml/kg	131	80	28	7
CS+sand	1ml/kg	162	94	68	51
CS+sand	1ml/kg	212	111	101	88
CS+sand	1ml/kg	122	66	56	38
CS+sand	1ml/kg	47	44	3	3
CS+sand	1.5ml/kg	131	109	22	20
CS+sand	1.5ml/kg	82	80	2	2
CS+sand	1.5ml/kg	178	102	76	52
CS+sand	1.5ml/kg	74	72	2	2
CS+sand	1.5ml/kg	99	59	40	12

CS+sand	1.5ml/kg	31	30	1	1
CS+sand	1.5ml/kg	105	92	13	13
CS+sand	1.5ml/kg	139	101	38	29
CS+sand	1.5ml/kg	78	76	2	2
CS+sand	1.5ml/kg	91	89	2	2
CS+sand	2ml/kg	185	151	34	33
CS+sand	2ml/kg	176	129	47	47
CS+sand	2ml/kg	50	49	1	1
CS+sand	2ml/kg	35	32	3	2
CS+sand	2ml/kg	55	54	1	1
CS+sand	2ml/kg	57	55	2	2
CS+sand	2ml/kg	111	105	6	5
CS+sand	2ml/kg	61	51	10	8
CS+sand	2ml/kg	27	25	2	2
CS+sand	2ml/kg	67	63	4	2

D Further Simulation Results

This appendix provides additional results from the simulation studies described in 4 and 5. Simulated small-sample power curves for longitudinal MCTs with Gaussian endpoints are shown in in D.1. Simulated type I error rates and power curves for asymptotic longitudinal MCTs with binomial and Poisson endpoints are presented in D.2 and D.3.

D.1 Gaussian Data

We show here additional simulation results for small-sample inference in longitudinal settings with a Gaussian endpoint.

Simulated powers for many-to-one, all-pairwise, and grand-mean comparisons of Gaussian means in setups involving $q = \{3, 4, 5\}$ treatment groups and $m = \{3, 4, 5\}$ time points are shown in Figure 60 (for comparisons of treatment groups at multiple occasions) and Figure 62 (for comparisons of occasions within multiple treatment groups).

Simulated powers for many-to-one, all-pairwise, and grand-mean comparisons of Gaussian means in setups involving $q = 3$ treatment groups and $m = 3$ time points under various longitudinal correlations are shown in Figure 61 (for comparisons of treatment groups at multiple occasions) and Figure 63 (for comparisons of occasions within multiple treatment groups).

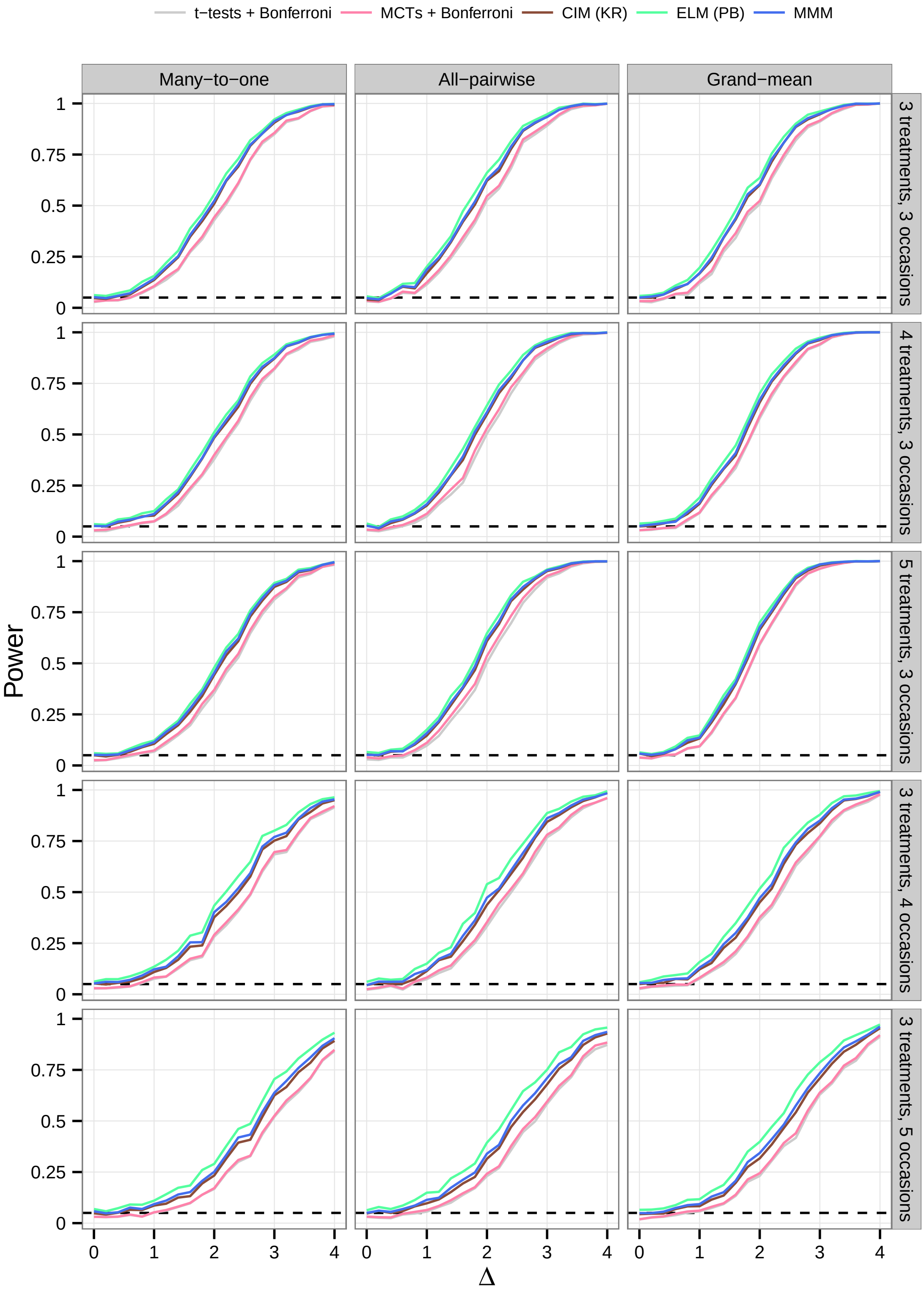


Figure 60: Simulated powers for many-to-one, all-pairwise, and grand-mean comparisons among $q = \{3, 4, 5\}$ Gaussian treatment means separately and simultaneously at $m = \{3, 4, 5\}$ occasions, with $n_k = 10$ independent subjects per treatment group, and different small-sample degrees of freedom (1000 simulation runs).

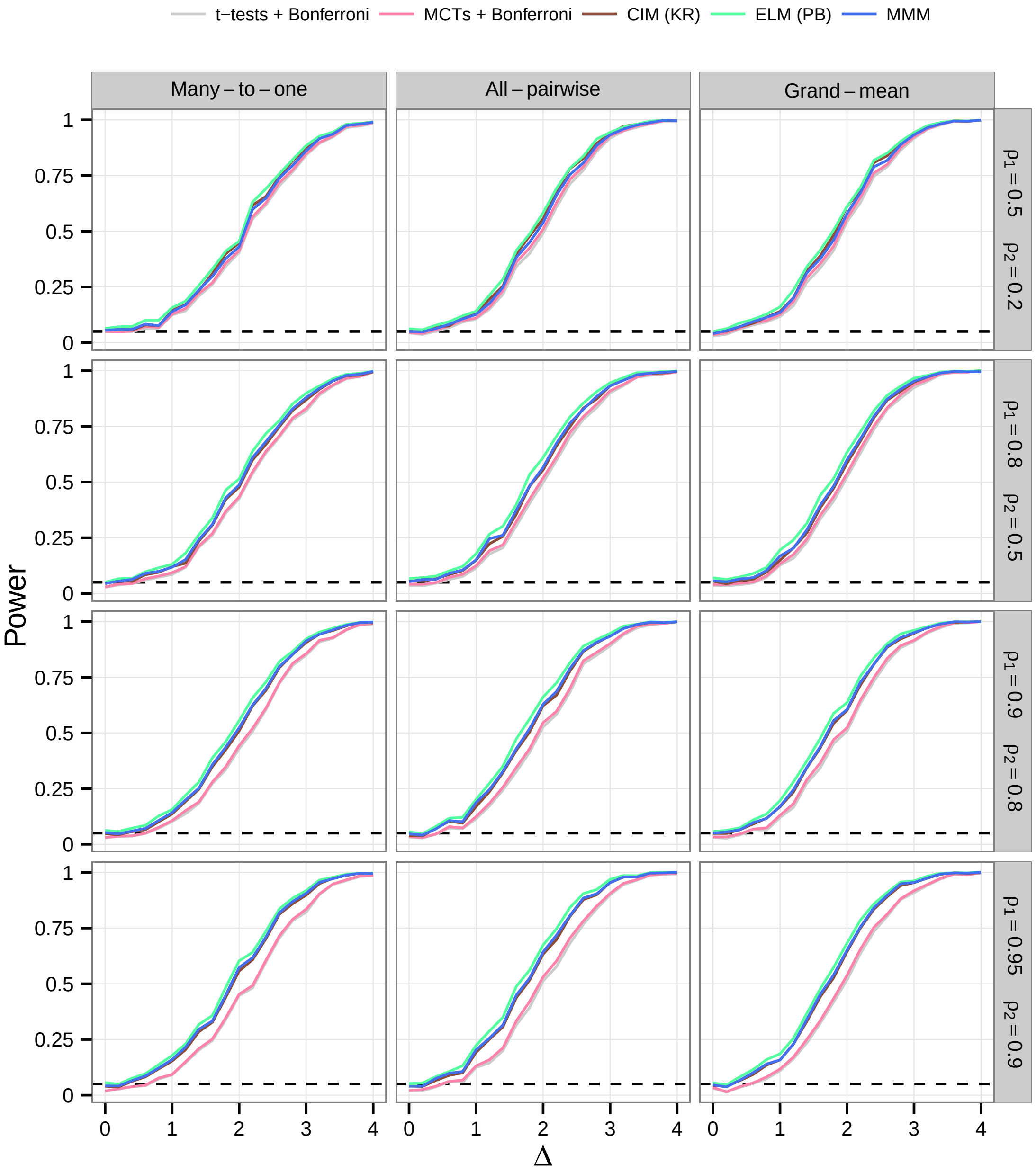


Figure 61: Simulated powers for many-to-one, all-pairwise, and grand-mean comparisons among $q = 3$ Gaussian treatment means separately and simultaneously at $m = 3$ occasions, with $n_k = 10$ independent subjects per treatment group, different longitudinal correlations, and different small-sample degrees of freedom (1000 simulation runs).

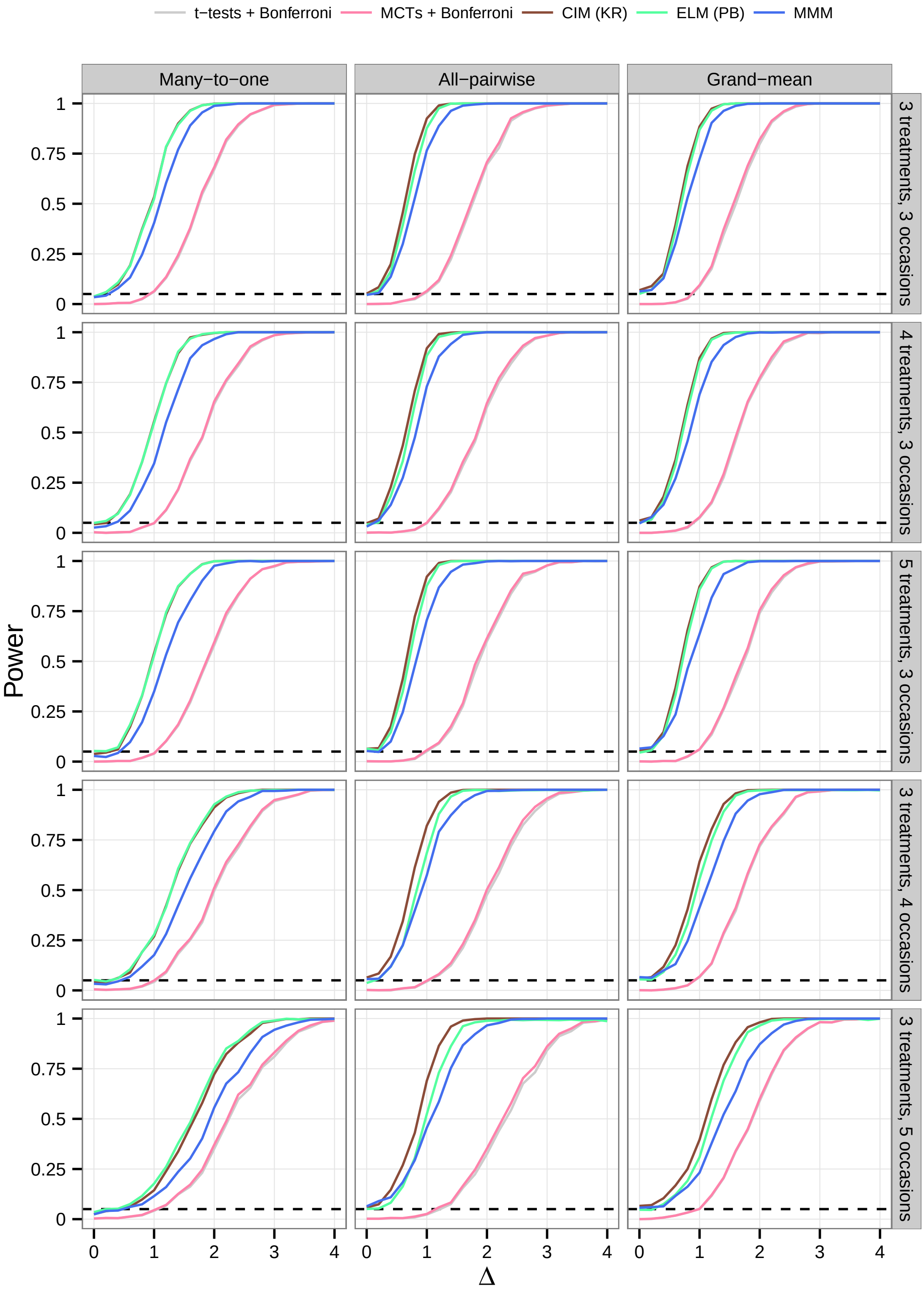


Figure 62: Simulated powers for many-to-one, all-pairwise, and grand-mean comparisons among $m = \{3, 4, 5\}$ Gaussian occasion means separately and simultaneously for $q = \{3, 4, 5\}$ treatments, with $n_k = 10$ independent subjects per treatment group, and different small-sample degrees of freedom (1000 simulation runs).

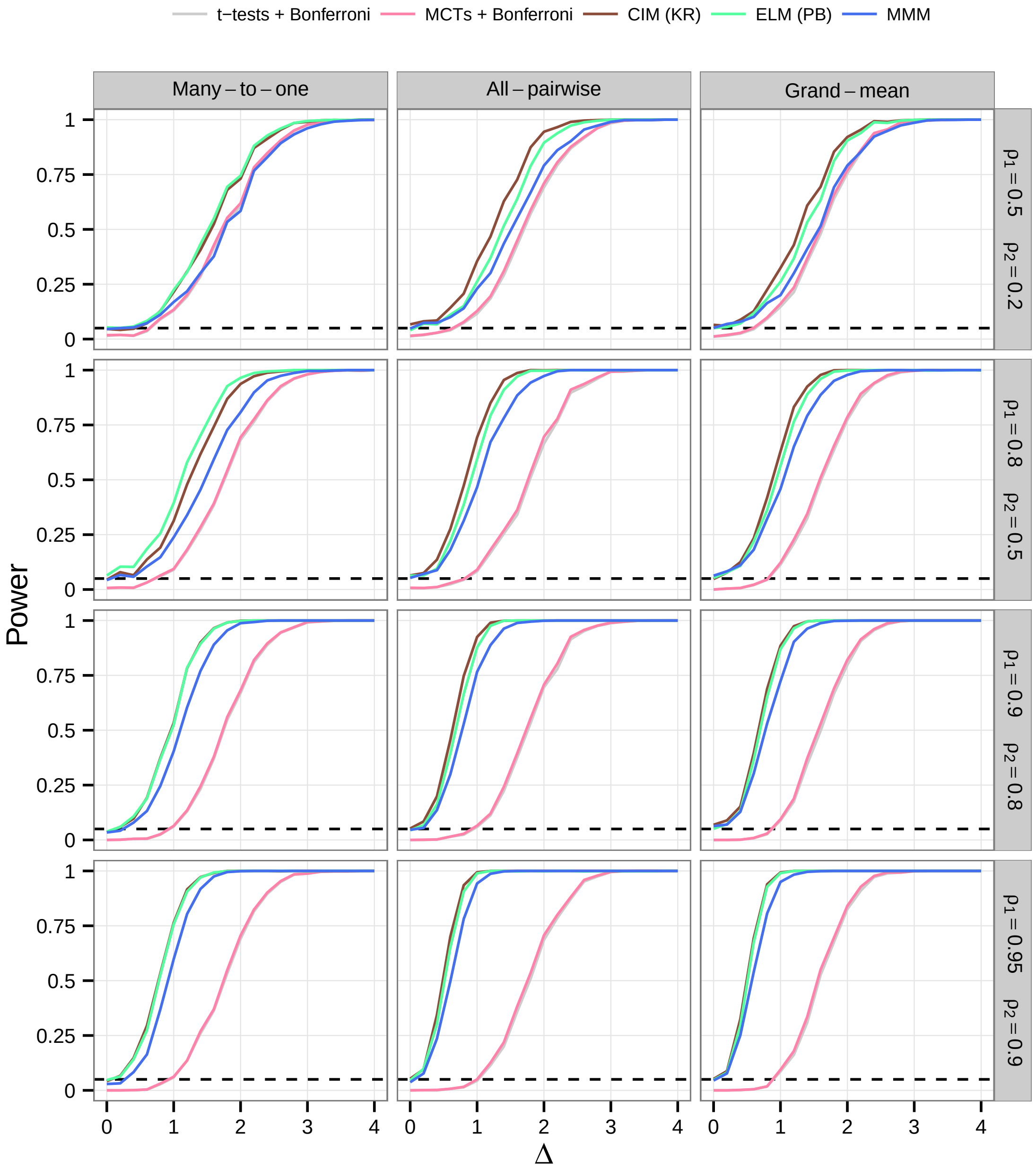


Figure 63: Simulated powers for many-to-one, all-pairwise, and grand-mean comparisons among $m = 3$ Gaussian occasion means separately and simultaneously for $q = 3$ treatments, with $n_k = 10$ independent subjects per treatment group, different longitudinal correlations, and different small-sample degrees of freedom (1000 simulation runs).

D.2 Binomial Data

We show here additional simulation results for asymptotic inference in longitudinal settings with a binomial endpoint.

Type I Error

Simulated type I error rates for many-to-one, all-pairwise, and grand-mean comparisons of binomial proportions with $\pi = \{0.5, 0.6, 0.7, 0.8\}$ in setups involving $q = \{3, 4, 5\}$ treatment groups and $m = \{3, 4, 5\}$ time points are shown in Figures 64, 65, and 66 for tests based on multiple marginal GLMs, and in Figures 67, 68, and 69 for tests based on a joint GEE.

Additional simulated type I error rates for the duplex procedure with binomial data are presented in Figure 70.

Power

Simulated powers for many-to-one comparisons of binomial proportions with $\pi = \{0.5, 0.6, 0.7, 0.8\}$ in setups involving $q = \{3, 4, 5\}$ treatment groups and $m = \{3, 4, 5\}$ time points are shown in Figure 71.

Simulated powers for many-to-one comparisons of binomial proportions with $\pi = \{0.5, 0.6, 0.7, 0.8\}$ in setups involving $q = 3$ treatment groups and $m = 3$ time points under various longitudinal correlations are shown in Figure 72.

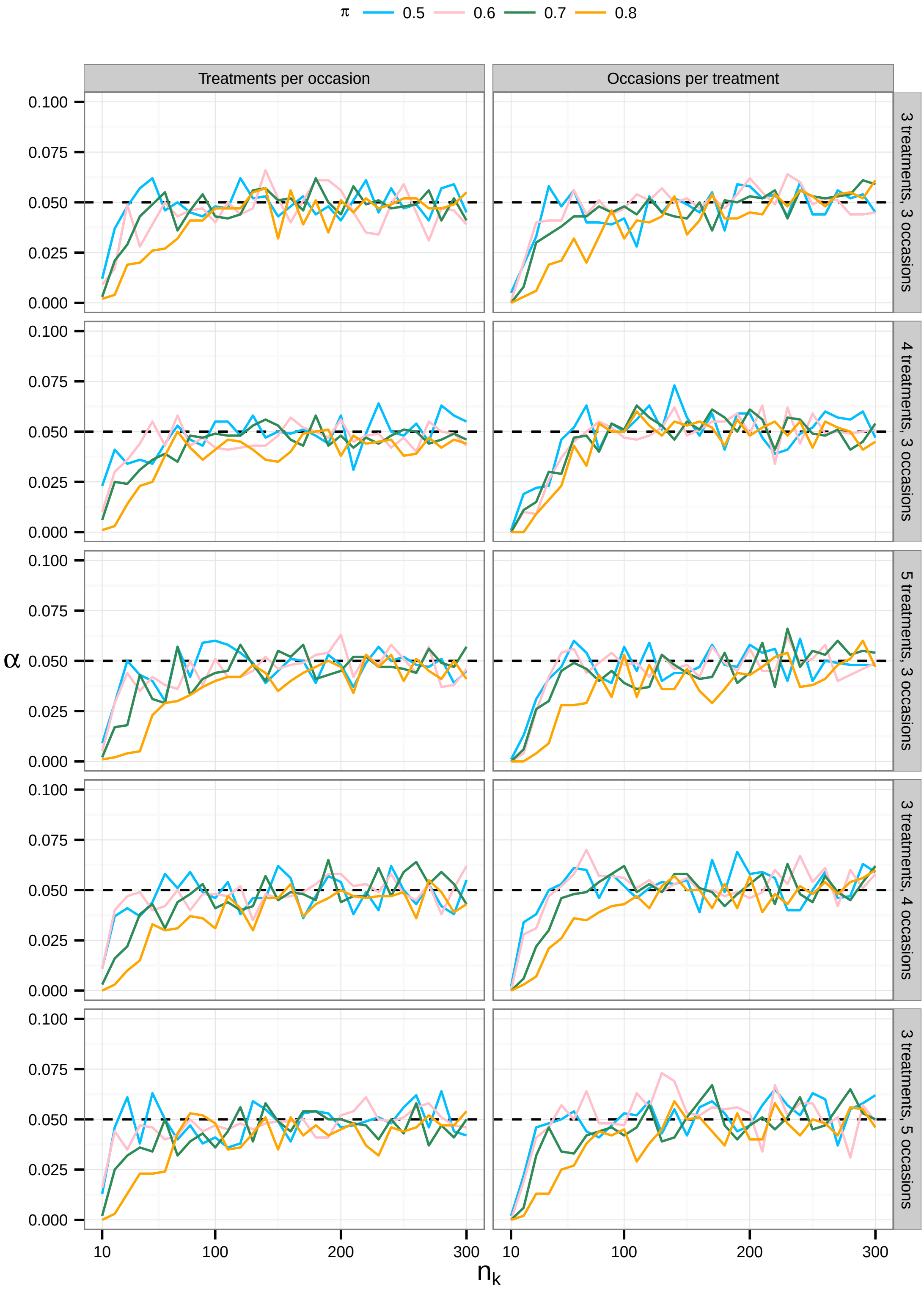


Figure 64: Simulated type I error rates for asymptotic many-to-one comparisons of binomial data ($\pi = \{0.5, 0.6, 0.7, 0.8\}$) involving $q = \{3, 4, 5\}$ treatment groups and $m = \{3, 4, 5\}$ occasions, with n_k independent subjects per treatment group, based on multiple marginal GLMs (1000 simulation runs). Left: comparisons of treatment means separately and simultaneously within occasions; right: comparisons of occasion means separately and simultaneously within treatment groups.

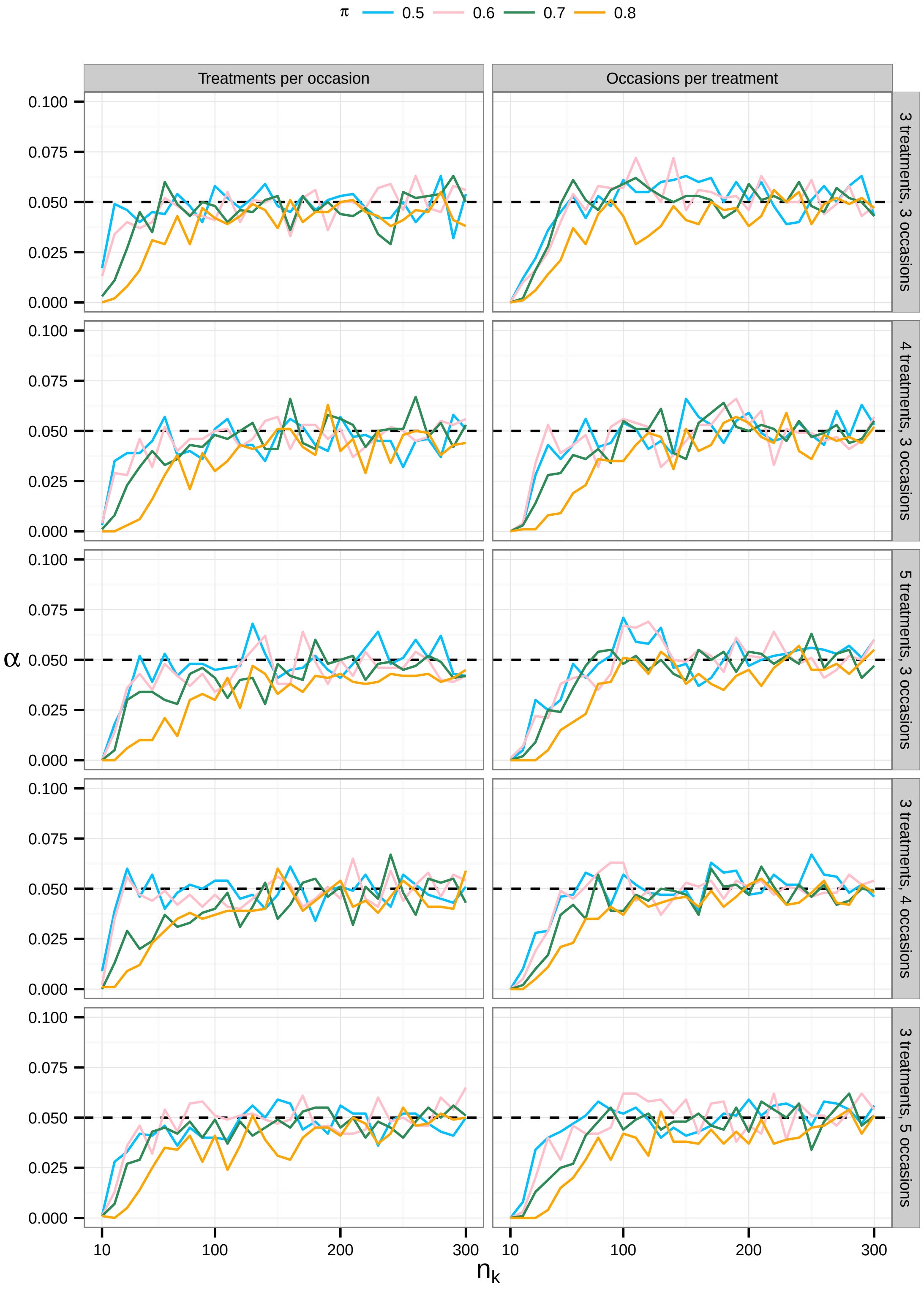


Figure 65: Simulated type I error rates for asymptotic all-pairwise comparisons of binomial data ($\pi = \{0.5, 0.6, 0.7, 0.8\}$) involving $q = \{3, 4, 5\}$ treatment groups and $m = \{3, 4, 5\}$ occasions, with n_k independent subjects per treatment group, based on multiple marginal GLMs (1000 simulation runs). Left: comparisons of treatment means separately and simultaneously within occasions; right: comparisons of occasion means separately and simultaneously within treatment groups.

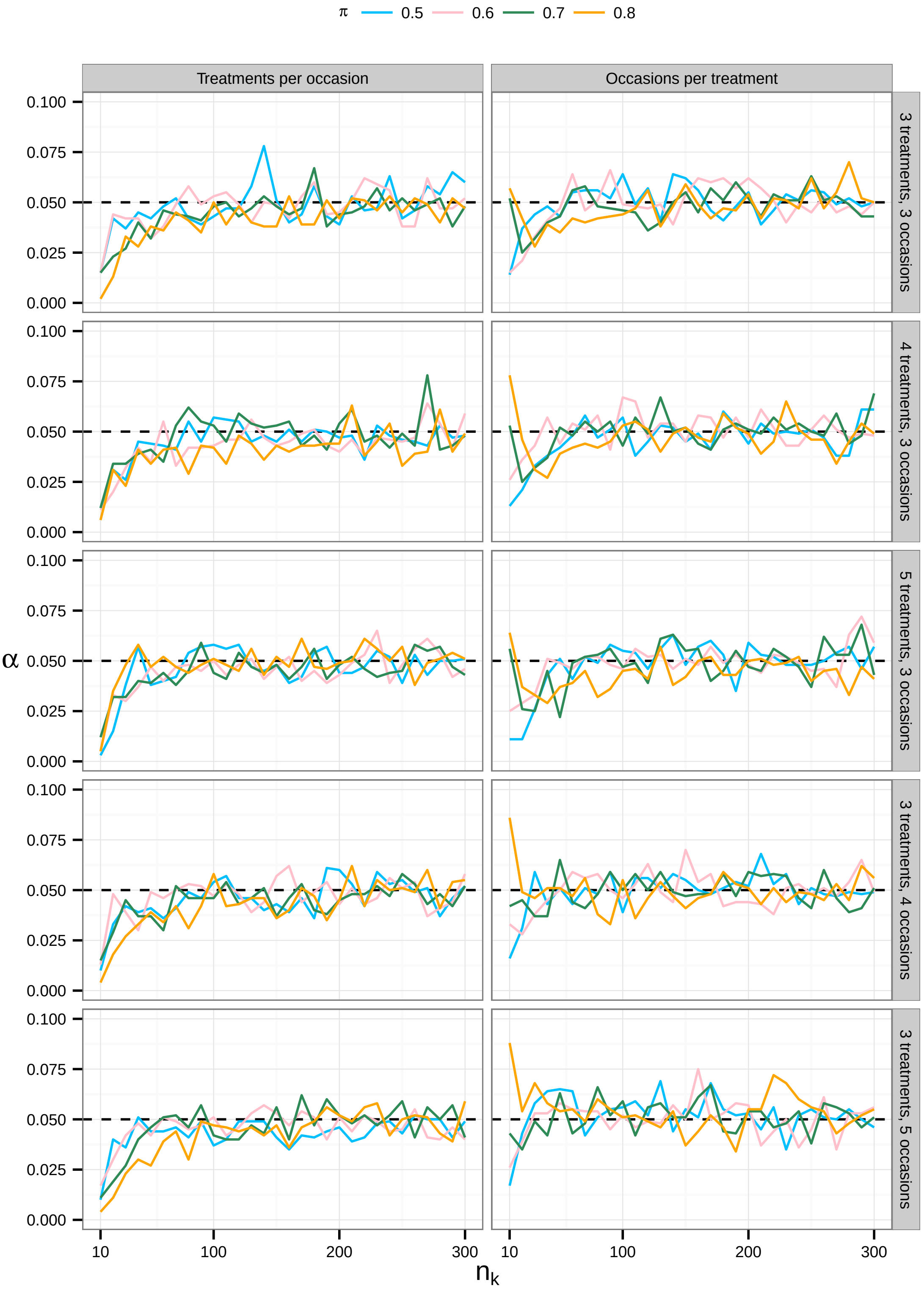


Figure 66: Simulated type I error rates for asymptotic grand-mean comparisons of binomial data ($\pi = \{0.5, 0.6, 0.7, 0.8\}$) involving $q = \{3, 4, 5\}$ treatment groups and $m = \{3, 4, 5\}$ occasions, with n_k independent subjects per treatment group, based on multiple marginal GLMs (1000 simulation runs). Left: comparisons of treatment means separately and simultaneously within occasions; right: comparisons of occasion means separately and simultaneously within treatment groups.

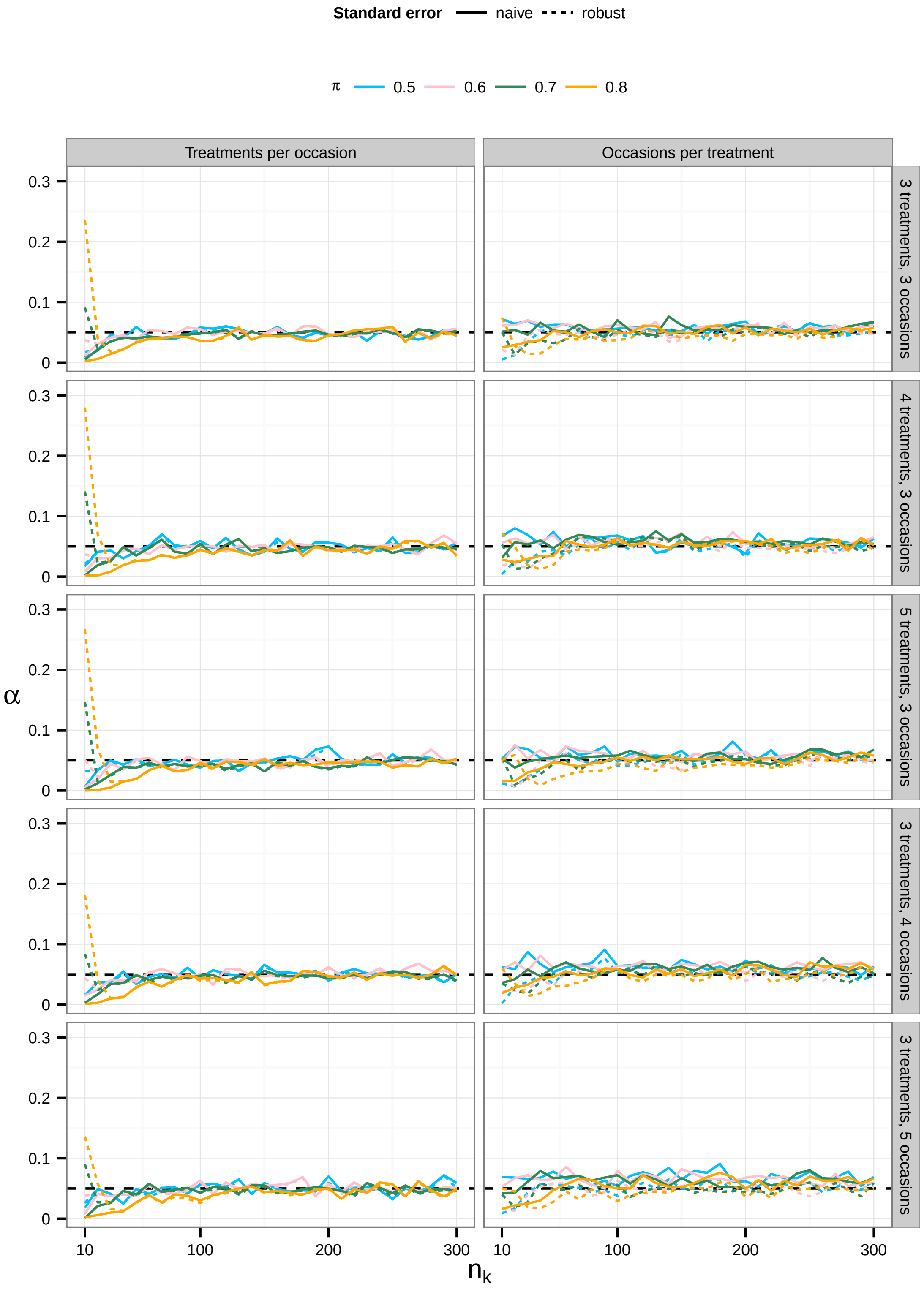


Figure 67: Simulated type I error rates for asymptotic many-to-one comparisons of binomial data ($\pi = \{0.5, 0.6, 0.7, 0.8\}$) involving $q = \{3, 4, 5\}$ treatment groups and $m = \{3, 4, 5\}$ occasions, with n_k independent subjects per treatment group, based on GEEs (1000 simulation runs). Left: comparisons of treatment means separately and simultaneously within occasions; right: comparisons of occasion means separately and simultaneously within treatment groups.

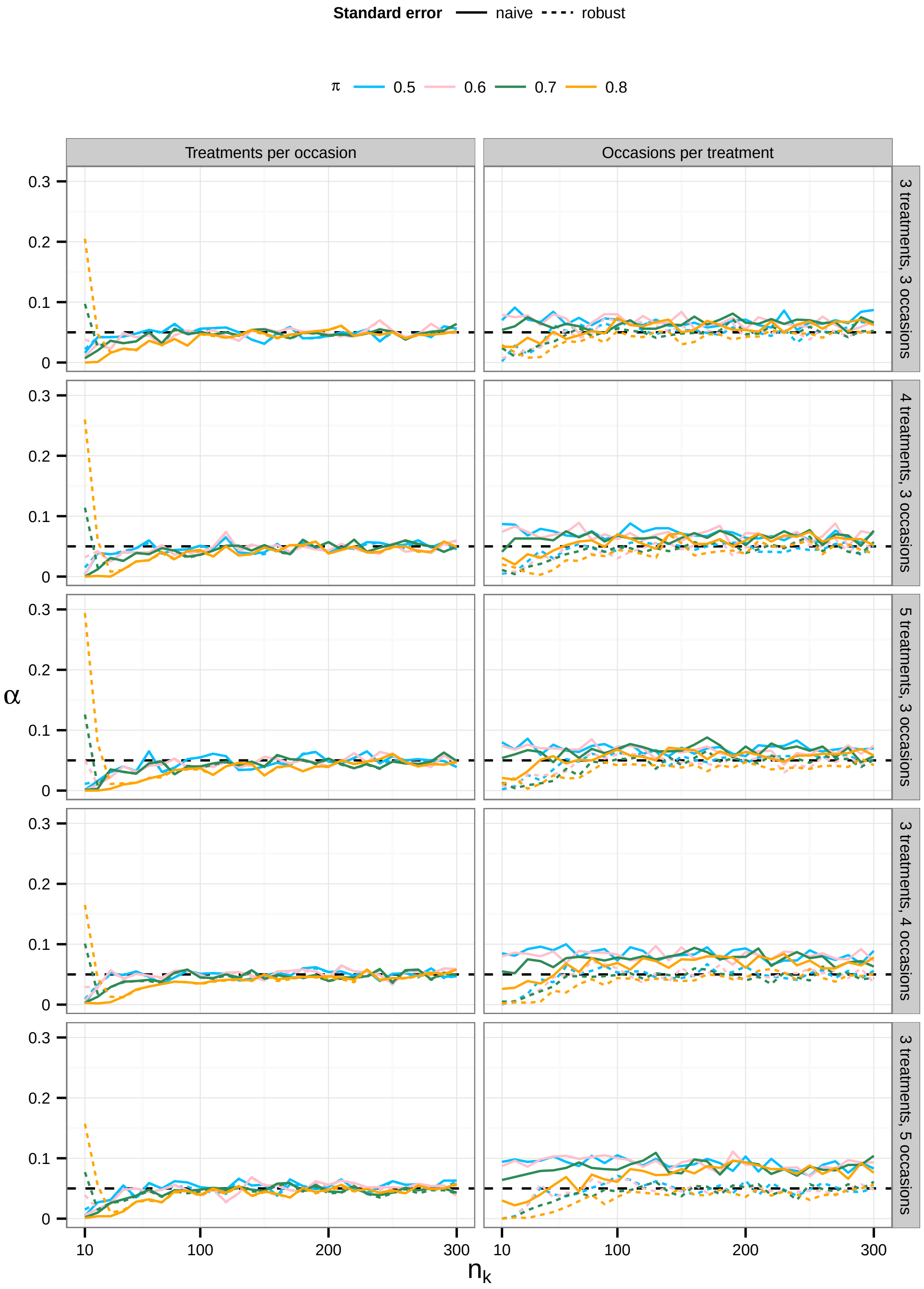


Figure 68: Simulated type I error rates for asymptotic all-pairwise comparisons of binomial data ($\pi = \{0.5, 0.6, 0.7, 0.8\}$) involving $q = \{3, 4, 5\}$ treatment groups and $m = \{3, 4, 5\}$ occasions, with n_k independent subjects per treatment group, based on GEEs (1000 simulation runs). Left: comparisons of treatment means separately and simultaneously within occasions; right: comparisons of occasion means separately and simultaneously within treatment groups.

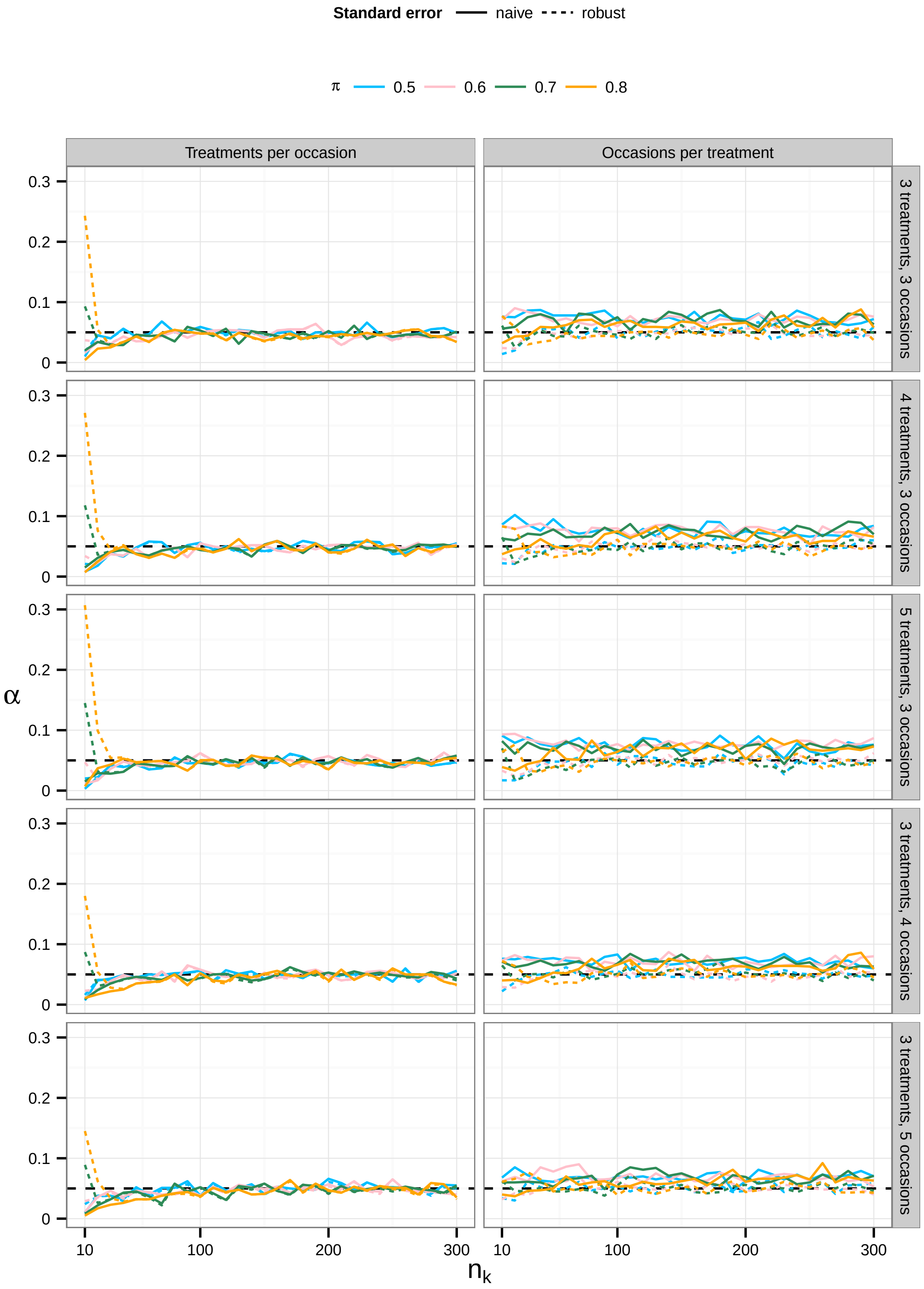


Figure 69: Simulated type I error rates for asymptotic grand-mean comparisons of binomial data ($\pi = \{0.5, 0.6, 0.7, 0.8\}$) involving $q = \{3, 4, 5\}$ treatment groups and $m = \{3, 4, 5\}$ occasions, with n_k independent subjects per treatment group, based on GEEs (1000 simulation runs). Left: comparisons of treatment means separately and simultaneously within occasions; right: comparisons of occasion means separately and simultaneously within treatment groups.

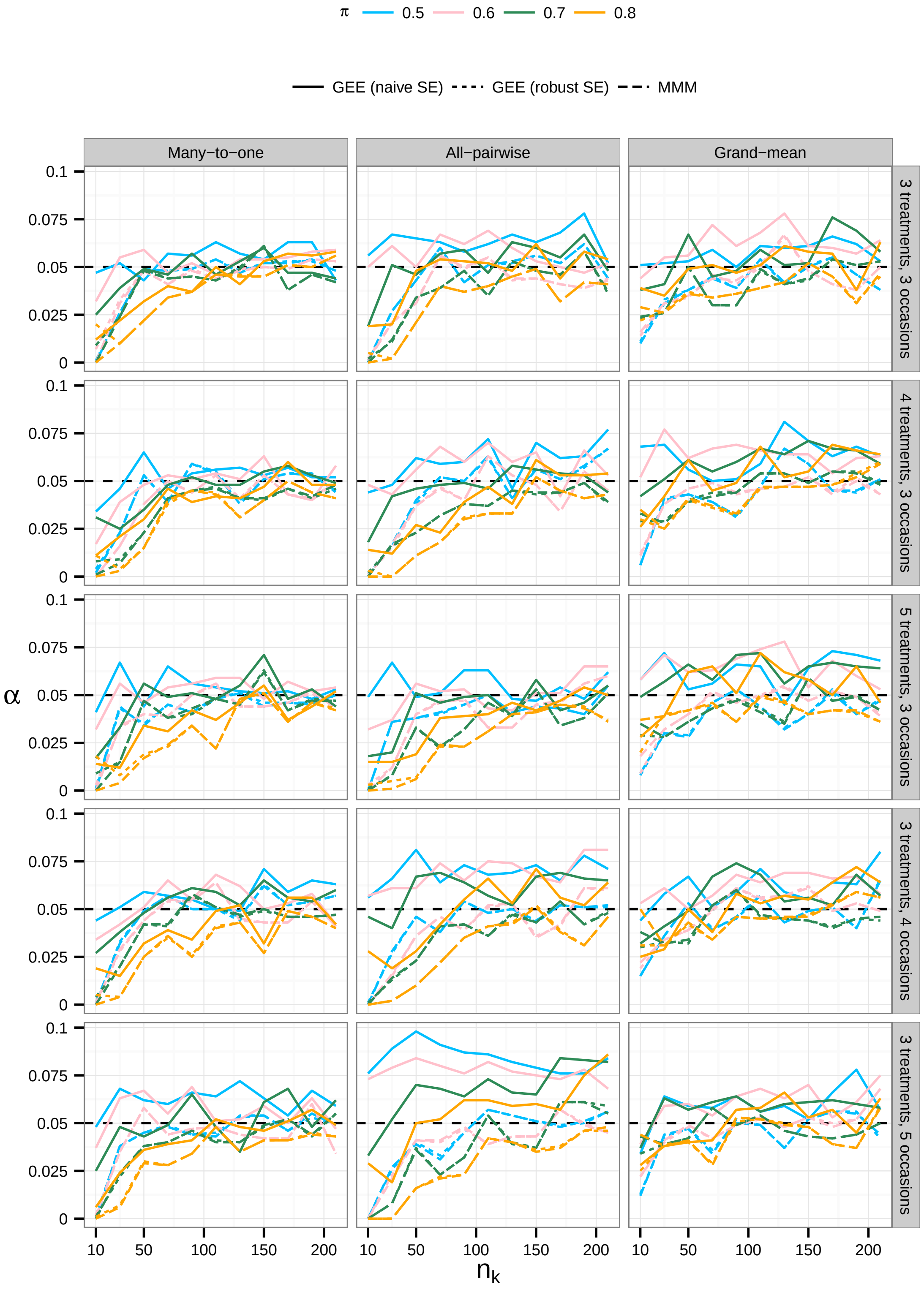


Figure 70: Simulated type I error rates for asymptotic many-to-one, all-pairwise, and grand-mean comparisons of binomial data ($\pi = \{0.5,\ 0.6,\ 0.7,\ 0.8\}$) among $q = \{3, 4, 5\}$ treatment means separately and simultaneously at $q = \{3, 4, 5\}$ occasions and among $q = \{3, 4, 5\}$ occasion means separately and simultaneously for $q = \{3, 4, 5\}$ treatments, with n_k independent subjects per treatment group, based on GEEs or multiple marginal GLMs (1000 simulation runs).

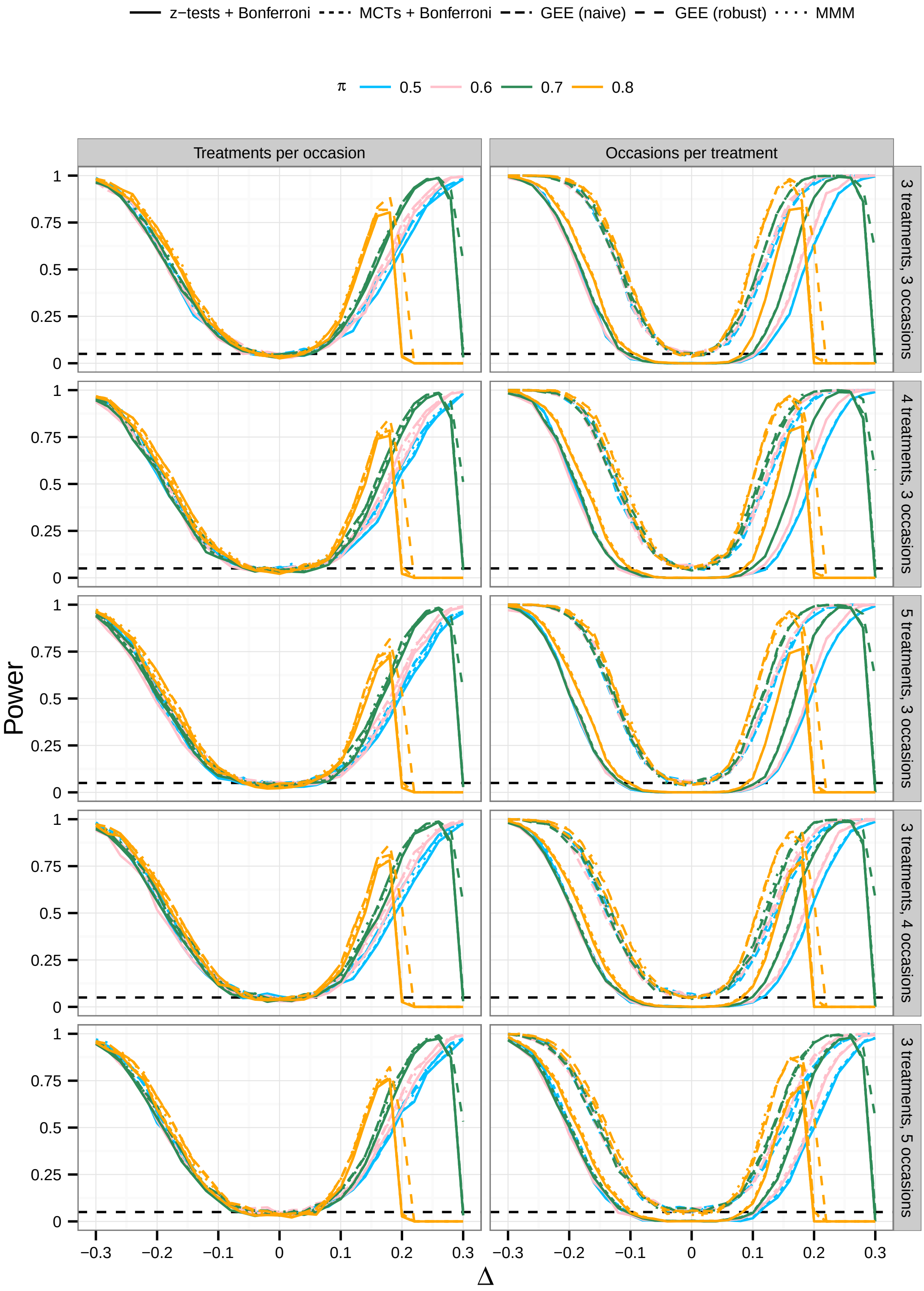


Figure 71: Simulated powers for asymptotic many-to-one comparisons of binomial data ($\pi = \{0.5,\ 0.6,\ 0.7,\ 0.8\}$) involving $q = \{3, 4, 5\}$ treatment groups and $m = \{3, 4, 5\}$ occasions, with $n_k = 100$ independent subjects per treatment group (1000 simulation runs). Left: comparisons of treatment means separately and simultaneously within occasions; right: comparisons of occasion means separately and simultaneously within treatment groups.

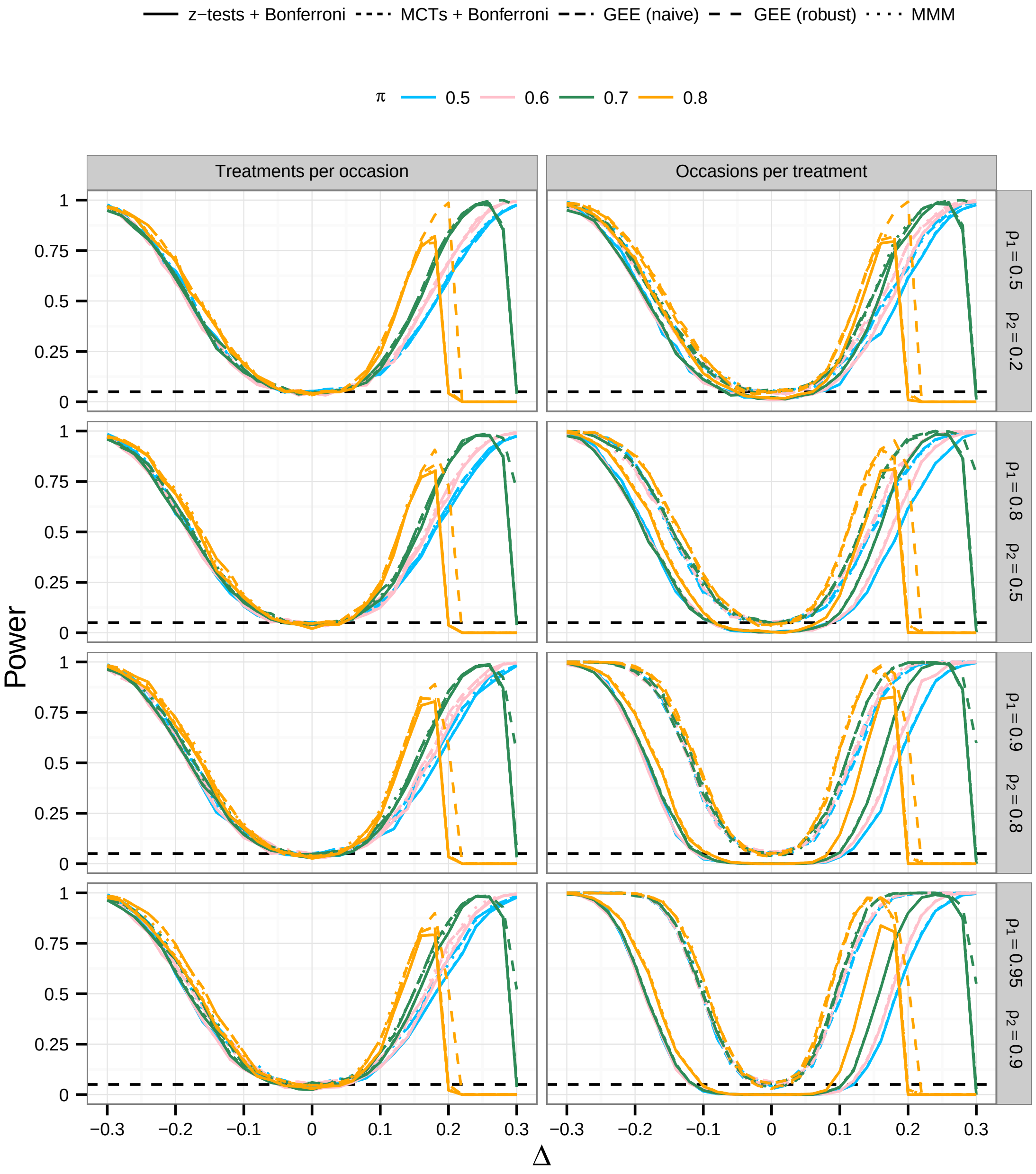


Figure 72: Simulated powers for asymptotic many-to-one comparisons of binomial data ($\pi = \{0.5,\ 0.6,\ 0.7,\ 0.8\}$) involving $q = 3$ treatment groups and $m = 3$ occasions, with $n_k = 100$ independent subjects per treatment group, and different longitudinal correlations (1000 simulation runs). Left: comparisons of treatment means separately and simultaneously within occasions; right: comparisons of occasion means separately and simultaneously within treatment groups.

D.3 Poisson Data

We show here additional simulation results for asymptotic inference in longitudinal settings with a Poisson endpoint.

Type I Error

Simulated type I error rates for many-to-one, all-pairwise, and grand-mean comparisons of Poisson rates with $\lambda = \{3, 5, 10, 20\}$ in setups involving $q = \{3, 4, 5\}$ treatment groups and $m = \{3, 4, 5\}$ time points are shown in Figures 73, 74, and 75 for tests based on multiple marginal GLMs, and in Figures 76, 77, and 78 for tests based on joint GEEs.

Additional simulated type I error rates for the duplex procedure with Poisson data are presented in Figure 79.

Power

Simulated powers for many-to-one comparisons of Poisson rates with $\lambda = \{3, 5, 10, 20\}$ in setups involving $q = \{3, 4, 5\}$ treatment groups and $m = \{3, 4, 5\}$ time points are shown in Figure 80.

Simulated powers for many-to-one comparisons of Poisson rates with $\lambda = \{3, 5, 10, 20\}$ in setups involving $q = 3$ treatment groups and $m = 3$ time points under various longitudinal correlations are shown in Figure 81.

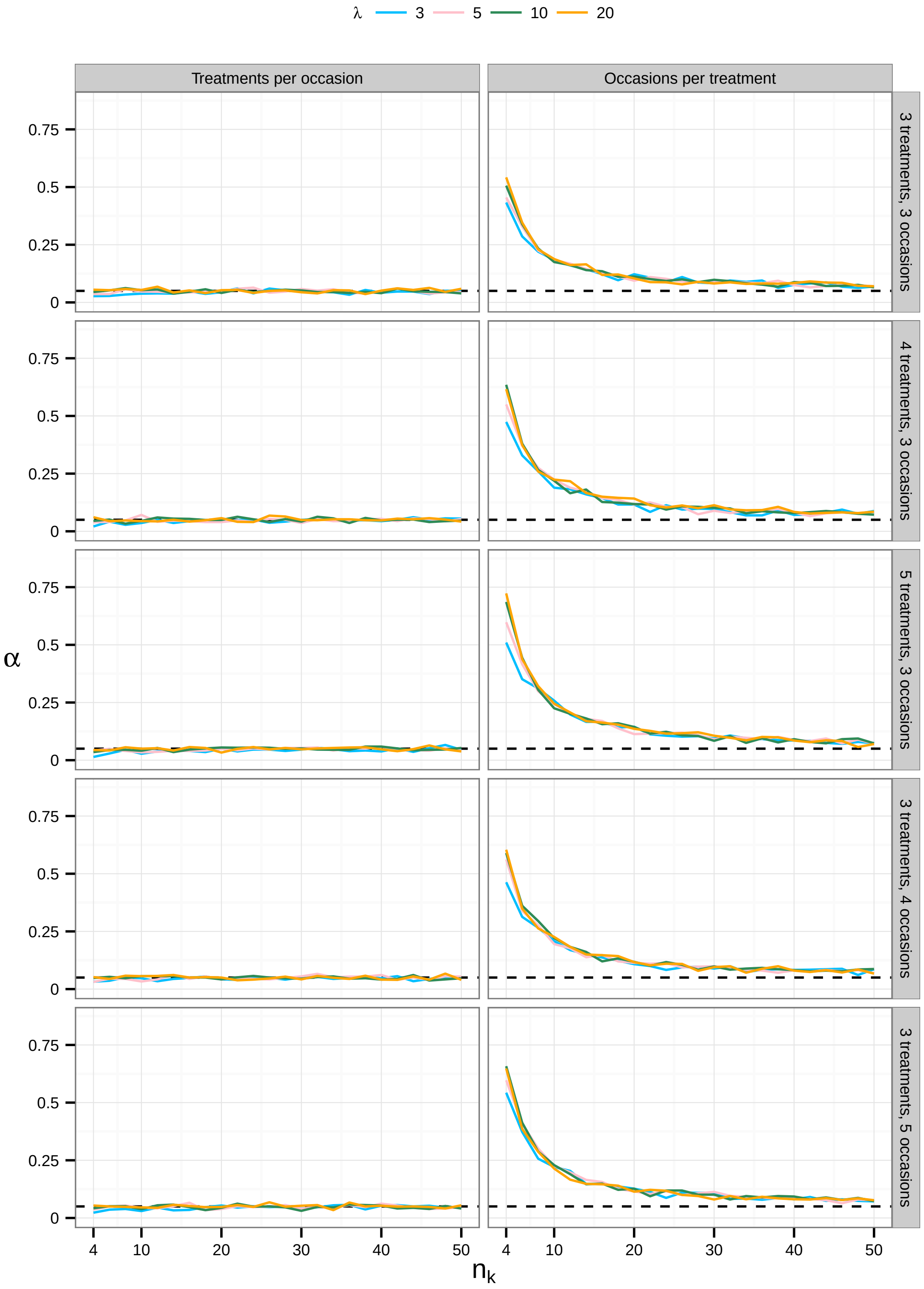


Figure 73: Simulated type I error rates for asymptotic many-to-one comparisons of Poisson data ($\lambda = \{3,\ 5,\ 10,\ 20\}$) involving $q = \{3, 4, 5\}$ treatment groups and $m = \{3, 4, 5\}$ occasions, with n_k independent subjects per treatment group, based on multiple marginal GLMs (1000 simulation runs). Left: comparisons of treatment means separately and simultaneously within occasions; right: comparisons of occasion means separately and simultaneously within treatment groups.

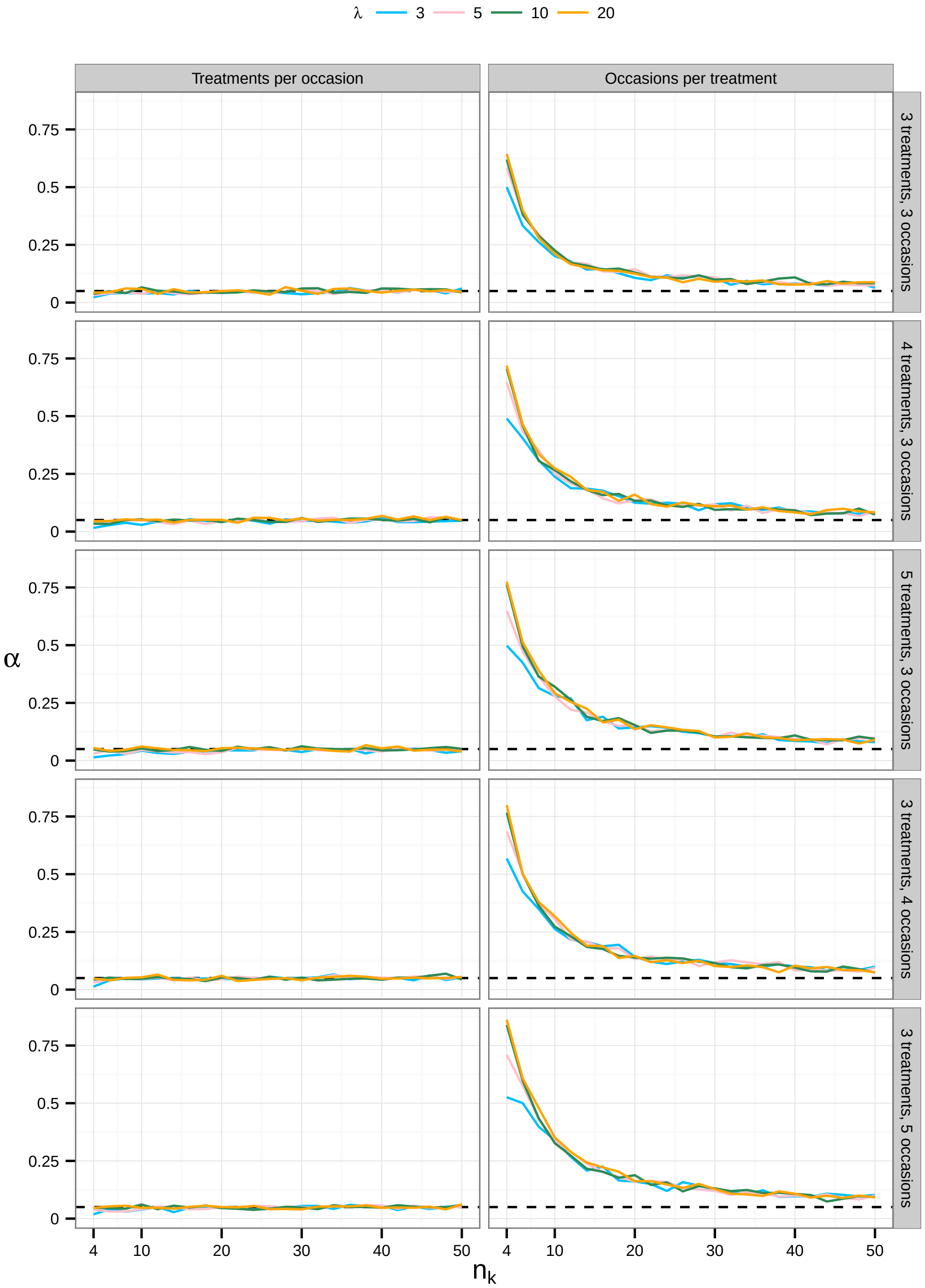


Figure 74: Simulated type I error rates for asymptotic all-pairwise comparisons of Poisson data ($\lambda = \{3,\ 5,\ 10,\ 20\}$) involving $q = \{3, 4, 5\}$ treatment groups and $m = \{3, 4, 5\}$ occasions, with n_k independent subjects per treatment group, based on multiple marginal GLMs (1000 simulation runs). Left: comparisons of treatment means separately and simultaneously within occasions; right: comparisons of occasion means separately and simultaneously within treatment groups.

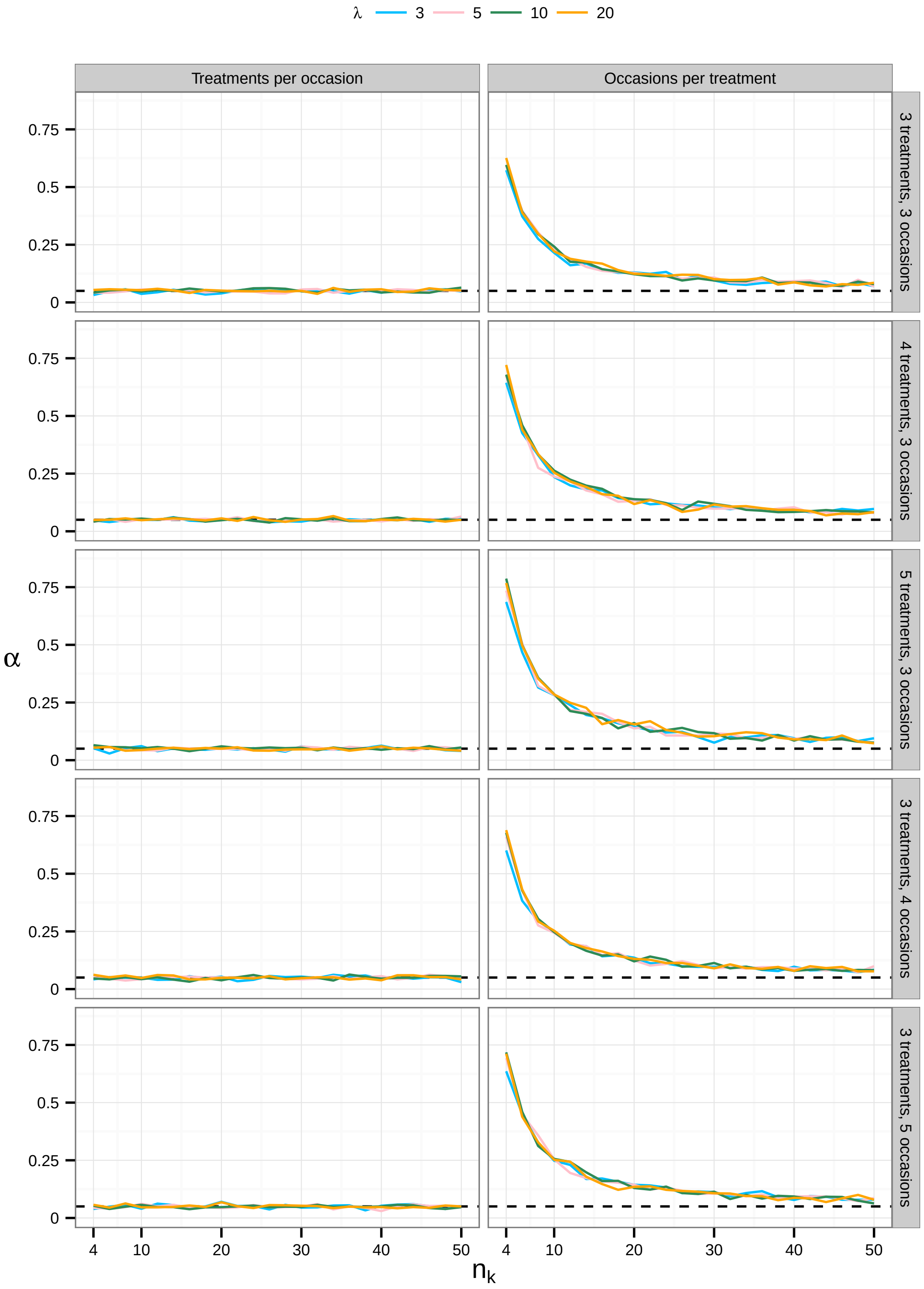


Figure 75: Simulated type I error rates for asymptotic grand-mean comparisons of Poisson data ($\lambda = \{3,\, 5,\, 10,\, 20\}$) involving $q = \{3, 4, 5\}$ treatment groups and $m = \{3, 4, 5\}$ occasions, with n_k independent subjects per treatment group, based on multiple marginal GLMs (1000 simulation runs). Left: comparisons of treatment means separately and simultaneously within occasions; right: comparisons of occasion means separately and simultaneously within treatment groups.

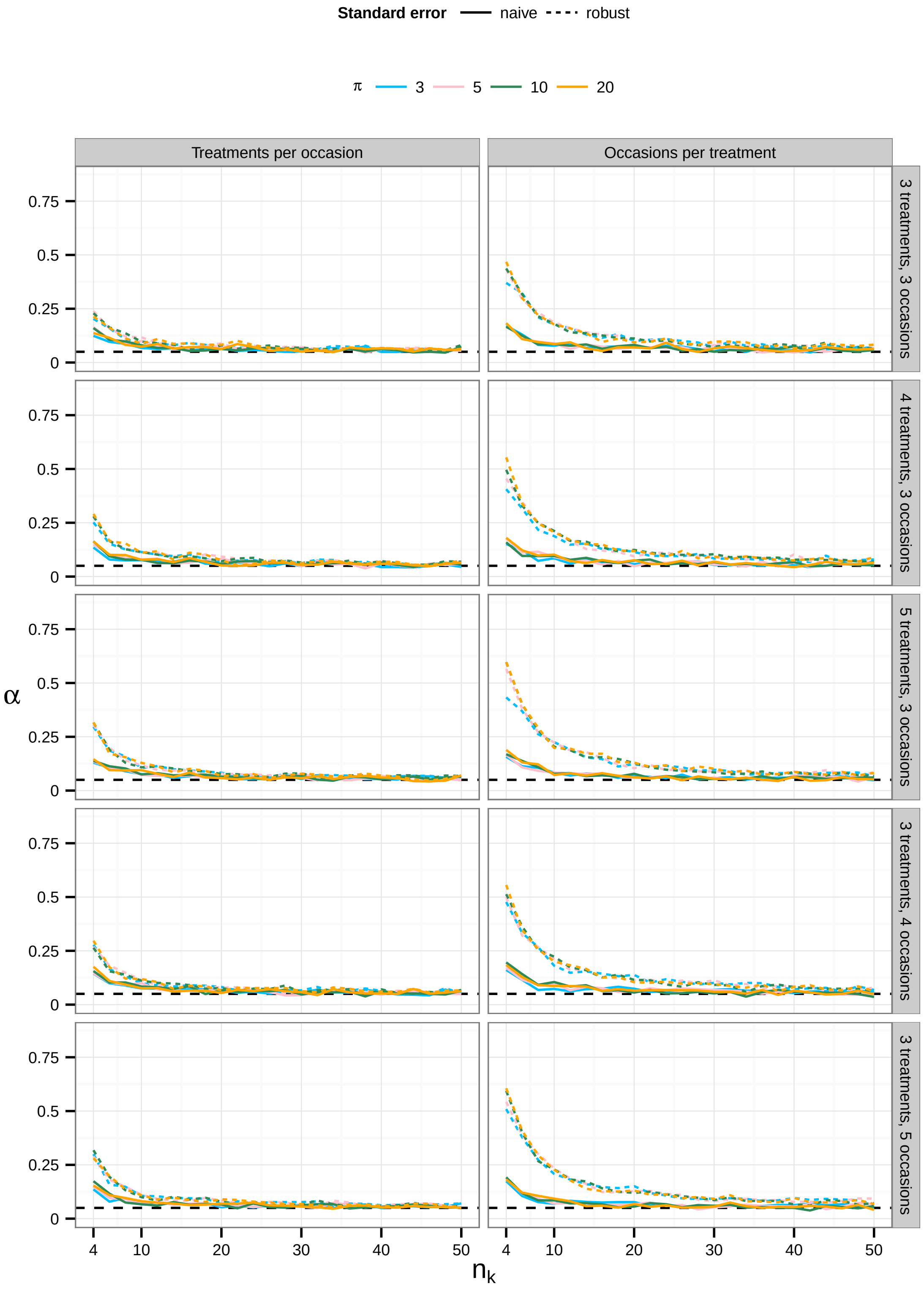


Figure 76: Simulated type I error rates for asymptotic many-to-one comparisons of Poisson data ($\lambda = \{3,\ 5,\ 10,\ 20\}$) involving $q = \{3, 4, 5\}$ treatment groups and $m = \{3, 4, 5\}$ occasions, with n_k independent subjects per treatment group, based on GEEs (1000 simulation runs). Left: comparisons of treatment means separately and simultaneously within occasions; right: comparisons of occasion means separately and simultaneously within treatment groups.

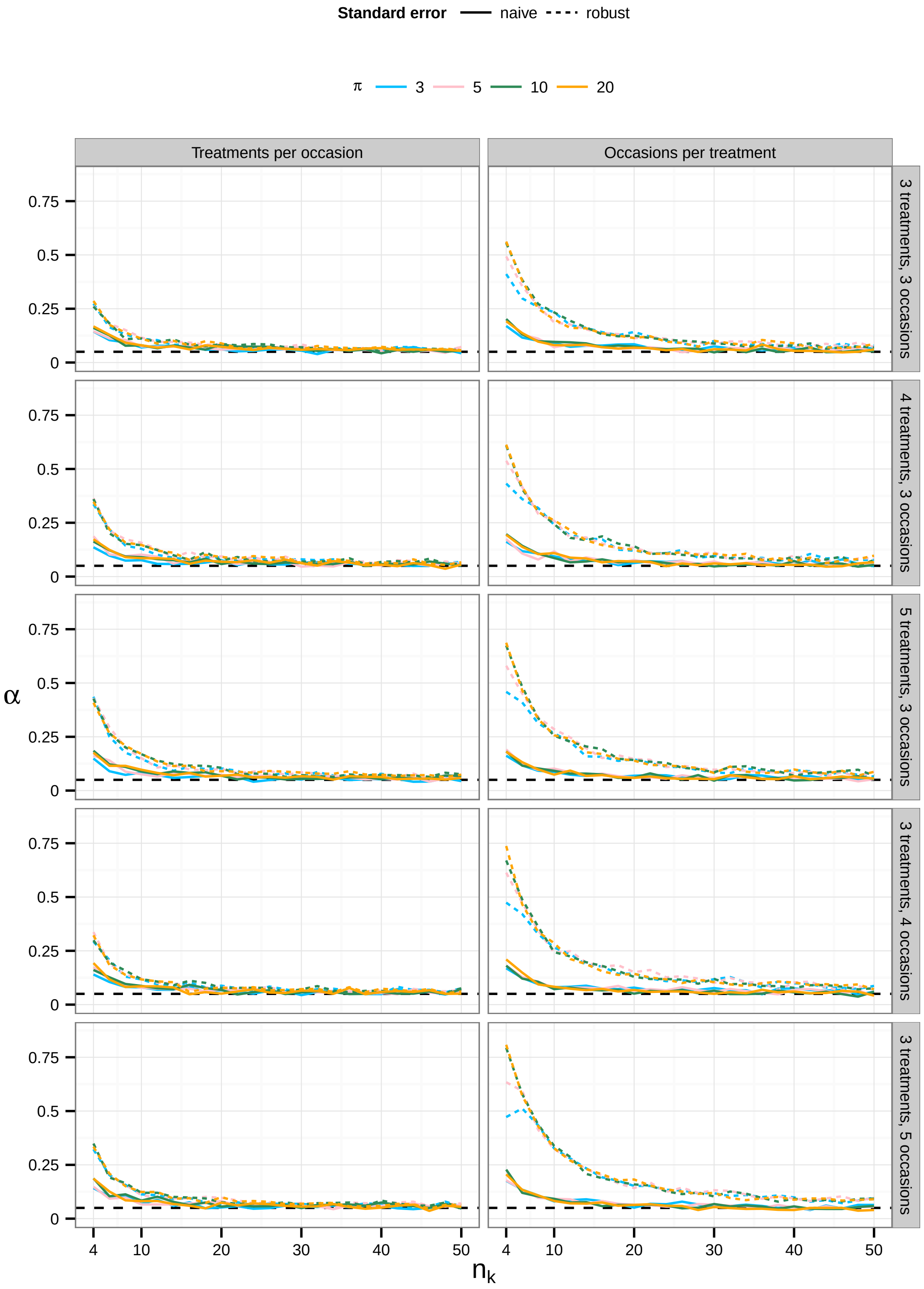


Figure 77: Simulated type I error rates for asymptotic all-pairwise comparisons of Poisson data ($\lambda = \{3, 5, 10, 20\}$) involving $q = \{3, 4, 5\}$ treatment groups and $m = \{3, 4, 5\}$ occasions, with n_k independent subjects per treatment group, based on GEEs (1000 simulation runs). Left: comparisons of treatment means separately and simultaneously within occasions; right: comparisons of occasion means separately and simultaneously within treatment groups.

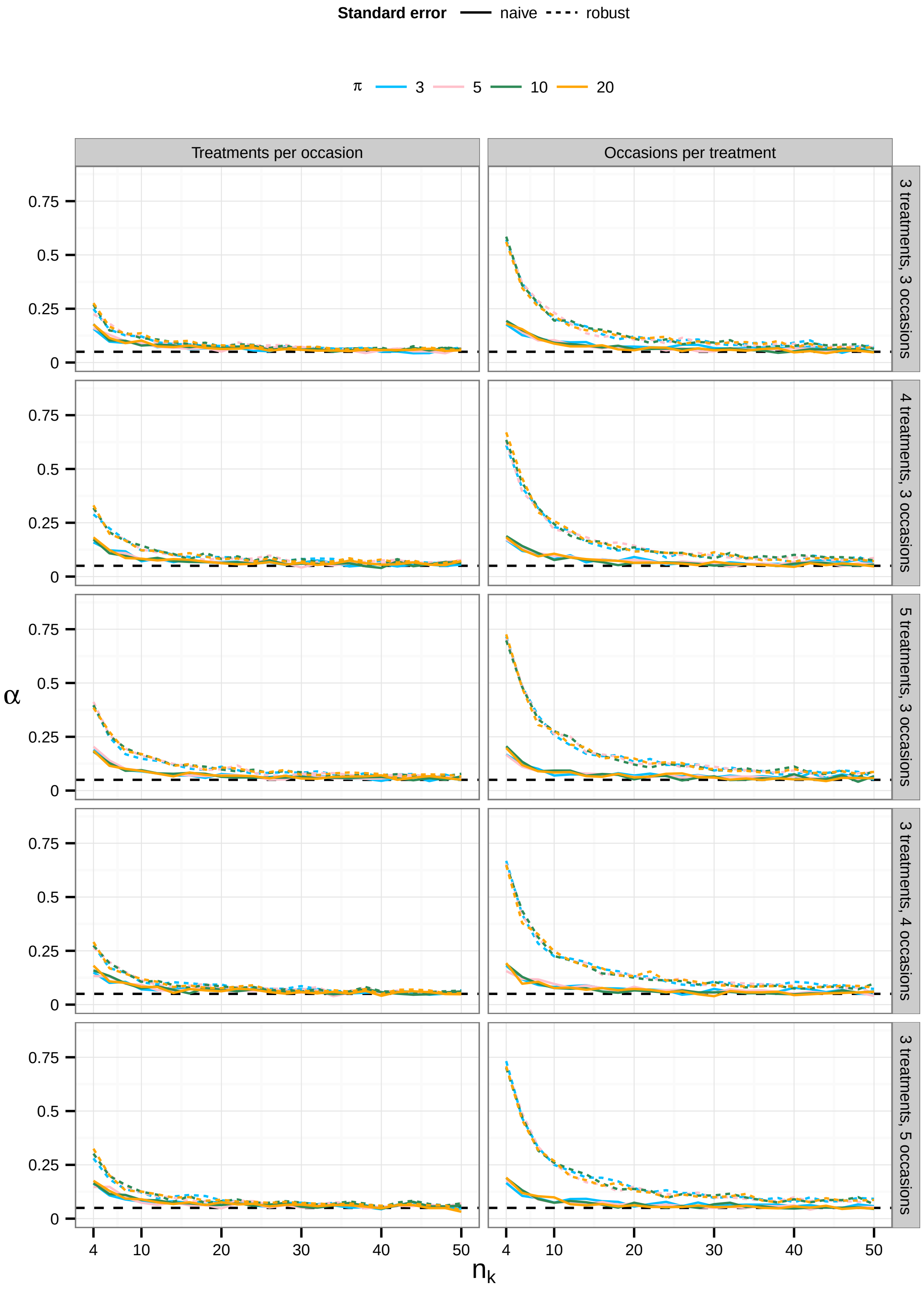


Figure 78: Simulated type I error rates for asymptotic grand-mean comparisons of Poisson data ($\lambda = \{3,\ 5,\ 10,\ 20\}$) involving $q = \{3, 4, 5\}$ treatment groups and $m = \{3, 4, 5\}$ occasions, with n_k independent subjects per treatment group, based on GEEs (1000 simulation runs). Left: comparisons of treatment means separately and simultaneously within occasions; right: comparisons of occasion means separately and simultaneously within treatment groups.

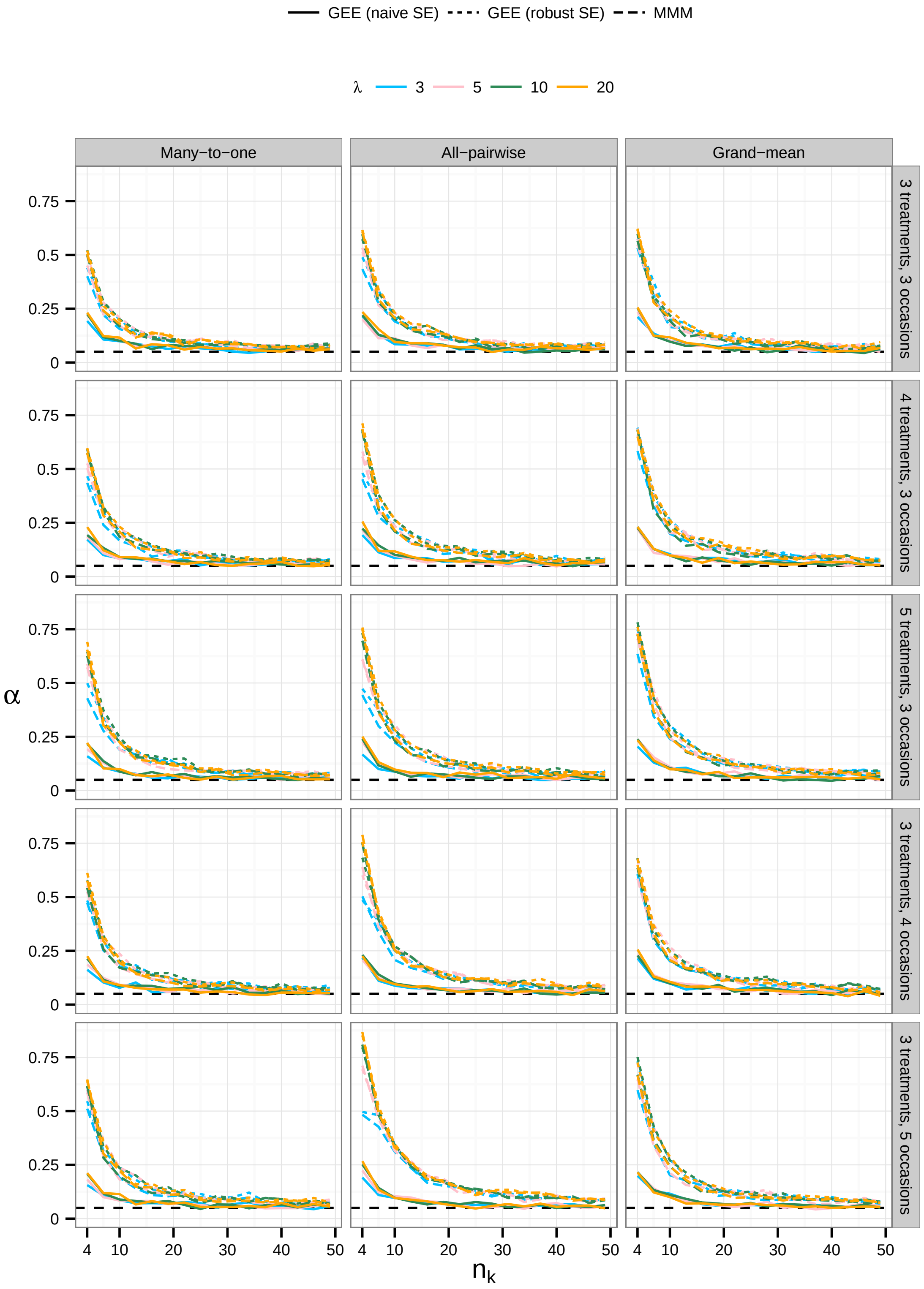


Figure 79: Simulated type I error rates for asymptotic many-to-one, all-pairwise, and grand-mean comparisons of Poisson data ($\lambda = \{3,\ 5,\ 10,\ 20\}$) among $q = \{3, 4, 5\}$ treatment means separately and simultaneously at $q = \{3, 4, 5\}$ occasions and among $q = \{3, 4, 5\}$ occasion means separately and simultaneously for $q = \{3, 4, 5\}$ treatments, with n_k independent subjects per treatment group, based on GEEs or multiple marginal GLMs (1000 simulation runs).

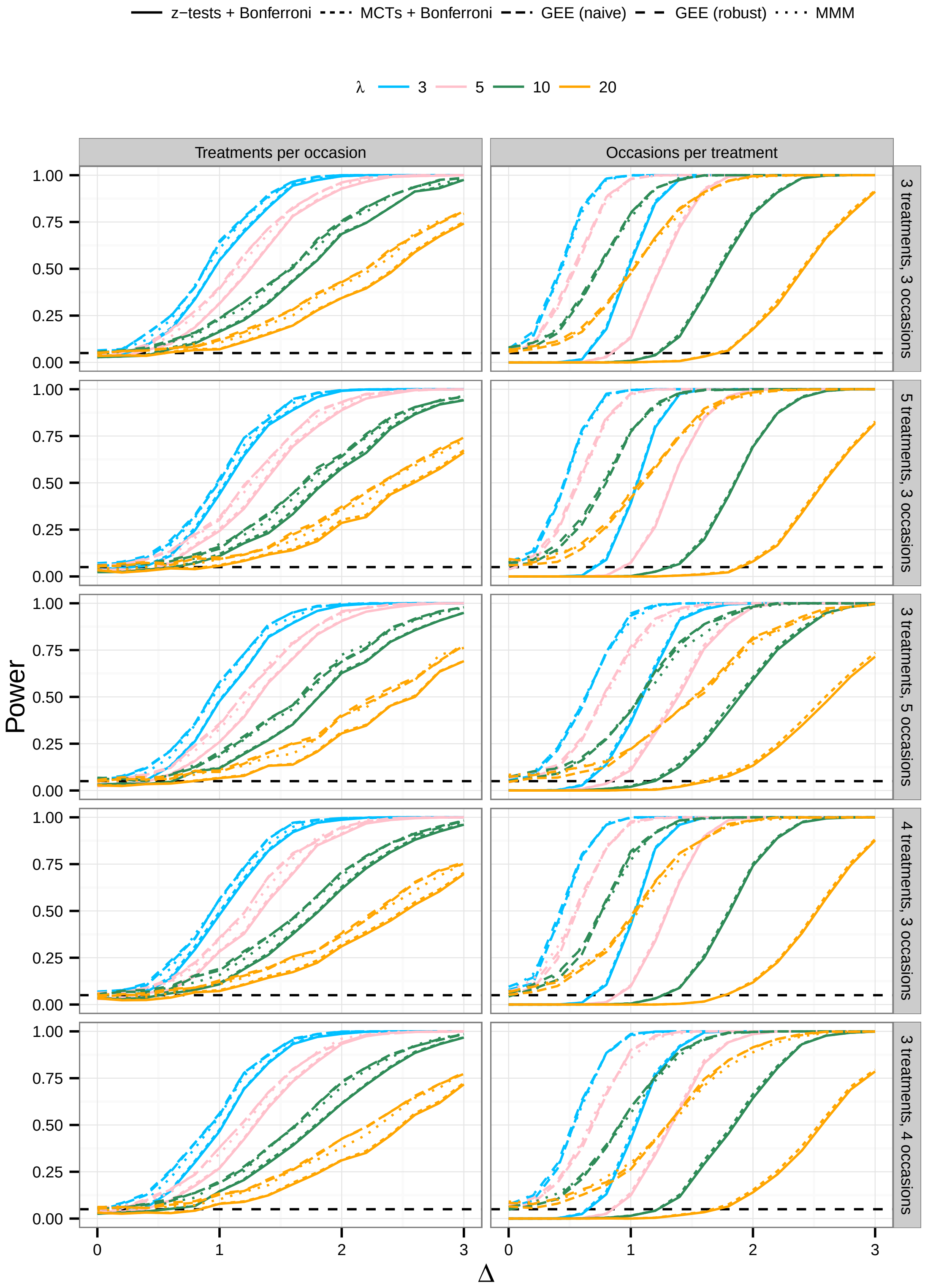


Figure 80: Simulated powers for asymptotic many-to-one comparisons of Poisson data ($\lambda = \{3, 5, 10, 20\}$) involving $q = \{3, 4, 5\}$ treatment groups and $m = \{3, 4, 5\}$ occasions, with $n_k = 100$ independent subjects per treatment group (1000 simulation runs). Left: comparisons of treatment means separately and simultaneously within occasions; right: comparisons of occasion means separately and simultaneously within treatment groups.

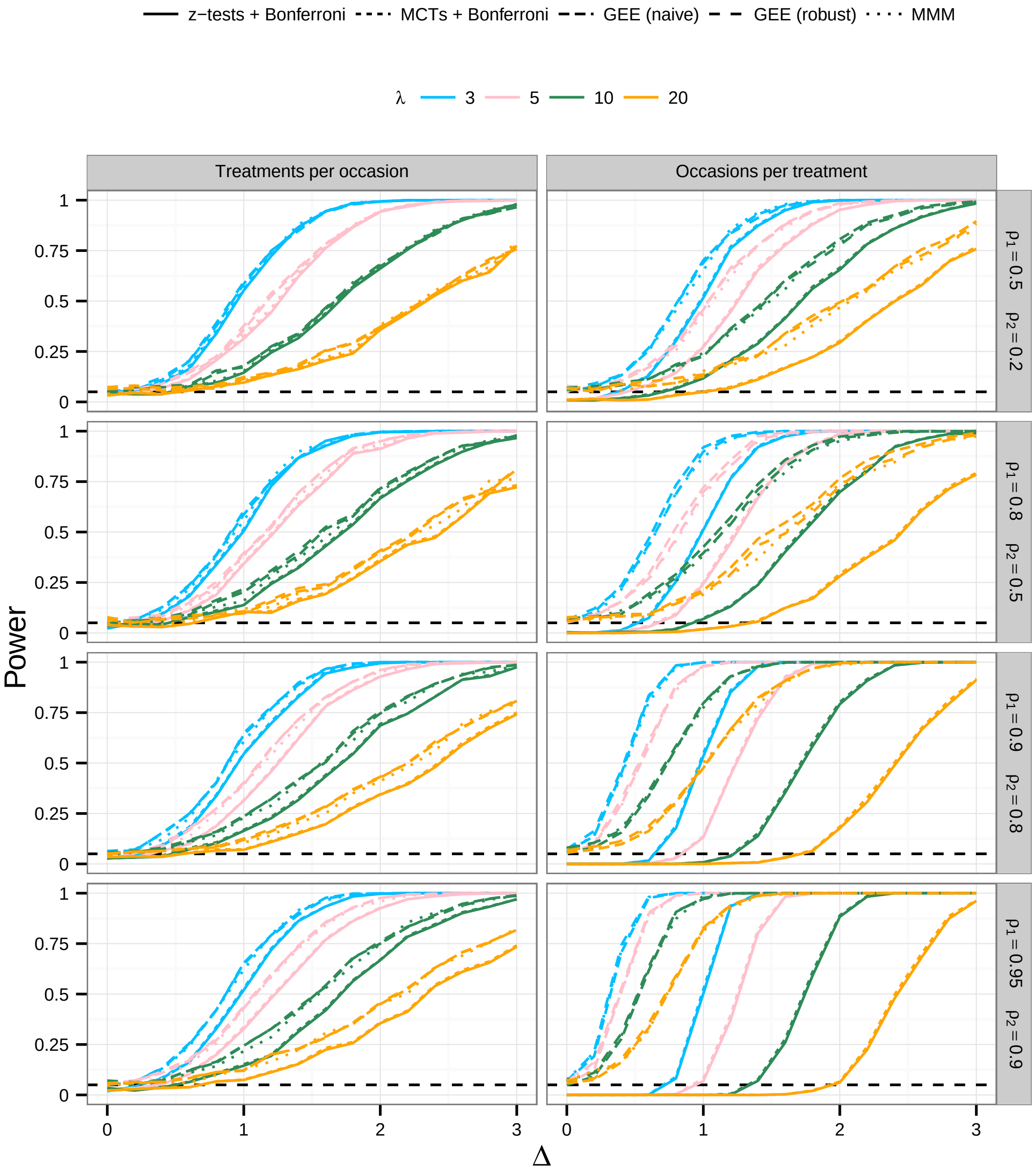


Figure 81: Simulated powers for asymptotic many-to-one comparisons of Poisson data ($\lambda = \{3, 5, 10, 20\}$) involving $q = 3$ treatment groups and $m = 3$ occasions, with $n_k = 100$ independent subjects per treatment group, and different longitudinal correlations (1000 simulation runs). Left: comparisons of treatment means separately and simultaneously within occasions; right: comparisons of occasion means separately and simultaneously within treatment groups.

E Implementation in R

E.1 Example R Code

All simulations and data analyses presented in this work were executed in R (R Core Team 2015) and some add-on packages. MCTs and the MMM procedure are implemented in multcomp (Hothorn et al. 2015, described in Hothorn et al. 2008). LMMs can be fitted with nlme (Pinheiro et al. 2015, described in Pinheiro and Bates 2000) and lme4 (Bates et al. 2015b, described in Bates et al. 2015a). The Kenward-Roger adjustment is available in pbkrtest (Halekoh and Højsgaard 2014b, described in Halekoh and Højsgaard 2014a). A package for GEEs is geepack (Højsgaard et al. 2014, described in Halekoh et al. 2006). An implementation of AICc-based model selection is provided by MuMIn (Barton 2015). Copulae are available in copula (Hofert et al. 2014, described in Yan 2007). All graphics were generated with ggplot2 (Wickham and Chang 2015, described in Wickham 2009).

The R code shown in the following contains the essential building blocks used for the simulations and data analyses.

Generating Correlated Gaussian Data

To obtain correlated normal data with heterogeneous variances $\sigma_j^2 = j$, $j = 1, \ldots, m$ and a Toeplitz correlation pattern with off-diagonal parameters par, construct an nt × nt covariance matrix:

```
co <- sqrt(diag(1:nt)) %*% toeplitz(c(1, par)) %*% sqrt(diag(1:nt))
```

Draw random values from an nt-dimensional normal distribution with mean me and covariance matrix co for ng treatment groups with ni subjects per treatment group:

```
dalist <- list()
for(g in 1:ng){
  dalist[[g]] <- data.frame(mvtnorm::rmvnorm(n = ni, mean = rep(me, nt),
                                             sigma = co))
}
dada <- as.data.frame(abind::abind(dalist, along = 1))
dada$group <- as.factor(rep(LETTERS[1:ng], each = ni))
dada$person <- as.factor(1:(ng * ni))
mda <- reshape2::melt(da, c("group", "person"))
mda$tg <- with(mda, variable:group)
```

Generating Correlated Discrete Data

To obtain correlated binary data, construct an nt-dimensional distribution with binomial margins that have probability pi and cluster size si using an nt-dimensional normal copula characterized by a Toeplitz correlation structure with off-diagonal parameters par:

```
cop <- copula::mvdc(copula = copula::normalCopula(param = par, dim = nt,
                                                  dispstr = "toep"),
                    margins = rep("binom", nt),
                    paramMargins = list(list(size = si, prob = pi)),
                    marginsIdentical = TRUE)
```

To obtain correlated count data, construct an nt-dimensional distribution with Poisson margins that have parameter lambda using an nt-dimensional normal copula characterized by a Toeplitz correlation structure with off-diagonal parameters par:

```
cop <- copula::mvdc(copula = copula::normalCopula(param = par, dim = nt,
                                                  dispstr = "toep"),
                    margins = rep("pois", nt),
                    paramMargins = list(list(lambda = lambda)),
                    marginsIdentical = TRUE)
```

Draw random values from the distribution cop for ng treatment groups with ni subjects per treatment group:

```
dalist <- list()
for(g in 1:ng){
  dalist[[g]] <- data.frame(copula::rMvdc(ni, cop))
}
da <- as.data.frame(abind::abind(dalist, along = 1))
da$group <- as.factor(rep(LETTERS[1:ng], each = ni))
da$person <- as.factor(1:(ng * ni))
mda <- reshape2::melt(da, c("group", "person"))
mda$tg <- with(mda, variable:group)
```

Contrast Coefficient Matrices

Construct a contrast matrix for many-to-one comparisons ("Dunnett") among ng treatment means separately and simultaneously at nt occasions, with ni subjects per treatment group:

```
K1 <- diag(nt) %x% multcomp::contrMat(rep(ni, ng), type = "Dunnett")
```

Construct a contrast matrix for many-to-one comparisons ("Dunnett") among nt occasion means separately and simultaneously within ng treatment groups, with ni subjects per treatment group:

```
K2 <- multcomp::contrMat(rep(ni, nt), type = "Dunnett") %x% diag(ng)
```

Unite the contrasts in K1 and K2 to carry out all comparisons simultaneously:

```
K12 <- rbind(K1, K2)
```

Other contrasts besides many-to-one are possible e.g., with type = "Tukey" for all-pairwise comparisons, or type = "GrandMean" for ANOM.

Simultaneous Inference from Multiple Marginal Models

For Gaussian data, fit one linear model per time point:

```
modlist <- apply(da[, 1:nt], 2, function(y) lm(y ~ da$group - 1))
```

For binary data, fit one binomial GLM with logit link function:

```
modlist <- apply(da[, 1:nt], 2, function(y)
                 glm(y ~ da$group - 1, family = binomial(link = "logit")))
```

For count data, fit one Poisson GLM with log link function:

```
modlist <- apply(da[, 1:nt], 2, function(y)
                 glm(y ~ da$group - 1, family = poisson(link = "log")))
```

Perform asymptotic multiple comparisons according to the contrast coefficients in `K`:

```
MMM <- glht(do.call(multcomp::mmm, modlist), linfct = K, df = 0)
```

Compute adjusted p-values and SCI bounds:

```
summary(MMM)$test$pvalues
confint(MMM)
```

For a small-sample adjustment with Gaussian data, choose an integer value greater than 0 for `df`.

Simultaneous Inference Based on a Joint CIM

Create a continuous time variable:

```
mda$timeC <- as.numeric(mda$variable)
```

Fit a CIM with time-stratified random effects:

```
CIM <- lme4::lmer(value ~ tg - 1 + (timeC|person), mda)
```

Perform asymptotic multiple comparisons according to the contrast coefficients in `K`:

```
multcomp::glht(CIM, linfct = K, df=0)
```

For a small-sample adjustment, choose an integer value greater than 0 for `df`. For example, apply the Kenward-Roger method and use the average DF across all comparisons.

Compute the Kenward-Roger adjusted covariance matrix:

```
krMat <- pbkrtest::vcovAdj(CIM)
```

Calculate the average Kenward-Roger DF, rounded down to the nearest integer:

```
krDF <- floor(mean(apply(K, 1, function(x)
                   pbkrtest::ddf_Lb(VVa = krMat, Lcoef = x))))
```

Perform Kenward-Roger adjusted multiple comparisons according to the contrast coefficients in `K`:

```
multcomp::glht(parm(coef = fixef(CIM), vcov = as.matrix(krMat)),
               linfct = K, df=krDF)
```

Compute adjusted p-values with `summary()` and SCI bounds with `confint()` as shown above.

Functionality for multiple tests using the Kenward-Roger method is also available in the packages `lsmeans` (Lenth 2015) and `lmerTest` (Kuznetsova et al. 2015a, described in Kuznetsova et al. 2015b).

Simultaneous Inference Based on a Joint ELM

Fit an ELM with occasion-specific variances σ_j^2 and AR(1) residual correlation:

```
ELM <- nlme::gls(value ~ tg - 1, mda,
                 weights = varIdent(form = ~1|variable),
                 correlation = corAR1(form = ~1|person))
```

Perform asymptotic multiple comparisons according to the contrast coefficients in `K`:

```
multcomp::glht(ELM, linfct = K, df = 0)
```

For a small-sample adjustment, choose an integer value greater than 0 for `df`.

Compute adjusted p-values with `summary()` and SCI bounds with `confint()` as shown above.

Simultaneous Inference Based on GEEs

Sort the data by clusters (necessary for working with `geepack::geeglm`):

```
mda <- mda[order(mda$person), ]
```

For binary data, fit binomial GEEs with AR(1) working correlation:

```
GEE <- geepack::geeglm(value ~ tg - 1, data = mda, id = person,
                       family = binomial, corstr = "ar1")
```

For count data, fit Poisson GEEs with AR(1) working correlation:

```
GEE <- geepack::geeglm(value ~ tg - 1, data = mda, id = person,
                       family = poisson, corstr = "ar1")
```

Perform asymptotic multiple comparisons according to the contrast coefficients in `K`, using the naive covariance matrix of the data:

```
multcomp::glht(parm(coef = coef(GEE), vcov = GEE$geese$vbeta.naiv), linfct = K)
```

Perform asymptotic multiple comparisons according to the contrast coefficients in `K`, using the robust (sandwich) covariance matrix:

```
multcomp::glht(parm(coef = coef(GEE), vcov = GEE$geese$vbeta), linfct = K)
```

Compute adjusted p-values with `summary()` and SCI bounds with `confint()` as shown above.

AICc Model Selection

Assemble a set of candidate models e.g., ELMs with occasion-specific variances σ_j^2 and CS, AR(1), or UN residual correlation:

```
ELM1 <- nlme::gls(value ~ tg - 1, mda,
                  weights = varIdent(form = ~1|variable),
                  correlation = corCompSymm(form = ~1|person))
ELM2 <- nlme::gls(value ~ tg - 1, mda,
                  weights = varIdent(form = ~1|variable),
                  correlation = corAR1(form = ~1|person))
ELM3 <- nlme::gls(value ~ tg - 1, mda,
                  weights = varIdent(form = ~1|variable),
                  correlation = corSymm(form = ~1|person))
```

Select the model that gets most support from the data using AICc:

```
MuMIn::model.sel(ELM1, ELM2, ELM3)
```

AICc model selection can also be applied to CIMs or other LMMs.

Approximative Power

The following function appPower calculates the approximative global, any-pair, and all-pairs power as in 4.5 for longitudinal MCTs. It takes as input a Gaussian mean vector mu, the associated covariance matrix cov, a vector of sample sizes n, degrees of freedom df, a matrix of contrast coefficients cmat, a type I error rate alpha, and specifications whether two-sided or directional one-sided hypotheses are to be tested and which type of power is to be calculated. The output provides the approximate power. The function is based on code from the R package MCPAN (Schaarschmidt et al. 2013).

```
appPower <- function(mu, cov, n, df, cmat, rhs = 0, alpha = 0.05,
                     alternative = c("two.sided", "less", "greater"),
                     power = c("global", "anypair", "allpairs")){

  MU <- matrix(mu, ncol = 1)
  COV <- cmat %*% (cov/n) %*% t(cmat)
  R <- cov2cor(COV)
  M <- nrow(cmat)
  RHS <- rep(rhs, M)
  ExpT <- (as.numeric(cmat %*% MU) - RHS) / sqrt(diag(COV))

  switch(alternative,
         two.sided = {
           crit <- mvtnorm::qmvt(p = 1 - alpha, tail = "both.tails",
                                 df = df, corr = R)[["quantile"]]
         },
         less = {
           crit <- mvtnorm::qmvt(p = 1 - alpha, tail = "upper.tail",
                                 df = df, corr = R)[["quantile"]]
         },
         greater = {
           crit <- mvtnorm::qmvt(p = 1 - alpha, tail = "lower.tail",
                                 df = df, corr = R)[["quantile"]]
```

```
       })

switch(power,
       global = {
         switch(alternative,
                two.sided = {
                  beta <- mvtnorm::pmvt(lower = rep(-crit, M),
                                        upper = rep(crit, M),
                                        delta = ExpT, df = df,
                                        corr = R)
                  whichHA <- which(ExpT != 0)
                },
                less = {
                  beta <- mvtnorm::pmvt(lower = rep(crit, M),
                                        upper = rep(Inf, M),
                                        delta = ExpT, df = df,
                                        corr = R)
                  whichHA <- which(ExpT < 0)
                },
                greater = {
                  beta <- mvtnorm::pmvt(lower = rep(-Inf, M),
                                        upper = rep(crit, M),
                                        delta = ExpT, df = df,
                                        corr = R)
                  whichHA <- which(ExpT > 0)
                })
       },
       anypair = {
         switch(alternative,
                two.sided = {
                  whichHA <- which(ExpT != 0)
                  MHA <- length(whichHA)
                  if(MHA < 1){
                    warning("All contrasts are under their corresponding
                            null hypotheses, thus any-pair power cannot be
                            calculated.")
                    beta <- 1 - alpha
                  }else{
                    beta <- mvtnorm::pmvt(lower = rep(-crit, MHA),
                                          upper = rep(crit, MHA),
                                          delta = ExpT[whichHA],
                                          df = df,
                                          corr = R[whichHA, whichHA])
                  }
                },
                less = {
                  whichHA <- which(ExpT < 0)
                  MHA <- length(whichHA)
```

```
           if(MHA < 1){
             warning("All contrasts are under their corresponding
                     null hypotheses, thus any-pair power cannot be
                     calculated.")
             beta <- 1 - alpha
           }else{
             beta <- mvtnorm::pmvt(lower = rep(crit, MHA),
                                   upper = rep(Inf, MHA),
                                   delta = ExpT[whichHA],
                                   df = df,
                                   corr = R[whichHA, whichHA])
           }
         },
         greater = {
           whichHA <- which(ExpT > 0)
           MHA <- length(whichHA)
           if(MHA < 1){
             warning("All contrasts are under their corresponding
                     null hypotheses, thus any-pair power cannot be
                     calculated.")
             beta <- 1 - alpha
           }else{
             beta <- mvtnorm::pmvt(lower = rep(-Inf, MHA),
                                   upper = rep(crit, MHA),
                                   delta = ExpT[whichHA],
                                   df = df,
                                   corr = R[whichHA, whichHA])
           }
         })
},
allpairs = {
  switch(alternative,
         two.sided = {
           whichHA <- which(ExpT != 0)
           MHA <- length(whichHA)
           if(MHA < 1){
             warning("All contrasts are under their corresponding
                     null hypotheses, thus all-pairs power cannot be
                     calculated.")
             beta <- 1 - alpha
           }else{
             nsim <- 100000
             RT <- mvtnorm::rmvt(n = nsim,
             delta = ExpT[whichHA],
                                 df = df,
                                 sigma = as.matrix(R[whichHA,
                                                     whichHA]),
                                 method = "svd")
```

```
                      nreject <- sum(apply(RT, 1, function(x){min(abs(x))})
                                     > abs(crit))
                      beta <- 1 - (nreject/nsim)
                    }
                  },
                  less = {
                    whichHA <- which(ExpT < 0)
                    MHA <- length(whichHA)
                    if(MHA < 1){
                      warning("All contrasts are under their corresponding
                              null hypotheses, all-pairs power cannot be
                              calculated.")
                      beta <- 1 - alpha
                    }else{
                      beta <- 1 - mvtnorm::pmvt(lower = rep(-Inf, MHA),
                                                upper = rep(crit, MHA),
                                                delta = ExpT[whichHA],
                                                df = df,
                                                corr = R[whichHA, whichHA])
                    }
                  },
                  greater = {
                    whichHA <- which(ExpT > 0)
                    MHA <- length(whichHA)
                    if(MHA < 1){
                      warning("All contrasts are under their corresponding
                              null hypotheses, all-pairs power cannot be
                              calculated.")
                      beta <- 1 - alpha
                    }else{
                      beta <- 1 - mvtnorm::pmvt(lower = rep(crit, MHA),
                                                upper = rep(Inf, MHA),
                                                delta = ExpT[whichHA],
                                                df = df,
                                                corr = R[whichHA, whichHA])
                    }
                  })
         })

  pow <- round(1 - beta[[1]], 3)
  return(pow)

}
```

Comparison-Specific Degrees of Freedom

The following function multDF computes adjusted p-values for MCTs with comparison-specific degrees of freedom. It takes as input a vector of Gaussian mean estimates Est, the associated covariance matrix Sig, a matrix of contrast coefficients K, a vector of contrast-specific degrees of freedom df, and a specification whether two-sided or directional one-sided hypotheses are to be tested. The output provides effect estimates, SEs, test statistics, and adjusted p-values. The function is based on code from the R package SimComp (Hasler 2014b).

```
multDF <- function(Est, Sig, K, df, alternative = "two.sided",
                   level = 0.95){

  ncomp <- nrow(K)
  MC <- summary(glht(parm(Est, Sig), K))
  Ests <- MC$test$coefficients
  Vars <- MC$test$sigma
  Stats <- MC$test$tstat
  R <- cov2cor(vcov(MC))

  pvalue <- upper <- lower <- numeric(ncomp)

  for(z in 1:ncomp){

    if(alternative=="two.sided"){
      pvalue[z] <- 1 - pmvt(lower = rep(-abs(Stats[z]), ncomp),
                            upper = rep(abs(Stats[z]), ncomp),
                            df = floor(df[z]),
                            corr = R)[1]
      quan <- qmvt(p = level,
                   tail = "both.tails",
                   df = floor(df[z]),
                   corr = R)$quantile
      upper[z] <- Ests[z] + quan * Vars[z]
      lower[z] <- Ests[z] - quan * Vars[z]
    }

    if(alternative=="greater"){
      pvalue[z] <- 1 - pmvt(lower = -Inf,
                            upper = rep(Stats[z], ncomp),
                            df = floor(df[z]),
                            corr = R)[1]
      quan <- qmvt(p = level,
                   tail = "lower.tail",
                   df = floor(df[z]),
                   corr = R)$quantile
      upper[z] <- Inf
      lower[z] <- Ests[z] - quan * Vars[z]
    }

    if(alternative=="less"){
      pvalue[z] <- 1 - pmvt(lower = rep(Stats[z], ncomp),
                            upper = Inf,
                            df = floor(df[z]),
                            corr = R)[1]
      quan <- qmvt(p = level,
                   tail = "upper.tail",
                   df = floor(df[z]),
                   corr = R)$quantile
      upper[z] <- Ests[z] - quan * Vars[z]
      lower[z] <- -Inf
    }

  }

  return(round(cbind(estimate = Ests, sigma = Vars, tstat = Stats,
                     lower, upper, p = pvalue), 4))

}
```

E.2 R Package SimLongi

A development version of an R package SimLongi is available from https://github.com/PhilipPallmann/SimLongi. It facilitates the implementation of the methods proposed in this thesis. In particular, it enables to user to perform longitudinal MCTs for Gaussian endpoints based on a joint ELM or LMM very easily with:

- specification of a set of candidate models,
- selection of the "best" model via AICc,
- construction of contrast coefficient matrices for comparisons of treatment groups and/or time points,
- choice among various DF approximations.

We take as an example the bradykinin dataset described in 2.1, which is available in SimLongi as object brady.

Example 1: Assume we want to carry out two-sided many-to-one comparisons of treatment groups per time point (direction="gpt") based on an AICc-selected ELM. The candidate set contains models with homo- and heteroscedastic variances (heteroscedastic across treatment groups or time points or both) and CS, AR(1), or UN correlation of the residuals, and the DF are to be approximated using the ESS:

```
SimLongi(data=brady, response="logConc", group="Drug", time="Time",
         id="ID", covariates=NULL, var=list("hom", "het", "hett", "hetg"),
         cor=list("CS", "AR1", "UN"), type="Dunnett", base=1,
         direction="gpt", alternative="two.sided", level=0.95, df="ess")
```

Example 2: Now suppose we want to perform all-pairwise comparisons of time points per group (`direction="tpg"`) based on an AICc-selected CIM. The candidate set contains models with different structures of the random effects, and a Kenward-Roger adjustment is to be used:

```
SimLongiMix(data=brady, response="logConc", group="Drug", time="Time",
            id="ID", covariates=NULL, rand=list("1|id", "time|id",
            "group|id", "timegroup|id"), type="Tukey", direction="tpg",
            alternative="greater", level=0.95, df="kr")
```

A `ggplot2`-style plotting function `PlotCI` for SCIs is provided by `SimLongi` as well.

This `R` package is far from being an all-purpose tool, simply because there is no way to cover all possible structures of error covariance matrices and random effects, especially when further covariates are involved. Instead it is meant to be an aid to get started and explore the options that come with our longitudinal MCTs.

In addition, `SimLongi` contains a function `PowApprox` to calculate the approximate powers given in 4.5, and two generic functions `ESSgls` and `ESSlme` that compute the ESS associated with parameter estimates of an ELM fitted with `gls` or an LMM fitted with `lme` of the package `nlme`.

Acknowledgements

I am indebted to many people for their support, guidance, and understanding.

Ludwig Hothorn for making me a biostatistician in the first place, for providing a work environment to thrive in, and for continuously believing in my progress.

Frank Schaarschmidt for teaching me some invaluable lessons about statistics and, perhaps unintentionally, about life in general.

Daniel Gerhard for constantly making me see things in a different light.

Hannelore Visser for always lending me an ear and sometimes a helping hand.

Clemens Buczilowski for highly significant contributions to my daily wellbeing.

Andreas Kitsche for his irrepressible optimism and sunny disposition.

Charlotte Vogel for making some cloudy days just that little bit brighter.

Everyone at the Institute of Biostatistics for making it a marvellous place to work at.

Thomas Jaki for convincing me that Lancaster is a marvellous place, too.

Mario Hasler for all the discussions that cemented my understanding of multiple inference.

Christian Ritz for many wise and witty comments on my work, and for co-refereeing this thesis.

Siegfried Kropf for co-refereeing this thesis as well.

Jutta Papenbrock for being the chair of my examination board.

Mias Pretorius for the bradykinin receptor data.

Elias Böckmann for the greenhouse whitefly data.

Josephine Karanja for the azadirachtin data.

All those whom I've forgotten to mention here for not being mad at me (hopefully).

My Ma and Grandma for everything.

Curriculum Vitae

Personal

Name: Philip Steffen Pallmann

Date of birth: 1 March 1987

Place of birth: Geislingen an der Steige, Germany

Citizenship: German

Education

Apr 2012 – April 2016

PhD, Biology (major: biostatistics), Leibniz Universität Hannover.

Thesis: *Multiple contrast tests with repeated and multiple endpoints—with biological applications.*

Advisor: Prof. Dr. Dr. Ludwig A. Hothorn

Oct 2010 – Mar 2012

MSc, Plant Biotechnology (major: biostatistics), Leibniz Universität Hannover.

Thesis: *Two-sample tests and multiple contrast tests of several diversity indices.*

Advisor: Prof. Dr. Dr. Ludwig A. Hothorn

Oct 2007 – Sep 2010

BSc, Plant Biotechnology (major: plant physiology), Leibniz Universität Hannover.

Thesis: *Root movement of Arabidopsis mutants.*

Advisor: Prof. Dr. Günther F. E. Scherer

Sep 1997 – Jun 2006

Michelberg-Gymnasium Geislingen (secondary school)

Sep 1993 – Jul 1997

Albert-Einstein-Schule Geislingen (primary school)

Employment

Since Apr 2015

Senior research associate.

Department of Mathematics and Statistics, Lancaster University, UK.

Mar 2013 – Jan 2014

Statistical consultant.

Prescos LLC—Preclinical Research & Scientific Consulting Services, San Diego, CA.

Apr 2012 – Mar 2015

Research assistant.

Institut für Biostatistik, Leibniz Universität Hannover.

Oct 2010 – Mar 2012

Student assistant.

Institut für Biostatistik, Leibniz Universität Hannover.

May 2009 – Apr 2010

Student assistant.

Institut für Pflanzenkrankheiten und Pflanzenschutz, Leibniz Universität Hannover.

Sep 2006 – May 2007

Community service.

Helfenstein-Klinik Geislingen.

Jun 2003 – Sep 2007

Freelance journalist & photographer.

Geislinger Zeitung / Neue Württembergische Zeitung Göppingen / Südwest Presse Ulm.

Awards

2013

Bernd Streitberg Award of the International Biometric Society, German Region.

2012

Second Prize for Best Student Oral Presentation at the 36th International Biometric Conference, Kobe, Japan.

Publications

Peer-Reviewed Articles

Philip Pallmann, Ludwig A. Hothorn (2016) Analysis of means (ANOM): a generalized approach using R. *Journal of Applied Statistics*, **43**(8), 1544–1563. doi:10.1080/02664763.2015.1117584.

Jacinter A. Otieno, **Philip Pallmann**, Hans-Michael Poehling (2016) Combination of soil-applied azadirachtin with entomopathogens for integrated management of western flower thrips. *Journal of Applied Entomology*, **140**(3), 174–186. doi:10.1111/jen.12242.

Philip Pallmann, Frank Schaarschmidt (2016) Common pitfalls when testing additivity of treatment mixtures with chi-square analyses. *Journal of Applied Entomology*, **140**(1–2), 135–141. doi:10.1111/jen.12258.

Philip Pallmann, Ludwig A. Hothorn (2015) Boxplots for grouped and clustered data in toxicology. *Archives of Toxicology*. doi:10.1007/s00204-015-1608-4.

Philip Pallmann, Mias Pretorius, Christian Ritz (2015) Simultaneous comparisons of treatments at multiple time points: combined marginal models versus joint modeling. *Statistical Methods in Medical Research*. doi:10.1177/0962280215603743

Josephine Karanja, Hans-Michael Poehling, **Philip Pallmann** (2015) Efficacy and dose response of soil-applied neem formulations in substrates with different amounts of organic matter, in the control of whiteflies, *Aleyrodes proletella* and *Trialeurodes vaporariorum* (Hemiptera: Aleyrodidae). *Journal of Economic Entomology*, **108**(3), 1182–1190. doi:10.1093/jee/tov047.

Annekathrin Weese, **Philip Pallmann**, Jutta Papenbrock, Anja Riemenschneider (2015) *Brassica napus* L. cultivars show a broad variability in their morphology, physiology and metabolite levels in response to sulfur limitations and to pathogen attack. *Frontiers in Plant Science*, **6**:9. doi:10.3389/fpls.2015.00009.

Philip Pallmann, Ludwig A. Hothorn, Gemechis D. Djira (2014) A Levene-type test of homogeneity of several variances against ordered alternatives. *Computational Statistics*, **29**(6), 1593–1608. doi:10.1007/s00180-014-0508-z.

Thomas Jaki, **Philip Pallmann**, Martin J. Wolfsegger (2013) Estimation in AB/BA cross-over trials with application to bioequivalence studies with incomplete and complete data settings. *Statistics in Medicine*, **32**(30), 5469–5483. doi:10.1002/sim.5886.

Philip Pallmann, Frank Schaarschmidt, Ludwig A. Hothorn, Christiane Fischer, Heiko Nacke, Kai U. Priesnitz, Nicholas J. Schork (2012) Assessing group differences in biodiversity by simultaneously testing a user-defined selection of diversity indices. *Molecular Ecology Resources*, **12**(6), 1068–1078. doi:10.1111/1755-0998.12004.

Other

Thomas Jaki, **Philip Pallmann**, Martin J. Wolfsegger (2013) Authors' reply to Comments on "Estimation in AB/BA cross-over trials with application to bioequivalence studies with incomplete and complete data settings". *Statistics in Medicine*, **32**(30), 5487-5488. doi:10.1002/sim.6000.

Currently Under Review

Thomas Jaki, **Philip Pallmann**, Dominic Magirr. The R package MAMS for designing multi-arm multi-stage clinical trials. Submitted to *Journal of Statistical Software*.

Jacinter A. Otieno, **Philip Pallmann**, Hans-Michael Poehling. Additive and synergistic interactions amongst *Orius laevigatus* (Heteroptera: Anthocoridae), entomopathogens and neem for controlling western flower thrips (Thysanoptera: Thripidae). Submitted to *Biocontrol*.

Book Reviews

Philip Pallmann (2015) Analyzing Baseball Data with R, by Max Marchi and Jim Albert. *Journal of the Royal Statistical Society, Series A: Statistics in Society*, **178**(4), 1099. doi:10.1111/rssa.3_12138.

Philip Pallmann (2015) Simultaneous Statistical Inference: With Applications in the Life Sciences, by Thorsten Dickhaus. *Biometrical Journal*, **57**(6), 1151–1152. doi:10.1002/bimj.201500129.

Philip Pallmann (2015) Applied Meta-Analysis with R, by Ding-Geng (Din) Chen and Karl E. Peace. *Journal of Applied Statistics*, **42**(4), 914–915. doi:10.1080/02664763.2014.989464.

R Packages

Philip Pallmann (2016) ANOM: Analysis of means. R package version 0.4.3. `http://CRAN.R-project.org/package=ANOM`.

Philip Pallmann (2016) simbe: Simultaneous assessment of bioequivalence on multiple pharmacokinetic parameters. R package version 0.1. `https://github.com/PhilipPallmann/simbe`.

Philip Pallmann (2016) toxbox: Boxplots for toxicological data. R package version 1.1.5. `https://github.com/PhilipPallmann/toxbox`.

Philip Pallmann, Amanda Turner (2015) BayesMAMS: Designing Bayesian multi-arm multi-stage studies. R package version 0.1. `http://CRAN.R-project.org/package=BayesMAMS`.

Thomas Jaki, Dominic Magirr, **Philip Pallmann** (2015) MAMS: Designing multi-arm multi-stage studies. R package version 0.7. `http://CRAN.R-project.org/package=MAMS`.

Ralph Scherer, **Philip Pallmann** (2014) simboot: Simultaneous inference for diversity indices. R package version 0.2-5. `http://CRAN.R-project.org/package=simboot`.

Conference Talks

Philip Pallmann, Thomas Jaki. Simultaneous confidence regions for multi-parameter bioequivalence. 62nd Biometrisches Kolloquium / DAGStat 2016, Göttingen, Germany. 18 March 2016.

Philip Pallmann, Thomas Jaki. The Adaptive Designs Working Group of the Network of Hubs for Trials Methodology Research. UKCRC Registered CTU Network Bi-annual Statisticians' Operational Group Meeting, Sheffield University, UK. 5 October 2015.

Philip Pallmann. Multiple contrast tests with longitudinal data. International Conference on Simultaneous Inference, Hannover, Germany. 25 September 2013.

Philip Pallmann. Two-sample and multiple contrast tests for several biodiversity indices. 59th Biometrisches Kolloquium / DAGStat 2013, Freiburg, Germany. 20 March 2013. *Bernd Streitberg Award Session.*

Philip Pallmann, Thomas Jaki, Martin J. Wolfsegger. Establishing bioequivalence in AB/BA crossover trials using the ratio of AUCs estimated from sparse sampling designs. 59th Biometrisches Kolloquium / DAGStat 2013, Freiburg, Germany. 19 March 2013.

Philip Pallmann. Comparing biodiversity by simultaneously testing a user-defined selection of diversity indices. 36th International Biometric Conference, Kobe, Japan. 30 August 2012. *Second Prize for Best Student Oral Presentation.*

Philip Pallmann. Comparing biodiversity by simultaneously testing a user-defined selection of diversity indices. 58th Biometrisches Kolloquium, Berlin, Germany. 14 March 2012.

Conference Posters

Philip Pallmann, Ludwig A. Hothorn. Analysis of means with the R package 'ANOM'. 62nd Biometrisches Kolloquium / DAGStat 2016, Göttingen, Germany. 16 March 2016. doi:10.13140/RG.2.1.4924.2006.

Philip Pallmann. Simultaneous small-sample inference in longitudinal settings using multiple contrasts. 60th Biometrisches Kolloquium, Bremen, Germany. 12 March 2014. doi:10.13140/RG.2.1.1795.1121.

Seminar & Workshop Talks

Philip Pallmann A Bayesian group-sequential design in trauma care research. Annual Meeting of the Adaptive Designs Working Group of the Medical Research Council Network of Hubs for Trials Methodology Research, Lancaster University, UK. 23 February 2016.

Philip Pallmann, Thomas Jaki. Estimation in AB/BA cross-over trials with application to bioequivalence studies with incomplete and complete data settings. Statistics Forum, Lancaster University, Lancaster, UK. 29 November 2012.