Semi-parametric marginal models for hierarchical data and their corresponding full models

Semi-parametric Marginal Models for Hierarchical Data and Their Corresponding Full Models

Geert Molenberghs [a,*], Michael G. Kenward [b], [a]

[a] *I-BioStat, Hasselt University, Diepenbeek, Belgium & Katholieke Universiteit Leuven, Leuven, Belgium*

[b] *Medical Statistics Unit, London School of Hygiene and Tropical Medicine*

Abstract

Semi-parametrically specified models for multivariate, longitudinal, clustered, multilevel, and other hierarchical data, particularly for non-Gaussian outcomes, are ubiquitous because their parameters can most often be conveniently estimated using the important class of generalized estimating equations (GEE). The focus here is on marginal models, to be understood as models that condition neither on random effects nor on other outcomes, but merely on fixed covariates. In spite of their well-deserved popularity, concern could be raised as to whether such models can always be viewed as a partially specified version of a model with full distributional assumptions, or rather whether such a *parent* simply does not exist. It is shown, through the use the hybrid marginal-conditional models, that the answer is affirmative. For conventional GEE with a working correlation structure, the Bahadur model is sometimes considered to be the natural parent candidate, but we show this is a misconception. The result presented here, which is conceptual in nature, is valid whenever the exponential family is used for the semi-parametric specification, or when a straightforward transformation to an exponential family member is possible, implying validity for a broad classes of binary, ordinal, nominal, and count data. The result is illustrated in the context of trivariate binary data. Further, as an illustration, many of the models considered are applied to data from a developmental toxicity study.

Key words: Bahadur model, Beta-binomial model, Conditional model, Dale model, Exponential family, Generalized estimating equations, Hybrid marginal-conditional Model; Marginal model

* Corresponding author: Geert Molenberghs, I-BioStat, Universiteit Hasselt, Agoralaan 1, B-3590 Diepenbeek, Belgium, Tel. +3211268205. Fax: +3211268299.
Email address: geert.molenberghs@uhasselt.be (Geert Molenberghs).

1 Introduction

The statistical literature boasts many modelling approaches for hierarchically organized data structures, including longitudinal, clustered, spatial, and multivariate data. Models for non-Gaussian data are usefully subdivided into marginal, conditional, and random-effects models. Here, we focus on marginal models, where a component of the outcome vector is modelled in terms of covariates but with no additional structure involving other outcomes or latent constructs, such as random effects. Examples in a categorical setting include the models formulated by Bahadur (1961), Dale (1986), Molenberghs and Lesaffre (1994), Lang and Agresti (1994), and Glonek and McCullagh (1995). An extensive review is provided by Molenberghs and Verbeke (2005). These authors point out that such fully parametrically specified models pose non-trivial computational challenges, largely explaining the popularity of semi-parametrically specified alternatives, of which generalized estimating equations (GEE; Liang and Zeger 1986) are a prominent example. In their original form, GEE's only require the correct specification of the first moment of the distribution to get consistent estimates. Correlation between outcomes is accommodated by means of a working correlation matrix. There are a number of variations to this theme, including the use of odds ratios rather than correlations to capture association, and second- and higher-order versions, in which not just the first, but rather the first few, moments of the joint distribution are specified. An excellent early review is Liang, Zeger, and Qaqish (1992), who also show that the set of moments specified by GEE coincide with these of the Bahadur model. When the pairwise correlations are modelled as well, it is customary to refer to GEE2, as opposed to the basic GEE, or GEE1. For further review, see Molenberghs and Verbeke (2005).

Rather than being a model, GEE is an estimation technique. It is usually used to fit marginal models, although Zeger, Liang, and Albert (1988) provided an early instance of GEE applied to random-effects models. In this paper, we are concerned with marginally-specified semi-parametric models.

There are obvious advantages to operating within the full likelihood paradigm, including the availability of a well stocked inferential toolkit, existence of goodness-of-fit machinery, access to the full data generating mechanism, the relative ease with which incomplete data can be analysed (Little and Rubin 2002, Molenberghs and Kenward 2007), an so on. Commonly quoted disadvantages are the aforementioned computational complexity, which can be prohibitive even for moderate-length outcome sequences, and a greater vulnerability to model misspecification, relative to GEE or other semi-parametric approaches.

There is another, sometimes overlooked but nevertheless important, consideration, which is the focus of this contribution. We explain this in terms of

GEE for binary data. Upon correctly specifying a semi-parametric model for an outcome sequence of, say, length n, all pairwise and bivariate models that can be formed from this sequence are coherent, in the sense that all pairwise probabilities belong to the unit interval. In other words, all pairwise models emanating from such a sequence are coherent. This, in itself, may not always be guaranteed. However, we start from the premise that the portions of the model actually specified are coherent, and then look beyond this, addressing the concern as to whether there then is a corresponding n-way model. This is an understandable consideration because, even in the absence of specified higher-order correlations, it may be esthetically and conceptually undesirable to estimate marginal regression relationships that cannot be placed within a fully specified model. We will provide examples where such non-existence arises. However, this ought not to be confused with the broader statement that *there is no compatible model*. We will establish that, whenever the lower-order marginal model, needed for GEE or other semi-parametric estimation technique such as pseudo-likelihood (Molenberghs and Verbeke 2005) is correctly specified, then there is always at least one fully specified compatible model, even in situations where no single compatible Bahadur model exists. There is no justification for a particular focus on the Bahadur model, other than that it is sometimes perceived as the 'natural' fully-specified counterpart of a semi-parametric model in terms of univariate logit links and bivariate correlations. This focus notwithstanding, our result hold generically and is tied neither to the Bahadur model nor to correlations as measures of association.

To establish this result, we will employ the mixed marginal-conditional, or hybrid, modelling framework, proposed by Fitzmaurice and Laird (1993); see also Molenberghs and Ritter (1996), Molenberghs and Danielson (1999), and Glonek (1996). Fitzmaurice and Laird (1993) model the marginal mean parameters, together with the canonical interaction parameters in the multivariate exponential family distribution of Cox (1972). Molenberghs and Ritter (1996) and Molenberghs and Danielson (1999) also incorporate the second-order association into the marginal part, leaving the third- and higher-order parameters conditional in nature. Their model is thus parameterized using marginal means, pairwise marginal odds ratios and higher order conditional odds ratios. It is our contribution here to expand a collection of marginally specified means and pairwise association parameters, resulting from a semi-parametric model, into a compatible, fully-specified model. Because the marginal and conditional sets of parameters are orthogonal in the sense of Cox and Reid (1987), validity is assured. A particularly simple completion is in terms of higher-order conditional independence, i.e., by setting all additional parameters equal to zero. Note that connections between fully and semi-parametrically specified models have also been discussed by Heagerty and Zeger (1996).

Such a fully specified model, compatible with a semi-parametrically specified one, will be termed a *parent*. Our result is constructive in nature, rather

than just one of existence. The relevance of the construction lies in the fact that the parent provides a natural description of the framework into which the semi-parametrically specified parameters fit. The implication is that such semi-parametric methods as GEE1, GEE2, ALR, etc. can always be applied because there is always a valid parent, and hence a probabilistic basis. The sole condition is that the parametrically specified portion of the model be valid, but this is no different to any other statistical modelling exercise.

It ought to be clear that a parent need not be unique, and examples of the reverse are provided. When one is interested in a partially specified model only, i.e., when a strict semi-parametric view-point is adopted, this is irrelevant. However, should one be interested in modelling additional higher-order moments, or even all moments, thereby shifting to a full likelihood approach, then a particular parent can be chosen in accordance with various considerations, both statistical and/or a substantive in nature. For example, the quality of fit can be considered, or rather the particular parent can be chosen that best relates to quantities of scientific relevance (e.g., correlations or odds ratios to capture association).

Our results hold for binary and categorical data, but more generally for the entire exponential family, including cases where the outcome vector does not even need to be homogeneous, but rather can amalgamate different data types.

The rest of the paper is organized as follows. Various modelling frameworks are presented in Section 2: the Bahadur model (Section 2.1), generalized estimating equations (Section 2.2), the beta-binomial model (Section 2.3), the hybrid marginal-conditional model (Section 2.4), and second-order generalized estimating equations (Section 2.5). The main result about compatible, fully-specified models is presented in Section 3. Conceptual examples and counterexamples are the subject of Section 4. Many of the models discussed are applied to data from a developmental toxicity study in Section 5.

2 Modelling Frameworks

Denote, for each individual, subject, or experimental unit $i = 1, \ldots, N$, assumed to be independent, a series of measurements by $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{n_i})'$, along with covariate information, usually grouped into a matrix X_i; wherever possible, the latter will be dropped from notation.

2.1 The Bahadur Model

We first present the general form of the Bahadur model, to then focus on the case of clustered, exchangeable data.

Bahadur (1961) introduced this model, with its elegant closed form. It does however have a number of computational problems surrounding it, which stem from the complicated and highly restrictive shape of its parameter space. The model is conceived for binary data.

Assume the marginal distribution of Y_{ij} to be Bernoulli with $E(Y_{ij}) = P(Y_{ij} = 1) \equiv \pi_{ij}$. Let the multinomial joint distribution $f(\boldsymbol{y})$ of $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})'$ with associated 2^{n_i} probability vector and let

$$\varepsilon_{ij} = \frac{Y_{ij} - \pi_{ij}}{\sqrt{\pi_{ij}(1-\pi_{ij})}} \quad \text{and} \quad e_{ij} = \frac{y_{ij} - \pi_{ij}}{\sqrt{\pi_{ij}(1-\pi_{ij})}}, \tag{1}$$

where y_{ij} is an actual value of the binary response variable Y_{ij}. Further, let $\rho_{ijk} = E(\varepsilon_{ij}\varepsilon_{ik})$, $\rho_{ijk\ell} = E(\varepsilon_{ij}\varepsilon_{ik}\varepsilon_{i\ell})$,... $\rho_{i12\ldots n_i} = E(\varepsilon_{i1}\varepsilon_{i2}\ldots\varepsilon_{in_i})$. Then, the general Bahadur model can be written as $f(\boldsymbol{y}_i) = f_1(\boldsymbol{y}_i) \cdot c(\boldsymbol{y}_i)$, where

$$f_1(\boldsymbol{y}_i) = \prod_{j=1}^{n_i} \pi_{ij}^{y_{ij}} (1-\pi_{ij})^{1-y_{ij}},$$

$$c(\boldsymbol{y}_i) = 1 + \sum_{j<k} \rho_{ijk} e_{ij} e_{ik} + \sum_{j<k<\ell} \rho_{ijk\ell} e_{ij} e_{ik} e_{i\ell} + \ldots + \rho_{i12\ldots n_i} e_{i1} e_{i2} \ldots e_{in_i}.$$

Consider the Bahadur model for the special case of exchangeable clustered data, as used by Aerts *et al* (2002) in an environmental context, that is for which the ordering of the measurements Y_{ij} within unit i is immaterial. This motivates motivating the choices: $\pi_{ij} = \pi_i$, $\rho_{ijk} = \rho_{i(2)}$, for $j < k$, $\rho_{ijk\ell} = \rho_{i(3)}$ for $j < k < \ell$, etc. We do not need the individual outcomes Y_{ij}; rather the summary statistic $Z_i = \sum_{j=1}^{n_i} Y_{ij}$, representing the number of successes within a unit, with realized value z_i, is sufficient. In this setting, the Bahadur model reduces to

$$f(\boldsymbol{y}_i) = \pi_i^{z_i}(1-\pi_i)^{n_i - z_i} \left[1 + \sum_{r=2}^{n_i} \rho_{i(r)} \sum_{s=0}^{r} \binom{z_i}{s} \binom{n_i - z_i}{r-s} (-1)^{s+r} \lambda_i^{r-2s} \right], \tag{2}$$

with $\lambda_i = \sqrt{\pi_i/(1-\pi_i)}$. In addition, setting all three- and higher-way correlations equal to zero, and rewriting the probability mass as a function of Z_i leads to:

$$f(z_i) \equiv f(z_i | \pi_i, \rho_{i(2)}, n_i) = \binom{n_i}{z_i} \pi_i^{z_i} (1-\pi_i)^{n_i - z_i}$$

$$\times \left[1 + \rho_{i(2)} \left\{ \binom{n_i - z_i}{2} \frac{\pi_i}{1-\pi_i} - z_i(n_i - z_i) + \binom{z_i}{2} \frac{1-\pi_i}{\pi_i} \right\} \right]. \quad (3)$$

We will refer to model (3) as $Bah(n_i, p = 2)$, the second-order Bahadur model for clusters of size n_i, higher-order versions $Bah(n, p)$ derived from (2). In Section 4, we will pay particular attention to $Bah(n, p = 2)$ and $Bah(n, p = 3)$. Note that, when $n = p$, the model is the fully specified, unrestricted model (2).

In spite of its simplicity, estimation with moderate to large cluster sizes is a more time-consuming endeavour than with the semi-parametrically specified GEE. More importantly for our purposes, the parameter space is highly constrained (Bahadur 1961, Kupper and Haseman 1978, Prentice 1988, Declerck, Aerts, and Molenberghs 1998, and Aerts *et al* 2002). The constraints are more severe with increasing n and, for given n, with decreasing p. For $Bah(n = 12, p = 2)$, for example, the pairwise correlation is bounded from above by a value in $[0.09; 0.18]$, the bound itself being a function of the marginal probability π. The severity of these constraints imply that, for large clusters, the low-order (exchangeable) Bahadur models, while simpler and seemingly closer to GEE, suffer from highly constrained correlation parameter spaces. This is troublesome because it makes it likely that for a given GEE a corresponding Bahadur model simply does not exist. This is a an observation crucial to our developments in the coming sections. We now introduce such a GEE.

2.2 Generalized Estimating Equations

A technique that enables a researcher to restrict modelling to the mean profile is based on so-called *generalized estimating equations* (GEEs, Liang and Zeger 1986, Diggle *et al* 2002, Molenberghs and Verbeke 2005). GEE's are conveniently introduced as the equations:

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{N} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}'} V_i^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_i) \;=\; \mathbf{0}, \quad (4)$$

where $V_i = A_i^{1/2} R_i A_i^{1/2}$, A_i is the matrix with the marginal variances on the main diagonal and zeros elsewhere, and $R_i = R_i(\boldsymbol{\alpha})$ is the marginal correlation matrix. Because the correlation requires a new parameter $\boldsymbol{\alpha}$, it is not assumed to be correctly modelled, but rather a working assumption is made.

Given that the marginal mean $\boldsymbol{\mu}_i$ has been correctly specified as $h(\boldsymbol{\mu}_i) = X_i \boldsymbol{\beta}$ and with mild regularity conditions holding, the estimator $\widehat{\boldsymbol{\beta}}$ obtained from

solving (4) is consistent and asymptotically normally distributed with mean $\boldsymbol{\beta}$ and asymptotic variance-covariance matrix

$$\begin{aligned} \mathrm{Var}(\widehat{\boldsymbol{\beta}}) &= I_0^{-1} I_1 I_0^{-1}, \qquad (5) \\ I_0 &= \sum_{i=1}^{N} \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} V_i^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}'}, \\ I_1 &= \sum_{i=1}^{N} \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} V_i^{-1} \mathrm{Var}(\boldsymbol{Y}_i) V_i^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}'}. \end{aligned}$$

Consistent estimates can be obtained by replacing all unknown quantities in (5) by consistent estimates. $\mathrm{Var}(\boldsymbol{Y}_i)$ in (5) is typically replaced by $(\boldsymbol{y}_i - \boldsymbol{\mu}_i)(\boldsymbol{y}_i - \boldsymbol{\mu}_i)'$. The working correlation parameters $\boldsymbol{\alpha}$ are based on the e_{ij} in (1). Independence, exchangeability, AR(1), and unstructured working correlation matrices are commonly used.

The need to specify third- and higher-order moments or, more precisely, third- and higher-order correlations, is thereby avoided, and two-way correlations can be be misspecified without affecting the consistency of the marginal parameter estimators. Should the two-way correlations be correctly specified, and should a set of appropriate third- and higher-order correlations be chosen, together with marginal logit links for binary outcomes, then the Bahadur model would follow. Thus, classical GEE can be viewed as a moment-based version of the Bahadur model. After choosing the marginal response functions, there is always at least one, trivial, Bahadur model that corresponds to the estimating equations, found by setting all correlations to zero. This sets independence apart as a special but trivial case. Thus, subtle compatibility issues arise only when at least one pairwise correlation is non-zero.

In general, the working correlations, found upon convergence of GEE, need not correspond to a valid compatible model *within the Bahadur family*, given the severe constraints on the Bahadur model parameters (Section 2.1). It remains to be shown that another family, specifically the hybrid model family (Section 2.4), will take on this role (Section 3).

There exist a number of variations to the above-sketched GEE theme. We will briefly mention three of these; they will be applied in the data analysis Section 5. A detailed overview of these and other GEE versions can be found in Molenberghs and Verbeke (2005). First, Prentice (1988) replaced the moment-based estimation for the working correlation parameters by a second set of estimating equations. Under the working assumption that both sets are independent, computational complexity is avoided and, again, the correlation model need not be correctly specified for the marginal regression parameters to be consistent and asymptotically normal. Second, one can set out by linearizing the link function, in the sense of Nelder and Wedderburn (1972),

to construct a working variate, so that iteratively reweighted least squares (IRLS) can be used (McCullagh and Nelder 1989). The third, while important in general, is particularly so in the current context, is due to Lipsitz, Laird, and Harrington (1991) who considered GEE for binary data with odds ratios rather than correlations. Essentially, the Bahadur-based correlation is replaced by

$$\psi_{ijk} = \frac{(\mu_{ijk})(1 - \mu_{ij} - \mu_{ik} + \mu_{ijk})}{(\mu_{ij} - \mu_{ijk})(\mu_{ik} - \mu_{ijk})}, \tag{6}$$

which leads to the following bivariate joint probabilities:

$$\mu_{ijk} = \begin{cases} \frac{1+[\mu_{ij}+\mu_{ik}](\psi_{ijk}-1)-S_{ijk}}{2(\psi_{ijk}-1)} & \text{if } \psi_{ijk} \neq 1, \\ \mu_{ij}\mu_{ik} & \text{otherwise,} \end{cases} \tag{7}$$

where

$$S_{ijk} = \sqrt{[1 + (\psi_{ijk} - 1)(\mu_{ij} + \mu_{ik})]^2 + 4\psi_{ijk}(1 - \psi_{ijk})\mu_{ij}\mu_{ik}}.$$

These are due to Plackett (1965) and form the basis of the bivariate (Dale 1986) and multivariate (Molenberghs and Lesaffre 1994) Dale models. Note that the relationship between correlation-GEE and the Bahadur model is mirrored by the relationship between the odds-ratio-GEE and the Dale model.

The odds-ratio formulation of GEE will play a key role when establishing our main result regarding the relationship between GEE and fully specified models. To this end, we will need an odds-ratio type model, but rather of a hybrid marginal-conditional nature (Section 2.4). For both the correlation-based and odds-ratio-based GEE, second-order versions have been formulated. These, too, will play a role in the developments to come (Section 2.5). Also, from the preceding it is clear that there is a conceptual link between GEE and the Bahadur model, when the outcomes are of a binary type and the dependence structure is captured via correlations. However, the Bahadur model is not unique in its family relationship with GEE. To discuss this issue, and to study its impact for the relationship between partially and fully specified models, it is helpful to introduce two further models.

2.3 The Beta-binomial Model

This model is conceived for clustered data, for which we operate in the setting, and with notation, of Section 2.1.

Assume, again, that $\boldsymbol{Y}_i$ is a n_i-dimensional vector of Bernoulli-distributed outcomes, with now cluster-specific success probability b_i. Assuming the elements in $\boldsymbol{Y}_i$ to be independent, conditionally on b_i, we have that the resulting conditional density of $\boldsymbol{Y}_i$ given b_i is binomial with n_i trials and success probability b_i. It is conveniently expressed in terms of $Z_i = \sum_j Y_{ij}$. The beta-binomial model (Skellam 1948, Kleinman 1973, Molenberghs and Verbeke 2005) assumes the parameters b_i to be sampled from a beta distribution with parameters α and β (which can depend on covariates, but this dependence is temporarily dropped from notation), i.e., the density of b_i equals

$$f(b_i|\alpha,\beta) = \frac{b_i^{\alpha-1}(1-b_i)^{\beta-1}}{B(\alpha,\beta)},$$

where $B(.,.)$ denotes the beta function. The mean and variance take the form:

$$\mu_i = \mathrm{E}(Z_i) \;=\; n_i\frac{\alpha}{\alpha+\beta}, \qquad \mathrm{Var}(Z_i) = n_i\pi(1-\pi)[1+(n_i-1)\rho], \tag{8}$$

and the density can be written as

$$f_i(z_i|\pi,\rho) = \binom{n_i}{\boldsymbol{z}_i} \frac{B[z_i+\pi(\rho^{-1}-1), n_i-z_i+(1-\pi)(\rho^{-1}-1)]}{B[\pi(\rho^{-1}-1),(1-\pi)(\rho^{-1}-1)]}, \tag{9}$$

in terms of the average proportion π of successes and the within-cluster correlation ρ. A purely moment-based view would merely require that the mean (8) and variance (8) be correctly specified, i.e., π must lie within the unit interval and $\rho \geq -1/(n-1)$.

2.4 A Hybrid Marginal-conditional Model

For each individual, subject, or experimental unit i in a study, a series of n categorical measurements Y_{ij}, grouped into a vector $\boldsymbol{Y}_i$ is recorded, together with covariate information $\boldsymbol{x}_i$. Assume that the parameters of primary interest are the first and second order marginal parameters. Extensions to more than two orders follow from similar logic.

Model building originates from the quadratic version of the joint distribution proposed by Cox (1972) and used by] Zhao and Prentice (1990) and Fitzmaurice and Laird (1993). Write

$$f(\boldsymbol{y}_i|\boldsymbol{\Psi}_i,\boldsymbol{\Omega}_i) = \exp\left\{\boldsymbol{\Psi}_i'\boldsymbol{v}_i + \boldsymbol{\Omega}_i'\boldsymbol{w}_i - A(\boldsymbol{\Psi}_i,\boldsymbol{\Omega}_i)\right\}, \tag{10}$$

with outcomes and pairwise cross-products thereof grouped into

$$\boldsymbol{v}_i = (\boldsymbol{y}_i^{'}; y_{i1}y_{i2}, \ldots, y_{i,n-1}y_{in})^{'}, \tag{11}$$

third- and higher-order cross-products collected in

$$\boldsymbol{w}_i = (y_{i1}y_{i2}y_{i3}, \ldots, y_{i1}y_{i2}\ldots y_{in})^{'}, \tag{12}$$

for $\boldsymbol{\Psi}_i$ and $\boldsymbol{\Omega}_i$ the corresponding canonical parameter vectors. Further, let $\boldsymbol{\mu}_i = E(\boldsymbol{V}_i)$ and $\boldsymbol{\nu}_i = E(\boldsymbol{W}_i)$. The distribution is fully parameterized by modelling $\boldsymbol{\Psi}_i$ and $\boldsymbol{\Omega}_i$. However, we choose to model $\boldsymbol{\mu}_i$ and $\boldsymbol{\Omega}_i$, enabling a direct representation of the marginal means and the pairwise marginal odds ratios. A model for $\boldsymbol{\mu}_i$ is specified via a vector of link functions $\boldsymbol{\eta}_i = \boldsymbol{\eta}_i(\boldsymbol{\mu}_i)$. From the covariate vector $\boldsymbol{x}_i$ a design matrix $\boldsymbol{X}_i$ is derived, such that $\boldsymbol{\eta}_i = \boldsymbol{X}_i\boldsymbol{\beta}$, with $\boldsymbol{\beta}$ a vector of parameters of interest.

Similarly, a model for the conditional higher order parameters needs to be constructed. As in Fitzmaurice and Laird (1993), and because the components of $\boldsymbol{\Omega}_i$ can be interpreted as conditional higher order log odds ratios, Molenberghs and Ritter (1996) assumed an identity link and specified the covariate dependence as $\boldsymbol{\Omega}_i = \boldsymbol{X}_i'\boldsymbol{\alpha}$, with $\boldsymbol{X}_i'$ another design matrix and $\boldsymbol{\alpha}$ a parameter vector. A simple model is found by setting the components of $\boldsymbol{\Omega}_i$ equal to constants.

Following derivations in Fitzmaurice and Laird (1993) and Fitzmaurice, Laird, and Rotnitzky (1993), the likelihood equations can be written as:

$$\frac{\partial \ell}{\partial(\boldsymbol{\beta}, \boldsymbol{\alpha})} = \sum_{i=1}^{N} \begin{pmatrix} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} & 0 \\ 0 & \frac{\partial \boldsymbol{\Omega}_i}{\partial \boldsymbol{\alpha}} \end{pmatrix}^{'} \begin{pmatrix} \boldsymbol{M}_i^{-1} & 0 \\ -\boldsymbol{N}_i\boldsymbol{M}_i^{-1} & I \end{pmatrix} \begin{pmatrix} \boldsymbol{v}_i - \boldsymbol{\mu}_i \\ \boldsymbol{w}_i - \boldsymbol{\nu}_i \end{pmatrix}, \tag{13}$$

with $\boldsymbol{M}_i = \mathrm{cov}(\boldsymbol{V}_i)$ and $\boldsymbol{N}_i = \mathrm{cov}(\boldsymbol{V}_i, \boldsymbol{W}_i)$. The form of the derivatives in the first matrix of (13) depends on the choice of link functions and linear predictors. Under the assumed linear model for $\boldsymbol{\Omega}_i$, the derivative reduces to $\boldsymbol{X}_i'$. Let

$$\left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}\right)^{'} = \boldsymbol{X}_i'(\boldsymbol{D}_i')^{-1}, \qquad \boldsymbol{D}_i = \left(\frac{\partial \boldsymbol{\eta}_i}{\partial \boldsymbol{\mu}_i}\right).$$

Given that the model is a mixed parameterization of an exponential family model (Barndorff-Nielsen 1978), the parameter vectors $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are orthogonal in the sense of Cox and Reid (1987). This implies that the maximum likelihood estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are asymptotically independent, with the inverse of the

expected information matrix:

$$\begin{pmatrix} \left[\sum_{i=1}^{N} \boldsymbol{X}_i'(\boldsymbol{D}_i')^{-1}\boldsymbol{M}_i^{-1}\boldsymbol{D}_i^{-1}\boldsymbol{X}_i\right]^{-1} & 0 \\ 0 & \left[\sum_{i=1}^{N}(\boldsymbol{X}_i')'(\boldsymbol{P}_i - \boldsymbol{N}_i\boldsymbol{M}_i^{-1}\boldsymbol{N}_i')\boldsymbol{X}_i'\right]^{-1} \end{pmatrix},$$

where $\boldsymbol{P}_i = \text{cov}(\boldsymbol{W}_i)$. It is this property that will play a key role in Section 3. Details on parameter estimation are reviewed in Molenberghs and Verbeke (2005).

2.5 Second-order Generalized Estimating Equations

Second-order GEE, referred to as GEE2, has been proposed by Zhao and Prentice (1990), using correlations, and by Liang, Zeger, and Qaqish (1992), using odds ratios. These explicitly model the pairwise association structure and require working assumptions for the third and fourth moments. Both are of a marginal type and therefore can be seen as lying in between GEE with correlations and the Bahadur model on the one hand, and GEE with odds ratios and the multivariate Dale model on the other hand.

However, a third, and for our purposes, very important set of GEE2, proposed also by Heagerty and Zeger (1996), can be derived directly from the hybrid model by specifying only its first and second moments:

$$\boldsymbol{U}(\boldsymbol{\beta}) = \sum_{i=1}^{N}\left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}\right)\boldsymbol{M}_i^{-1}(\boldsymbol{v}_i - \boldsymbol{\mu}_i) = \boldsymbol{0}. \tag{14}$$

Observe that these estimating equations assume the same form for any fixed value of $\boldsymbol{\Omega}_i$, with $\boldsymbol{\Omega}_i = \boldsymbol{0}$ as a special but important case. However, this leaves $\boldsymbol{M}_i$ partly unspecified. A standard procedure is to replace it by a working covariance matrix, depending on a set of (nuisance) parameters $\boldsymbol{\alpha}$. Heagerty and Zeger (1996) advocated setting the higher order conditional association parameters equal to zero (or, more generally, to a fixed constant). This particular set of GEE2 does not require estimation of extra parameters. It can also be seen as score equations for the likelihood specified by the following member of the quadratic exponential family of Zhao and Prentice (1990):

$$f(y_i|\boldsymbol{\Psi}_i) = \exp\left\{\boldsymbol{\Psi}_i'\boldsymbol{v}_i - A(\boldsymbol{\Psi}_i)\right\}. \tag{15}$$

Computing the covariance $\boldsymbol{M}_i$ in (14) involves the third- and fourth-order probabilities.

A technique, related to but slightly different from GEE, is *alternating logistic regressions* (Carey, Zeger, and Diggle 1993, Diggle *et al* 2002, Molenberghs and

Verbeke 2005). It is intended for binary data and is based on marginal logistic regressions combined with the odds ratio as a measure of association. The appeal is that it allows for formal inferences, not only about the marginal mean regressions, but also about the association parameters, as in GEE2, whereas the computational complexity does not exceed that of GEE1. The method has been implemented in standard statistical software. For example, along with GEE1, it can be fitted using the SAS procedure GLIMMIX. Important related models, to which our ideas also apply, can be found in Heagerty (2002) and Ilk and Daniels (2007).

With μ_{ijk} the pairwise success probabilities, as in (7), and $\alpha_{ijk} = \ln(\psi_{ijk})$ the marginal log odds ratio, the method is based on specifying marginal and conditional logistic regressions simultaneously:

$$\text{logit } \Pr(Y_{ij} = 1|\boldsymbol{x}_{ij}) = \boldsymbol{x}_{ij}\boldsymbol{\beta}, \tag{16}$$

$$\text{logit } \Pr(Y_{ij} = 1|Y_{ik} = y_{ik}) = \alpha_{ijk}y_{ik} + \ln\left(\frac{\mu_{ij} - \mu_{ijk}}{1 - \mu_{ij} - \mu_{ik} + \mu_{ijk}}\right), \tag{17}$$

for $j = 1, \ldots, n_i$ and $1 \leq k < j$. Whereas (16) is a conventional logistic regression, (17) has no intercept and the second term is treated as an offset. Carey, Zeger, and Diggle (1993) established compatibility of (16) and (17). Estimation is based on alternating between fitting the sets (16) and (17).

2.6 Categorical Outcomes

Upon noting that a categorical outcome variable with c modalities can be represented by $c-1$ (dependent) binary indicator variables, Molenberghs and Ritter (1996) and Molenberghs and Danielson (1999) showed how the hybrid model can easily be extended to both nominal as well as ordinal outcome vectors. The same is true for the marginal GEE and GEE2 models, with both correlations and odds ratios (Liang, Zeger, and Qaqish 1992, Molenberghs and Lesaffre 1999). In addition, the multivariate Dale model has been formulated for ordinal outcomes (Molenberghs and Lesaffre 1994), while the beta-binomial model extends by transforming itself into the Dirichlet-multinomial model (Aerts *et al* 2002). Only the Bahadur model has not been formulated for categorical outcomes, even though this is possible, in principle.

3 Every Valid Semi-parametric Model Corresponds to At Least One Full Model

A concern that might arise when formulating and fitting semi-parametric marginal models, in the sense of limiting modelling to lower order moments

only, is that the model may be incompatible with a fully specified model, in the sense that no model exists with the prescribed low-order moments. Such a compatible model will be termed here, for brevity, a *parent*.

There are cases where a parent trivially exists. For example, consider a semi-parametric model for binary outcomes with logistic regression models for the means and independence working assumptions. Then, the Bahadur model with the same logistic regressions and with all second- and higher-order correlations equal to zero is such a parent. Obviously, there is nothing particular here about the outcome being binary or the choice for the Bahadur model. It is evident that for independence working assumptions no problem arises. Note also that such a parent will not be unique. Indeed, given that zero correlations specify a valid joint distribution and provided that the joint probabilities are all different from 0 and 1, valid Bahadur distributions will arise in a sufficiently small open neighbourhood of the zero-correlation vector, showing that there are uncountably many parents. Now, the condition of non-boundary probabilities is satisfied as soon as all regression parameters and all covariate values are finite, since then the mean probabilities lie in the interior of the unit interval. The same then holds for the joint probabilities under independence, since these are merely products of mean probabilities and their complements.

In Section 4, we will present a semi-parametric model without a Bahadur parent that does have, at the same time, a hybrid marginal-conditional parent. This example and other developments notwithstanding, it is important to reiterate that there is no particular focus on either the Bahadur model or the use of correlation coefficients to quantify pairwise and higher-order association. In particular, these considerations apply to the commonly encountered marginal odds ratios as well (Molenberghs and Verbeke 2005).

The result presented here is for the case of a correctly specified semi-parametric model, in the sense that all lower-order moments, up to order $p < n$ are specified and valid. For example, a first-order model will have provided valid bivariate distributions for all possible pairs from the n outcomes. In such a case, $p = 2$. While the results hold for any p, we focus, without loss of generality, on the pairwise case.

Our assumption implies that all

$$f(y_{ij}, y_{ik}|\boldsymbol{\theta}), \qquad (1 \leq j < k \leq n), \tag{18}$$

where $\boldsymbol{\theta}$ is the parameter vector describing the lower-order moments, are valid. In case one starts from a value for $\boldsymbol{\theta}$ that results from fitting a model, it would be proper to write $\widehat{\boldsymbol{\theta}}$. However, this is immaterial for our purposes and we will merely assume $\boldsymbol{\theta}$ is any parameter corresponding to valid lower-order densities.

We claim that there always exists a valid $f(y_{i1}, \ldots, y_{in}|\widehat{\boldsymbol{\theta}}, \boldsymbol{\psi})$, where $\boldsymbol{\psi}$ is additionally needed to parameterize the n-way distribution. Note first that the set (18) by its very construction is always internally compatible, in the sense that for any choice of $j \neq k \neq \ell \neq j$, $f(y_{ij}, y_{ik}|\boldsymbol{\theta})$ and $f(y_{ij}, y_{i\ell}|\boldsymbol{\theta})$ produce the same univariate margin $f(y_{ij}|\boldsymbol{\theta}')$, for an appropriate sub-vector $\boldsymbol{\theta}'$ of $\boldsymbol{\theta}$.

We will show that one can, almost trivially, make use of the hybrid marginal-conditional model to establish existence of a parent. The essence of the argument is that the semi-parametrically specified marginal model can be completed, using conditional specifications for the higher orders.

To enable progress, we first need to remove a stumbling block. The hybrid model is specified in terms of logits and (log) odds ratios. This is sufficient in cases where the same holds for the semi-parametrically specified model such as, for example GEE1 or GEE2 with odds ratios. However, a problem arises when the Bahadur-related GEE1 or GEE2 with correlations is employed. Fortunately, one can rewrite the correlation-based bivariate probabilities in odds-ratio form. To see this, note that, given the univariate probabilities π_{ij} and π_{ik}, together with the correlation ρ_{ijk}, the bivariate probabilities easily follow. Then, a 2×2 contingency table can be constructed as:

$$\begin{array}{|c|c|}\hline \mu_{11} = \pi_{ij}\pi_{ik} + \vartheta & \mu_{12} = \pi_{ij}(1-\pi_{ik}) - \vartheta \\ \hline \mu_{21} = (1-\pi_{ij})\pi_{ik} - \vartheta & \mu_{22} = (1-\pi_{ij})(1-\pi_{ik}) + \vartheta \\ \hline \end{array}, \tag{19}$$

where $\vartheta = \rho_{ijk}\sqrt{\pi_{ij}(1-\pi_{ij})\pi_{ik}(1-\pi_{ik})}$. Switching to the odds ratio,

$$\psi_{ijk} = \frac{[\pi_{ij}\pi_{ik} + \vartheta] \cdot [(1-\pi_{ij})(1-\pi_{ik}) + \vartheta]}{[\pi_{ij}(1-\pi_{ik}) - \vartheta] \cdot [(1-\pi_{ij})\pi_{ik} - \vartheta]}, \tag{20}$$

while the univariate probabilities remain intact.

A few remarks are appropriate here. First, one has to establish validity of this parameter combination. However, it has been shown by several authors that in the bivariate case every triple $(\pi_{ij}, \pi_{ik}, \psi_{ijk})$ with $0 \leq \pi_{ij}, \pi_{ij} \leq 1$ and $0 \leq \psi_{ijk}$, specifies valid second-order probabilities (Fréchet 1951, Dale 1986, Palmgren 1989). But even beyond that, given that a valid two-by-two table is specified, the marginal probabilities and odds ratio derived from it form a valid combination. Second, the above procedure applies, when covariates are recorded along with the outcomes, separately for every covariate level. The rest of the derivation is then conducted separately for every covariate level. Third, while the transformation towards logits and odds ratios is specific for Bahadur-type semi-parametric models, such as GEE with correlations, it is important to realize that the construction is general. For example, a semi-parametrically

specified model in terms of kappa (agreement) parameters could be used as input, too. This statement is clearly valid, upon noting that all that is required is the input μ_{11}, μ_{12}, μ_{21}, and μ_{22}, as in (19), to apply the transformation to an odds-ratio representation.

Given this lower-order (second-order, say) specification, a hybrid model is then easily constructed. The validity follows from the orthogonality, derived by Barndorff-Nielsen (1978) for this type of hybrid model placed within the exponential family, where the lower-order, marginal (mean) parameters are dual to the corresponding conditional parameters. Thus, in principle, one can choose an arbitrary set of higher-order parameters. However, in the interest of simplicity and elegance, and since we merely aim to illustrate that a parent exists, it is convenient to complete the model in a conditionally specified GEE2 way, specified by the equations (14). The corresponding parent then reduces to (15).

We will now discuss four important extensions, thereby broadening the scope of our result.

First, the above result applies not only to a pairwise but also to a higher-order semi-parametric specification. Assume that the moments up to order p have been specified and derived using GEE or any other semi-parametric method. This produces a contingency table of order p, for which the univariate marginal probabilities and the pairwise and up to the pth order marginal odds ratios are specified, as in the multivariate Dale model (Molenberghs and Lesaffre 1994). The moments of orders $p+1$ to n are then conditionally specified. This comes down to adopting model (10), with (11) now containing outcomes and cross-products of orders 1 to p, and (12) containing cross-products of orders $p+1$ to n. It is again a convenient choice to set the $\mathbf{\Omega}_i$ parameters equal to zero, producing a p-dimensional analog of (15), i.e., a GEEp with conditional higher-order assumptions.

Second, it has been assumed implicitly that the semi-parametric specification is purely marginal in nature. This need not be a limitation because, in the case that some or all of the lower-order moments are conditionally specified, the transformation is needed only for the marginal part. This is executed by placing the outcomes and cross-products corresponding to the marginally specified part in (11) and relegating the remaining ones to (12).

Third, the results also apply to categorical data, using the argument of Section 2.6.

Fourth, and building on the previous point, the result is applicable to any exponential family setting. To see this note that, even though in Section 2.4 the categorical case has been emphasized, the general results of Fitzmaurice and Laird (1993) and Molenberghs and Ritter (1996) makes use only of (10),

the split of the outcome and cross-product vector into the sub-vectors (11) and (12), and the orthogonality between the dual sets of (conditional) canonical and (marginal) mean parameters (Barndorff-Nielsen 1978). For example, the likelihood equations (13) are generic for the entire multivariate exponential family.

In the light of these considerations, it follows that the result presented here applies to encompassing classes such as models for binary, nominal, ordinal, and count data. In addition, outcome vectors that amalgamate responses of various types can be placed under the umbrella of this result.

4 Examples and Counterexamples in the Clustered Data Case

Table 1 contains a number of illustrations for the clustered data case, with cluster size $n = 3$. It is assumed that the first and second moments are specified through the mean π, common to the cluster and the second-order correlation $\rho_{(2)}$, common to all pairs within a cluster. Such a specification corresponds, for example, to GEE with exchangeable (working) correlation structure. The given method is valid if and only if all pairwise distributions that can be formed from the cluster are valid. This is the same as requiring the corresponding Bahadur model, $Bah(2, 2)$ to be valid. Note that $Bah(2, 2)$ is not a fully specified model in this case, because clusters are of size 3 but the model specifies the three pairs of outcomes only. For each parameter combination, the full models $Bah(3, 2)$ and $Bah(3, 3)$ are considered, only the latter of which is unconstrained in the sense that the additional parameters are allowed to be free. Of course, in this case with $n = 3$, there is only one additional parameter, $\rho_{(3)}$, the three-way correlation coefficient. Furthermore, the marginal beta-binomial model is considered, as well as the hybrid model.

Given our general result, the hybrid model is always valid, whenever the specification for π and $\rho_{(2)}$ is valid. In other words, the validity of the hybrid model coincides with the validity of the original GEE. Because the validity of GEE derives, by construction, from the existence of the $Bah(2, 2)$ model, effectively these three situations coincide. All other models considered exhibit restrictions. We discuss the Bahadur and beta-binominal restrictions in turn.

Declerck, Aerts, and Molenberghs (1998) and Aerts *et al* (2002) present correlation bounds on $\rho_{(2)}$ for, among other models, the $Bah(n, p = 2)$ and $Bah(n, p = 3)$ distributions. For the case of $p = 2$, Bahadur (1961) provided the following bounds:

$$-\frac{2}{n(n-1)} \min\left(\frac{\pi}{1-\pi}, \frac{1-\pi}{\pi}\right) \leq \rho_{(2)} \leq \frac{2\pi(1-\pi)}{(n-1)\pi(1-\pi) + 0.25 - \gamma_0}, \quad (21)$$

Table 1
Examples and counterexamples of the existence of valid models, within a particular family, when the pairwise model have been pre-specified. A clustered-data example is considered, implying that there is a common univariate probability π and a common pairwise correlation $\rho_{(2)}$. Bah(n,p) stands for the Bahadur model for a cluster of size n, where correlations strictly higher than p are set equal to zero. $BB(n)$ is the beta-binomial model for clusters of size n. Data cluster size is $n = 3$. The cases of $\pi = 0.5$ and $\pi = 0.1$ are considered. ($\surd$: a valid model exists; $\times$: no valid model exists.)

Valid pairwise-correlation ranges per model		
Model	$\pi = 0.5$	$\pi = 0.1$
GEE/$Bah(2,2)$; Hybrid	$[-1; 1]$	$[-1/9; 1]$
$Bah(3,2)$	$[-1/3; 1]$	$[-1/27; 18/34]$
$Bah(3,3)$	$[-1/3; 1]$	$[-1/10; 7/10]$
$BB(2)$	$[-1; 1]$	$[-1; 1]$
$BB(3)$	$[-1/2; 1]$	$[-1/2; 1]$

Validity of models by pairwise-correlation interval							
			Hybrid				
Sit.	π	$\rho_{(2)}$	GEE/$Bah(2,2)$	$Bah(3,2)$	$Bah(3,3)$	$BB(2)$	$BB(3)$
1	0.5	$[-1.000; -0.500]$	$\surd$	$\times$	$\times$	$\surd$	$\times$
2	0.5	$[-0.500; -0.333]$	$\surd$	$\times$	$\times$	$\surd$	$\surd$
3	0.5	$[-0.333;\ \ 1.000]$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$
4	0.1	$[-1.000; -0.500]$	$\times$	$\times$	$\times$	$\surd$	$\times$
5	0.1	$[-0.500; -0.111]$	$\times$	$\times$	$\times$	$\surd$	$\surd$
6	0.1	$[-0.111; -0.100]$	$\surd$	$\times$	$\times$	$\surd$	$\surd$
7	0.1	$[-0.100; -0.037]$	$\surd$	$\times$	$\surd$	$\surd$	$\surd$
8	0.1	$[-0.037;\ \ 0.529]$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$
9	0.1	$[\ \ 0.529;\ \ 0.700]$	$\surd$	$\times$	$\surd$	$\surd$	$\surd$
10	0.1	$[\ \ 0.700;\ \ 1.000]$	$\surd$	$\times$	$\times$	$\surd$	$\surd$

where

$$\gamma_0 = \min_{z=0}^{n} \left\{ [z - (n-1)\pi - 0.5]^2 \right\}.$$

Applying these bounds for $n = 2$ and $n = 3$, respectively, provides the valid ranges for $Bah(2,2)$ and $Bah(3,2)$, respectively, and, therefore, also for GEE and the hybrid model. Declerck, Aerts, and Molenberghs (1998) derived the corresponding bounds for the three-way Bahadur model, $Bah(n, p = 3)$. These

are more involved than (21), and some additional notation is required. Let the coefficient of $\rho_{i(r)}$ in expression (2) be denoted by $g_r(\lambda, n, z)$. Hence, the three-way Bahadur model under exchangeability can be written as

$$f(\mathbf{y}) = \pi^z(1-\pi)^{n-z}\left[1 + \sum_{r=2}^{3} \rho_{(r)} g_r(\lambda, n, z)\right].$$

Let the values for n and λ (or equivalently π) be arbitrary but fixed and drop them from notation. Hence, $g_r(\lambda, n, z)$ is abbreviated as $g_r(z)$. Let $\boldsymbol{z}_P$, $\boldsymbol{z}_Z$, and $\boldsymbol{z}_N$ be the vectors containing the values of z for which $g_3(z)$ is positive, zero and negative, respectively. Denote a general element of $\boldsymbol{z}_P$, $\boldsymbol{z}_Z$, and $\boldsymbol{z}_N$ by z_P, z_Z and z_N, respectively. Further, let $\Delta(z_P, z_N) = g_3(z_N) - g_3(z_P)$ and $\tau(z_P, z_N) = g_2(z_N)g_3(z_P) - g_2(z_P)g_3(z_N)$. Straightforward but tedious algebra then produces the following bounds:

$$\max\left[\max_{(z_P,z_N):\tau>0}\left(\frac{\Delta}{\tau}\right), \max_{z_P:g_2>0}\left(-\frac{1+g_3}{g_2}\right),\right. \tag{22}$$

$$\left.\max_{z_N:g_2>0}\left(-\frac{1-g_3}{g_2}\right), \max_{z_Z:g_2>0}\left(-\frac{1}{g_2}\right), -1\right] \tag{23}$$

$$\leq \rho_{(2)} \leq \tag{24}$$

$$\min\left[\min_{(z_P,z_N):\tau<0}\left(\frac{\Delta}{\tau}\right), \min_{z_P:g_2<0}\left(-\frac{1+g_3}{g_2}\right),\right. \tag{25}$$

$$\left.\min_{z_N:g_2<0}\left(-\frac{1-g_3}{g_2}\right), \min_{z_Z:g_2<0}\left(-\frac{1}{g_2}\right), 1\right]. \tag{26}$$

In contrast, the bounds for the beta-binomial model are much simpler. First, when a hierarchical view is taken, i.e., preserving the random-effects interpretation of the model, and hence the connection with the beta distribution, then it is clear from density (9) that the correlation must be non-negative, exactly as one would expect from this and any other model with a random-effects interpretation and the correlation parameter governing the distribution of the random effects. However, when a marginal, moment-based view is taken, it is sufficient for the mean and variance to be valid. It then follows immediately from (8) that the model is valid if and only if ρ belongs to $[-1/(n-1); 1]$. When there are more than two cluster members this is, again, a non-fully specified model, relaxing restrictions on the parameter space, but preventing in some cases the existence of a fully specified model *of the beta-binomial type*.

The hybrid model can also be used here to specify a valid model. The use of (26) still is straightforward, even though a bit more involved than the two-way counterpart (21).

Note that there are two crucial differences between $\pi = 0.5$ and other values,

such as $\pi = 0.1$. First, none of the fully and partially specified models have an upper bound. Second, the pairwise models, including therefore GEE and $BB(2)$, have no lower bound neither. Thus, when cluster sizes are $n = 2$, or when pairwise models are fitted to clusters of size $n = 3$, the correlation is not restricted. By implication, in the interval $[-1, -1/2]$, examples can be found of models for which either $Bah(3, 2)$ and $Bah(3, 3)$ are non-existent, or $BB(3)$ combined with these two does not provide a valid parameter combination. In the rest of the correlation interval, all models are valid.

For $\pi = 0.1$, the allowable interval is different for every model and the Bahadur models for clusters of size $n = 3$ also exclude values close to 1. This is the more generic situation, with bounds becoming tighter as n increases. Thus, outside the interval $[-1/10; 7/10]$, there is no valid fully specified model for clusters of size $n = 3$ in this case, even though GEE spans $[-0.111; 1.000]$. While the three-way beta-binomial model enjoys a broader range, it still is not possible to formulate one with correlations below -0.5. Recall also that, when a beta-binomial model with hierarchical interpretation is envisaged, then no negative correlations are allowable. Of course, given our general result, it is possible to formulate a hybrid model whenever a GEE is correctly formulated.

5 Analysis of Developmental Toxicity Data

This developmental toxicity study investigates the dose-response relationship in mice of the potentially hazardous chemical compound DEHP, standing for di(2-ethylhexyl)phthalate, used in vacuum pumps (Windholz 1983) and as plasticizers for numerous plastic devices made of polyvinyl chloride. DEHP provides the finished plastic products with desirable flexibility and clarity (Shiota, Chou, and Nishimura 1980). It has been well documented that small quantities of phthalic acid esters, of which DEHP is an instance, may leak out of polyvinyl chloride plastic containers in the presence of food, milk, blood, or various solvents. Due to their ubiquitous distribution and presence in human and animal tissues, considerable concern has developed as to the possible toxic effects of the phthalic acid esters (Autian 1973). The developmental toxicity study, conducted in timed-pregnant mice during the period of major organogenesis and described by Tyl *et al* (1988), has attracted much interest in the toxicity of DEHP. The doses selected for the study were 0, 44, 91, 191, and 292 mg/kg/day, respectively. For analysis, they are rescaled to range in the unit interval, i.e., 0, 0.125, 0.25, 0.5, and 1. The dams were sacrificed, just before normal delivery, and the status of uterine implantation sites recorded. A total of 1082 live foetuses were dissected from the uterus, anesthetized, and examined for external, visceral, and skeletal malformations. Foetuses are clustered within mothers; hence the implied association needs to be accommodated in the analysis.

The univariate marginal dose-response takes the form

$$Y_{ij} \sim \text{Bernoulli}(\pi_{ij}), \qquad \text{logit}(\pi_{ij}) = \beta_0 + \beta_d\, d_i, \tag{27}$$

where d_i ranges through the rescaled dose levels. The outcome is coded as $Y_{ij} = 1$ if any malformation occurs and 0 when the animal is malformation-free.

To illustrate our results, we fit sixteen different models to these data. Parameter estimates and standard errors are presented in Table 2. All marginal and hybrid models share the univariate marginal regression function (27). For the conditional models, the corresponding first-order logit takes the form:

$$\text{logit}[P(Y_{ij} = 1 | d_i, Y_{ik} = 0, k \neq j)] = \beta_0 + \beta_d\, d_i.$$

Here, conditioning is on all other fetuses in the same litter. The association parameter is the conditional odds ratio, ψ_c, for a pair of outcomes Y_{ij} and Y_{ik}, given all others are failures.

Among the marginal models, there are two full-likelihood models, fitted with maximum likelihood: the Bahadur (Model 1) and beta-binomial models (Model 9). For GEE 1 with correlations (Models 3–8), three estimation methods are considered, as discussed in Section 2.2: the original approach of Liang and Zeger(1986), Prentice's (1988) modification, and the quasi-likelihood-based linearization method. Moreover, both exchangeable as well as independence working assumptions are considered. For all other GEE approaches, exchangeable working assumptions (for GEE1), or simply an exchangeable association structure (for GEE2), is assumed. A further GEE1 is reported as Model 13, where now the odds ratio is used (Section 2.2), as proposed by Lipsitz, Laird, and Harrington (1991). Switching to GEE2 (Section 2.5), there are three instances. First, Model 2 corresponds to GEE2 with correlations, in the spirit of Zhao and Prentice (1988). Second, Model 12 features odds ratios, as proposed by Liang, Zeger, and Qaqish (1992). Third, the hybrid Model 14 can be considered both GEE2 with marginal exchangeable second-order structure and conditional independence higher-order assumptions, as well as full likelihood. Next to conventional GEE2, also alternating logistic regressions, described in Section 2.5 as well, are presented as Model 10.

A further non-likelihood method is pseudo-likelihood, also termed composite likelihood or pairwise likelihood for this case (Model 11; Molenberghs and Verbeke 2005, Ch. 9). Essentially, the likelihood function for a sequence of n measurements is replaced by the product of pairwise contributions for all possible $n(n-1)/2$ pairs. These pairwise contributions are written as the Dale probabilities (7). As with GEE, consistency and asymptotic normality follows, and an information sandwich is used for precision estimation.

Table 2
Developmental Toxicity Data. Any malformation in the DEHP study. Parameter estimates (standard errors) from analyses based on marginal models, conditional models, and the hybrid model β_0 and β_d are the intercept and dose effect, respectively; the association parameter is either the correlation, ρ, or the odds ratio, ψ. (exch: exchangeable working correlation; ind: independence working correlation.)

Number	Model	Estimation	β_0	β_d	Association	
Marginal models						
1	Bahadur	ML	-3.83(0.27)	5.38(0.47)	ρ	0.06(0.01)
2	GEE2(exch)	ZP	-5.23(0.40)	5.35(0.60)	ρ	0.09(0.01)
3	GEE1(exch)	LZ	-4.05(0.31)	5.84(0.61)	ρ	0.11
4	GEE1(ind)	LZ	-3.98(0.30)	5.56(0.61)		
5	GEE1(exch)	Prentice	-4.06(0.31)	5.89(0.61)	ρ	0.15(0.05)
6	GEE1(ind)	Prentice	-3.98(0.30)	5.56(0.61)		
7	GEE1(exch)	Lineariz.	-4.04(0.31)	5.82(0.61)	ρ	0.11
8	GEE1(ind)	Lineariz.	-3.98(0.30)	5.56(0.61)		
9	Beta-binomial	ML	-3.83(0.31)	5.59(0.56)	ρ	0.16(0.05)
10	ALR	CZD	-3.8(0.31)	5.59(0.56)	ψ	3.22(0.93)
11	Pairwise	PL	-3.98(0.30)	5.57(0.61)	ψ	3.00(0.81)
12	GEE2(exch)	LZQ	-3.69(0.25)	5.06(0.51)	ψ	2.64(0.61)
13	GEE1(exch)	LLH	-4.02(0.31)	5.79(0.62)	ψ	1.51(0.51)
Hybrid models						
14	GEE2(exch)/full	MR	-3.69(0.25)	5.06(0.51)	ψ	2.64(0.61)
Conditional models						
15	Quadr. loglin.	ML	-2.04(0.42)	2.98(0.66)	ψ_c	1.17(0.04)
16	Quadr. loglin.	PL	-1.80(0.35)	2.95(0.56)	ψ_c	1.22(0.04)

(Estimation: CZD: Carey, Zeger, and Diggle; LLH: Lipsitz, Laird, and Harrington; LZQ: Liang, Zeger, and Qaqish; ML: maximum likelihood; MR: Molenberghs and Ritter; PL: pseudo-likelihood; ZP: Zhao and Prentice;)

As a useful basis for comparison, the basic exponential family model (10) is considered, with the elements of $\boldsymbol{w}_i$ removed, i.e., the quadratic approximation, but without the transformation from the natural parameters to the marginal mean parameters. This means not the hybrid model results but rather a model with fully conditionally interpreted parameters. The model is fitted using either maximum likelihood (Model 15) or pseudo-likelihood (Model 16). The latter pseudo-likelihood, described in Molenberghs and Verbeke (2005, Ch. 12), is similar in spirit, yet different from, the one employed in Model 11. Here, the likelihood contribution for a subject is replaced by the product of n full conditionals $f(y_{ij}|\{y_{ik}, k \neq j\})$.

The regression parameters β_0 and β_d are very similar among all marginal models. Note that, while the beta-binomial model is generated hierarchically, its mean and variance have a marginal interpretation. Furthermore, the hybrid model is marginal in the mean and pairwise association parameters; thus, what is reported in Table 2 for the hybrid has got a marginal interpretation. Only Models 15 and 16 features regression and association parameters with a conditional interpretation. The regression parameters, therefore, do not describe the marginal odds but rather the odds of a success, conditional on all other outcomes within a cluster being failures. Evidently, this odds is lower than the corresponding marginal odds. For a detailed discussion, see Molenberghs and Verbeke (2005, Part III).

Turning to the models with correlation parameters, note that the correlation is smallest for the Bahadur model, followed first by GEE2, then by GEE1, with finally the beta-binomial model at the other end. This is entirely natural in view of the constraints operating. Indeed, (1) the Bahadur model is valid only if all higher-order probabilities are valid; (2) GEE2 is based on validity of probabilities in the orders 1 to 4; (3) for GEE1, only the first and second orders are required; (4) the beta-binomial model has no constraints in the range of positive correlations.

The marginal odds ratios, resulting from Models 10–14 are all similar, especially when compared against the background of the precision with which they are estimated. The noticeable exception is Model 13. This is the only GEE1-based model, with the others either of a GEE2 (Models 12 and 14) or of a related (Models 10 and 11) nature. Since the null hypothesis of no association corresponds to $\psi = 1$, the odds ratio is non-significant in Model 13, while it is significant in the other odds ratio model. Given the odds ratio is considerably less constrained in general and even unconstrained for a pair of outcomes (Palmgren 1989, Molenberghs and Lesaffre 1994), constraints are less of a plausible explanation than would be inefficiency associated with GEE1.

Finally, GEE2 Models 12 and 14 are similar but slightly different, in the sense that the higher-order structure is set to zero in a marginal way in Model 12

and in a conditional way in the hybrid Model 14. In terms of the first and second order marginal parameters they coincide. These considerations lead to the conclusion that the parameter and precision estimates are equal, up to two decimal places.

6 Concluding Remarks

In this paper, we have shown that a broad class of semi-parametrically specified models always admits a fully-specified parent model. The result is shown by construction, using a mixed marginally and conditionally parameterized multivariate exponential family model. First, the result is valid for a wide class of semi-parametric models where specification is done in terms of (parts of) the exponential-family formulation, including binary, nominal, ordinal, and Poisson outcomes. Second, it is also valid when the outcome vector combines outcomes of different types. Third, using transformation results, the result can be applied as well when the semi-parametric specification is not directly in terms of the exponential family, such as logistic regressions for binary data coupled with pairwise correlation, as in classical generalized estimating equations.

While the hybrid model provides an elegant vehicle to derive the existence of a fully-specified parent model, it may not be the most obvious choice *a priori*. For example, for the aforementioned correlation-based GEE, the Bahadur model may come to mind first as a candidate parent. However, this family is not always able to provide a parent. This underscores the use of the result derived. Examples and counterexamples have been given. The appeal of the existence of a valid parent, *even when not explicitly constructing it*, is that the type of semi-parametric modelling considered here can always be seen as describing a portion of a joint distribution. If this would be the case, it would have been impossible to provide an entirely natural description as to what framework GEE-based parameters, for example, fit in. The implication is that such semi-parametric methods as GEE1, GEE2, ALR, etc. can always be applied because there always is at least one valid parent. Thus, for every application of such semi-parametric models, there always is a probabilistic basis. The sole condition is that the parametrically specified portion of the model be valid, but this is not different from any other statistical modelling exercise.

Clearly, our results do not imply uniqueness of a parent. In fact, the applications have demonstrated that there can be more than one. As long as one is interested in a partially specified model only, i.e., so long as one takes a semi-parametric view, this is of neither use nor concern. Should one be interested in modelling more moments, or even all moments so as to move towards full

likelihood, then a parent can be chosen in accordance with statistical criteria (e.g., quality of model fit) and/or scientific criteria (e.g., which additional parameters are best suited to provide insight in the problem at hand).

The suite of models used in this paper to derive and illustrate our main result, jointly with a few additional models, have been fitted to clustered data from a developmental toxicity study conducted in mice. As such, the impact of the constraints operating on the association parameters, in particular when it takes the form of a correlation, is clearly demonstrated.

Acknowledgment

The authors gratefully acknowledge support from IAP research Network P6/03 of the Belgian Government (Belgian Science Policy).

References

Aerts, M., Geys, H., Molenberghs, G., and Ryan, L.M. (2002). *Topics in Modeling of Clustered Data.* London: Chapman & Hall/CRC.

Autian, J. (1973). Toxicity and health threats of phthalate esters: Review of the literature. *Environmental Health Perspectives*, **4**, 3–26.

Bahadur, R.R. (1961). A representation of the joint distribution of responses to n dichotomous items. In: *Studies in Item Analysis and Prediction,*, H. Solomon (Ed.). Stanford Mathematical Studies in the Social Sciences VI. Stanford, CA: Stanford University Press.

Barndorff-Nielsen, O.E. (1978). *Information and Exponential Families in Statistical Theory.* Chichester: John Wiley.

Carey, V.C., Zeger, S.L., and Diggle, P.J. (1993). Modelling multivariate binary data with alternating logistic regressions. *Biometrika* **80**, 517–526.

Cox, D.R. (1972). The analysis of multivariate binary data. *Applied Statistics* **21**, 113–120.

Cox, D.R., and Reid, N. (1987). On the stability of maximum-likelihood estimators of orthogonal parameters. *Canadian Journal of Statistics* **49**, 1–39.

Dale, J. R. (1986). Global cross-ratio models for bivariate, discrete, ordered responses. *Biometrics* **42**, 909–917.

Declerck, L., Aerts, M., and Molenberghs, G. (1998). Behaviour of the likelihood ratio test statistic under a Bahadur model for exchangeable binary data. *Journal of Statistical Computation and Simulation* **61**, 15–38.

Diggle, P.J., Heagerty, P.J., Liang, K.-Y., and Zeger, S.L. (2002). *Analysis of Longitudinal Data (2nd ed.).* Oxford Science Publications. Oxford: Clarendon Press.

Fitzmaurice, G.M. and Laird, N.M. (1993). A likelihood-based method for analysing longitudinal binary responses. *Biometrika* 80, 141–151.

Fitzmaurice, G.M., Laird, N.M., and Rotnitzky A. (1993). Regression models for discrete longitudinal responses. *Statistical Science*, **8**, 284–309.

Fréchet, M. (1951). Sur les tableaux de corrélation dont les marges sont données. *Annals Université Lyon, Section A, Series 3* **14**, 53–77.

Glonek, G.F.V (1996). A class of regression models for multivariate categorical responses. *Biometrika* **83**, 15–28.

Glonek, G.F.V. and McCullagh, P. (1995). Multivariate logistic models. *Journal of the Royal Statistical Society, Series B* **81**, 477–482.

Heagerty, P.J. (2002). Marginalized transition models and likelihood inference for longitudinal categorical data. *Biometrics*, **58**, 342–351.

Heagerty, P.J. and Zeger, S.L. (1996). Marginal regression models for clustered ordinal measurements. *Journal of the American Statistical Association* **91**, 1024–1036.

Ilk O. and Daniels M.J. (2007). Marginalized transition random effects models for multivariate longitudinal binary data. *The Canadian Journal of Statistics*, **35**, 105–123.

Kleinman, J. (1973). Proportions with extraneous variance: single and independent samples. *Journal of the American Statistical Association* **68**, 46–54.

Kupper, L.L. and Haseman, J.K. (1978). The use of a correlated binomial model for the analysis of certain toxicology experiments. *Biometrics* **34**, 69–76.

Lang, J.B. and Agresti, A. (1994). Simultaneously modelling joint and marginal distributions of multivariate categorical responses. *Journal of the American Statistical Association* **89**, 625–632.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

Liang, K.-Y., Zeger, S. L., and Qaqish, B. (1992). Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society, Series B* **54**, 3–40.

Lipsitz, S.R., Laird, N.M., and Harrington, D. P. (1991). Generalized estimating equations for correlated binary data: using the odds ratio as a measure of association. *Biometrika* **78**, 153-160.

Little, R.J.A. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data (2nd ed.)*. New York: John Wiley.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*. London: Chapman & Hall/CRC.

Molenberghs, G. and Danielson, L. (1999). Simple methods for the analysis of multivariate and longitudinal categorical data. In: *Proceedings of the 7th International Conference on Probability Theory and Mathematical Statistics and Vilnius Conference (1998)*, B. Grigelionis *et al* (Eds.). TEV, Vilnius/VSP, Utrecht, pp. 499–514.

Molenberghs, G. and Kenward, M.G. (2007). *Missing Data in Clinical Stud-*

ies. Chichester: John Wiley.

Molenberghs, G. and Lesaffre, E. (1994). Marginal modelling of correlated ordinal data using a multivariate Plackett distribution. *Journal of the American Statistical Association* **89**, 633–644.

Molenberghs, G. and Lesaffre, E. (1999). Marginal modelling of multivariate categorical data. *Statistics in Medicine* 18, 2237–2255.

Molenberghs, G. and Ritter, L. (1996). Likelihood and quasi-likelihood based methods for analysing multivariate categorical data, with the association between outcomes of interest. *Biometrics* **52**, 1121–1133.

Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data.* New York: Springer.

Nelder, J.A. and Wedderburn, R.W.M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series B* 135, 370–384.

Palmgren, J. (1989). *Regression Models for Bivariate Responses.* Technical Report 101, Department of Biostatistics, Seattle, WA.

Plackett, R.L. (1965). A class of bivariate distributions. *Journal of the American Statistical Association* **60**, 516–522.

Prentice, R.L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics* **44**, 1033–1048.

Shiota, K., Chou, M.J., and Nishimura, H. (1980). Embryotoxic effects of di-2-ethylhexyl phthalate (DEHP) and di-*n*-butyl phthalate (DBP) in mice. *Environmental Research* **22**, 245–253.

Skellam, J.G. (1948). A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials. *Journal of the Royal Statistical Society, Series B* **10**, 257–261.

Tyl, R.W., Price, C.J., Marr, M.C., and Kimmel, C.A. (1988). Developmental toxicity evaluation of dietary di(2-ethylhexyl)phthalate in Fischer 344 rats and CD-1 mice. *Fundamental and Applied Toxicology* **10**, 395–412.

Windholz, M. (1983). *The Merck Index: An Encyclopedia of Chemicals, Drugs, and Biologicals (10th ed.)* Rahway, NJ: Merck and Co.

Zeger, S.L., Liang, K.-Y., and Albert, P.S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics* **44**, 1049–1060.

Zhao, L.P. and Prentice, R.L. (1990). Correlated binary regression using a quadratic exponential model. *Biometrika* **77**, 642–648.