# Doubly robust pseudo-likelihood for incomplete hierarchical binary data

Peer-reviewed author version

HERMANS, Lisa; IVANOVA, Anna; SOTTO, Cristina; MOLENBERGHS, Geert; VERBEKE, Geert & Kenward, Michael G. (2020) Doubly robust pseudo-likelihood for incomplete hierarchical binary data. In: Statistical modelling, 20 (1) , p. 42 -57.

DOI: 10.1177/1471082X18808611
Handle: http://hdl.handle.net/1942/34576

# Doubly Robust Pseudo-likelihood for Incomplete Hierarchical Binary Data


**Lisa Hermans [1], Anna Ivanova [2], Cristina Sotto [3], Geert Molenberghs [1, 2], Geert Verbeke [2, 1] and Michael G. Kenward [4]**

[1] I-BioStat, Universiteit Hasselt, B-3590 Diepenbeek, Belgium

[2] I-BioStat, KU Leuven, B-3000 Leuven, Belgium

[3] Janssen Pharmaceutica; B-2340 Beerse, Belgium

[4] Ashkirk, United Kingdom

---

**Address for correspondence:** Lisa Hermans, The Interuniversity Institute for Biostatistics and statistical Bioinformatics, Universiteit Hasselt, Agoralaan gebouw D, 3590 Diepenbeek, Belgium.

**E-mail:** `lisa.hermans@uhasselt.be`.

**Phone:** (+32) 11 268 268.

**Fax:** (+1) 999 888 666.


---


**Abstract:** Missing data is almost inevitable in correlated-data studies. For non-Gaussian outcomes with moderate to large sequences, direct-likelihood methods can involve complex, hard-to-manipulate likelihoods. Popular alternative approaches, like generalized estimating equations, that are frequently used to circumvent the computational complexity of full likelihood, are less suitable when scientific interest, at least in part, is placed on the association structure; pseudo-likelihood methods are then a viable alternative. When the missing data are missing at random, Molenberghs et al. (2011) proposed a suite of corrections to the standard form of pseudo-likelihood, taking the form of singly and doubly robust estimators. They provided the basis, and exemplified it in insightful yet primarily illustrative examples. We here consider the important case of marginal models for hierarchical binary data, provide an effective implementation and illustrate it using data from an analgesic trial. Our doubly robust estimator is more convenient than the classical doubly robust estimators. The ideas are illustrated using a marginal model for a binary response, more specifically a Bahadur model.

---

**Key words:** Bahadur model; double robustness; inverse probability weighting; missing at random; pairwise likelihood; pseudo-likelihood

# 1 Introduction

Incomplete data has become an important concern for applied statisticians, especially in longitudinal and otherwise hierarchical outcome data. When the vector $\boldsymbol{Y}_i$ of planned measurements may contain missing values, the process behind these, as well as its impact on inference, needs to be addressed.

The choice of inferential framework for analyzing incomplete data will depend largely upon the nature of missingness. Conventionally, the process driving the latter is classified according to the terminology of Little and Rubin (2002, Chap. 6). When missingness is independent of both the observed and unobserved outcomes, it is called *missing completely at random* (MCAR), while when the missingness is independent of the unobserved measurements, conditional on the observed ones, the process is said to be *missing at random* (MAR). When neither MCAR nor MAR holds, missingness is termed *missing not at random* (MNAR).

Very commonly, direct likelihood is used as the basis for analyzing correlated outcomes under MAR. The unified modeling framework provided by the linear mixed model, yielding both random-effects as well as marginally interpretable regression parameters, is the dominant choice for Gaussian outcomes, while generalized linear mixed models remain popular for non-Gaussian outcomes, though marginalization is not always straightforward. Other likelihood-based options for marginal inference exist, such as the Bahadur (1961) model and the multivariate Dale or global odds ratio model (Molenberghs and Lesaffre, 1994, 1999) for binary data, but these involve complex likelihoods, can be computationally prohibitive in moderate to large studies, and are vulnerable to misspecification.

These issues have motivated the development of alternatives to likelihood, perhaps the most popular of which being generalized estimating equations or GEE (Liang and Zeger, 1986; Diggle et al., 2002; Molenberghs and Verbeke, 2005), along with variations or extensions such as GEE2 (Liang et al., 1992) and alternating logistic regressions (Carey et al., 1993), when association parameters are also of scientific interest. Standard GEE is valid only under MCAR, but a weighted version (WGEE; Robins et al., 1995) has been developed, using Horvitz-Thompson ideas (Cochran, 1977), to allow valid use of GEE under MAR. The WGEE approach, however, tends to be biased when the model for the weights is misspecified (Beunckens et al., 2008;

Molenberghs and Kenward, 2007). To this end, doubly robust approaches (Scharfstein et al., 1999; Van der Laan and Robins, 2003; Bang and Robins, 2005; Rotnitzky, 2009; Birhanu et al., 2011), which further supplement the use of weights with a predictive model for the unobserved responses, given the observed ones, have been constructed. This not only removes or at least alleviates bias, but also increases efficiency.

Pseudo-likelihood (PL) methods (le Cessie and van Houwelingen, 1991; Geys et al., 1998, 1999; Aerts et al., 2002) comprise yet another alternative to full likelihood. While the term 'pseudo-likelihood' has various meanings in the literature, we take it here to mean the replacement of a likelihood function by a simpler function that still allows a consistent and asymptotically normal estimator of the model parameter vector, albeit with potentially reduced precision (Arnold and Strauss, 1991). This is in contrast to GEE methods, where the score equations are replaced with alternative, simpler functions.

Pseudo-likelihood is different to full likelihood and is therefore not guaranteed to be valid under MAR. Rubin (1976) provided conditions for ignorability that are sufficient but not always necessary. Yi et al. (2011) provide an example, using a pairwise (pseudo-)likelihood method for incomplete longitudinal binary data, that is ignorable under MAR, even though it is not a full likelihood approach. Molenberghs et al. (2011), on the other hand, propose a suite of corrections to pseudo-likelihood in its standard form, also to ensure its validity under MAR. These corrections hold for pseudo-likelihood in general and follow both single and double robustness ideas. They showed that, in contrast to the GEE case and in particular for both robust versions, PL-based estimating equations admit very convenient simplifications.

Molenberghs et al. (2011) applied the methodology to multivariate Gaussian responses and to a conditional model for clustered binary data. They provided a general outline with predominantly illustrative examples using normal and binary data. However, the marginal modeling of longitudinal binary data is very common in practice. Molenberghs et al. (2011) only sketched the methodology using a marginal Bahadur model for the binary responses; they did not pursue it in detail. The further development of doubly robust pseudo-likelihood for incomplete hierarchical binary data under MAR is the central theme of this paper.

The theoretical part, estimating equations and precision estimators, are calculated and reported for the first time Application is shown through a case study and easy-to-use SAS code is provided.

It should be clear that we are not fitting the full Bahadur model. In fact, we use its first and second moments only, because this allows us to describe the marginal mean function, whilst providing the vehicle to take correlations and incompleteness into account. Note that there is a similar connection between standard and weighted GEE for binary data on the one hand and the Bahadur model on the other. The latter connection was studied in detail by Molenberghs and Kenward (2010). Note that apart from very simple settings, the Bahadur model is prohibitive to fit (Aerts et al., 2002).

The rest of the paper is organized as follows. Section 2 introduces the necessary background and concepts from PL and incomplete data. Our contribution, i.e., PL based on the Bahadur model, is the subject of Section 3. Analysis of the case study can be found in Section 4

# 2 Background on Pseudo-likelihood and Incomplete Data

Let the random variable $Y_{ij}$ denote the response for the $i$th study subject at the $j$th occasion ($i = 1, \ldots, N; j = 1, \ldots, n_i$). Independence across subjects is assumed. We group the outcomes into a vector $\boldsymbol{Y}_i = (y_{i1}, \ldots, y_{in_i})'$. Throughout, we will allow for covariates; where possible, they will be suppressed from notation. Let $f(\boldsymbol{y}_i|\boldsymbol{\theta})$ be a posited model. We will focus on binary responses.

The principal idea behind pseudo-likelihood is to replace a numerically challenging joint density (and hence likelihood) by a simpler function assembled from suitable factors. Arnold and Strauss (1991) gave a formal, general definition and studied its statistical properties.

Our attention will be confined to so-called pairwise marginal likelihood, in which the conventional log-likelihood

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{N} f(y_{i1}, \ldots, y_{in_i}|\boldsymbol{\theta}),$$

is replaced by

$$p\ell(\boldsymbol{\theta}) = \sum_{i=1}^{N} \sum_{1 \leq j < k \leq n_i} f(y_{ij}, y_{ik}|\boldsymbol{\theta}). \tag{2.1}$$

Maximization of Eq. (2.1) can be done, subject to adequate regularity conditions, by solving the pseudo-likelihood (score) equations, which are obtained by differentiating the logarithmic pseudo-likelihood and equating the resulting derivative to zero. Suppose that $\boldsymbol{\theta}$ is the true parameter. Under suitable regularity conditions (Arnold and Strauss, 1991; Geys et al., 1999; Aerts et al., 2002), it can be shown that maximizing the log of the pseudo-likelihood produces a consistent and asymptotically normal estimator $\widetilde{\boldsymbol{\theta}}$ so that $\sqrt{N}(\widetilde{\boldsymbol{\theta}}_N - \boldsymbol{\theta})$ converges in distribution to

$$N_p \left[ \mathbf{0} \, , \; I_0(\boldsymbol{\theta})^{-1} I_1(\boldsymbol{\theta}) I_0(\boldsymbol{\theta})^{-1} \right] . \tag{2.2}$$

Where $I_0$ and $I_1$ are information matrices based on the first ($I_1$) and second ($I_0$) derivatives of the estimating equation. However, in the case of incomplete data, PL is valid when the missingness is MCAR, but this validity does not generally extend to MAR mechanisms. Therefore, Molenberghs et al. (2011) developed general forms of estimating equations for incomplete data, applied these to the specific case of PL, and established their validity. They did not, however, consider in full detail the important case of marginal models for repeated binary data, which is the focus of this paper.

# 3 Pseudo-likelihood for Incomplete Binary Data

## 3.1 General Formulation

We begin this section by introducing some further notation. The response vector $\boldsymbol{Y}_i$ is divided into its observed ($\boldsymbol{Y}_i^o$) and missing ($\boldsymbol{Y}_i^m$) components. We further define a vector of missingness indicators $\boldsymbol{R}_i = (R_{i1}, R_{i2}, \ldots, R_{in_i})'$, with $R_{ij} = 1$ if $Y_{ij}$ is observed and 0 otherwise. In the specific case of dropout in longitudinal studies, the vector $\boldsymbol{R}_i$ can be replaced by the dropout indicator $D_i = 1 + \sum_{j=1}^{n_i} R_{ij}$, denoting the time at which subject $i$ drops out.

Molenberghs et al. (2011) considered three classes of estimating equations for pairwise likelihood, respectively naive, singly robust ('sr'), and doubly robust ('dr'). For each of these three, the original authors further considered: complete cases (CC; using only subjects will all planned measurements observed), complete pairs (CP; where all complete pairs from incomplete sequences are also added), and available cases (AC; where additionally single observations from incomplete pairs are used), leading to nine sets of estimating equations. The word 'naive' refers to the fact that these estimating equations would generally lead to biased estimators under MAR. Here only the response is modelled with a Bahadur model. For the single robust setting a weight model is introduced, using a logistic structure. For the double robust version the model was further extended with a predictive model for the unobserved outcomes using again a Bahadur model. All these estimating equations are presented in Table 1.

In this table, $\widetilde{R}_i = 1$ if subject $i$ is fully observed and 0 otherwise. In the robust cases, the probability for subject $i$ to be completely observed and to be observed up to and including occasion $j$ are respectively denoted as

$$\pi_i = \prod_{\ell=2}^{n_i}(1 - p_{i\ell}) \qquad \text{and} \qquad \pi_{ij} = \prod_{\ell=2}^{j}(1 - p_{i\ell}),$$

where $p_{i\ell} = P(D_i = \ell | D_i \geq \ell, \boldsymbol{y}_{i\bar{\ell}}, \boldsymbol{x}_{i\bar{\ell}})$ are the component probabilities of dropping out at occasion $\ell$, given the subject is still in the study, the covariate history $\boldsymbol{x}_{i\bar{\ell}}$ and the outcome history $\boldsymbol{y}_{i\bar{\ell}}$. $p_{i\ell}$ can be modeled using a logistic regression. Further, $R_{ijk}$ and $\pi_{ijk}$ are the indicator and probability, respectively, for observing both $Y_{ij}$ and $Y_{ik}$. Note that for the case of dropout, whenever $j < k$,

$$R_{ijk} \equiv R_{ik} \qquad \text{and} \qquad \pi_{ijk} \equiv \pi_{ik} = \prod_{\ell=2}^{k}(1 - p_{i\ell}),$$

in which case, e.g. the single robust version of the CP estimating equation can be re-expressed as:

$$\boldsymbol{U}_{\text{CP,sr}} = \sum_{i=1}^{N} \sum_{j<k<d_i} \frac{R_{ik}}{\pi_{ik}} \boldsymbol{U}_i(y_{ij}, y_{ik}).$$

An important result is that all three doubly robust versions coincide (Molenberghs

Table 1: *Estimating equations for pairwise pseudo-likelihood. Abbreviations used: CC: complete cases; CP: complete pairs; AC: available pairs; sr: singly robust; dr: doubly robust.*

| type | $\boldsymbol{U}_{*,\text{naive}}$ | $\boldsymbol{U}_{*,\text{sr}}$ | $\boldsymbol{U}_{*,\text{dr}}$ |
|---|---|---|---|
| $\boldsymbol{U}_{\text{CC},*}$ | $\sum_{i=1}^{N} \widetilde{R}_i \sum_{j<k} \boldsymbol{U}_i(y_{ij}, y_{ik})$ | $\sum_{i=1}^{N} \frac{\widetilde{R}_i}{\pi_i} \sum_{j<k} \boldsymbol{U}_i(y_{ij}, y_{ik})$ | $\sum_{i=1}^{N} \sum_{j<k} \left[ \frac{\widetilde{R}_i}{\pi_i} \boldsymbol{U}_i(y_{ij}, y_{ik}) + \left(1 - \frac{\widetilde{R}_i}{\pi_i}\right) E_{\boldsymbol{Y}^m \mid \boldsymbol{y}^o} \boldsymbol{U}_i(y_{ij}, y_{ik}) \right]$ |
| $\boldsymbol{U}_{\text{CP},*}$ | $\sum_{i=1}^{N} \sum_{j<k<d_i} \boldsymbol{U}_i(y_{ij}, y_{ik})$ | $\sum_{i=1}^{N} \sum_{j<k<d_i} \frac{R_{ijk}}{\pi_{ijk}} \boldsymbol{U}_i(y_{ij}, y_{ik})$ | $\sum_{i=1}^{N} \sum_{j<k<n_i} \left[ \frac{R_{ijk}}{\pi_{ijk}} \boldsymbol{U}_i(y_{ij}, y_{ik}) + \left(1 - \frac{R_{ijk}}{\pi_{ijk}}\right) E_{\boldsymbol{Y}^m \mid \boldsymbol{y}^o} \boldsymbol{U}_i(y_{ij}, y_{ik}) \right]$ |
| $\boldsymbol{U}_{\text{AC},*}$ | $\sum_{i=1}^{N} \left[ \sum_{j<k<d_i} \boldsymbol{U}_i(y_{ij}, y_{ik}) + \sum_{j=1}^{d_i-1} (n_i - d_i + 1) \boldsymbol{U}_i(y_{ij}) \right]$ | $\sum_{i=1}^{N} \left[ \sum_{j=1}^{d_i-1} \frac{R_{ij}}{\pi_{ij}} \boldsymbol{U}_i(y_{ij}) + \sum_{j<k} \frac{R_{ik}}{\pi_{ik}} \boldsymbol{U}_i(y_{ik} \mid y_{ij}) \right]$ | $\sum_{i=1}^{N} \left[ \sum_{j<k} \frac{R_{ik}}{\pi_{ik}} \boldsymbol{U}_i(y_{ik} \mid y_{ij}) + \sum_{j=1}^{d_i-1} \frac{R_{ij}}{\pi_{ij}} \boldsymbol{U}_i(y_{ij}) + \sum_{j<k} \left(1 - \frac{R_{ik}}{\pi_{ik}}\right) E_{\boldsymbol{Y}^m \mid \boldsymbol{y}^o} \boldsymbol{U}_i(y_{ik} \mid y_{ij}) + \sum_{j=1}^{d_i-1} \left(1 - \frac{R_{ij}}{\pi_{ij}}\right) E_{\boldsymbol{Y}^m \mid \boldsymbol{y}^o} \boldsymbol{U}_i(y_{ij}) \right]$ |

et al., 2011), i.e.,

$$\begin{aligned}
\boldsymbol{U}_{\text{CC,dr}} &= \boldsymbol{U}_{\text{CP,dr}} = \boldsymbol{U}_{\text{AC,dr}} = \\
&= \sum_{i=1}^{N} \left\{ \sum_{j<k<d_i} \boldsymbol{U}_i(y_{ij}, y_{ik}) + \sum_{j=1}^{d_i-1} (n_i - d_i + 1) \boldsymbol{U}_i(y_{ij}) \right. \\
&\qquad \left. + \sum_{j<d_i\leq k} E\left[\boldsymbol{U}_i(y_{ik}|y_{ij})\right] + \sum_{d_i\leq j<k} E\left[\boldsymbol{U}_i(y_{ij}, y_{ik})\right] \right\}. \qquad (3.1)
\end{aligned}$$

It is thus not necessary to explicitly model the missing-data mechanism. Further, under exchangeability, Molenberghs et al. (2011) showed that the expectations in $\boldsymbol{U}_{\text{AC,dr}}$ vanish, making Eq. (3.1) essentially equivalent to $\boldsymbol{U}_{\text{AC,naive}}$, which is very convenient for implementation, as this reduces to an observed data analysis. More information on this can be found in Appendix 1.1.

## 3.2 Full Bahadur Model

To introduce the Bahadur model (Bahadur, 1961) we follow the developments in Molenberghs et al. (2011). Denote $\nu_{ij} = P(Y_{ij} = 1)$, $\nu_{ijk} = P(Y_{ij} = 1, Y_{ik} = 1)$, and $\nu_{ik|j} = P(Y_{ik} = 1|y_{ij} = \ell)(\ell = 0, 1)$. Pairwise Bahadur probabilities take the form

$$\nu_{ijk} = \nu_{ij}\nu_{ik}\left[1 + \rho_{ijk}\frac{1-\nu_{ij}}{\sqrt{\nu_{ij}(1-\nu_{ij})}}\frac{1-\nu_{ik}}{\sqrt{\nu_{ik}(1-\nu_{ik})}}\right]. \qquad (3.2)$$

The multivariate Bahadur probabilities are $f(\boldsymbol{y}_i) = f_1(\boldsymbol{y}_i)c(\boldsymbol{y}_i)$, with:

$$\begin{aligned}
f_1(\boldsymbol{y}_i) &= \prod_{j=1}^{n_i} \nu_{ij}^{y_{ij}} (1-\nu_{ij})^{1-y_{ij}}, \qquad (3.3) \\
c(\boldsymbol{y}_i) &= 1 + \sum_{j_1<j_2} \rho_{ij_1j_2} e_{ij_1} e_{ij_2} + \sum_{j_1<j_2<j_3} \rho_{ij_1j_2j_3} e_{ij_1} e_{ij_2} e_{ij_3} + \\
&\quad \cdots + \rho_{ij_1j_2\ldots j_{n_i}} e_{ij_1} e_{ij_2} \cdots e_{ij_{n_i}}, \qquad (3.4)
\end{aligned}$$

where $e_{ij} = \dfrac{y_{ij} - \nu_{ij}}{\sqrt{\nu_{ij}(1-\nu_{ij})}}$.

### 3.3 Pairwise Bahadur Model for the Outcome

Based on the definitions made in Section 3.2 the log-likelihood terms from a pairwise Bahadur model take the following form:

$$\begin{aligned} p\ell_{ijk} &= y_{ij}y_{ik}\ln\nu_{ijk} + y_{ij}(1-y_{ik})\ln(\nu_{ij}-\nu_{ijk}) + (1-y_{ij})y_{ik}\ln(\nu_{ik}-\nu_{ijk}) \\ &+ (1-y_{ij})(1-y_{ik})\ln(1-\nu_{ij}-\nu_{ik}+\nu_{ijk}). \end{aligned} \tag{3.5}$$

Starting from pseudo-likelihood contribution (3.5), pairwise and conditional contributions to the score equation take the form as follows

$$\begin{aligned} \boldsymbol{U}_{ijk} &= \frac{y_{ij}y_{ik}}{\nu_{ijk}}\frac{\partial}{\partial\boldsymbol{\theta}}\nu_{ijk} + \frac{y_{ij}(1-y_{ik})}{\nu_{ij}-\nu_{ijk}}\frac{\partial}{\partial\boldsymbol{\theta}}(\nu_{ij}-\nu_{ijk}) + \frac{(1-y_{ij})y_{ik}}{\nu_{ik}-\nu_{ijk}}\frac{\partial}{\partial\boldsymbol{\theta}}(\nu_{ik}-\nu_{ijk}) \\ &+ \frac{(1-y_{ij})(1-y_{ik})}{1-\nu_{ij}-\nu_{ik}+\nu_{ijk}}\frac{\partial}{\partial\boldsymbol{\theta}}(1-\nu_{ij}-\nu_{ik}+\nu_{ijk}), \end{aligned} \tag{3.6}$$

$$\begin{aligned} \boldsymbol{U}_{ik|j} &= \frac{y_{ij}y_{ik}\nu_{ij}}{\nu_{ijk}}\frac{\partial}{\partial\boldsymbol{\theta}}\left(\frac{\nu_{ijk}}{\nu_{ij}}\right) + \frac{y_{ij}(1-y_{ik})\nu_{ij}}{\nu_{ij}-\nu_{ijk}}\frac{\partial}{\partial\boldsymbol{\theta}}\left(\frac{\nu_{ij}-\nu_{ijk}}{\nu_{ij}}\right) \\ &+ \frac{(1-y_{ij})y_{ik}(1-\nu_{ij})}{\nu_{ik}-\nu_{ijk}}\frac{\partial}{\partial\boldsymbol{\theta}}\left(\frac{\nu_{ik}-\nu_{ijk}}{1-\nu_{ij}}\right) \\ &+ \frac{(1-y_{ij})(1-y_{ik})(1-\nu_{ij})}{1-\nu_{ij}-\nu_{ik}+\nu_{ijk}}\frac{\partial}{\partial\boldsymbol{\theta}}\left(\frac{1-\nu_{ij}-\nu_{ik}+\nu_{ijk}}{1-\nu_{ij}}\right), \end{aligned} \tag{3.7}$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\alpha}')'$, and $\nu_{ij} = \nu_{ij}(\boldsymbol{\beta})$ and the association parameters are functions of $\boldsymbol{\alpha}$. Hence, $\nu_{ijk} = \nu_{ijk}(\boldsymbol{\beta}, \boldsymbol{\alpha})$.

The expectations of these over the conditional distribution of the unobserved outcomes given the observed ones are further required. Evidently, because Eqs. (3.6)–(3.7) are linear in the triplet $y_{ij}, y_{ik}$ and $y_{ij}y_{ik}$, it suffices to calculate the expectations over these. Their corresponding probabilities are

$$\nu_{ij|\overline{d}} = \frac{\nu_{i\overline{d}j}}{\nu_{i\overline{d}}} \qquad \text{and} \qquad \nu_{ijk|\overline{d}} = \frac{\nu_{i\overline{d}jk}}{\nu_{i\overline{d}}}, \tag{3.8}$$

where $\overline{d}$ refers to the set of indices $(1, 2, \ldots, d-1)$, corresponding to the observed portion of $\boldsymbol{y}$.

Combining Eqs. (3.6) and (3.7) with Eq. (3.8) leads to

$$
\begin{aligned}
E(\boldsymbol{U}_{ijk}) &= \frac{\nu_{i\bar{d}jk}}{\nu_{i\bar{d}}\nu_{ijk}}\frac{\partial}{\partial\boldsymbol{\theta}}\nu_{ijk} + \frac{\nu_{i\bar{d}j}-\nu_{i\bar{d}jk}}{\nu_{i\bar{d}}(\nu_{ij}-\nu_{ijk})}\frac{\partial}{\partial\boldsymbol{\theta}}(\nu_{ij}-\nu_{ijk}) \\
&\quad + \frac{\nu_{i\bar{d}k}-\nu_{i\bar{d}jk}}{\nu_{i\bar{d}}(\nu_{ik}-\nu_{ijk})}\frac{\partial}{\partial\boldsymbol{\theta}}(\nu_{ik}-\nu_{ijk}) \\
&\quad + \frac{\nu_{i\bar{d}}-\nu_{i\bar{d}j}-\nu_{i\bar{d}k}+\nu_{i\bar{d}jk}}{\nu_{i\bar{d}}(1-\nu_{ij}-\nu_{ik}+\nu_{ijk})}\frac{\partial}{\partial\boldsymbol{\theta}}(1-\nu_{ij}-\nu_{ik}+\nu_{ijk}) \qquad (3.9)
\end{aligned}
$$

and

$$
\begin{aligned}
E(\boldsymbol{U}_{ik|j}) &= \frac{y_{ij}\nu_{i\bar{d}k}\nu_{ij}}{\nu_{i\bar{d}}\nu_{ijk}}\frac{\partial}{\partial\boldsymbol{\theta}}\left(\frac{\nu_{ijk}}{\nu_{ij}}\right) + \frac{y_{ij}(\nu_{i\bar{d}}-\nu_{i\bar{d}k})\nu_{ij}}{\nu_{i\bar{d}}(\nu_{ij}-\nu_{ijk})}\frac{\partial}{\partial\boldsymbol{\theta}}\left(\frac{\nu_{ij}-\nu_{ijk}}{\nu_{ij}}\right) \\
&\quad + \frac{(1-y_{ij})\nu_{i\bar{d}k}(1-\nu_{ij})}{\nu_{i\bar{d}}(\nu_{ik}-\nu_{ijk})}\frac{\partial}{\partial\boldsymbol{\theta}}\left(\frac{\nu_{ik}-\nu_{ijk}}{1-\nu_{ij}}\right) \\
&\quad + \frac{(1-y_{ij})(\nu_{i\bar{d}}-\nu_{i\bar{d}k})(1-\nu_{ij})}{\nu_{i\bar{d}}(1-\nu_{ij}-\nu_{ik}+\nu_{ijk})}\frac{\partial}{\partial\boldsymbol{\theta}}\left(\frac{1-\nu_{ij}-\nu_{ik}+\nu_{ijk}}{1-\nu_{ij}}\right). \qquad (3.10)
\end{aligned}
$$

## 3.4 Predictive Bahadur model in the Doubly Robust Estimating Equations

Many of the probabilities in the predictive model, i.e., the ones involving $\bar{d}$, in (3.9)–(3.10) are of dimension 3 or higher. The calculation of the probabilities in the multivariate Bahadur model is cumbersome because of the very constrained parameter space. Pairwise PL is used exactly to circumvent this problem. In the spirit of, among others, Bang and Robins (2005), we follow a more pragmatic route and propose a convenient and sufficiently rich predictive model. An attractive option is the pairwise Bahadur model, pertaining to response at occasions $j$ and $k$, but where the history, corresponding to $\overline{d}$, is included as a set of predictor variables. This amounts to using

$$
\begin{aligned}
E(\boldsymbol{U}_{ijk}) &\equiv E\left[\boldsymbol{U}_i(y_{ij}, y_{ik})\right] \\
&= \sum_{y_{ij}=0}^{1} \sum_{y_{ik}=0}^{1} \boldsymbol{U}_i(y_{ij}, y_{ik}) q(y_{ij}, y_{ik}), && (3.11) \\
E(\boldsymbol{U}_{ik|j}) &\equiv E\left[\boldsymbol{U}_i(y_{ik}|y_{ij})\right] = \sum_{y_{ik}=0}^{1} \boldsymbol{U}_i(y_{ik}|y_{ij}) q(y_{ik}|y_{ij}), && (3.12)
\end{aligned}
$$

where $q(y_{ij}, y_{ik}) = P(Y_{ij} = y_{ij}, Y_{ik} = y_{ik}|Y_{i\overline{d}} = y_{i\overline{d}})$ and $\boldsymbol{U}_i(y_{ij}, y_{ik})$ and $\boldsymbol{U}_i(y_{ik}|y_{ij})$ are as defined in Eqs. (3.6) and (3.7). Evidently, modeling the $q(\cdot)$ terms, will imply the need for an additional parameter vector, $\boldsymbol{\phi}$, say.

## 3.5 Precision Estimation

In the naive case, uncertainty stems from the $\boldsymbol{\theta}$ parameter only. The asymptotic variance-covariance matrix in Eq. (2.2) can then be consistently estimated by $\widehat{I_0}^{-1}\widehat{I_1}\widehat{I_0}^{-1}$, with

$$
I_0 = \frac{1}{N}\sum_{i=1}^{N} \frac{\partial \boldsymbol{V}_i}{\partial \boldsymbol{\theta}} \quad \text{and} \quad I_1 = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{S}_i(\widehat{\boldsymbol{\theta}})\boldsymbol{S}_i'(\widehat{\boldsymbol{\theta}}), \tag{3.13}
$$

where $\boldsymbol{U} = \sum_{i=1}^{N} \boldsymbol{V}_i(\boldsymbol{\theta})$ and $\boldsymbol{S}_i(\widehat{\boldsymbol{\theta}}) = \boldsymbol{V}_i$ is the corresponding estimating function, i.e., shorthand notation for the formulas in Table 1.

In the singly robust case, we must also take into account uncertainty coming from estimating the $\boldsymbol{\psi}$ parameters in the weight model. The entire score for subject $i$ is $\boldsymbol{S}_i = (\boldsymbol{V}_i', \boldsymbol{W}_i')'$, with $\boldsymbol{W} = \sum_{i=1}^{N} \boldsymbol{W}_i(\boldsymbol{\psi})$ the estimating equations coming from the weight model, and the asymptotic variance-covariance is based on the following matrices:

$$I_0 = \frac{1}{N}\sum_{i=1}^{N}\begin{pmatrix} \frac{\partial \boldsymbol{V}_i}{\partial \boldsymbol{\theta}} & \frac{\partial \boldsymbol{V}_i}{\partial \boldsymbol{\psi}} \\ \boldsymbol{0} & \frac{\partial \boldsymbol{W}_i}{\partial \boldsymbol{\psi}} \end{pmatrix} \quad \text{and} \quad I_1 = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{S}_i(\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\psi}})\boldsymbol{S}_i'(\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\psi}}). \tag{3.14}$$

In the doubly robust case, for the general expression, the weight model is complemented with a predictive model. The score function for this conditional Bahadur model is $\boldsymbol{T}(\boldsymbol{\phi})$, with an extra set of parameters $\boldsymbol{\phi}$. The precision of the parameters can be estimated using the matrices as follows:

$$I_0 = \frac{1}{N}\sum_{i=1}^{N}\begin{pmatrix} \frac{\partial \boldsymbol{V}_i}{\partial \boldsymbol{\theta}} & \frac{\partial \boldsymbol{V}_i}{\partial \boldsymbol{\psi}} & \frac{\partial \boldsymbol{V}_i}{\partial \boldsymbol{\phi}} \\ \boldsymbol{0} & \frac{\partial \boldsymbol{W}_i}{\partial \boldsymbol{\psi}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \frac{\partial \boldsymbol{T}_i}{\partial \boldsymbol{\phi}} \end{pmatrix} \quad \text{and} \quad I_1 = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{S}_i(\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\psi}},\widehat{\boldsymbol{\phi}})\boldsymbol{S}_i'(\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\psi}},\widehat{\boldsymbol{\phi}}), \tag{3.15}$$

From Eq. (3.1), (3.15) can be simplified to the following expressions

$$I_0 = \frac{1}{N}\sum_{i=1}^{N}\begin{pmatrix} \frac{\partial \boldsymbol{V}_i}{\partial \boldsymbol{\theta}} & \frac{\partial \boldsymbol{V}_i}{\partial \boldsymbol{\phi}} \\ \boldsymbol{0} & \frac{\partial \boldsymbol{T}_i}{\partial \boldsymbol{\phi}} \end{pmatrix} \quad \text{and} \quad I_1 = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{S}_i(\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\phi}})\boldsymbol{S}_i'(\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\phi}}). \tag{3.16}$$

More detailed calculations and complete formulas can be found in Appendix B. See also Bang and Robins (2005) and Rotnitzky (2009).

# 4 Results

In this section, we apply the proposed methodology to data from a clinical trial designed to investigate an analgesic drug. All analyses have been performed with SAS (version 9.4). First, the Bahadur model, using three different estimating equations for CC, CP and AC, was fitted with an NLMIXED procedure. To make use of NLMIXED's functionality, an objective function is formulated of which the first derivative coincides with the estimating function under consideration. For optimization, the default Quasi-Newton technique was applied. Further, to estimate the precision, a sandwich-type robust variance estimator was used and, to perform the calculations, the IML procedure was implemented. The Bahadur model, based on the full likelihood, was again fitted in an NLMIXED procedure with similar settings. For more details, see Appendix C.

## 4.1 Analgesic Clinical Trial

The analgesic trial was a single-arm clinical trial involving 395 patients who were given analgesic treatment for pain caused by chronic non-malignant disease. Treatment was to be administered for 12 months and assessed by means of a five-point 'Global Satisfaction Assessment' (GSA) scale: (1) very good; (2) good; (3) indifferent; (4) bad; (5) very bad. As it is frequently of interest to physicians to classify a patient's status as either improving or worsening, some analyses have considered a dichotomized version, GSABIN, which is 1 if GSA $\leq 3$ and 0 otherwise; this outcome will be adopted for our analysis as well. Apart from the outcome of interest, a number of covariates are available, such as age, sex, weight, duration of pain in years prior to the start of the study, type of pain, physical functioning, psychiatric condition, respiratory problems,

Table 2: *The Analgesic Trial. Absolute and relative frequencies of the five GSA categories for each of the four follow-up times.*

| GSA | Month 3 | | Month 6 | | Month 9 | | Month 12 | |
|---|---|---|---|---|---|---|---|---|
| 1 | 55 | 14.3% | 38 | 12.6% | 40 | 17.6% | 30 | 13.5% |
| 2 | 112 | 29.1% | 84 | 27.8% | 67 | 29.5% | 66 | 29.6% |
| 3 | 151 | 39.2% | 115 | 38.1% | 76 | 33.5% | 97 | 43.5% |
| 4 | 52 | 13.5% | 51 | 16.9% | 33 | 14.5% | 27 | 12.1% |
| 5 | 15 | 3.9% | 14 | 4.6% | 11 | 4.9% | 3 | 1.4% |
| Total | 385 | | 302 | | 227 | | 223 | |

etc.

GSA was rated by each person four times during the trial: at months 3, 6, 9, and 12. An overview of the frequencies per follow-up time is given in Table 2. Inspection of Table 2 reveals varying totals per column, due to missingness. At three months, 10 subjects lack a measure, with these numbers being 93, 168, and 172 at subsequent times.

An overview of the extent of missingness (Table 3) indicates that only around 40% of the subjects have complete data. Both dropout and intermittent patterns of missingness occur – the former amounting to roughly 40%, with less than 20% for the latter.

Table 3: *The Analgesic Trial. Overview of missingness patterns and the frequencies with which they occur. 'O' indicates observed and 'M' indicates missing.*

| | Measurement Occasion | | | | $N$ | % |
|---|---|---|---|---|---|---|
| | Month 3 | Month 6 | Month 9 | Month 12 | | |
| Completers | O | O | O | O | 163 | 41.2 |
| Dropouts | O | O | O | M | 51 | 12.91 |
| | O | O | M | M | 51 | 12.91 |
| | O | M | M | M | 63 | 15.95 |
| Non-Monotone Missingness | O | O | M | O | 30 | 7.59 |
| | O | M | O | O | 7 | 1.77 |
| | O | M | O | M | 2 | 0.51 |
| | O | M | M | O | 18 | 4.56 |
| | M | O | O | O | 2 | 0.51 |
| | M | O | O | M | 1 | 0.25 |
| | M | O | M | O | 1 | 0.25 |
| | M | O | M | M | 3 | 0.76 |

## 4.2 Analysis of the Case Study

For all ensuing analyses of the analgesic trial data, we consider only completers and dropouts, i.e., a subset of 328 patients from the original data set, and the dichotomized outcome (GSABIN). We first build a logistic regression for the dropout indicator, in terms of the previous outcome ($y_{i,j-1}$) and pain control assessment at baseline ($x_i$), i.e.,

$$\text{logit } P(D_i = j | D_i \geq j, x_i, y_{i,j-1}) = \psi_0 + \psi_x x_i + \psi_{prev} y_{i,j-1}.$$

The highly significant p-value ($p < .0001$) for the parameter $\psi_{prev}$ corresponding to the previous outcome provides evidence against MCAR in favor of MAR. Weights are then calculated based on predicted probabilities from this logistic model.

Preliminary analyses have indicated that, among a set of potential covariates, the linear and quadratic effects of time $t_{ij}$, as well as the effect of baseline pain control assessment ($\text{PCA}_0$, denoted $x_i$) are of importance. The marginal regression model for the dichotomized GSA score, GSABIN, denoted as $Y$, is thus specified as

$$\text{logit } P(Y_{ij} = 1 | t_{ij}, x_i) = \beta_0 + \beta_1 t_{ij} + \beta_2 t_{ij}^2 + \beta_3 x_i. \tag{4.1}$$

For the correlation across the within-subject outcomes, we posit a Toeplitz type correlation structure:

$$\begin{pmatrix} 1 & \rho^{(1)} & \rho^{(2)} & \rho^{(3)} \\ \rho^{(1)} & 1 & \rho^{(1)} & \rho^{(2)} \\ \rho^{(2)} & \rho^{(1)} & 1 & \rho^{(2)} \\ \rho^{(3)} & \rho^{(2)} & \rho^{(1)} & 1 \end{pmatrix}, \tag{4.2}$$

where $\rho^{(k)}$, $k = 1, 2, 3$ denotes the correlation between outcomes that are $k$ time points apart. Hence, the Bahadur density is $f(\boldsymbol{y}_i) = f_1(\boldsymbol{y}_i)c(\boldsymbol{y}_i)$, with $f_1(\boldsymbol{y}_i)$ as in Eq. (3.3) with $n_i = 4$ and Eq. (3.4) taking the specific form:

$$\begin{aligned} c(\boldsymbol{y}_i) &= 1 + \sum_{j_1 < j_2; j_2 - j_1 = k} \rho_{ij_1j_2}^{(k)} e_{ij_1} e_{ij_2}, \\ &= 1 + \rho^{(1)} \left( e_{i1}e_{i2} + e_{i2}e_{i3} + e_{i3}e_{i4} \right) + \rho^{(2)} \left( e_{i1}e_{i3} + e_{i2}e_{i4} \right) + \rho^{(3)} e_{i1}e_{i4}. \end{aligned}$$

The resulting parameter estimates, along with corresponding standard errors, for model specification Eq. (4.1), with a Toeplitz correlation structure (Eq. 4.2), using full likelihood and estimating equations from Table 1 are presented in Table 4. The variability of the estimated weights, or additionally the variability of the estimated parameters of the predictive model, is incorporated in the computation of the standard errors. The high degree of similarity with the results of full likelihood indicate that the extra variability induced by the weights, or additionally by the parameters of the predictive model, does not seem to have a large impact on either the estimates or their standard errors.

Similar results are observed throughout the whole table, but in particular for the parameter estimates under full likelihood, naive AC and the doubly robust cases. Moreover, substantial efficiency over full likelihood seems to be gained under the naive AC and doubly robust approaches. Whereas these observations are not surprising for the doubly robust case, precisely because of their property, the relatively good performance of the naive AC case seems counterintuitive. However, under exchangeability, as shown before, the naive AC can be seen as a doubly robust estimator, given that then the expectation in these estimation equations can be removed because observed and unobserved components from a subject's history are interchangeable. To this effect, we assessed the plausibility of the Toeplitz correlation structure of the analgesic trial data, using full likelihood (approximate F-test in NLMIXED), and determined that the three correlation parameters $\rho^{(k)},\ k = 1, 2, 3$, were not significantly different ($p = 0.9078$), which implies that the underlying correlation structure might very well be exchangeable. This explains the excellent behaviour of the naive AC estimator.

Next, we consider the CP versions, both single and doubly robust. The estimates for the parameters seem reasonably close to those under full likelihood. In addition, the standard errors under the singly robust case seem comparable, but those of the doubly robust case are generally larger than those from full likelihood, a result that could be attributed to the fact of single robustness. The estimates for the $\beta$ parameters

Table 4: *Analgesic Trial. Parameter estimates (empirically-corrected standard errors) for naive, singly and doubly robust pairwise likelihood and for full likelihood.*

| Effect | Par. | $\boldsymbol{U}_{\text{CC,naive}}$ | $\boldsymbol{U}_{\text{CP,naive}}$ | $\boldsymbol{U}_{\text{AC,naive}}$ | $\boldsymbol{U}_{\text{full.lik.}}$ |
|---|---|---|---|---|---|
| Inter. | $\beta_0$ | 3.131 (0.678) | 2.962 (0.563) | 2.590 (0.493) | 2.626 (0.509) |
| Time | $\beta_1$ | -0.913 (0.492) | -0.908 (0.401) | -0.675 (0.354) | -0.602 (0.362) |
| $\text{Time}^2$ | $\beta_2$ | 0.170 (0.096) | 0.177 (0.081) | 0.151 (0.074) | 0.120 (0.076) |
| $\text{PCA}_0$ | $\beta_3$ | -0.130 (0.132) | -0.125 (0.113) | -0.186 (0.099) | -0.209 (0.106) |
| $\text{corr}_1$ | $\rho^{(1)}$ | 0.217 (0.069) | 0.244 (0.055) | 0.259 (0.057) | 0.297 (0.063) |
| $\text{corr}_2$ | $\rho^{(2)}$ | 0.199 (0.075) | 0.234 (0.068) | 0.250 (0.069) | 0.293 (0.074) |
| $\text{corr}_3$ | $\rho^{(3)}$ | 0.224 (0.102) | 0.232 (0.104) | 0.240 (0.104) | 0.337 (0.117) |
| Effect | Par. | $\boldsymbol{U}_{\text{CC,sr}}$ | $\boldsymbol{U}_{\text{CP,sr}}$ | $\boldsymbol{U}_{\text{AC,sr}}$ | |
| Inter. | $\beta_0$ | 3.090 (0.637) | 2.712 (0.552) | 1.718 (0.560) | |
| Time | $\beta_1$ | -0.997 (0.468) | -0.775 (0.389) | -0.280 (0.347) | |
| $\text{Time}^2$ | $\beta_2$ | 0.193 (0.090) | 0.151 (0.078) | 0.092 (0.070) | |
| $\text{PCA}_0$ | $\beta_3$ | -0.195 (0.133) | -0.167 (0.113) | -0.196 (0.115) | |
| $\text{corr}_1$ | $\rho^{(1)}$ | 0.263 (0.079) | 0.295 (0.062) | 0.333 (0.064) | |
| $\text{corr}_2$ | $\rho^{(2)}$ | 0.257 (0.086) | 0.273 (0.076) | 0.303 (0.076) | |
| $\text{corr}_3$ | $\rho^{(3)}$ | 0.295 (0.115) | 0.298 (0.112) | 0.299 (0.108) | |
| Effect | Par. | $\boldsymbol{U}_{\text{CC,dr}}$ | $\boldsymbol{U}_{\text{CP,dr}}$ | $\boldsymbol{U}_{\text{AC,dr}}$ | |
| Inter. | $\beta_0$ | 3.577 (1.136) | 2.736 (0.874) | 1.533 (0.692) | |
| Time | $\beta_1$ | -1.333 (0.851) | -0.785 (0.647) | -0.104 (0.480) | |
| $\text{Time}^2$ | $\beta_2$ | 0.241 (0.164) | 0.149 (0.132) | 0.052 (0.108) | |
| $\text{PCA}_0$ | $\beta_3$ | -0.196 (0.220) | -0.153 (0.193) | -0.197 (0.147) | |
| $\text{corr}_1$ | $\rho^{(1)}$ | 0.255 (0.118) | 0.305 (0.088) | 0.366 (0.108) | |
| $\text{corr}_2$ | $\rho^{(2)}$ | 0.247 (0.165) | 0.281 (0.139) | 0.338 (0.158) | |
| $\text{corr}_3$ | $\rho^{(3)}$ | 0.305 (0.276) | 0.329 (0.275) | 0.350 (0.243) | |

from the CC cases are somewhat higher, whereas those from the AC cases are lower than those for full likelihood. The CP results seem to fall in between the CC and AC results, suggesting a comprise between the latter two. This can be inferred from the incremental nature of the contributions in expressions in Table 1. However, as AC case uses more information than the CP case, this one is generally to be preferred.

# 5 Discussion

Pseudo-likelihood approaches have become a practical alternative to full likelihood methods, particularly for applications involving complex likelihood forms. In view of the various issues arising from marginally modelling incomplete non-Gaussian longitudinal data, we move away from conditional pseudo-likelihood, and focus on *marginal pseudo-likelihood*, considering the specific case of incomplete longitudinal binary data, as proposed in Molenberghs et al. (2011). While the numerical and computational issues accompanying the likelihood expressions of the marginal model for the binary longitudinal responses are circumvented by means of substituting pairwise pseudo-likelihood expressions for their full likelihood counterparts, the incompleteness in the data is addressed using concepts of inverse probability weighting and predictive terms in the form of expectations, thereby yielding singly and doubly robust estimators. This expands the set of tools available for fitting marginal models to incomplete non-Gaussian longitudinal data.

In this paper, we assessed the performance of pseudo-likelihood approaches proposed in Molenberghs et al. (2011), in order to provide practical insight into alternative strategies for marginal models for non-Gaussian incomplete longitudinal data. The analysis of the case study demonstrates the feasibility and adequacy of the proposed methodology. Singly robust estimators with correctly specified dropout model and our doubly robust estimators were found to be at least as efficient as direct likelihood methods. Moreover, under full or near exchangeability, the naive available case version is as efficient as the doubly robust estimators, allowing one to invoke double robustness without having to use weights or expectations.

While the situation examined in this paper focuses on dropout, in principle, the general methodology applies for non-monotone missingness as well; one then has to pay particular attention to the construction of both weights and predictions, and some non-trivial algebraic challenges will emerge. Other possibilities include imputing all missing cases or imputing only non-monotone missing cases to render the missingness monotone and subsequently using pseudo-likelihood on the imputed data sets. Also, while multiple imputation approaches generally prescribe Gaussian type data, variations for non-Gaussian data can be utilized and seem reasonably stable even with model misspecification; see, for instance, Beunckens et al. (2008).

## Supplementary Materials

More detailed information on calculations, data and SAS code can be found in the accompanying Supplementary Materials through the link: http://www.statmod.org/smij/archive.html.

## Acknowledgements

The authors gratefully acknowledge support from the 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 Modelling of Clustered Data*. Chapman & Hall/CRC, London.

Arnold, B. C. and Strauss, D. (1991). Pseudolikelihood estimation: Some examples. *Sankhya B*, **53**, 233–243.

Bahadur, R. R. (1961). A representation of the joint distribution of responses of n dichotomous items. In Solomon, H., editor, *Studies in Item Analysis and Prediction*, pages 158–168. Stanford University Press, California.

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, **61**, 962–972.

Beunckens, C., Sotto, C., and Molenberghs, G. (2008). A simulation study comparing weighted estimating equations with multiple imputation based estimating equations for longitudinal binary data. *Computational Statistics and Data Analysis*, **52**, 1533–1548.

Birhanu, T., Molenberghs, G., Sotto, C., and Kenward, M. G. (2011). Doubly robust and multiple-imputation-based generalized estimating equations. *Journal of Biopharmaceutical Statistics*, **21**, 202–225.

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

Cochran, W. G. (1977). *Sampling Techniques.* John Wiley & Sons, New York.

Diggle, P. J., Heagerty, P., Liang, K. Y., and Zeger, S. L. (2002). *Analysis of Longitudinal Data, 2nd ed.* Oxford University Press, New York.

Geys, H., Molenberghs, G., and Lipsitz, S. R. (1998). A note on the comparison of pseudo-likelihood and generalized estimating equations for marginal odds ratio models. *Journal of Statistical Computation and Simulation*, **62**, 45–72.

Geys, H., Molenberghs, G., and Ryan, L. (1999). Pseudo-likelihood modelling of multivariate outcomes in developmental toxicology. *Journal of the American Statistical Association*, **94**, 734–745.

le Cessie, S. and van Houwelingen, J. C. (1991). A goodness-of-fit test for binary regression models, based on smoothing methods. *Biometrics*, **47**, 1267–1282.

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.

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

Molenberghs, G. and Kenward, M. G. (2007). *Missing Data in Clinical Studies.* John Wiley & Sons, New York.

Molenberghs, G. and Kenward, M. G. (2010). Semi-parametric marginal models for hierarchical data and their corresponding full models. *Computional Statistics & Data Analysis*, **54**, 585–597.

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

Molenberghs, G., Kenward, M. G., Verbeke, G., and Birhanu, T. (2011). Pseudo-likelihood estimation for incomplete data. *Statistica Sinica*, **21**, 187–206.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, **90**, 106–121.

Rotnitzky, A. (2009). Inverse probability weighted methods. In Fitzmaurice, G., D. M. V. G. . M. G., editor, *Longitudinal Data Analysis*, pages 453–476. Chapman & Hall/CRC, Florida.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, **63**, 581–592.

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Adjusting for nonignorable drop-out using semi-parametric nonresponse models. *Journal of the American Statistical Association*, **94**, 1096–1120 (with Rejoinder, 1135–1146).

Van der Laan, M. J. and Robins, J. M. (2003). *Unified Methods for Censored Longitudinal Data and Causality.* Springer-Verlag, New York.

Yi, G. Y., Zeng, L., and Cook, R. J. (2011). A robust pairwise likelihood method for incomplete longitudinal binary data arising in clusters. *The Canadian Journal of Statistics*, **39**, 34–51.