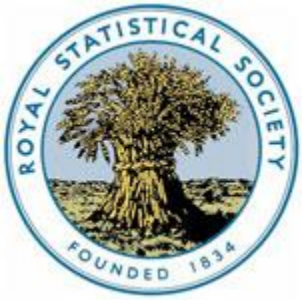



A versatile method for confirmatory evaluation of the effects of a covariate in multiple models

Author(s): Christian Bressen Pipper, Christian Ritz and Hans Bisgaard

Source: *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, Vol. 61, No. 2 (MARCH 2012), pp. 315-326
Published by: Wiley for the Royal Statistical Society
Stable URL: http://www.jstor.org/stable/41430965
Accessed: 08-10-2017 22:13 UTC



*Appl. Statist.* (2012)
**61**, *Part* 2, *pp.* 315–326

# A versatile method for confirmatory evaluation of the effects of a covariate in multiple models

Christian Bressen Pipper, Christian Ritz and Hans Bisgaard

*University of Copenhagen, Denmark*

[Received August 2010. Final revision July 2011]

**Summary.** Modern epidemiology often requires testing of the effect of a covariate on multiple end points from the same study. However, popular state of the art methods for multiple testing require the tests to be evaluated within the framework of a single model unifying all end points. This severely limits their use in applications where there are different types of end point, e.g. binary, continuous or time to event. We use an asymptotic representation of parameter estimates to combine multiple models without additional constraints. This result enables the use of established tools for multiple testing to provide a fine-tuned control of the overall type I error in a wide range of epidemiological experiments where in reality no other useful alternative exists. The methodology proposed is applied to a multiple-end-point study of the effect of neonatal bacterial colonization on development of childhood asthma.

*Keywords*: Epidemiology; Multiple end points; Multiple models; Multiple testing; Type I error

## 1. Introduction

In biomedical and epidemiological studies multiple end points are often identified as relevant (Zhang *et al.*, 1997). For example, in a recent epidemiological study on childhood asthma (Bisgaard *et al.*, 2007) that motivated this work, five clinical markers of recurrent wheeze were identified as the primary end points, on the basis of clinical judgement that these end points as closely as possible reflect the asthmatic condition of the children in the study. The secondary end points encompassed five additional physiological variables such as lung function and blood counts.

The primary goal of such studies is often to establish which if any end points are affected by the target risk factor. In the childhood asthma study the specific goal was to investigate the effect of bacterial colonization (the absence or presence) on the 10 end points. We shall examine this in Section 4 and strengthen the exploratory conclusions that were reached in Bisgaard *et al.* (2007) to a confirmatory conclusion.

The perils of testing for potential effects in multiple end points in terms of finding false effects has been pointed out by many researchers (e.g. Altman (1991), pages 453–454) and various approaches have been proposed to control the familywise error rate, i.e. the risk of committing a type I error.

One popular method is to consider adjusted $p$-values that can be compared directly with any chosen level of significance (Wright, 1992). An equivalent approach is to consider a corrected level of significance based on a given prespecified nominal level and to refer the original unadjusted $p$-values to this corrected level of significance. However, many such approaches

*Address for correspondence*: Christian Bressen Pipper, Statistics Group, Department of Basic Sciences and Environment, Faculty of Life Sciences, University of Copenhagen, Thorvaldsensvej 40, Frederiksberg C, DK 1871, Denmark.
E-mail: pipper@life.ku.dk

 0035–9254/12/61315

for adjusting $p$-values or correcting the significance level are conservative because they do not take into account the correlation between test statistics (D'Agostino and Russell, 2005) (the Bonferroni adjustment is a prominent example).

A very practical and popular solution to this problem is provided in current state of the art procedures (Hothorn *et al.*, 2008; Herberich *et al.*, 2010; Bretz *et al.*, 2010). The idea here is to base the adjustment of $p$-values or correction of the level of significance on simultaneous asymptotic normality of the commonly used $z$-statistics derived within the standard inference framework of a statistical model. As noted in for instance Bretz *et al.* (2010), page 5, this approach can be appreciably less conservative than the Bonferroni method if the test statistics are substantially correlated.

Even though these procedures are very general and flexible they require the test statistics to be evaluated in a single simultaneous model to establish simultaneous asymptotic normality. As such they are of limited practical use in multiple-end-point studies where it is often not feasible to encompass fundamentally different end points in a single simultaneous model.

In this paper we establish simultaneous asymptotic normality of the $z$-statistics for the effect of a covariate from models for different end points. We do this without inducing any similarity restrictions between end point models. The basic idea is to exploit the fact that the test statistics can be decomposed into sums of independent and identically distributed zero-mean random variables, which can then be combined to establish simultaneous asymptotic normality.

This result, combined with the established single-model procedures, enables a confirmatory evaluation of the effect of a covariate on multiple end points that

(a) is appreciably less conservative but just as operational as Bonferroni-type corrections for correlated end points,
(b) is applicable to multiple end points of different types (e.g. continuous, binary and event times),
(c) is capable of handling missing values among end points and explanatory variables and
(d) can deal with different models for different end points (e.g. different explanatory variables).

Consequently, this paper offers a method for improving confirmatory evaluation in complex situations like the childhood asthma example that is analysed in Section 4 where in reality the only other established operational alternative would be a Bonferroni-type correction (Holm, 1979). We have implemented the method in the publicly available package `multmod` in the statistical computing environment R (R Development Core Team, 2010).

The structure of the paper is as follows. In Section 2 we establish simultaneous asymptotic normality of the $z$-statistics. This result is extended to handle missing values in Section 3. Section 4 is dedicated to the analysis of the childhood asthma study. The performance of the proposed methodology in a scenario which is similar to that of this epidemiological study is evaluated by simulation in Section 5. Finally, a discussion is provided in Section 6. Appendix A provides the technical details of how to obtain the confirmatory evaluation proposed.

## 2. Combining asymptotic properties of estimators from multiple models

Suppose that the $j$th end point ($j = 1, \ldots, J$) follows a statistical model in which the covariate effect of interest can be identified as a real-valued parameter $\beta_j$ such that $\beta_j = 0$ corresponds to no effect. Then, in most standard models, the resulting maximum likelihood estimator $\hat{\beta}_j$ will have the asymptotic property

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

where $i = 1, \ldots, n$ corresponds to the $i$th independent observation and $\Psi_{ij}$ is the sum of score function co-ordinates for observation $i$ weighted by the appropriate row of minus the inverse Fisher information matrix (Bickel *et al.* (1998), theorem 2, page 44).

Now define the vectors $\beta = (\beta_1, \ldots, \beta_J)$, $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_J)$ and $\Psi_i = (\Psi_{i1}, \ldots, \Psi_{iJ})$. Then

$$(\hat{\beta} - \beta)\sqrt{n} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Psi_i + o_P(1).$$

Consequently, since the $\Psi_i$s are independent realizations of a zero-mean variable with finite variance, the multivariate central limit theorem yields

$$(\hat{\beta} - \beta)\sqrt{n} \rightsquigarrow N(0, \Sigma), \tag{2}$$

where by the law of large numbers $\Sigma$ is the limit in probability of

$$\frac{1}{n} \sum_{i=1}^{n} \Psi_i^{\mathrm{T}} \Psi_i.$$

As a consequence of the above characterization of $\Sigma$, a consistent estimator of this variance–covariance matrix is

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} \hat{\Psi}_i^{\mathrm{T}} \hat{\Psi}_i \tag{3}$$

where $\hat{\Psi}_{ij}$ are the empirical counterparts of $\Psi_{ij}$ obtained by plugging in the parameter estimates from the $j$th end point model.

For the childhood asthma study we apply linear regression, Cox regression and logistic regression to analyse the individual end points. If for these models we assume that the effect of interest appears as the first co-ordinate in the parameter vector $\gamma_j$ which is linked to the $j$th end point through the design vector $X_i^j = (X_{i1}^j, \ldots, X_{ip_j}^j)^{\mathrm{T}}$, then the $\Psi_{ij}$s have the specific forms that are given below.

(a) *Linear regression*: $\Psi_{ij}$ is the first co-ordinate of the vector

$$-E\left(\frac{1}{\sigma^2} X_i^j X_i^{j\mathrm{T}}\right)^{-1} X_i^j (Y_i^j - X_i^{j\mathrm{T}} \gamma_j),$$

where $E$ denotes expected value, $Y_i^j$ denotes the $j$th end point and $\sigma^2$ is the residual variance.

(b) *Logistic regression*: $\Psi_{ij}$ is the first co-ordinate of the vector

$$-E\{p_i(1 - p_i) X_i^j X_i^{j\mathrm{T}}\}^{-1} X_i^j (Y_i^j - p_i),$$

where $Y_i^j$ denotes the $j$th end point and

$$p_i = \frac{\exp(X_i^{j\mathrm{T}} \gamma_j)}{1 + \exp(X_i^{j\mathrm{T}} \gamma_j)}$$

is the probability of $Y_i^j = 1$.

(c) *Cox regression*: based on Cox partial likelihood $\Psi_{ij}$ is the first co-ordinate of the vector

$$\left\{\int_0^\infty v(s, \gamma_j) s^0(s, \gamma_j) \lambda_0(s)\, \mathrm{d}s\right\}^{-1} \int_0^\infty \left\{X_i^j - \frac{s^1(s, \gamma_j)}{s^0(s, \gamma_j)}\right\} \mathrm{d}M_i(s),$$

where $M_i(t)$ denotes the counting process martingale, $s^r(t, \gamma_j)$ is the limit in probability of

$$\frac{1}{n}\sum_{i=1}^{n} Y_i(t)\exp(X_i^{j\mathrm{T}}\gamma_j)(X_i^j)^{\otimes r}$$

with $Y_i(t)$ denoting the at-risk process, $a^{\otimes 0}=1$, $a^{\otimes 1}=a$, $a^{\otimes 2}=aa^{\mathrm{T}}$,

$$v(t,\gamma_j)=\frac{s^2(t,\gamma_j)}{s^0(t,\gamma_j)}-\frac{s^1(t,\gamma_j)^{\otimes 2}}{s^0(t,\gamma_j)^2}$$

and $\lambda_0(s)$ denotes the baseline hazard. For more details on large sample properties in Cox regression see Andersen *et al.* (1993), section VII.2.

The results above on the simultaneous asymptotic normality of $\hat{\beta}$ in combination with the variance–covariance estimator (3) facilitate confirmatory evaluation of the covariate effect across the individual end point models by using standard multiple-testing tools based on simultaneous normality such as those described in Bretz *et al.* (2010), chapter 3.

In particular we focus on a confirmatory evaluation of the hypotheses $\beta_j=0$ by means of the commonly used $z$-statistics $Z_j=\hat{\beta}\sqrt{n_j}/\widehat{\mathrm{se}(\hat{\beta}_j)}$, where $\widehat{\mathrm{se}(\hat{\beta}_j)}$ is the square root of the corresponding diagonal element of the inverse observed Fisher information for the $j$th end point model. This is accomplished by utilizing the above results to control the familywise error rate in terms of calculating a corrected level of significance or equivalently adjusted $p$-values for evaluation of the $z$-statistics. The details and theoretical properties of this approach are given in Appendix A.

## 3. Handling missing values

In practice, some of the end points considered may have missing values and/or relevant explanatory variables may not always have been recorded. A simple and popular solution to this complication is to perform a complete-case analysis for each of the end points. If, for each end point, we can make the standard assumption of missingness completely at random (Little and Rubin (2002), page 12) valid complete-case maximum likelihood inference and consequently the validity of property (1) for the resulting estimator $\hat{\beta}_j$ is ensured.

Specifically, let $M_{ij}$ denote the indicator of whether the $i$th observation is used in the analysis of the $j$th end point and put $n_j=\Sigma_{i=1}^{n} M_{ij}$. We assume that $M_{ij}$, $i=1,\ldots,n$, are independent identically distributed variables. In this set-up, and with $\Psi_{ij}$ as defined in Section 2, it was shown in Nielsen (1997) that the assumption of missingness completely at random implies

$$(\hat{\beta}_j-\beta_j)\sqrt{n_j}=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Xi_{ij}+o_P(1),$$

where $\Xi_{1j},\ldots,\Xi_{nj}$ are independent replications of a zero-mean variable with finite variance and are given by

$$\Xi_{ij}=\begin{cases}\{\sqrt{P(M_{1j}=1)}\}^{-1}\Psi_{ij} & \text{if } M_{ij}=1,\\ 0, & \text{if } M_{ij}=0.\end{cases}$$

This asymptotic result may then be used as in Section 2 on noting that $P(M_{1j}=1)$ can be consistently estimated by $n_j/n$.

With the more flexible assumption of missingness at random (Little and Rubin (2002), page 12), inference based on complete-case maximum likelihood is no longer valid (Little and Rubin (2002), chapter 3). However, as also stated in Little and Rubin (2002), definition 6.4, one may

obtain valid inference based on the observed data likelihood ignoring the missing data mechanism. This also entails an asymptotic property of the resulting maximum likelihood estimator that is similar to equation (1) to which the machinery of Section 2 may be applied. For this empirical counterparts of the $\Psi_{ij}$s involved in equation (1) can be obtained as in Section 2, based on the observed data score function, ignoring the missing data mechanism, and the corresponding observed Fisher information (Little and Rubin (2002), chapter 9).

In the missingness completely at random situation above we are restricting ourselves to using complete-case information for each end point. Thus we still apply information from cases with partially missing end points or covariates. This is in contrast with the global tests that were developed in Bregenzer and Lehmacher (1998), where only cases that are complete for all end points and covariates are used.

## 4. Childhood asthma after bacterial colonization

In Bisgaard *et al.* (2007) a variety of asthma-related markers were shown to be affected by neonatal bacterial colonization. In the original analysis the effects of bacterial colonization were assessed separately for each end point without controlling the familywise error rate. The analyses were based on the first 5 years of follow-up in the 'Copenhagen studies on asthma in childhood' cohort in which 411 children of asthmatic mothers were followed from birth (Bisgaard, 2004). Results for the following 10 end points were reported:

(a) age at onset of the first wheezy episode (days);
(b) age at onset of persistent wheeze (days);
(c) age at first acute severe exacerbation of wheeze (days);
(d) age at first hospitalization for wheeze (days);
(e) current asthma status at 5 years of age;
(f) allergen-specific immunoglobulin E above or below 0.35 kilounits $l^{-1}$ at 4 years of age;
(g) post-bronchodilator-specific airway resistance after $\beta_2$ agonist inhalation (kilopascals seconds per litre) at 5 years of age;
(h) reversibility of specific airway resistance after $\beta_2$ agonist inhalation (per cent) at 5 years of age;
(i) blood eosinophil count (times $10^{-9}$ per litre) at 4 years of age;
(j) total immunoglobulin E (kilounits per litre) at 4 years of age.

At approximately 4 weeks of age, bacterial colonization status was recorded as the presence or absence of one or more of the bacteria *Streptococcus Pneumoniae*, *Haemophilus Influenzae* and *Moraxella catarrhalis*. Bacterial colonization was recorded in 321 children with 61 children being colonized. Moreover, the numbers of children with complete information for effect assessment in each of the 10 end points were 279, 279, 279, 279, 260, 251, 213, 213, 239 and 248. For each of the end points a complete-case analysis was conducted as specified below. Arguments to support the assumption of missingness completely at random and consequently a complete-case analysis were given in Bisgaard *et al.* (2007).

The effects of bacterial colonization on age-at-onset end points were assessed by Cox regression adjusting for gender, gestational age, maternal smoking during the third trimester, maternal use of antibiotics during the third trimester, breast-feeding (yes or no), neonatal bronchial responsiveness and presence or absence of older siblings at home. The effect of bacterial colonization on current asthma and specific immunoglobulin E was assessed by logistic regression. The effects of bacterial colonization on reversibility of specific airway resistance, log(post-bronchodilator-specific airway resistance), log(blood eosinophil count) and log(total immunoglobulin

**Table 1.** Adjusted $p$-values based on the method proposed or the Bonferroni method

| *Outcome* | *p-values for the following methods:* | |
|---|---|---|
| | *Bonferroni* | *Proposed* |
| First wheezy episode | 0.15 | 0.13 |
| Persistent wheeze | 0.17 | 0.15 |
| Exacerbation of wheeze | 0.013 | 0.012 |
| Hospitalization for wheeze | 0.028 | 0.026 |
| Current asthma | 0.00058 | 0.00055 |
| Specific immunoglobulin E | 1.00 | 1.00 |
| Post-bronchodilator | 0.51 | 0.45 |
| Reversibility | 0.42 | 0.37 |
| Blood eosinophil count | 0.058 | 0.053 |
| Total immunoglobulin E | 0.36 | 0.32 |

E) were assessed by means of analysis of variance. For each of the 10 end points $z$-tests for effect of bacterial colonization yielded the $p$-values 0.016, 0.018, 0.0013, 0.0028, 0.000058, 0.47, 0.068, 0.052, 0.0060 and 0.044.

If we wish to make a confirmatory conclusion on the bacterial colonization effect for each end point by using the Bonferroni-type correction that is explained in detail in Appendix A we should evaluate the obtained $p$-values at a 0.0051 level of significance. In this context we conclude that there is a significant effect of bacterial colonization on age at first acute severe exacerbation of wheeze, age at first hospitalization of wheeze and current asthma at 5 years of age. The proposed correction method that is described in detail in Appendix A suggests that we should evaluate the $p$-values at a less conservative 0.0056 level of significance which for the $p$-values obtained leads to the same conclusion as when applying the Bonferroni correction. Equivalently we could base our conclusions on the adjusted $p$-values evaluated at a 5% level. These $p$-values are presented in Table 1.

From Table 1 we note that the adjusted $p$-values based on the method proposed are consistently smaller than those based on the Bonferroni correction, confirming that our procedure is less conservative in this application where the asthma markers are naturally expected to be linked.

As stated in Bisgaard *et al.* (2007) the primary end points of that paper were clinical diagnoses of various stages of asthma development during childhood, which comprise the first five end points in the above list. A confirmatory conclusion based on these five end points would correspond to evaluating at a 0.010 level of significance by using the Bonferroni-type correction and at a 0.012 level of significance by using our proposed method. In both instances we conclude that bacterial colonization relates to clinical diagnoses of severe or late stages of asthma development. Adjusted $p$-values for the primary end point analysis are given in Table 2. Again the procedure proposed provides a less conservative adjustment for multiple testing than the Bonferroni correction.

## 5. Simulation study

To evaluate the properties of the proposed adjustment for multiple testing in a scenario that is similar to that of the application and also to show the potential gain from using the method we

**Table 2.** Adjusted $p$-values for the primary end point analysis

| *Outcome* | *p-values for the following methods:* | |
|---|---|---|
| | *Bonferroni* | *Proposed* |
| First wheezy episode | 0.077 | 0.064 |
| Persistent wheeze | 0.088 | 0.074 |
| Exacerbation of wheeze | 0.0063 | 0.0057 |
| Hospitalization for wheeze | 0.014 | 0.012 |
| Current asthma | 0.00029 | 0.00023 |

performed a small simulation study. Simulations were conducted using the statistical programming environment R (R Development Core Team, 2010).

We consider a number of scenarios where the adjustment is evaluated on the basis of 10 simulated end points constructed to minic the end points in the asthma study. In accordance with the nature of the end points in the asthma study the simulated data are constructed so that end points 1–5, end points 6, 9 and 10, and end points 7 and 8 are correlated. The specific recipe for generating the simulated data is as follows:

$$\begin{aligned}
\text{Endpoint}_{i1} &= \{\min(\tilde{T}_{i1}, 5), I(\tilde{T}_{i1} \leqslant 5)\} \qquad \tilde{T}_{i1} = -\log\{\Phi(\varepsilon_{i1})\}\exp(\beta\,\text{Colonized}_i),\\
\text{Endpoint}_{i2} &= \{\min(\tilde{T}_{i2}, 5), I(\tilde{T}_{i2} \leqslant 5)\}\\
\tilde{T}_{i2} &= [-\log\{\Phi(\varepsilon_{i1})\} - w\log\{\Phi(\varepsilon_{i2})\}]\exp(\beta\,\text{Colonized}_i),\\
\text{Endpoint}_{i3} &= \{\min(\tilde{T}_{i3}, 5), I(\tilde{T}_{i3} \leqslant 5)\}\\
\tilde{T}_{i3} &= [-\log\{\Phi(\varepsilon_{i1})\} - w\log\{\Phi(\varepsilon_{i2})\} - w\log\{\Phi(\varepsilon_{i3})\}]\ \exp(\beta\,\text{Colonized}_i),\\
\text{Endpoint}_{i4} &= \{\min(\tilde{T}_{i4}, 5), I(\tilde{T}_{i4} \leqslant 5)\}\\
\tilde{T}_{i4} &= [-\log\{\Phi(\varepsilon_{i1})\} - w\log\{\Phi(\varepsilon_{i2})\} - w\log\{\Phi(\varepsilon_{i3})\}\\
&\qquad - w\log\{\Phi(\varepsilon_{i4})\}]\exp(\beta\,\text{Colonized}_i),\\
\text{Endpoint}_{i5} &= I\{\tilde{T}_{i4} \leqslant \min(0.25 + 3w, 4)\},\\
\text{Endpoint}_{i6},\ P(\text{Outcome}_{i6} = 1) &= \frac{\exp(\gamma\,\text{Colonized}_i + \delta\,\text{Outcome}_{i10})}{1 + \exp(\gamma\,\text{Colonized}_i + \delta\,\text{Outcome}_{i10})},\\
\text{Endpoint}_{i7} &= \gamma\,\text{Colonized}_i + \varepsilon_{i5},\\
\text{Endpoint}_{i8} &= \gamma\,\text{Colonized}_i + \delta \times \varepsilon_{i5} + \varepsilon_{i6},\\
\text{Endpoint}_{i9} &= \gamma\,\text{Colonized}_i + \varepsilon_{i7},\\
\text{Endpoint}_{i10} &= \gamma\,\text{Colonized}_i + \delta \times \varepsilon_{i7} + \varepsilon_{i8},
\end{aligned}$$

where $i = 1, \ldots, 321$, $I(A)$ denotes the indicator of an event $A$, $\varepsilon_{i1}, \ldots, \varepsilon_{i8}$ are independent identically distributed standard normal variables, $\delta$ and $w$ determine the degree of correlation between end points and the parameters $\beta$ and $\gamma$ determine the effect of bacterial colonization. The variable $\text{Colonized}_i$ is an indicator of whether or not the $i$th child is colonized, sampled from a Bernoulli distribution with colonization frequency 61/321 corresponding to that of the application.

To align further with the application we induce a pattern of missingness completely at random by generating the missingness indicators $M_{ij}$ for the $i$th individual and $j$th end point according to a Bernoulli distribution with frequencies 279/321, 279/321, 279/321, 279/321, 260/321, 251/321, 213/321, 213/321, 239/321 and 248/321 for each of the 10 end points.

**Table 3.** Summary of the simulation study†

| $\beta$ | $(w,\delta)$ | $\bar{\alpha}_{\text{corrected}}$ | *Observed familywise error rate (proposed method)* | *Observed familywise error rate (Bonferroni)* |
|---|---|---|---|---|
| 0 | (5, 0.1) | 0.0053 | 0.049 | 0.046 |
| 0 | (1, 0.5) | 0.0057 | 0.051 | 0.046 |
| 0 | (0.1, 5) | 0.0067 | 0.050 | 0.040 |
| 1 | (5, 0.1) | 0.0053 | 0.029 | 0.028 |
| 1 | (1, 0.5) | 0.0057 | 0.029 | 0.026 |
| 1 | (0.1, 5) | 0.0069 | 0.032 | 0.024 |

†The sample mean of the corrected significance level based on the proposed method is denoted $\bar{\alpha}_{\text{corrected}}$.

We consider the scenarios $\beta = 0, 1$ corresponding to no effect or effect of colonization on the first five end points combined with $(w, \delta) = (5, 0.1), (1, 0.5), (0.1, 5)$, corresponding to increasing correlation between end points. In all scenarios $\gamma = 0$, corresponding to no effect of colonization on the last five end points.

For each scenario we generate 5000 data sets according to the specifications above. For each data set the effect of colonization is evaluated on the basis of complete-case analyses in terms of Cox regression for the first four end points, logistic regression for end points 5 and 6, and standard linear regression for the remaining end points. For each data set we use the method proposed and the Bonferroni method that is described in Appendix A to evaluate the resulting $z$-test statistics controlling the familywise error rate at a 5% level. Table 3 summarizes the performance of adjustment for multiple testing.

From Table 3 we conclude that in all scenarios the method proposed provides adequate control of the familywise error rate at a 5% level. Furthermore it is consistently less conservative than the Bonferroni correction for which the corrected level of significance is 0.0051. This becomes more pronounced as the correlation between end points increases and results in appreciable gains in terms of the familywise error rate. In terms of magnitude of correlation, the asthma study corresponds to the configuration $(w, \delta) = (1, 0.5)$.

## 6. Discussion

We have proposed a simple and flexible procedure for confirmatory evaluation of effects in parallel models for multiple end points. The procedure is based on simultaneous asymptotic normality of effect estimates combined with established procedures for multiple-testing adjustment. We have shown both theoretically and by simulations that the procedure can provide a fine-tuned control of the familywise error rate with a minimum of non-restrictive assumptions and thus can deal with a wide range of complex real life data scenarios.

In our comparison with other methods we have focused on a Bonferroni-type correction as it is in reality the only established type of procedure providing a sensible corrected level of significance with as few assumptions and in as complex situations as those which we consider.

One could argue that the gains when applying the procedure proposed instead of a standard Bonferroni procedure are negligible in the asthma study in Section 4, where the same conclusion is reached by both approaches. However, we draw attention to the fact that the corrected level of significance of the method proposed is somewhat (10%) higher than the Bonferroni-corrected

level of significance which could potentially lead to a materially different conclusion for another configuration of $p$-values. To exemplify, consider the age-at-onset outcomes persistent wheeze, severe exacerbation of wheeze and hospitalization for wheeze from the asthma study. A conclusion based on these three outcomes would yield corrected $p$-values 0.044, 0.0032 and 0.0070 with the procedure proposed whereas corrected $p$-values by using the Bonferroni correction would be 0.054, 0.0038 and 0.0084. Thus in this example conclusions would differ.

As also indicated by the simulation study that is presented in this paper the potential gain from using the method proposed becomes larger when the correlations between the outcomes and consequently the correlations between the $z$-tests increase. Thus we should expect more substantial gains in applications with larger correlations than in the asthma study, where the average absolute correlation between $z$-tests is 0.18.

Furthermore, if we are willing to control the familywise error rate at the cost of one corrected level of significance to which all $p$-values are related, Bonferroni-type procedures can be made less conservative by adopting a sequential strategy like the Holm procedure described by Holm (1979). In fact, we may apply the same strategy to our procedure by recomputing the corrected level of significance that we propose for various subsets of tests. One can reuse the arguments in Holm (1979) to conclude that this results in a procedure that also provides strong control of the familywise error rate, but less conservative than our proposed procedure. However, as noted in for instance Strassburger and Bretz (2008), adopting the sequential procedure above makes it notoriously difficult to provide standard inference products like a sensible simultaneous confidence band of the estimated effects.

Also note that even though the focus of Section 2 is on maximum likelihood estimation the procedure does not require estimates of effects to be obtained by maximum likelihood as long as property (1) holds, i.e. as long as the estimators are asymptotically linear (Bickel *et al.* (1998), page 19). Moreover, the procedure is easily extended to deal with more effects within each end point model.

One restriction, however, is that the procedure is partially model driven in the sense that the tests for each end point need to be derived from a statistical model fitted to the original data. Thus, for instance, the procedure is not applicable to an arbitrary collection of $p$-values that are obtained by using Fisher's exact test (Westfall and Young, 1989; Westfall and Troendle, 2008). The procedure is also not applicable if only aggregate data are available (in which case the $\chi^2$-test may still be applicable) because the decomposition according to independent identically distributed variables is based on the original data. However, we consider this to be a minor disadvantage as statistical models based on the original data are most often needed to obtain estimated effects that typically also are of practical interest.

Another issue is the simultaneous asymptotic normality that is stated in expression (2), the quality of which is essential for the confirmatory evaluation performed. Clearly this approximation may not be appropriate with small sample sizes or a large number of end points. One potential way of checking this assumption is by bootstrapping the distribution of $\hat{\beta}$ and using this to evaluate the normal approximation. For instance we may calculate the $z$-scores from the bootstrapped samples of $\hat{\beta}$. The resulting $Z_{\max}$-quantities as defined in Appendix A may then be applied to check the validity of the normal approximation as it is used in equation (4) in Appendix A. Specifically, if the approximation holds, the empirical cumulative distribution function of the bootstrapped $Z_{\max}$-values evaluated at $z_{1-\alpha_{\text{corrected}}/2}$ as defined in Appendix A should be close to 0.95. In the application, with $z_{1-\alpha_{\text{corrected}}/2} = 2.77$ and 10000 bootstrap samples, we obtain a value of 0.94. However, as noted by Bretz *et al.* (2010), page 140, bootstrap evaluation should be applied with caution as it may also work poorly in situations with many end points and small sample sizes.

The current implementation of our method uses the R package `mvtnorm` (Genz *et al.*, 2009) based on the methodology that was described in Genz and Bretz (2009) to compute multivariate normal probabilities. In terms of precision and efficiency this approach is ideal for most applications. However, with very large numbers of tests the approach becomes computationally very time consuming. In such cases we suspect that easy-to-calculate approximations like that presented in Efron (1997) may substantially reduce the computational burden.

## Acknowledgements

We thank the Joint Editor, Associate Editor and the reviewers for their constructive comments that substantially improved the readability and focus of an earlier version of this paper.

## Appendix A: Corrected significance level

Assume that there is no effect, i.e. $\beta_j = 0$ for all $j$; then according to expression (2) the simultaneous distribution of the $(Z_1, \ldots, Z_J)$ is asymptotically normal with zero mean and variance–covariance matrix $C = \mathrm{diag}(\Sigma)^{-1/2}\Sigma\,\mathrm{diag}(\Sigma)^{-1/2}$. This means that the familywise error rate when assuming no effect can be consistently approximated as

$$P(Z_{\max} > z_{1-\alpha/2}) \to 1 - \int_{-z_{1-\alpha/2}}^{z_{1-\alpha/2}} \cdots \int_{-z_{1-\alpha/2}}^{z_{1-\alpha/2}} \phi(s_1, \ldots, s_J, 0, C)\,\mathrm{d}s_1 \ldots \mathrm{d}s_J \qquad \text{as } n \to \infty \tag{4}$$

where $Z_{\max} = \max_{j=1,\ldots,J} |Z_j|$. Furthermore $\alpha$ is a prespecified significance level to be used for assessing the multiple tests, $z_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile in the standard normal distribution and $\phi$ is the density function of the $J$-dimensional normal distribution with mean 0 and variance–covariance matrix $C$.

On the basis of equation (3), a consistent estimator of $C$ can be obtained as

$$\hat{C} = \mathrm{diag}(\hat{\Sigma})^{-1/2}\hat{\Sigma}\,\mathrm{diag}(\hat{\Sigma})^{-1/2}.$$

To simplify the notation in what follows, we denote by $f(\alpha, C)$ the right-hand side of expression (4). The function $f$ is continuous and thus $f(\alpha, C)$ can be consistently approximated by $f(\alpha, \hat{C})$. Moreover, $f$ is strictly increasing as a function of $\alpha$ on $(0, 1)$ and hence for fixed $C$ the inverse $f_C^{-1}$ is well defined. Note that we can easily calculate $f_C^{-1}$ at a given value by using for instance the method of bisection.

With this notation established the corrected significance level corresponding to a nominal level $\alpha_{\text{nominal}}$ is defined as

$$\alpha_{\text{corrected}} = f_{\hat{C}}^{-1}(\alpha_{\text{nominal}}).$$

In the following two propositions we first establish weak control of the familywise error rate and then as the next step strong control.

*Proposition 1.* Under the intersection of null hypotheses, the familywise error rate of the procedure proposed converges to the prespecified nominal level. Consequently, the procedure provides weak control of the familywise error rate asymptotically.

*Proof.* Since $C \to f_C^{-1}(\alpha)$ is continuous as a function of $C$ the continuous mapping theorem yields $\alpha_{\text{corrected}} \to^{\mathrm{P}} f_C^{-1}(\alpha_{\text{nominal}})$ as $n \to \infty$. If $\beta_j = 0$ for $j = 1, \ldots, J$ then, by Slutsky's theorem, we find that

$$P(Z_{\max} > z_{1-\alpha_{\text{corrected}}/2}) \to f\{f_C^{-1}(\alpha_{\text{nominal}}), C\} = \alpha_{\text{nominal}} \qquad \text{as } n \to \infty.$$

In other words applying the asymptotic correction proposed implies that the familywise error rate when assuming no effect converges to the desired nominal level $\alpha_{\text{nominal}}$. As a direct consequence we have established weak control of the family wise error rate asymptotically.

Next, we establish that the method provides strong control of the familywise error rate asymptotically as defined by Hochberg and Tamhane (1987), page 3.

*Proposition 2.* Asymptotically, the procedure proposed provides strong control of the familywise error rate.

*Proof.* We apply the proposed correction to the subset of tests $\{j : \beta_j = 0\}$ to obtain a corrected significance level $\tilde{\alpha}_{\text{corrected}}$. By construction we have $\tilde{\alpha}_{\text{corrected}} \geqslant \alpha_{\text{corrected}}$ and thus, by proposition 1,

$$P(Z_{\max,0} > z_{1-\alpha_{\text{corrected}}/2}) \leqslant P(Z_{\max,0} > z_{1-\tilde{\alpha}_{\text{corrected}}/2}) \to \alpha_{\text{nominal}} \qquad \text{as } n \to \infty$$

with the definition $Z_{\max,0} = \max_{j=1,\ldots,J:\beta_j=0} |Z_j|$, showing that in general the familywise error rate is bounded asymptotically by $\alpha_{\text{nominal}}$ and thus proves our claim of strong control.

*Remark 1.* The method above coincides with the Bonferroni-type correction $\alpha_{\text{bonf}} = 1 - (1 - \alpha_{\text{nominal}})^{1/J}$ (Hsu (1996), page 58) for independent test statistics $Z_j$s since in this case $f(\alpha, C) = f(\alpha, I) = 1 - (1 - \alpha)^J$ with $I$ denoting the identity matrix. This correction is explicitly termed the Slepian correction and is slightly less conservative than the standard Bonferroni correction that is given by $\alpha_{\text{nominal}}/J$. However, as can be appreciated from the formulae the discrepancies between the two corrections are minute and in practice inconsequential. In the application and simulations of this paper we have abused the terminology slightly by explicitly using the Slepian correction while referring to it as the Bonferroni correction. The method proposed is less conservative than both of these corrections because of the inequality $f(\alpha, C) \leqslant f(\alpha, I)$, which was established by Šidák (1968).

*Remark 2.* As an alternative to correcting the level of significance we could have adjusted the original $p$-values by transforming them with $f_{\hat{C}}$ and evaluating them on the basis of the nominal level $\alpha_{\text{nominal}}$. This would clearly lead to the same conclusions in terms of what effects are significant.

## References


Altman, D. G. (1991) *Practical Statistics for Medical Research*. London: Chapman and Hall.

Andersen, P. K., Borgan, Ø., Gill, R. D. and Keiding, N. (1993) *Statistical Models based on Counting Processes*. New York: Springer.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y. and Wellner, J. A. (1998) *Efficient and Adaptive Estimation for Semiparametric Models*, 1st edn. New York: Springer.

Bisgaard, H. (2004) The Copenhagen Prospective Study on Asthma in Childhood (COPSAC): design, rationale, and baseline data from a longitudinal birth cohort study. *Ann. All. Asth. Immunol.*, **93**, 381–389.

Bisgaard, H., Hermansen, M. N., Buchvald, F., Loland, L., Halkjaer, L. B., Bønnelykke, K., Brasholt, M., Heltberg, A., Vissing, N. H., Thorsen, S. V., Stage, M. and Pipper, C. B. (2007) Childhood asthma after bacterial colonization of the airway in neonates. *New. Engl. J. Med.*, **357**, 1487–1495.

Bregenzer, T. and Lehmacher, W. (1998) Directional tests for the analysis of clinical trials with multiple endpoints allowing for incomplete data. *Biometr. J.*, **40**, 911–928.

Bretz, F., Hothorn, T. and Westfall, P. (2010) *Multiple Comparisons using R*. Boca Raton: Chapman and Hall–CRC.

D'Agostino, Sr, R. B. and Russell, H. K. (2005) Multiple endpoints, multivariate global tests. In *Encyclopedia of Biostatistics*, 2nd edn, vol. 5 (eds P. Armitage and T. Colton). Chichester: Wiley.

Efron, B. (1997) The length heuristic for simultaneous hypothesis tests. *Biometrika*, **84**, 143–157.

Genz, A. and Bretz, F. (2009) *Computation of Multivariate Normal and t Probabilities*. Berlin: Springer.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F. and Hothorn, T. (2009) mvtnorm: multivariate normal and t distributions. *R Package Version 0.9-7*.

Herberich, E., Sikorski, J. and Hothorn, T. (2010) A robust procedure for comparing multiple means under heteroscedasticity in unbalanced designs. *PLOS ONE*, **5**, article e9788.

Hochberg, Y. and Tamhane, A. C. (1987) *Multiple Comparison Procedures*, 1st edn. New York: Wiley.

Holm, S. (1979) A simple sequentially rejective multiple test procedure. *Scand. J. Statist.*, **6**, 65–70.

Hothorn, T., Bretz, F. and Westfall, P. (2008) Simultaneous inference in general parametric models. *Biometr. J.*, **50**, 346–363.

Hsu, J. C. (1996) *Multiple Comparisons Theory and Methods*. London: Chapman and Hall.

Little, R. and Rubin, D. (2002) *Statistical Analysis with Missing Data*, 2nd edn. New York: Wiley.

Nielsen, S. (1997) Inference and missing data: asymptotic results. *Scand. J. Statist.*, **24**, 261–274.

R Development Core Team (2010) *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Šidák, Z. (1968) On multivariate normal probabilities of rectangles: their dependence on correlations. *Ann. Math. Statist.*, **39**, 1425–1434.

Strassburger, K. and Bretz, F. (2008) Compatible simultaneous lower confidence bounds for the holm procedure and other bonferroni-based closed tests. *Statist. Med.*, **27**, 4914–4927.

Westfall, P. H. and Troendle, J. F. (2008) Multiple testing with minimal assumptions. *Biometr. J.*, **50**, 745–755.

Westfall, P. H. and Young, S. S. (1989) $p$ value adjustments for multiple tests in multivariate binomial models. *J. Am. Statist. Ass.*, **84**, 780–786.
Wright, S. P. (1992) Adjusted $p$-values for simultaneous inference. *Biometrics*, **48**, 1005–1013.
Zhang, J., Quan, H., Ng, J. and Stepanavage, M. E. (1997) Some statistical methods for multiple endpoints in clinical trials. *Contr. Clin. Trials*, **18**, 204–221.