KU Leuven
Biomedical Sciences Group
Faculty of Medicine
Department of Public Health and Primary Care
L-BioStat

KU LEUVEN

DOCTORAL SCHOOL
BIOMEDICAL SCIENCES

A Joint Model for Longitudinal Outcomes and Longitudinal Covariates

Margaux DELPORTE

Supervisor:	Prof. Dr. Geert Verbeke	
Co-supervisor:	Prof. Dr. Geert Molenberghs	
Co-supervisor:	Dr. Steffen Fieuws	
Chair:	Prof. Dr. Jeroen Vanoirbeek	
Secretary:	Prof. Dr. Francis Tuerlinckx	Dissertation presented
Jury members:	Prof. Dr. Ingrid Van Keilegom	in partial fulfilment of
	Prof. Dr. Christel Faes	the requirements for
	Prof. Dr. Edmund Njeru Njagi	the degree of Doctor in
	Prof. Dr. Roula Tsonaka	Biomedical Sciences (Biostatistics)

22 October 2024

List of Papers

The contents of this manuscript are based on the following original publications:

Chapter 3: Delporte, M., Fieuws, S., Molenberghs, G., Verbeke, G., Wanyama, S.S., Hatziagorou, E., & De Boeck, C. (2022). A joint normal-binary (probit) model. *International Statistical Review*. **90(S1)**, S37-S51.

Chapter 4: Delporte, M., Fieuws, S., Molenberghs, G., Verbeke, G, De Coninck D., & Hoorens, V. (2023). A Joint Normal-Binary (Probit) Model for High-Dimensional Longitudinal Data. *Statistical Modelling*, Accepted.

Chapter 5: Delporte, M., Molenberghs, G., Fieuws, S., & Verbeke, G. (2023) A Joint Normal-Ordinal(Probit) Model for Ordinal and Continuous Longitudinal Data. *Biostatistics*, Accepted.

Chapter 6: Delporte, M., Aerts, M., Verbeke, G., & Molenberghs, G. Analysing matched continuous longitudinal data: A review. Submitted.

Chapter 7: Delporte, M., Verbeke, G., Fieuws, S., & Molenberghs, G. Accelerating Computation: A Faster Pairwise Fitting Technique for Multivariate Probit Models. Submitted.

The author has also been involved in the following original publications:

Vandekerckhove, I., van den Hauwe, M., De Beukelaer, N., Stoop, E., Goudriaan, M., **Delporte, M.**, Molenberghs, G., Van Campenhout, A., De Waele, L., Goemans, N., De Groote, F., & Desloovere, K. (2022). Longitudinal Alterations in Gait Features in Growing Children With Duchenne Muscular Dystrophy. *Frontiers in Human Neuroscience*, **16 (861136)**.

De Witte, D., **Delporte, M.**, Molenberghs, G., Verbeke, G., Demarest, S., & Hoorens, V. (2022). Self-uniqueness beliefs and adherence to recommended precautions. A 5-wave longitudinal COVID-19 study. *Social Science & Medicine*, **317 (115595)**.

Van Eijgen, J., Heintz, A., Van der Pluijm, C., **Delporte, M.**, De Witte, D., Molenberghs, G., Barbosa-Breda, J., & Stalmans, I. (2023). Normal tension glaucoma: A dynamic optical coherence tomography angiography study. *Frontiers in Medicine*, **9 (1037471)**.

Delporte, M., De Coninck, D., d'Haenens, L., Delporte, M., Verbeke, G., Molenberghs, G., & Matthys, K. (2023). A longitudinal perspective on perceived vulnerability to disease during the COVID-19 pandemic in Belgium. *Health Promotion International*, **38 (2)**, 1-10.

Delporte, M., Delporte, M., Molenberghs, G., Verbeke, G., Demarest, S., & Hoorens, V (2023). Do optimism and moralization predict vaccination? A five-wave longitudinal study. *Health Psychology*, **42 (8)**, 603-614.

Delporte, M., De Witte, D., Demarest, S., Verbeke, G., Molenberghs, G., & Hoorens, V. (2023). Do health beliefs about COVID-19 predict morbidity? A longitudinal study. *Social and Personality Psychology Compass*, **17 (11)**.

Natalia, Y.A., **Delporte, M.**, De Witte, D., Beutels, P., Dewatripont, M., & Molenberghs, G. (2023). Assessing the impact of COVID-19 passes and mandates on disease transmission, vaccination intention, and uptake: a scoping review. *BMC Public Health*, **23 (1)**.

Table of Contents

Abbreviations

Here, we give a list of the most often used abbreviations in the thesis.

ABPA	:	Allergic bronchopulmonary aspergillosis.
ADL	:	Activities of daily living.
BIL	:	Highest bilirubin value since birth.
CGM	:	Congenital malformation.
DME	:	Diabetic Macular Edema.
FEV	:	Forced expiratory volume.
FCCM	:	Full Covariate Conditional Mean.
GEE	:	Generalised estimating equations.
GLM	:	Generalised linear model.
GLMM	:	Generalised linear mixed model.
LMM	:	Linear mixed model.
MANOVA	:	Multivariate analysis of variance.
MAR	:	Missing at random.
MCAR	:	Missing completely at random.
MMSE	:	Mini-Mental State Examination
MNAR	:	Missing not at random.
ML	:	Maximum likelihood.
NSZ	:	Neonatal seizures.
SE	:	Standard error.
SD	:	Standard deviation.
TDC	:	Time-dependent covariate.

List of Tables

List of Figures

Chapter 1

General introduction

1.1 Introduction

In many life-science studies, participants are followed over time with multiple measurements taken. While many statistical models assume independence of observations, this assumption breaks down when individuals are measured repeatedly. Laird and Ware (1982) introduced random-effects models to address this issue, capturing the correlation induced by clustered responses with random effects. Initially, these linear mixed models (LMM) could only handle continuous responses, but later, generalised linear mixed models (GLMM) were developed to accommodate non-continuous responses (e.g., binary responses) (Breslow and Clayton, 1993; Wolfinger and O'Connell, 1993; Engel and Keen, 1994).

Still, in many cases, multiple responses are recorded, allowing research questions about the association between responses and how this association evolves over time. These questions can be addressed by the use of joint models. Verbeke et al. (2014) offers an overview of different methodologies for jointly modelling multivariate longitudinal data. These approaches are also briefly discussed in Section 1.4. One approach is to assume generalised linear mixed models for all responses and allow correlation between the random effects of the responses. Note that this does not necessarily imply the inclusion of shared parameters. The correlations between the random effects of different models provide a glimpse of the association between the responses. However, this latent correlation is calculated between the underlying random effects, in contrast to the manifest correlation, which is on the scale of the observed responses.

A more complex situation arises when longitudinal responses are paired, such as analysing the effect of early antibiotic use on growth in twins where one twin receives antibiotics and the other does not. Another instance is the analysis of longitudinal data including both eyes of the patients. A third example are questionnaires repeatedly administered to mother-child dyads. In these examples,

two sources of correlations are combined: correlations between the members of a pair, and correlations induced by the repeated measurements. Up to this point, a general overview of the methodology for the analysis of paired longitudinal data has been lacking.

Some research questions involve predicting one longitudinal response based on another. One approach is to consider one of the responses as a predictor and treat it as a time-dependent covariate (TDC). A TDC is a predictor whose value can change over time within a study. They offer a straightforward solution, but this method comes with drawbacks. Firstly, the lag relationship needs to be accurately characterised. A famous example is provided by Rizopoulos (2012) in which a study identified a positive but statistically insignificant effect of smoking on the survival of coronary artery disease patients (Cavender et al., 1992). However, this illogical result was caused by not accounting for lagged effects, thus only capturing the immediate influence of smoking on mortality. Most smokers had ceased smoking by the final follow-up before their death, while many surviving patients continued to smoke. Secondly, the TDC must be categorised as endogenous or exogenous. If a response at time t predicts the value of the covariate at a later time $s > t$, the covariate is categorised as endogenous (Diggle, 2002). The presence of an endogenous TDC has important implications in terms of choosing meaningful targets of inference and valid methods of estimation. For example, Qian et al. (2020) have found that in the presence of endogeneous TDC the marginal interpretation of the parameters in a LMM does not hold anymore. Standard methods cannot provide causal summaries and alternative methods should be considered, such as g-computation and marginal structural models with inverse probability of treatment weights. A detailed overview regarding longitudinal data analysis with endogeneous time-dependent covariates can be found in Section 12.5 of Diggle (2002). Thirdly, missing data is almost inevitable, posing challenges particularly when missing values occur for covariates, since ignorability under missingness at random only holds for the responses (Rubin, 1976). Fifth, when the responses, as well as TDC's, are not measured at regular intervals, implementing TDC with lags is not possible.

Due to developments in data storage, large datasets with many longitudinal responses are frequently encountered. This has necessitated new methodologies for analysing high-dimensional datasets. Fieuws and Verbeke (2006) and Fieuws et al. (2006, 2008) developed a pairwise approach for fitting high-dimensional longitudinal data, focusing on latent correlations between the random effects. However, in some applications, the manifest correlations on the scale of the observed responses may be of greater interest. Additionally, deriving predictive models based on previously observed trends in outcomes and/or covariates may be scientifically valuable.

1.2 Motivating case studies

Our research is inspired and motivated by several case studies, spanning the fields of medicine, psychology and sociology. In these fields, longitudinal research is very common. In what follows, these motivating studies are briefly introduced.

1.2.1 Cystic fibrosis society patient registry

The Cystic Fibrosis Society Patient Registry contains data from 43,786 cystic fibrosis patients from 35 countries. The registry has collected data from 2008 until 2016, and contains repeated measurements within patients. Once a year multiple medical parameters from these patients are recorded, next to demographic variables such as gender and age. Medical parameters that are of special interest are the occurrence of allergic bronchopulmonary aspergillosis (ABPA) during the last year, and the forced expiratory volume in one second as a percentage of the average value for healthy people of the same age, height and sex (FEV). The aim is to jointly model FEV and ABPA and investigate the association between both responses. This will be done with both a correlation function and the conditional model of FEV given previous binary ABPA values.

1.2.2 COVID-19 data

Between March 2020 and March 2021, data on COVID-19 was collected at five crucial time points. Thousand Flemish adults were asked about their media habits and how they felt about their vulnerability to disease during the COVID-19 pandemic (De Coninck et al., 2022). This perceived vulnerability was divided into two aspects: "perceived infectability", which gauges beliefs about susceptibility to infections, and "germ aversion", which measures emotional distress in high-risk situations for pathogen transmission (Duncan et al., 2009). In addition, they could rate on a 5-point scale how often they received COVID-related news on multiple media channels. The main interest was to investigate the link between the media consumption and the perceived vulnerability to disease.

1.2.3 Vaccination data

The vaccination data was collected during a six-month period from December 2020 to May 2021 in Belgium, involving approximately 5000 participants. They were surveyed five times about their own vaccination intentions and their perceptions of others' intentions of the same age and gender. They could rate the degree of willingness to vaccinate on a five-point scale. Next, they were also questioned about their optimism about outcomes of a COVID-19 infection and

the perceived effectiveness of vaccination. The main research question is to explore the association between the vaccination intention and various measures of optimism regarding COVID outcomes.

1.2.4 Hip Fracture data

This dataset provides information on the cognitive and functional status of 60 elderly patients who have suffered from a hip fracture. This information is tracked from the moment they are admitted to the hospital until the twelfth day after their surgical procedure, as outlined in Milisen et al. (1998). Our primary goal is to examine the link between cognitive status and functional capacity and how this association evolves over time.

1.2.5 Ophthalmology data

This dataset was collected in the context of a clinical trial comparing the efficacy and safety of three treatments for central-involved Diabetic Macular Edema (DME). With 660 study eyes (220 per group) exhibiting DME, the study spans two years and follow-up visits occur every four weeks. Our main research question involves comparing the evolution of the visual acuity in the eye exhibiting DME with the unaffected eye.

1.3 Univariate models for correlated data

When researchers encounter correlated data, such as repeated measurements on the same subject or data clusters, they often choose random-effects models, also known as mixed models. These models are able to account for variability at multiple levels by incorporating random effects, which are variables that capture the effect of random variation within groups or clusters. For instance, in a study measuring patients' test scores across different hospitals, a mixed model can account for the fact that scores may be more similar within the same hospital compared to different hospitals due to shared procedures or environments. Similarly, in a longitudinal study, these models can capture patient characteristics on an individual level that lead to, for example, systematically higher scores or a steeper progression of the studied characteristic. This approach takes into account the fact that there are intra-cluster correlations and hence the independence assumption is violated. Mixed models are especially useful in longitudinal studies, where ignoring the clustered nature of the data can lead to incorrect conclusions about the relationships being studied.

1.3.1 Linear mixed models

Linear mixed models (LMMs) have gained popularity for analysing longitudinal and hierarchical Gaussian data. Let's assume we have N subjects, and the j-th measurement for subject i is represented by Y_{ij}. The vector $(Y_{i1}, ..., Y_{in_i})$ of all n_i measurements for subject i is denoted by $\boldsymbol{Y}_i$. Using this notation, the model can be written as:

$$\begin{aligned} \boldsymbol{Y_i}|\boldsymbol{b}_i &\sim N(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i, \boldsymbol{\Sigma}_i), \\ \boldsymbol{b}_i &\sim N(\boldsymbol{0}, \boldsymbol{D}), \end{aligned} \tag{1.1}$$

where $\boldsymbol{\beta}$ and $\boldsymbol{b}_i$ are the respectively p- and q-dimensional vectors of unknown regression coefficients and random effects, whereas $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ represent the known corresponding design matrices with dimensions $n_i \times p$ and $n_i \times q$, respectively. $\boldsymbol{D}$ constitutes the $q \times q$ variance-covariance matrix of the random effects. The random effects are often assumed to follow a multivariate normal distribution, but other distributions can also be considered.
The matrix $\boldsymbol{\Sigma}_i$ denotes the $n_i \times n_i$ covariance matrix of the residuals, which depends i through its dimension n_i. $\boldsymbol{\Sigma}_i$ is often chosen to be $\sigma^2\boldsymbol{I}_{n_i}$, implying that all responses are independent given $\boldsymbol{b}_i$ and $\boldsymbol{\beta}$; this is the "conditional independence assumption". However, when this assumption is unrealistic, the vector $\boldsymbol{\epsilon}_i$ can be modelled with a more general residual covariance structure $\boldsymbol{\Sigma}_i$. In this context, the residuals are, for example, decomposed as $\boldsymbol{\epsilon}_i = \boldsymbol{\epsilon}_{(1)i} + \boldsymbol{\epsilon}_{(2)i}$, where $\boldsymbol{\epsilon}_{(2)i}$ represents the component of serial correlation (Diggle, 2002).

A property of the LMM is that the parameters of the mean of the conditional model and the marginal model are exactly the same. This is because $E[Y_{ij}] = E[E(Y_{ij}|\boldsymbol{b}_i)] = \boldsymbol{X}'_{ij}\boldsymbol{\beta}$. Nevertheless, the marginal model is defined as:

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}^*_i, \tag{1.2}$$

where the residuals $\boldsymbol{\epsilon}^*_i$ are, by definition, correlated and normally distributed around $\boldsymbol{0}$ with variance $\boldsymbol{V}^*_i$. Consequently, the distribution of the response is:

$$\boldsymbol{Y}_i \sim N(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{V}^*_i),$$

with

$$\boldsymbol{V}^*_i = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}'_i + \boldsymbol{\Sigma}_i.$$

For more detailed information about linear mixed models, see Verbeke and Molenberghs (2000).

1.3.2 Generalised linear mixed models

The GLMM can be seen as the generalisation of the LMM for non-continuous data (Breslow and Clayton, 1993; Wolfinger and O'Connell, 1993; Engel and Keen, 1994). We make here the conditional independence assumption, i.e., that the Y_{ij} are independent conditional on the random effects. In addition, we assume that the conditional distribution of each Y_{ij} given the random effects belongs to the exponential familiy of distributions and

$$\text{Var}(Y_{ij}|\boldsymbol{b}_i) = v\{E(Y_{ij}|\boldsymbol{b}_i)\}\phi,$$

where $v(.)$ is a known variance function, a function of the conditional mean $E(Y_{ij}|\boldsymbol{b}_i)$. Further, the conditional mean of Y_{ij} depends on the fixed and random effects via the following linear predictor:

$$g\{E(Y_{ij}|\boldsymbol{b}_i)\} = \eta_{ij} = \boldsymbol{X}_{ij}\boldsymbol{\beta} + \boldsymbol{Z}_{ij}\boldsymbol{b}_i$$

for some known link function $g(.)$. Again the random effects are assumed to follow a multivariate distribution, which is often the multivariate normal distribution with zero mean and $q \times q$ covariance matrix $\boldsymbol{D}$. $\boldsymbol{X}_{ij}$ is the $1 \times p$ vector that contains values for the covariates of the p-dimensional fixed effects vector $\boldsymbol{\beta}$. Next, $\boldsymbol{Z}_{ij}$ is the $1 \times q$ design vector for the q random effects $\boldsymbol{b}_i$. This specification can be used for a broad class of generalised linear models. It is easy to see that the LMM is a special case of the GLMM, using the identity link as link function. We will now discuss two special cases for binary and ordinal data, both with the probit link function.

Random-effects probit model for binary outcomes

The random-effects probit model is specified as

$$\Phi^{-1}[P(Y_{ij} = 1|\boldsymbol{b}_i)] = \boldsymbol{X}_{ij}\boldsymbol{\beta} + \boldsymbol{Z}_{ij}\boldsymbol{b}_i, \tag{1.3}$$

where $\Phi(\cdot)$ is the cumulative normal distribution function and $\boldsymbol{b}_i \sim N(\boldsymbol{0}, \boldsymbol{D})$. As stated in the previous subsection, in the linear mixed model, the conditional parameters are equal to the marginal parameters. In the generalised linear mixed model, this is not the case. Molenberghs et al. (2010) derived closed-form expressions for the marginal generalised linear mixed model, shown in their supplementary materials (their Appendix D). The marginal density is the following:

$$P(Y_{i1} = 1, Y_{i2} = 1, \ldots, Y_{ip_i} = 1) = \Phi(\boldsymbol{X}_i\boldsymbol{\beta}; \boldsymbol{L}_i^{-1}), \tag{1.4}$$

with

$$\begin{aligned} \boldsymbol{L}_i &= \boldsymbol{I} - \boldsymbol{Z}_i\boldsymbol{M}_i^{-1}\boldsymbol{Z}_i', \\ \boldsymbol{M}_i &= \boldsymbol{D}^{-1} + \boldsymbol{Z}_i'\boldsymbol{Z}_i. \end{aligned}$$

Random-effects probit model for ordinal outcomes

A random-effects ordinal regression model with a probit link function can be implemented for the analysis of clustered or repeated measures of an ordinal response. A threshold concept is used, under the assumption of the presence of an underlying continuous response that determines the observed ordered response categories. For the d categories, a series of threshold values $\gamma_1, \gamma_2, ..\gamma_{d-1}$ are assumed. A response falls in category c if the latent response Y_{ik}^* exceeds the threshold value γ_{c-1}, but not γ_c. The hierarchical mixed model for the latent response at time j of this latent response of subject i is

$$Y_{ij}^* = \boldsymbol{X}_{ij}\boldsymbol{\beta} + \boldsymbol{Z}_{ij}\boldsymbol{b}_i + \epsilon_{ij},$$

where ϵ_{ij} denotes the residual and is assumed to be independently normally distributed with mean 0 and variance σ^2.
The probability that a response at time j for subject i falls into category c can be derived from the model for Y_{ik}^* and equals

$$P(Y_{ij} = c) = \Phi\left(\frac{\gamma_c - \zeta_{ij}}{\sigma}\right) - \Phi\left(\frac{\gamma_{c-1} - \zeta_{ij}}{\sigma}\right),$$

where $\zeta_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{b}_i$. In analogy, the probability that response j of subject i is less than or equal to category c is equal to

$$P(Y_{ij} \leq c) = \Phi\left(\frac{\gamma_c - \zeta_{ik}}{\sigma}\right).$$

The choice of the unit and the origin of ζ is arbitrary (Hedeker and Gibbons, 1994), but is often set to 1 and 0, respectively.

1.3.3 Time-dependent covariates

Time-dependent covariates (TDC's) are powerful means to assess whether the "current" value (depending on the lag) of a covariate is linked to the current value of the outcome of interest. However, we already touched on several issues with TDC's in Section 1.1. The lag relationship between the covariate and the response should be carefully reviewed, as this has important implications. Even more, when the data is collected at irregular intervals, the use of lags hinders the application of TDC's.
Second, it is important to examine whether there are feedback loops at play. When the outcome at time t, denoted with Y_{it}, predicts the covariate value at time $s > t$, the TDC is called endogeneous (Diggle, 2002). In contrast, exogeneity implies that the Y_{it} is conditionally independent of all future covariates given

both past outcomes and covariates. In the situation of endogeneous TDC's, many statistical properties do not hold anymore, such as the marginal interpretation of parameters in a LMM (Qian et al., 2020).
Another important consideration relates to the distinction between partly conditional and fully conditional regression models. While both conditional expectations might be of interest, it is important to distinguish between them as their parameters have different interpretations and require specific assumptions for valid covariance-weighted estimation, such as with LMM or GEE models. In what follows, we will assume a common set of discrete follow-up times, $t = 1, 2, ..., T$, with a clearly defined last study measurement timepoint T.
For instance, if we are interested in the relationship between a response on time t and a single time-dependent covariate X at the same moment, we use $E(Y_{it}|X_{it})$ to investigate if the average response varies with simultaneous TDC values. Alternatively, if we hypothesise a lag between the TDC and the response, we might analyse $E(Y_{it}|X_{i,t-k})$ for some lag k. We can also consider the entire exposure history by modelling $E(Y_{it}|X_{i1}, X_{i2}, \ldots, X_{i,t-1})$, possibly focusing on cumulative covariate values such as, for example, pack years. Finally, the average response given the entire covariate process can be modelled as $E(Y_{it}|X_{is}, s = 1, 2, \ldots, T)$.
Pepe and Couper (1997) classified $E(Y_{it}|X_{is}, s = 1, 2, \ldots, T)$ as the full covariate conditional mean (FCCM) and $E(Y_{it}|\text{subset}(X_{i1}, \ldots, X_{in_i}))$ for non-exhaustive subsets as partly conditional means. The cross-sectional mean $E(Y_{it}|X_{it})$ and expectations involving single lagged covariates or entire covariate histories are examples of partly conditional means.
When the covariate process is endogenous, the FCCM $E(Y_{it}|X_{is}, s = 1, 2, \ldots, T)$ can depend on some or all covariates X_{is}. However, for exogenous covariates $E(Y_{it}|X_{is}, s = 1, 2, \ldots, T) = E(Y_{it}|X_{i,t-1}, X_{i,t-2}, ..., X_{i1})$. Assuming only the k most recent covariate values predict the response alters this further to $E(Y_{it}|X_{i,t-1}, X_{i,t-2}, \ldots, X_{i,t-k})$. Thus, under assumptions of exogeneity and a finite covariate lag, the partly conditional mean may equal the FCCM. However, if the covariate process is endogenous, the FCCM does not equal the partly conditional mean. Section 12.5 of Diggle (2002) discusses methods of analysis in the presence of these endogeneous TDC's, such as g-computation and marginal structural models using inverse probability of treatment weights.
Related to this topic is the full covariate conditional mean assumption (Pepe and Anderson, 1994): in order to make sure that the estimates are unbiased it is necessary to assume

$$\mu_{it} \equiv E(Y_{it}|X_{it}) = E(Y_{it}|X_{i1}, X_{i2}, \ldots, X_{iT}).$$

If this assumption is not valid, but interest lies in the concurrent relationship of Y_{it} and X_{it} then GEE with the working independence assumption should be

used in order to avoid biased regression estimates. While Pepe and Anderson (1994) focus on GEE, this assumption is important for all longitudinal data analysis methods including likelihood-based methods such as LMM and GLMM, and when lags are used. A more detailed discussion can be found in Diggle (2002).

1.4 Multivariate models for correlated data

Up to this point, we have discussed models that are useful for analysing a single outcome that is measured repeatedly for each individual. However, in practice, it is common to encounter situations where multiple outcomes are simultaneously recorded and measured repeatedly over time within each subject. These outcomes can be of the same or different types, and various scientific questions may arise depending on the specific application. Several approaches have been developed for the joint analysis of multiple longitudinal outcomes. A more in-depth review of these methods can be found in Verbeke et al. (2014). We will denote the two longitudinal outcome vectors with $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$. The multivariate approach for multiple outcomes is then a straightforward extension.

A first approach is to specify directly the joint density $f(\boldsymbol{y}_1, \boldsymbol{y}_2)$ of $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$. In this approach, assumptions are required for the marginal association of $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$, which becomes cumbersome when the outcomes are of different types, the data is highly unbalanced or more than two outcomes are jointly modelled. However, advantages of the method are the availability of direct inferences and the symmetric treatment of the outcomes. A more detailed discussion can be found in Molenberghs and Verbeke (2005).

A second approach that avoids the direct specification of the joint distribution is the conditional modelling approach. Here the joint density is factorised as a product of the marginal and conditional densities

$$\begin{aligned} f(\boldsymbol{y}_1, \boldsymbol{y}_2) &= f(\boldsymbol{y}_1|\boldsymbol{y}_2) f(\boldsymbol{y}_2) \\ &= f(\boldsymbol{y}_2|\boldsymbol{y}_1) f(\boldsymbol{y}_1). \end{aligned}$$

A drawback of this method is that the conditional density requires careful consideration of the association between $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$, where one response acts as a time-dependent covariate, discussed in Section 1.3.3. Unlike the previous approach, the conditional models do not yield marginal inferences. Additionally, when dealing with high-dimensional data, numerous possible factorizations exist, unlike in the two-dimensional case.

Shared parameter models are a third option, where the random effects generate an association structure between the repeated measurements within the outcome as well as the repeated measurements between the outcomes. The joint density of the responses is then equal to

$$\begin{aligned} f(\boldsymbol{y}_1, \boldsymbol{y}_2) &= \int f(\boldsymbol{y}_1, \boldsymbol{y}_2|\boldsymbol{b}) f(\boldsymbol{b}) d\boldsymbol{b} \\ &= \int f(\boldsymbol{y}_1|\boldsymbol{b}) f(\boldsymbol{y}_2|\boldsymbol{b}) f(\boldsymbol{b}) d\boldsymbol{b}, \end{aligned} \tag{1.5}$$

in which $f(\boldsymbol{b})$ represents the random-effects density, which is often assumed to be (multivariate) normal. This model assumes that the random effect $\boldsymbol{b}$ represents underlying characteristics of the subject that governs both outcome processes. As a consequence, the responses are conditionally independent given the random effects. An advantage of this approach is that the responses do not have to be of the same nature and extension to the high-dimensional case is straightforward. In addition, the interpretation of the parameters in the submodels of the individual responses is not affected. When the responses have different measurement units or are of different types, an inflation factor can be included. A limitation of the shared parameter model is that it can enforce a restrictive association between the responses (Iddi and Molenberghs, 2012). For instance, if the model includes only a random intercept, this single random intercept must capture the associations within the sequences of the first response as well as the associations within the sequences of the second response. When the number of analysed responses increases, the association becomes more restrictive.

The shared-parameter approach can be generalised to the random-effects model approach. In the shared-parameter model, the same random effect is used in the different submodels. This is equivalent to using different random effects for each response, but where the random effects are perfectly correlated. In essence, this assumption of a perfect correlation can be relaxed. The mixed model then uses a q-dimensional random effects vector $\boldsymbol{b}$ which contains all response-specific random effects, and is assumed to follow a multivariate normal distribution with a zero mean and a covariance matrix $\boldsymbol{D}$. The matrix $\boldsymbol{D}$ reflects the correlation among repeated measures of the same response and between the measurements of different responses. In analogy with the shared parameter model, we assume that the responses are independent given the random effects, implying that the random effects entirely account for the correlation between the responses. As a result, the joint density of the responses, conditional on the random effects, is the product of the conditional densities of each response. A more detailed overview of shared parameter models and joint random effects models can be found in Chapter 13 of Fitzmaurice et al. (2008).

1.5 Estimation strategies

1.5.1 Maximum likelihood

Suppose we have L longitudinal outcomes, possibly of different types. For each outcome, a separate random-effects model can be specified. As outlined in the previous section, the joint density can be constructed by defining a joint distribution of random-effects, with $\boldsymbol{b}_i$ representing the vector of all random effects for all models.
Assuming the subjects are independent, it follows from the conditional independence of $Y_{1,i}, Y_{2,i}, \ldots, Y_{L,i}$ given $\boldsymbol{b}_i$ that the log-likelihood contribution for subject i to the full joint mixed model is:

$$l_i(\boldsymbol{\theta}^* \mid Y_{1i}, Y_{2i}, \ldots, Y_{Li}) = \log \left(\int \prod_{l=1}^{L} f_{li}(Y_{li} \mid \boldsymbol{b}_i) f(\boldsymbol{b}_i) d\boldsymbol{b}_i \right)$$

where $\boldsymbol{\theta}^*$ is the vector of all parameters (both fixed effects and covariance parameters). Except for special cases (e.g., with linear models), the integral in this expression cannot be calculated analytically and requires numerical approaches. In this thesis, we will use numerical integration, specifically adaptive Gaussian quadrature, which has been implemented in the SAS procedure NLMIXED. Due to the potentially high dimension of the random effects in $\boldsymbol{b}_i$, computation time can increase considerably with an increasing number of outcomes.

1.5.2 Pairwise fitting technique

As mentioned previously in Section 1.5.1, fitting joint random-effects models by maximising the likelihood becomes cumbersome and for large numbers of outcomes even infeasible. In these cases, an alternative approach based on pseudo-likelihood methodology can be used. The principal idea of this approach is to replace a computationally challenging joint density by a simpler function of appropriate factors (Molenberghs et al., 2011b)
Fieuws and Verbeke (2006) introduced the pairwise fitting approach, a special case of the pseudo-likelihood estimation. This approach is tailored to parameter estimation in multivariate data, with possibly outcomes of different types. In this method, each bivariate model is fitted separately and independently, rather than optimising the likelihood of the entire multivariate model simultaneously. This is done by maximising the likelihoods $l_{rs}(\boldsymbol{\theta}_{r,s} | \boldsymbol{Y}_r, \boldsymbol{Y}_s)$ for each pair (r, s), where $\boldsymbol{\theta}_{r,s}$ is the parameter vector for that specific pair. This approach significantly reduces the computational burden in comparison to maximum likelihood estimation since not all parameters are estimated simultaneously. Some parameters of the original

multivariate model are estimated multiple times due to their presence in more than one pairwise model, for example the fixed intercepts of the responses. The final parameter estimates are then obtained by averaging the estimates from all response pairs.

1.6 Thesis contribution

Researchers often face datasets with multiple responses and are interested in the association between these responses. In the situation of multiple longitudinal responses, the use of time-dependent covariates (TDC's) to investigate the association between the responses has considerable limitations, for example when endogenous TDC's are present or when the data is measured at irregular intervals. When the longitudinal responses are paired, the analysis becomes even more complex, and multiple methods, each with their own drawbacks are possible. In the case of multivariate binary data, understanding the correlations between different responses is hindered by the lack of computational tools. We will address these gaps in the research by formulating alternative models based on joint modelling and exploring pseudo-likelihood an alternative estimating approach.
The thesis contribution is based on five parts:

–1– Joint longitudinal modelling of binary and continuous data

The first major contribution is in the development of methodologies for jointly modelling longitudinal binary and continuous responses using generalised mixed models. This approach models the association between the different responses by allowing correlations between the subject-specific random effects. Conclusions about the association between the different responses are traditionally limited to examining these latent correlations. This work extends this by deriving closed-form formulas for manifest correlations, which reflect the correlations between the responses on the same scale as they appear in the data. Additionally, a marginal conditional model is constructed, which leans itself to predictions of one response based on another and potentially a subvector of the history of the predicted response. Prediction and confidence intervals are also derived for inference.

–2– Joint longitudinal modelling of multiple binary and continuous responses

Building on the foundational work of the previous chapter, this chapter extends the approach to high-dimensional datasets where multiple binary and continuous responses are analysed. In addition, we apply pseudo-likelihood methodology since computational issues are inevitable in the high-dimensional case.

–3– Joint longitudinal modelling of ordinal and continuous responses

This chapter involves the formulation of a normal-ordinal (probit) joint model to analyse continuous and ordinal longitudinal responses simultaneously. Closed-form formulas are derived to estimate the model-based correlations between responses on their original scale, providing a more intuitive interpretation of the correlations. In addition, we derived conditional expected values and probabilities of one response conditional on the other. The model is also extended to high-dimensional case.

–4– Analysis of paired longitudinal data

This chapter addresses the scenario of paired continuous longitudinal data, such as the analysis of longitudinal data from both eyes of patients. The complexity arises from combining correlations between the members of a pair and the correlations induced by repeated measurements within the pair. The chapter presents an overview and comparison of the methodology for analysing such paired longitudinal data, underscoring the importance of accounting for intra-pair correlations and missing data mechanisms.

–5– Analysis of multivariate probit models

The final chapter presents a faster computational approach for fitting multivariate probit models, in comparison to maximum likelihood. This is a potential solution to computational issues in fitting multivariate categorical data. This chapter introduces a pairwise fitting technique within the pseudo-likelihood framework, significantly improving computational efficiency and convergence.

Summary

In summary, this thesis advances the field of longitudinal data analysis by developing techniques to better address research questions based on multivariate longitudinal data. These contributions are particularly impactful in biomedical research, facilitating a deeper understanding of the complex relationships between various health outcomes over time.

1.7 Outline of thesis

This dissertation is composed of 7 chapters. Chapter 1 provides a general introduction of the objectives of this research, alongside the motivating case studies and a brief overview of existing longitudinal data methodology.
The next five chapters present the thesis' contribution. Chapter 2 explores the binary-continuous model, which is extended to the high-dimensional case in Chapter 3. Chapter 4 alters the model to the ordinal-continuous case. 3. Chapter 5 investigates the existing methodology for paired continuous longitudinal data. Lastly, Chapter 6 examines a faster computational approach for multivariate probit models.
The final chapter, Chapter 7, summarises the main findings, outlines the clinical implications, acknowledges study limitations, and proposes directions for future research.

Chapter 2

A joint normal-binary (probit) model

This chapter is based upon:

Delporte, M., Fieuws, S., Molenberghs, G., Verbeke, G., Wanyama, S.S., Hatziagorou, E.,& De Boeck, C. (2022). A joint normal-binary (probit) model. *International Statistical Review*. **90(S1)**, S37-S51.

The appendix is available via `margauxdelporte.github.io/Chapter2.pdf`.

Abstract

In biomedical research, often hierarchical binary and continuous responses need to be jointly modelled. In joint generalised linear mixed models this can be done with correlated random effects, which allows examining the association structure between the various responses and the evolution of this association over time. In addition, the effect of covariates on all outcomes can be assessed simultaneously. Still, investigating this association is often limited to examining the correlations between the responses on an underlying scale. In addition, the interpretation of this hierarchical model is conditional on the subject-specific random effects. This paper extends this approach and shows how manifest correlations can be computed, i.e., the associations between the observed responses. Further, a marginal model is formulated, in which the interpretation is no longer conditional on the random effects. In addition, prediction intervals are derived of one subvector of responses conditional on the other. These methods are applied in a case study of the lung function and allergic bronchopulmonary aspergillosis in patients with cystic fibrosis.

2.1 Introduction

Modelling longitudinal responses is of great interest in biomedical research. In many areas of (bio)-medical sciences longitudinal studies are conducted, which measure attributes of participants repeatedly over time. Many models are developed to model these kind of clustered responses, but currently the most popular amongst them are random-effects models. The first model in this approach was introduced by Laird and Ware (1982). This model was a linear mixed model to model continuous clustered responses. This approach was later extended to noncontinuous data with generalised linear mixed models (Breslow and Clayton, 1993; Wolfinger and O'Connell, 1993; Engel and Keen, 1994). In these articles the focus was on a separate analysis per response.

In longitudinal research often more than one response is recorded. Hence, a joint analysis of multiple responses is also of interest. A first possible approach is to include one or more responses as a time-dependent covariate in the model of one of the responses. An advantage is the ease of programming, but this approach has several drawbacks. A first drawback is that the lag relationship needs to be correctly characterized (Diggle, 2002). This also implies that the effect of the time-dependent covariate at a certain time point can be assessed on only one time point of the response. A second drawback is the correct specification of the time-dependent covariate as endogenous or exogenous, which has important implications. Endogenous covariates can be influenced by the previous response values (Diggle, 2002) and impose additional difficulties. Qian et al. (2020) found, for example, that the marginal interpretation of coefficients in a linear mixed

model does not hold anymore when endogenous covariates are present. Next, missing data in the responses is likely to occur in longitudinal studies due to drop-out, amongst other reasons. Treating one of the responses as a covariate imposes additional problems in terms of missing data. The principle of ignorability holds for direct-likelihood inferences under missingness at random Rubin (1976). However, this principle is only true for missingness in the response, but not for missingness in the covariates. A last drawback is the possibility that a time-dependent covariate serves as an intermediate variable, which is a link in the causal pathway between a covariate and the response (Diggle, 2002). When the model corrects for an intermediate variable by including it in the model, the effect of the covariate mediated through these variables is lost (Diggle, 2002).
In contrast, joint models offer a different approach with several advantages. These advantages include the ability to answer research questions that take several or all responses simultaneously into account or assess the effect of a covariate on all responses simultaneously. In addition, joint models can not only assess simultaneous effects of covariates on the repeated instances of the same sequence, but also across sequences. The association between various responses and the evolution of this association over time can also be scrutinized with joint models. The extension from two univariate random-effects models to a joint model is done with the random effects, which are allowed to be correlated for this purpose. In this way, the association between the different responses can be captured. For example, Chakraborty et al. (2003) fitted a joint model for two continuous responses; HIV-1 RNA concentration in blood and semen. The joint model allowed the researchers to obtain the correlations for the two responses. In addition, it enabled the comparison of the latter correlations between the antiretroviral and the no-antiretroviral group.
The different longitudinal responses in biomedical research are often of different types. There are a lot of different techniques available to tailor the approach to a specific case, which results in a variety of modelling applications in the literature. Ivanova et al. (2016) give an overview of joint generalised mixed models for different types of responses. They present, amongst others, a case study for the combination of a continuous and ordinal response. Efendi et al. (2013) formulate joint models for longitudinal continuous and time-to-event responses. They also include a marginal model in which the parameters have a marginal interpretation.
Molenberghs and Verbeke (2005) formulate the joint generalised linear mixed model for a binary and continuous response. In an article by Iddi and Molenberghs (2012) a case study is conducted where the longitudinal continuous visual acuity and the binary vision-loss are jointly modelled. In this article, different methods are applied. Firstly two hierarchical models are fitted; the shared-parameter model and the correlated intercepts model. In addition, the latter models are

fitted with the joint marginal multilevel model (JOMM), in which the parameters have a marginal interpretation. In addition, those authors conducted a case study with two longitudinal binary responses: the occurence of HIV and HCV in serological data.
In some studies, more than two responses are included in the joint model. Morrell et al. (2012) for example combined three longitudinal continuous responses in a joint mixed-effects model; the longitudinal changes of prostate-specific antigen, a free testosterone index and body mass index are used to predict prostate cancer. In a high-dimensional context, many more than three responses are included in the joint model, which can cause computational problems. Fieuws and Verbeke (2006) tackle this issue with a pairwise bivariate modelling approach for continuous data. They extended this pairwise approach to binary data (Fieuws et al., 2006) and the combination of various data types (Fieuws et al., 2008).
An important benefit of joint models is that they allow the examination of the correlations within and between the response vectors. It is in this context, important to distinguish the two kinds of correlations. Firstly, latent correlation is the correlation between responses at a latent scale. In this context the latent correlations are the correlations between the random effects of the different responses. In contrast, the manifest correlation is the correlation between the response values that are actually observed. The latter measure imposes a more convenient measure of the associations and can hence be of more scientific interest. The latent and manifest correlations are not equal to the manifest correlations for neither normal or non-normal responses. Hence, a formula for the manifest correlation should be derived.
The organisation of this paper is as follows. In Section 2.2 a dataset is introduced, which will be analysed in Section 2.5. Methodology for existing marginal and hierarchical models for continuous and binary longitudinal responses and joint binary-continuous responses are reviewed, and our marginal model for the binary-normal case is presented in Section 2.3. Lastly, the estimation method is formulated in Section 2.4 and concluding remarks are offered in Section 2.6.

2.2 Cystic fibrosis data

The dataset at hand contains data from Belgian cystic fibrosis patients from 2008 to 2016. The patients ($n = 1291$) were measured once a year, for at most 9 years. At one time point multiple medical parameters are obtained, next to demographic variables such as age. The medical parameters are, amongst others, whether allergic bronchopulmonary aspergillosis (ABPA) occurred during the last year and the forced expiratory volume (FEV) in one second as a percentage of the average value for healthy people of the same age, height and sex. It is important

to note that 28% of the Belgian patients had minimum one occurrence of ABPA over the course of the study. In addition, the disease is difficult to diagnose and can evolve from a acute diagnosis to a chronic disease. The prevalence of ABPA in the Belgian subjects is shown in Table 2.1. Table 2.1 also shows the large amount of drop-out in the study, which is amongst other reasons due to transplanted patients. The descriptive statistics of FEV from the Belgian patients at each time point are shown in Table 2.2. Time is defined here as the time the subject is included in the study, as opposed to the years passed since 2008. The aim is to jointly model FEV and ABPA and investigate the association between both responses. This will be done with both a correlation function and the conditional model of FEV given previous binary ABPA values.

Table 2.1*: Prevalence of ABPA at each time point in the study of the Belgian patients.*

	0	1	2	3	4	5	6	7	8
ABPA present	245	83	66	69	45	57	64	52	45
ABPA absent	1046	975	909	837	786	713	637	570	480

Table 2.2*: Descriptive statistics of FEV at each time point in the study of the Belgian patients.*

	0	1	2	3	4	5	6	7	8
Mean	84.01	83.0	82.1	81.5	80.8	79.2	77.4	76.5	75.6
Median	87.1	87.2	85.4	84.8	83.7	82.3	80.7	79.5	77.7
SD	23.5	23.2	22.9	23.4	23.2	23.1	22.9	22.7	22.4

2.3 Methodology

2.3.1 Model for a single longitudinal continuous response

The linear mixed model is an appropriate statistical model for clustered continuous responses. Random effects with a normal distribution are added to the model to account for the correlation induced by the clustering of responses. These random effects induce the intracluster correlation in repeated measured continuous responses. Consider a vector of continuous responses $\boldsymbol{Y}_i$. For the sequence in measurements in cluster $i, j = 1, ..., n_i$,

$$Y_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{b}_i + \epsilon_{ij} \tag{2.1}$$

is the hierarchical linear mixed model, where $\boldsymbol{\beta}$ is the p_i-dimensional vector of fixed-effects parameters. Two sets of covariates are incorporated into the model,

$\boldsymbol{x}_{ij}$ and $\boldsymbol{z}_{ij}$. The covariates $\boldsymbol{X}_i$ is $n_i \times p_i$ dimensional and is associated with the p_i-dimensional vector of fixed-effects parameters, $\boldsymbol{\beta}$. The other $n_i \times q$ dimensional set of covariates, $\boldsymbol{Z}_i$, is associated with the q-dimensional random effects $\boldsymbol{b_i} \sim \mathrm{N}(\boldsymbol{0},\boldsymbol{D})$. The covariance matrix $\boldsymbol{D}$ denotes the level of heterogeneity of the subjects. The residuals $\boldsymbol{\epsilon}_i \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_i)$ can be assumed independent given the random effects $\boldsymbol{b_i}$. Alternatively, correlation between the residuals can be allowed by decomposing $\boldsymbol{\epsilon}_i = \boldsymbol{\epsilon}_{(1)i} + \boldsymbol{\epsilon}_{(2)i}$, resulting in a different model for $\boldsymbol{\Sigma}_i$:

$$\begin{aligned} \boldsymbol{Y}_i &= \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_{(1)i} + \boldsymbol{\epsilon}_{(2)i} \qquad (2.2) \\ \boldsymbol{b_i} &\sim N(\boldsymbol{0}, \boldsymbol{D}) \\ \boldsymbol{\epsilon}_{(1)i} &\sim N(\boldsymbol{0}, \sigma_i^2 \boldsymbol{I}_{n_i}) \\ \boldsymbol{\epsilon}_{(2)i} &\sim N(\boldsymbol{0}, \tau^2 \boldsymbol{H}_i) \\ \boldsymbol{b_1}, ..., \boldsymbol{b_{n_i}}, \boldsymbol{\epsilon}_{(1)1}, ..., \boldsymbol{\epsilon}_{(1)n_i} &\quad \text{independent,} \end{aligned}$$

where the correlation matrix $\boldsymbol{H}_i$ of $\boldsymbol{\epsilon}_{(2)i}$ is presumed to have (j, k) elements of the form $h_{ijk} = g(|t_{ij} - t_{ik}|)$ for a decreasing function g$(\cdot)$ with g$(0) = 1$.
Since $E(\boldsymbol{Y}_{ij}) = E[E(\boldsymbol{Y}_{ij}|\boldsymbol{b}_i)] = \boldsymbol{x}_{ij}\boldsymbol{\beta}$, the marginal and conditional parameters are exactly equal to each other. Still, one can postulate the following marginal model:

$$Y_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \epsilon^*_{ij}. \qquad (2.3)$$

In contrast to the hierarchical model, the residuals $\boldsymbol{\epsilon}^*_i \sim N(\boldsymbol{0}, \boldsymbol{V}^*_i)$ are by definition correlated. In this way the covariance parameters in $\boldsymbol{V}^*_i$ induce the intracluster correlation due to the repeated measurements. As a result, the distribution of the response is the following: $\boldsymbol{Y}_i \sim N(\boldsymbol{X\beta}, \boldsymbol{V}^*_i)$. Note that the marginal model is a special case of the hierarchical model above in (2.1) and (2.2). From the hierarchical linear mixed model the marginal model is deduced with $\boldsymbol{\epsilon}^*_i \sim N(\boldsymbol{0}, \boldsymbol{V}^*_i)$ and $\boldsymbol{V}^*_i = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}'_i + \boldsymbol{\Sigma}_i$. Verbeke and Molenberghs (2003) state that the marginal model is less restrictive in the covariance parameters compared to the hierarchical model. The reason is that the marginal model removes the restriction of positive definiteness on the $\boldsymbol{D}$ and $\boldsymbol{\Sigma}_i$ matrices. Instead, the only restriction is that the $\boldsymbol{V}^*_i$ matrix is positive definite.

The estimation for the mixed model parameters is often done with maximum likelihood and restricted maximum likelihood (Verbeke and Molenberghs, 2000). Both methods are implemented in statistical software such as SAS and R.

2.3.2 Model for a single longitudinal binary response

For repeated measures of binary responses, a generalised linear mixed model is in common practical use. This model extends the classical generalised linear model with random effects. Both the logit link or the probit link are appropriate, but in this paper only the probit link will be considered. The latter lends itself to closed form expressions and back-transformations exist when interest is in the former. In analogy with Section 2.3.1, the responses Y_{ij} are conditional upon the q-dimensional random effects $\boldsymbol{b_i} \sim N(\mathbf{0},\boldsymbol{D})$ independently distributed as follows:

$$f_i(Y_{ij}|\boldsymbol{b}_i, \boldsymbol{\beta}, \phi) = \exp\{\phi^{-1}[Y_{ij}\lambda_{ij} - \psi(\lambda_{ij})] + c(Y_{ij}, \phi)\}, \tag{2.4}$$

with

$$\begin{aligned} \Phi^{-1}[\psi'(\lambda_{ij})] &= \Phi^{-1}(\mu_{ij}) = \Phi^{-1}(P(Y_{ij} = 1)) \\ &= \Phi^{-1}(E[Y_{ij}|\boldsymbol{b}_i, \boldsymbol{\beta}]) \\ &= \boldsymbol{x}_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{b}_i, \end{aligned} \tag{2.5}$$

where $\Phi(\cdot)$ is the cumulative normal distribution. In this case the scale parameter, ϕ, is equal to 1. Unlike in the linear mixed model, the conditional parameters are not equal to the marginal parameters. To obtain the parameters with a marginal interpretation direct marginal specification by generalised estimating equations (GEE) can be used. This method was introduced by Zeger et al. (1988). In addition, Molenberghs et al. (2010) computed closed-form expressions for the marginal generalised linear mixed model with a probit link from the conditional density. The derivation can be found in their supplementary materials (their Appendix D). The marginal model is the following:

$$P(Y_{i1} = 1, Y_{i2} = 1, \ldots, Y_{ip_i} = 1) = \Phi(\boldsymbol{X}_i\boldsymbol{\beta}; \boldsymbol{L}_i^{-1}), \tag{2.6}$$

with

$$\boldsymbol{L}_i = \boldsymbol{I} - \boldsymbol{Z}_i(\boldsymbol{D}^{-1} + \boldsymbol{Z}'_i\boldsymbol{Z}_i)^{-1}\boldsymbol{Z}'_i. \tag{2.7}$$

2.3.3 Joint model for continuous-binary responses

Consider a continuous response Y_{1ij} and a binary response Y_{2ik}, where the subscripts denote the j^{th} and k^{th} measurement on the i^{th} subject for respectively the continuous and binary measurement. The goal of a joint model is to describe the joint density of $(\boldsymbol{Y}_{1i}, \boldsymbol{Y}_{2i})$. A flexible approach to attain this goal is to model both responses with a mixed model and let the random effects correlate. In this way it is not necessary that both responses are of the same nature. In addition, the interpretation of the parameters in the submodels of the individual

responses is not altered. Furthermore, the covariates of both submodels may, but do not have to be the same. Joint mixed models are described in Fieuws et al. (2006). In addition, this approach has been applied in the context of two responses of an ordinal and continuous nature (Ivanova et al., 2016). The setting of a hierarchical joint model for a combined binary and continuous response is described by Molenberghs and Verbeke (2005). However, they did not formulate the marginal probit-normal model. In the hierarchical model there are two vectors of random effects, $\boldsymbol{b}_{1i}$ and $\boldsymbol{b}_{2i}$ for, respectively, the continuous response Y_{1ij} and the binary response Y_{2ik}. In addition, it is assumed that the random effects $\boldsymbol{b_i} = (\boldsymbol{b_{1i}}, \boldsymbol{b_{2i}}) \sim N(\boldsymbol{0}, \boldsymbol{D})$. Hence, the variance-covariance matrix of the random effects, $\boldsymbol{D}$, provides the means to model the correlation within the repeated measurements of the same response as well as the correlations between the vectors of measurements of the different responses. It is assumed that correlation between the random effects completely accounts for the association between both responses. As a consequence, it is assumed that the responses are independent conditional on the random effects. Hence, the conditional joint density of the two responses becomes the product of the density of the individual responses conditional on the random effects.
In most cases, no specific structure is assumed for the variance-covariance matrix of the random effects. Still, a special case of the joint longitudinal model is the shared-parameter model. In this class of models a restriction is imposed on the $\boldsymbol{D}$ matrix; the random effects should be perfectly correlated (Molenberghs and Verbeke, 2005). It is important to ensure the meaningfulness of the models when there are shared parameters between models for responses of different types, which also have different measurement units. When a random effect is shared across responses, whether or not the measurement units are different, an inflation factor can be useful. A second drawback of the shared parameter model is that it can impose a more restrictive kind of association between the responses (Iddi and Molenberghs, 2012). When, for example, only a random intercepts is included in the shared-parameter model, a single random intercept has to capture both the associations between the sequences of the first response and the association between the sequences of the second response.

Above the joint density of an n_i-dimensional continuous vector $\boldsymbol{Y}_{1i}$ and the p_i-dimensional binary vector $\boldsymbol{Y}_{2i}$ conditional on the random effects is described. Integrating out the random effects results in the marginal joint density. As shown in Online Appendix A, the marginal joint density for the joint probit-normal model is the following:

$$f(\boldsymbol{y}_{1i}, \boldsymbol{y}_{2i} = \boldsymbol{1}) = \phi(\boldsymbol{X}_{1i}\boldsymbol{\beta}; \boldsymbol{V}_i)\Phi(\boldsymbol{X}_{2i}\boldsymbol{\beta} - \boldsymbol{\alpha}_i; \boldsymbol{B}_i), \tag{2.8}$$

with

$$\begin{aligned}
\boldsymbol{V}_i &= \boldsymbol{\Sigma}_i + \boldsymbol{Z}_{1i}\boldsymbol{D}\boldsymbol{Z}'_{1i}, \\
\boldsymbol{B}_i^{-1} &= \boldsymbol{I} - \boldsymbol{Z}_{2i}\boldsymbol{K}_i\boldsymbol{Z}'_{2i}, \\
\boldsymbol{K}_i^{-1} &= \boldsymbol{D}^{-1} + \boldsymbol{Z}'_{1i}\boldsymbol{\Sigma}_i^{-1}\boldsymbol{Z}_{1i} + \boldsymbol{Z}'_{2i}\boldsymbol{Z}_{2i}, \\
\boldsymbol{\alpha}_i &= \boldsymbol{H}_i(\boldsymbol{Y}_i - \boldsymbol{X}_{1i}\boldsymbol{\beta}), \\
\boldsymbol{H}_i &= -\boldsymbol{B}_i\boldsymbol{Z}_{2i}\boldsymbol{K}_i\boldsymbol{Z}'_{1i}\boldsymbol{\Sigma}_i^{-1}. \\
\boldsymbol{L} &= \boldsymbol{I}_{p_i} - \boldsymbol{Z}_{2i}\boldsymbol{M}^{-1}\boldsymbol{Z}'_{2i}
\end{aligned}$$

Note that $\boldsymbol{y}_{2i} = \boldsymbol{1}$ denotes a vector of successes of the binary response. It is possible to alter this to an arbitrary vector of successes and failures via the use of contrasts, which is explained in Online Appendix A.

2.3.4 Conditional distributions of the joint model

Next, conditional distributions and expected values can be derived from (2.8), where subvectors of the response vector are modelled. Let $\widetilde{\boldsymbol{Y}}_{1i}$ be the $\widetilde{n}_i$-dimensional subvector of the continuous response vector $\boldsymbol{Y}_{1i}$ and $\widetilde{\boldsymbol{Y}}_{2i}$ the $\widetilde{p}_i$-dimensional subvector of the binary response vector $\boldsymbol{Y}_{2i}$. In addition, $\widetilde{\boldsymbol{X}}_{1i}$ and $\widetilde{\boldsymbol{X}}_{2i}$ are the corresponding submatrices of respectively $\boldsymbol{X}_{1i}$ and $\boldsymbol{X}_{2i}$. The conditional distribution of a subvector of the continuous response given a subvector of the binary response is the ratio of (2.8) and (2.6). The expected value of a subvector of continuous responses given a subvector of successes of the binary response equals

$$\begin{aligned}
E[\widetilde{\boldsymbol{Y}}_{\boldsymbol{1i}}|\widetilde{\boldsymbol{Y}}_{\boldsymbol{2i}} = \boldsymbol{1}] = \quad & \\
& \frac{e^{-\frac{1}{2}G_i}}{\Phi(\widetilde{\boldsymbol{X}}_{2i}\boldsymbol{\beta}; \boldsymbol{L}^{-1})}\sqrt{\frac{|\boldsymbol{E}_i||\boldsymbol{T}_i|}{|\boldsymbol{V}_i||\boldsymbol{B}_i|}}\Phi(\widetilde{\boldsymbol{X}}'_{2i}\boldsymbol{\beta} + \boldsymbol{H}_i\widetilde{\boldsymbol{X}}'_{1i}\boldsymbol{\beta}, \boldsymbol{F}_i, \boldsymbol{T}_i) \\
& \Big(\boldsymbol{E}_i(\boldsymbol{V}_i^{-1}\widetilde{\boldsymbol{X}}'_{1i}\boldsymbol{\beta} + \boldsymbol{H}'_i\boldsymbol{B}_i^{-1}\boldsymbol{F}_i) + \boldsymbol{E}_i\boldsymbol{H}'_i\boldsymbol{B}_i^{-1} \\
& \boldsymbol{T}_i\big[-F_1(o_1) \quad -F_2(o_2) \quad \ldots \quad -F_p(o_p)\big]\Big),
\end{aligned} \tag{2.9}$$

where

$$
\begin{aligned}
G_i &= -\boldsymbol{F}_i'\boldsymbol{T}_i^{-1}\boldsymbol{F}_i + (\widetilde{\boldsymbol{X}}_{1i}\boldsymbol{\beta})'\boldsymbol{V}_i^{-1}(\widetilde{\boldsymbol{X}}_{1i}\boldsymbol{\beta}) - (\boldsymbol{V}_i^{-1}\widetilde{\boldsymbol{X}}_{1i}\boldsymbol{\beta})'\boldsymbol{E}_i(\boldsymbol{V}_i^{-1}\widetilde{\boldsymbol{X}}_{1i}\boldsymbol{\beta}),\\
\boldsymbol{E}_i^{-1} &= \boldsymbol{H}_i'\boldsymbol{B}_i^{-1}\boldsymbol{H}_i + \boldsymbol{V}_i^{-1},\\
\boldsymbol{T}_i^{-1} &= \boldsymbol{B}_i^{-1} - (\boldsymbol{H}_i'\boldsymbol{B}_i^{-1})'\boldsymbol{E}_i(\boldsymbol{H}_i'\boldsymbol{B}_i^{-1}),\\
\boldsymbol{F}_i &= \boldsymbol{T}_i \cdot (\boldsymbol{H}_i'\boldsymbol{B}_i^{-1})'\boldsymbol{E}_i(\boldsymbol{V}_i^{-1}\widetilde{\boldsymbol{X}}_{1i}\boldsymbol{\beta}),\\
o &= \widetilde{\boldsymbol{X}}_{2i}'\boldsymbol{\beta} + \boldsymbol{H}_i\widetilde{\boldsymbol{X}}_{1i}'\boldsymbol{\beta} - \boldsymbol{F}_i\\
F_i(x_i) &= \int_{-\infty}^{o_1}\dots\int_{-\infty}^{o_{i-1}}\int_{-\infty}^{o_{i+1}}\dots\int_{-\infty}^{o_{\widetilde{p}_i}}\varphi(x_1,\dots x_{i-1},x,x_{i+1},\dots x_{\widetilde{p}_i})\\
&\quad dx_{\widetilde{n}_i},..dx_{i+1}dx_{i-1}...dx_1\\
\varphi(x) &= \begin{cases} \frac{\phi(x,\boldsymbol{T}_i)}{\Phi(\widetilde{\boldsymbol{X}}_{2i}'\boldsymbol{\beta}+\boldsymbol{H}_i\widetilde{\boldsymbol{X}}_{1i}'\boldsymbol{\beta}-\boldsymbol{F}_i,\boldsymbol{T}_i)}, & \text{for } \boldsymbol{x} \le \widetilde{\boldsymbol{X}}_{2i}'\boldsymbol{\beta} + \boldsymbol{H}_i\widetilde{\boldsymbol{X}}_{1i}'\boldsymbol{\beta} - \boldsymbol{F}_i,\\ 0, & \text{otherwise.}\end{cases}
\end{aligned}
$$

The derivation of this formula can be found in Online Appendix B1. In addition, the formula and derivation of the prediction interval of the conditional expected value can be found in Online Appendix B2.
The probability of a subvector of successes of the binary response conditional on a subvector of the continuous response equals the corresponding density, which is

$$P(\widetilde{\boldsymbol{y}}_{\boldsymbol{2i}} = \mathbf{1}|\widetilde{\boldsymbol{y}}_{\boldsymbol{1i}}) = \frac{\phi(\widetilde{\boldsymbol{X}}_{1i}\boldsymbol{\beta};\boldsymbol{V}_i)\Phi(\widetilde{\boldsymbol{X}}_{2i}\boldsymbol{\beta}-\boldsymbol{\alpha}_i;\boldsymbol{B}_i)}{\phi(\widetilde{\boldsymbol{X}}_{1i}\boldsymbol{\beta};\boldsymbol{V}_i)} = \Phi(\widetilde{\boldsymbol{X}}_{2i}\boldsymbol{\beta}-\boldsymbol{\alpha}_i;\boldsymbol{B}_i). \tag{2.10}$$

The derivation of the confidence interval of this conditional probability can be found in Online Appendix C.

Analogously, it is possible to derive the conditional distribution of a subvector of continuous responses $\widetilde{\boldsymbol{Y}}^{a}_{1i}=(Y^{a_1}_{1i}Y^{a_2}_{1i}...Y^{a_{n_a}}_{1i})$ given a subvector of continuous responses $\widetilde{\boldsymbol{Y}}^{b}_{1i}=(Y^{b_1}_{1i}Y^{b_2}_{1i}...Y^{b_{n_b}}_{1i})$ and a subvector of successes of the binary response $\widetilde{\boldsymbol{Y}}_{2i}$. The marginal conditional expectation is the following:

$$
\begin{aligned}
E[\widetilde{\boldsymbol{Y}}^{\boldsymbol{a}}_{\boldsymbol{1i}}|\widetilde{\boldsymbol{Y}}^{\boldsymbol{b}}_{\boldsymbol{1i}}=\widetilde{\boldsymbol{y}}^{\boldsymbol{b}}_{\boldsymbol{1i}},\widetilde{\boldsymbol{y}}_{\boldsymbol{2i}}=\boldsymbol{1}] \;=\; & \\
& \frac{e^{-0.5G_i}}{(2\pi)^{\frac{n_b}{2}}f(\widetilde{\boldsymbol{y}}^{\boldsymbol{b}}_{\boldsymbol{1i}},\widetilde{\boldsymbol{y}}_{\boldsymbol{2i}}=\boldsymbol{1})}\frac{\sqrt{|\boldsymbol{E}_i||\boldsymbol{T}_i|}}{\sqrt{|\boldsymbol{V}_i||\boldsymbol{B}_i||\boldsymbol{E}^{bb}_i|}} \\
& \Phi(\widetilde{\boldsymbol{X}}^{'}_{2i}\boldsymbol{\beta}+\boldsymbol{H}_i\widetilde{\boldsymbol{X}}^{'}_{1i}\boldsymbol{\beta},\boldsymbol{F}_i,\boldsymbol{T}_i) \\
& \Bigg\{\Bigg((\boldsymbol{E}_i\boldsymbol{V}_i^{-1}\widetilde{\boldsymbol{X}}_{1i}\boldsymbol{\beta}_1)^a+\boldsymbol{E}^{ab}_i(\boldsymbol{E}^{bb}_i)^{-1}(\widetilde{\boldsymbol{y}}^{\boldsymbol{b}}_{\boldsymbol{1i}}-(\boldsymbol{E}_i\boldsymbol{V}_i^{-1}\widetilde{\boldsymbol{X}}_{1i}\boldsymbol{\beta}_1)^b)\Bigg) \\
+ \quad & \Bigg((\boldsymbol{E}_i\boldsymbol{H}_i'\boldsymbol{B}_i^{-1})^a-\boldsymbol{E}^{ab}_i(\boldsymbol{E}^{bb}_i)^{-1}(\boldsymbol{E}_iH_i'\boldsymbol{B}_i^{-1})^b\Bigg) \\
\times \quad & \Bigg(\boldsymbol{T}_i\big[-F_1(o_1)\quad -F_2(o_2)\quad ...\quad -F_p(o_p)\big]+\boldsymbol{F}_i)\Bigg)\Bigg\},
\end{aligned}
\tag{2.11}
$$

with

$$
\begin{aligned}
\boldsymbol{E}_i^{-1} &= \boldsymbol{H}_i'\boldsymbol{B}_i^{-1}\boldsymbol{H}_i+\boldsymbol{V}_i^{-1}, \\
\boldsymbol{T}_i^{-1} &= (\boldsymbol{E}_i\boldsymbol{H}_i'\boldsymbol{B}_i^{-1})^{b'}(\boldsymbol{E}^{bb}_i)^{-1}(\boldsymbol{E}_i\boldsymbol{H}_i'\boldsymbol{B}_i^{-1})^b+\boldsymbol{B}_i^{-1}-(\boldsymbol{H}_i'\boldsymbol{B}_i^{-1})'\boldsymbol{E}_i(\boldsymbol{H}_i'\boldsymbol{B}_i^{-1}) \\
\boldsymbol{F}_i &= \boldsymbol{T}_i\cdot\Bigg((\boldsymbol{E}_i\boldsymbol{H}_i'\boldsymbol{B}_i^{-1})^{b'}(\boldsymbol{E}^{bb}_i)^{-1}(\widetilde{\boldsymbol{y}}^{\boldsymbol{b}}_{\boldsymbol{1i}}-(\boldsymbol{E}_i\boldsymbol{V}_i^{-1}\widetilde{\boldsymbol{X}}_{1i}\boldsymbol{\beta}_1)^b)+(\boldsymbol{H}_i'\boldsymbol{B}_i^{-1})'\boldsymbol{E}_i(\boldsymbol{V}_i^{-1}\widetilde{\boldsymbol{X}}_{1i}\boldsymbol{\beta})\Bigg) \\
G_i &= \Bigg(\widetilde{\boldsymbol{y}}^{\boldsymbol{b}}_{\boldsymbol{1i}}-(\boldsymbol{E}_i\boldsymbol{V}_i^{-1}\widetilde{\boldsymbol{X}}_{1i}\boldsymbol{\beta}_1)^b\Bigg)'(\boldsymbol{E}^{bb}_i)^{-1}\Bigg(\widetilde{\boldsymbol{y}}^{\boldsymbol{b}}_{\boldsymbol{1i}}-(\boldsymbol{E}_iV^{-1}\widetilde{\boldsymbol{X}}_{1i}\boldsymbol{\beta}_1)^b\Bigg)-\boldsymbol{F}_i'\boldsymbol{T}_i^{-1}\boldsymbol{F}_i+ \\
& (\widetilde{\boldsymbol{X}}_{1i}\boldsymbol{\beta})'\boldsymbol{V}_i^{-1}(\widetilde{\boldsymbol{X}}_{1i}\boldsymbol{\beta})-(\boldsymbol{V}_i^{-1}\widetilde{\boldsymbol{X}}_{1i}\boldsymbol{\beta})'\boldsymbol{E}_i(\boldsymbol{V}_i^{-1}\widetilde{\boldsymbol{X}}_{1i}\boldsymbol{\beta})
\end{aligned}
$$

where the superscript a denotes the rows a_1 until a_{n_a} from the matrix and superscript b denotes the rows b_1 until b_{n_b} from the matrix. Analogously, the superscript bb denotes the submatrix with rows b_1 until b_{n_b} and columns b_1 until b_{n_b}. The superscript ab denotes the submatrix with rows a_1 until a_{n_a} and columns b_1 until b_{n_b}. The derivation of the latter formula and the corresponding prediction interval is shown in Online Appendix D.

The conditional probability of a subvector of successes $\widetilde{\boldsymbol{Y}}^{\boldsymbol{a}}_{\boldsymbol{2i}}$ given both a subvector of the continuous response $\widetilde{\boldsymbol{Y}}_{\boldsymbol{1i}}$ and a subvector of successes of the binary response $\widetilde{\boldsymbol{Y}}^{\boldsymbol{b}}_{\boldsymbol{2i}}$ equals

$$P(\widetilde{\boldsymbol{y}}^{\boldsymbol{a}}_{\boldsymbol{2i}} = \mathbf{1} | \widetilde{\boldsymbol{y}}_{\boldsymbol{1i}}, \widetilde{\boldsymbol{y}}^{\boldsymbol{b}}_{\boldsymbol{2i}} = \mathbf{1}) = \frac{\Phi(\widetilde{\boldsymbol{X}}_{2i}\boldsymbol{\beta} - \boldsymbol{H}_i(\widetilde{\boldsymbol{Y}}_{\boldsymbol{1i}} - \widetilde{\boldsymbol{X}}_{1i}\boldsymbol{\beta}); \boldsymbol{B}_i)}{\Phi(\widetilde{\boldsymbol{X}}^{b}_{2i}\boldsymbol{\beta} - \boldsymbol{H}^{b}_i(\widetilde{\boldsymbol{Y}}_{\boldsymbol{1i}} - \widetilde{\boldsymbol{X}}_{1i}\boldsymbol{\beta}); B^{bb}_i)}, \quad (2.12)$$

where in analogy with (2.11), the superscript b denotes the rows b_1 until b_{p_b} and the superscript bb denotes the submatrix with rows b_1 until b_{p_b} and columns b_1 until b_{p_b}. The derivation of the confidence interval can be found in Online Appendix E.

2.3.5 Correlation function

From (2.8) it is possible to derive a correlation function of the manifest correlation $\rho_{Y_{1ij},Y_{2ik}}$. The latter represents the association between the continuous response at time j and the binary response at time k. Manifest correlation indicates the correlation between the observed scores. In contrast, the latent correlation refers to the correlation between the random effects. The latter is easier to calculate, but scientific interest can lie more in the manifest correlation than in the latent correlation. The formula of the manifest correlation function is the following:

$$\begin{aligned}\rho_{Y_{1ij},Y_{2ik}} = & \left\{ \left(\frac{1}{|\boldsymbol{D}|^{1/2}} \frac{1}{|\boldsymbol{M}_i|^{1/2}} \frac{1}{L_i^{1/2}} - 1 \right) \boldsymbol{x}'_{1ij}\boldsymbol{\beta}\Phi(L_i^{1/2}\boldsymbol{x}'_{2ik}\boldsymbol{\beta}) + \right. \quad (2.13) \\ & \left. \frac{1}{|\boldsymbol{D}|^{1/2}} \frac{1}{|\boldsymbol{M}_i|^{1/2}} \frac{1}{L_i} \boldsymbol{Z}'_{1ij}\boldsymbol{M}_i^{-1}\boldsymbol{Z}_{2ik}\phi(L_i^{1/2}\boldsymbol{x}'_{2ik}\boldsymbol{\beta}) \right\} \\ & \frac{1}{\sqrt{\left(\boldsymbol{Z}'_{1ij}\boldsymbol{D}\boldsymbol{Z}_{1ij} + \Sigma_{1ij}\right)\Phi(L_i^{1/2}\boldsymbol{x}'_{2ik}\boldsymbol{\beta})(1 - \Phi(L_i^{1/2}\boldsymbol{x}'_{2ik}\boldsymbol{\beta}))}},\end{aligned}$$

with $\boldsymbol{M}_i = \boldsymbol{D}^{-1} + \boldsymbol{Z}'_{2ik}\boldsymbol{Z}_{2ik}$. Note that the correlation function depends on values of the covariates $\boldsymbol{X}_{1ij}$ and $\boldsymbol{X}_{2ik}$. The derivation of this function is shown in Online Appendix F.

In a more general case, the manifest variance-covariance matrix between subvectors of the two responses $\widetilde{\boldsymbol{Y}}_{1i}$ and $\widetilde{\boldsymbol{Y}}_{2i}$ is the following:

$$\rho_{\widetilde{\boldsymbol{Y}}_{1i},\widetilde{\boldsymbol{Y}}_{2i}} = \frac{\widetilde{\boldsymbol{Z}_{1i}}}{\sqrt{|\boldsymbol{D}|}} \begin{bmatrix} \sqrt{|\boldsymbol{E}_1|}\boldsymbol{E}_1\widetilde{\boldsymbol{Z}}_{2i}^{1'}\phi\left(\widetilde{\boldsymbol{X}}_{2i}^{1}\boldsymbol{\beta}\cdot\sqrt{1-\widetilde{\boldsymbol{Z}}_{2i}^{1}\boldsymbol{E}_1\widetilde{\boldsymbol{Z}}_{2i}^{1'}}\right)\left(1-\widetilde{\boldsymbol{Z}}_{2i}^{1}\boldsymbol{E}_1\widetilde{\boldsymbol{Z}}_{2i}^{1'}\right)^{-1} \\ \sqrt{|\boldsymbol{E}_2|}\boldsymbol{E}_2\widetilde{\boldsymbol{Z}}_{2i}^{2'}\phi\left(\widetilde{\boldsymbol{X}}_{2i}^{2}\boldsymbol{\beta}\cdot\sqrt{1-\widetilde{\boldsymbol{Z}}_{2i}^{2}\boldsymbol{E}_2\widetilde{\boldsymbol{Z}}_{2i}^{2'}}\right)\left(1-\widetilde{\boldsymbol{Z}}_{2i}^{2}\boldsymbol{E}_2\widetilde{\boldsymbol{Z}}_{2i}^{2'}\right)^{-1} \\ .. \\ \sqrt{|\boldsymbol{E}_{\widetilde{p}_i}|}\boldsymbol{E}_{\widetilde{p}_i}\widetilde{\boldsymbol{Z}}_{2i}^{\widetilde{p}_i'}\phi\left(\widetilde{\boldsymbol{X}}_{2i}^{\widetilde{p}_i}\boldsymbol{\beta}\cdot\sqrt{1-\widetilde{\boldsymbol{Z}}_{2i}^{\widetilde{p}_i}\boldsymbol{E}_{\widetilde{p}_i}\widetilde{\boldsymbol{Z}}_{2i}^{\widetilde{p}_i'}}\right)\left(1-\widetilde{\boldsymbol{Z}}_{2i}^{\widetilde{p}_i}\boldsymbol{E}_{\widetilde{p}_i}\widetilde{\boldsymbol{Z}}_{2i}^{\widetilde{p}_i'}\right)^{-1} \end{bmatrix} \tag{2.14}$$

where the superscript denotes the row of the submatrix and $\boldsymbol{E}_{ik}^{-1} = \widetilde{\boldsymbol{Z}}_{2i}^{k'}\widetilde{\boldsymbol{Z}}_{2i}^{k} + \boldsymbol{D}^{-1}$.The derivation of this function is shown in Online Appendix G.

2.4 Parameter estimation

The estimation of the parameters in random-effects models is done with maximization of the marginal likelihood Molenberghs and Verbeke (2005). This maximization is based on

$$L(\boldsymbol{\beta},\boldsymbol{D}) = \prod_{i=1}^{N} f(\boldsymbol{Y}_{1i}=\boldsymbol{y}_{1i},\boldsymbol{Y}_{2i}=\boldsymbol{y}_{2i}). \tag{2.15}$$

It is possible to maximize this analytical joint marginal likelihood, but the manipulation is cumbersome. In general numerical techniques are applied prior to maximizing the likelihood, instead of integrating out the random effects distribution analytically. This numerical approximation of the integral is executed with gaussian quadrature and adaptive gaussian quadrature. In the latter methods the order of approximation, or equivalently the accuracy, can be pre-specified (Pinheiro and Bates, 1995). In practice the order of approximation, Q, is increased until numerical stability is obtained in the parameter estimates and approximated likelihood. The SAS procedures GLIMMIX and NLMIXED, next to other major statistical tools, are readily available for fitting the model in this paper.

2.5 Analysis of the cystic fibrosis data

In this section, the cystic fibrosis data is analysed, with focus on the association between the two responses: FEV and ABPA. The response sequence of both the binary response ABPA and the continuous response FEV will be modelled jointly. The advantage is that both the association between the responses and the association within each response vector can be examined. Both responses will be predicted by the same covariates: time, the age at the first measurement (X_{1i}), the female sex (X_{2i}), the presence of meconium ileus at birth (X_{3i}) and the interaction between the time and age at the first measurement. In addition, a linear relation between the response and the continuous covariates is presumed. The hierarchical model has the following form:

$$\begin{aligned}
Y_{1ij} &= b_{10i} + \beta_{10} + (b_{11i} + \boldsymbol{\beta}_{11})t_{ij} + \beta_{12}X_{1i} + \beta_{13}X_{2i} + \beta_{14}X_{3i} \\
&\quad + \beta_{15}t_{ij} \times X_{1i} + \epsilon_{ij}, \\
\Phi^{-1}(P(Y_{2ik} = 1)) &= b_{20i} + \beta_{20} + (b_{21i} + \beta_{21})t_{ik} + \beta_{22}X_{1i} + \beta_{23}X_{2i} + \beta_{24}X_{3i} \\
&\quad + \beta_{25}t_{ij} \times X_{1i}, \\
(b_{10i}, b_{11i}, b_{20i}, b_{21i})' &\sim N\left[\begin{pmatrix}0\\0\\0\\0\end{pmatrix}, \begin{pmatrix} d_{11} & d_{21} & d_{13} & d_{14} \\ d_{21} & d_{22} & d_{23} & d_{24} \\ d_{13} & d_{23} & d_{33} & d_{34} \\ d_{14} & d_{24} & d_{34} & d_{44}\end{pmatrix}\right], \\
\epsilon_i &\sim N(0, \boldsymbol{\Sigma}_i),
\end{aligned}$$

where t_{ij} is the time point at which response j from person i is measured. In both responses a random intercept and a random slope is included. The parameter estimates and standard errors of the joint model are shown in Table 2.3. For FEV, a likelihood ratio test indicated that random slopes are significant ($\chi^2_{1:2} = 1343.7$, $p < 0.001$). No assumptions are made about the structure of the variance-covariance matrix of the random effects. The advantage is that if the D matrix is overly simple, incorrect inferences can result. In addition, it is assumed that the observations are independent given the random effects, and hence no serial correlation is included.

The latent correlations indicate that there is a moderate positive correlation between the random intercepts of both response values ($\hat{\rho} = 0.281, p < 0.001$). Hence, a better lung function at baseline than expected under the model is associated with a higher probability of absence of ABPA at baseline on the probit-scale than expected. In addition, the correlation between the random intercept of FEV and the random slope of ABPA is not significant ($\hat{\rho} = 0.069, p = 0.480$). Further, a higher increase in lung function than expected under the model is positively related to a higher probability of absence of ABPA than expected at baseline on the probit scale ($\hat{\rho} = 0.236, p = 0.005$); there is a significant association between the random slope of FEV and the random intercept of ABPA.

Table 2.3*: The parameter estimates of the joint hierarchical model.*

Response	Effect	Par.	Est. Model (se)
FEV	Intercept	β_{10}	$106.580\ (1.292)$
	Time	β_{11}	$-1.496\ (0.131)$
	Age	β_{12}	$-1.217\ (0.056)$
	Female	β_{13}	$-3.039\ (1.168)$
	Meconium	β_{14}	$-2.344\ (1.743)$
	Time $\times$ Age	β_{15}	$0.002\ (0.007)$
	Residual variance	σ^2	$32.525\ (0.626)$
ABPA	Intercept	β_{20}	$4.829\ (0.438)$
	Time	β_{21}	$-0.064\ (0.091)$
	Age	β_{22}	$-0.103\ (0.012)$
	Female	β_{23}	$-0.096\ (0.178)$
	Meconium	β_{24}	$-0.554\ (0.268)$
	Time $\times$ Age	β_{25}	$0.010\ (0.003)$
		d_{11}	$370.130\ (16.516)$
		d_{21}	$3.040\ (1.410)$
		d_{22}	$2.876\ (0.191)$
		d_{31}	$12.588\ (2.802)$
		d_{32}	$0.933\ (0.334)$
		d_{33}	$5.412\ (1.017)$
		d_{41}	$0.400\ (0.566)$
		d_{42}	$0.006\ (0.053)$
		d_{43}	$-0.146\ (0.115)$
		d_{44}	$0.090\ (0.017)$

Lastly, the random slopes of the responses are not significanly correlated ($\hat{\rho} = 0.013, p = 0.902$). In this case, the interpretation of the latter correlations on a probit scale and in terms of latent processes is less straightforward compared to the interpretation of the manifest correlations.
Table 2.4 shows the manifest correlations between FEV and the absence of ABPA at different time points. Note that the correlation between FEV and the presence of ABPA equals -1 times the correlation between FEV and the absence of ABPA. Since the correlations depend on the covariates, the covariates are kept fixed at the modus for the binary variables and the median for continuous variables; the age is set at 15.630, the sex is male and there was no meconium ileus at birth. In addition, a Wald test is performed to test whether the latent correlations are equal

to zero ($H_0 : d_{31} = d_{31} = d_{41} = d_{42} = 0$, H_α: minimum one parameter does not equal 0). If the latent correlations equal zero, the manifest correlations would subsequently also equal zero. The test indicates the correlations are significantly different from 0 ($\chi^2_4 = 43.309$, $p < 0.001$). As apparent from Table 2.4, the correlation is slightly stronger for earlier measurements of the binary ABPA and later measurements of FEV. A possible explanation is that having ABPA in an early stage shows an overall weakness and subsequently results in deterioration of the lungs. Alternatively, patients with ABPA can develop bronchiectasis, which is a form of airway damage that both causes a higher risk of different infections and worse lung functioning (Gothe et al., 2017).

Table 2.4*: The manifest correlations between lung function (FEV) and the absence of allergic bronchopulmonary aspergillosis (ABPA) at different time points. Time is defined as the years a participant is included in the study.*

	year(ABPA)								
year(FEV)	0	1	2	3	4	5	6	7	8
0	0.146	0.148	0.150	0.151	0.151	0.151	0.150	0.149	0.147
1	0.155	0.157	0.159	0.159	0.159	0.159	0.158	0.157	0.155
2	0.163	0.165	0.166	0.167	0.167	0.166	0.165	0.163	0.161
3	0.169	0.171	0.172	0.173	0.172	0.171	0.170	0.168	0.166
4	0.174	0.177	0.177	0.177	0.177	0.176	0.174	0.172	0.170
5	0.179	0.180	0.181	0.181	0.180	0.179	0.177	0.175	0.173
6	0.182	0.183	0.184	0.184	0.183	0.181	0.180	0.177	0.175
7	0.184	0.185	0.186	0.185	0.184	0.183	0.181	0.178	0.176
8	0.185	0.186	0.187	0.186	0.185	0.184	0.181	0.179	0.176

Second, the conditional distribution of FEV given ABPA is examined. Table 2.5 shows the expected value of FEV at a given time conditional on the knowledge that person i had no ABPA at three given time point within the study. Again, the covariates are set to their modes and medians.
Having no ABPA at an early stage results in slightly higher expected FEV values at a certain time compared to having no ABPA at a later stage. For example, when a person has no ABPA at time point zero, one, and two the expected value of FEV at time point eight is 78.3 $[42.3, 144.9]$. When the person had no ABPA at time point five, six and seven the predicted value of time point eight is 78.1 $[42.0, 145.0]$. The slightly more powerful effect of earlier values of ABPA on the FEV was also visible in the manifest correlations. Still, since the prediction intervals are quite large, it is fruitful to include previous measures of the lung function in the prediction of future lung function.

Table 2.5: *Expected value and prediction interval of the FEV value at a given time point in the study (year(FEV)), conditional on the fact that the person had no allergic bronchopulmonary aspergillosis at the given years in the study (year(ABPA)). The time indicates the years a person is participating in the study.*

year(FEV)	year(ABPA)	E$[Y_{1ij}]$	PI Y_{1ij}
3	$0, 1, 2$	85.1	$[52.8; 137.0]$
4	$0, 1, 2$	83.7	$[50.9; 137.7]$
5	$0, 1, 2$	82.4	$[48.9; 138.7]$
6	$0, 1, 2$	81.0	$[46.8; 140.3]$
7	$0, 1, 2$	79.6	$[44.6; 142.4]$
8	$0, 1, 2$	78.3	$[42.3; 144.9]$
4	$1, 2, 3$	83.6	$[50.8; 137.7]$
5	$1, 2, 3$	82.3	$[48.9; 138.6]$
6	$1, 2, 3$	80.9	$[46.7; 140.2]$
7	$1, 2, 3$	79.6	$[44.5; 142.3]$
8	$1, 2, 3$	78.2	$[42.3; 144.7]$
5	$2, 3, 4$	82.2	$[48.8; 138.7]$
6	$2, 3, 4$	80.9	$[46.6; 140.4]$
7	$2, 3, 4$	79.5	$[44.4; 142.4]$
8	$2, 3, 4$	78.2	$[42.2; 144.8]$
6	$3, 4, 5$	80.8	$[46.6; 140.2]$
7	$3, 4, 5$	79.5	$[44.4; 142.2]$
8	$3, 4, 5$	78.1	$[42.2; 144.8]$
7	$4, 5; 6$	79.4	$[44.3; 142.3]$
8	$4, 5, 6$	78.1	$[42.1; 144.8]$
8	$5, 6, 7$	78.1	$[42.0; 145.0]$

Note that to ensure that lower limits of the prediction intervals are not lower than zero a log-transformation is applied to compute the prediction intervals, after which the values are transformed back to the original scale.

The expected value of FEV was calculated given the FEV and ABPA values of the three preceding time points. In Table 2.6, the predictions are conditional on the quartiles of FEV and the occurence of ABPA preceding the prediction. In addition, the expected value is conditional whether or not there is absence of an ABPA diagnosis two and three years preceding the prediction. Unsurprisingly, having high FEV values preceding the prediction, results in a higher expected value. In addition, the presence of ABPA at the three preceding time point results in slightly lower predicted values compared to an acute ABPA diagnosis

preceding the prediction. Nevertheless, the ABPA values preceding the prediction does not modify the expected value much when the FEV values are included in the prediction. The covariance parameters between the random effects of FEV and the random slope of ABPA were not significantly different from 0, as indicated by a Wald test ($H_0 : d_{14} = d_{24} = 0, H_a$: minimum one parameter does not equal 0, $\chi^2_2 = 0.499, p = 0.779$). A sensitivity analysis was conducted where both parameters were restricted to equal 0 and the model was refitted. The differences between the results of the latter model and the model with unconstrained covariance parameters were minor.

Table 2.6*: Expected FEV value and prediction interval of the FEV value at time j in the study. This predicted value is conditional on the three preceding FEV measurements ($Y_{1i(j-3)}, Y_{1i(j-2)}, Y_{1i(j-1)}$) and presence of acute ABPA (ABPA at j-1 and no ABPA diagnosis at $j-2$ and $j-3$) or chronic ABPA (presence of ABPA at the three foregoing time points). The time indicates the years a person is participating in the study.*

				Acute ABPA		Chronic ABPA	
j	$Y_{1i(j-3)}$	$Y_{1i(j-2)}$	$Y_{1i(j-1)}$	$E[Y_{1ij}]$	PI Y_{1ij}	$E[Y_{1ij}]$	PI Y_{1ij}
3	64.9	64.9	64.9	62.67	[49.9; 78.7]	62.43	[49.2;79.3]
4	64.9	64.9	64.9	62.38	[49.7; 78.4]	62.18	[48.9;79.1]
5	64.9	64.9	64.9	62.14	[49.4; 78.1]	61.98	[48.7;78.9]
6	64.9	64.9	64.9	61.96	[49.3; 77.9]	61.82	[48.6;78.7]
7	64.9	64.9	64.9	61.82	[49.1; 77.8]	61.72	[48.5;78.6]
8	64.9	64.9	64.9	61.73	[49.1; 77.6]	61.66	[48.4;78.5]
3	84	84	84	81.65	[68.5; 97.3]	81.48	[67.5;98.4]
4	84	84	84	81.60	[68.6; 97.1]	81.47	[67.4;98.4]
5	84	84	84	81.59	[68.5; 97.2]	81.48	[67.4;98.5]
6	84	84	84	81.60	[68.6; 97.1]	81.51	[67.4;98.5]
7	84	84	84	81.64	[68.6; 97.1]	81.57	[67.5;98.6]
8	84	84	84	81.71	[68.7; 97.1]	81.65	[67.6;98.6]
3	97.7	97.7	97.7	95.33	[82.1; 110.7]	95.21	[80.7;112.4]
4	97.7	97.7	97.7	95.45	[82.2; 110.9]	95.36	[80.8;112.6]
5	97.7	97.7	97.7	95.61	[82.3; 111.1]	95.52	[80.9;112.8]
6	97.7	97.7	97.7	95.75	[82.5; 111.1]	95.70	[81.0;113.0]
7	97.7	97.7	97.7	95.92	[82.7; 111.3]	95.88	[81.2;113.2]
8	97.7	97.7	97.7	96.08	[83.0; 111.2]	96.05	[81.4;113.3]

Note that to ensure that lower limits of the prediction intervals are not lower than zero a log-transformation is applied to compute the prediction intervals, after which the values are transformed back to the original scale.

Models for repeated measurements with time dependent covariates provide an alternative to the random-effects joint model approach. In contrast to random-effects joint models, exogeneity has to be assumed and ignorability does not hold for the response treated as a time-dependent covariate (i.e., ABPA). As a consequence, multiple imputation needed to be conducted for both responses. Imputed values were generated via fully conditional specification (van Buuren, 2007). We performed 10 imputations, where each variable was imputed using all other variables. Logistic regression was applied for the binary variables, while linear regression was implemented for the continuous variables. A 'wide' data format was used to impute the longitudinal variables at each time point separately, incorporating information from other time points. Along with the longitudinal variables FEV and ABPA, we included Age, Sex, and Meconium in the imputation model. No interaction terms or polynomial terms were included.
Pepe and Anderson (1994) show that the parameter estimates for time-dependent covariates will likely be biased unless the "full covariate conditional model" (FCCM) holds. This means that the conditional mean of Y_{ij} given the full time-dependent covariate sequence $X_{1i}, ..., X_{in}$ depends only on the time-dependent covariate value(s) in the model. Their article focuses on the GEE approach, but this applies equally to all longitudinal data analysis methods including likelihood-based methods such as linear and generalised linear mixed models (Diggle, 2002). Pepe and Anderson (1994) state that in absence of the full covariate conditional mean (FCCM) assumption, unbiased estimation can be obtained with working independence generalised estimation equations. The time-dependent covariates model is the following:

$$\begin{aligned} E[Y_{1ij}] &= \beta_0 + \beta_1 t_{ij} + \beta_2 X_{1i} + \beta_3 X_{2i} + \beta_4 X_{3i} + \beta_5 t_{ij} X_{3i} + \\ &\quad \alpha_1 APBA_{i(j-1)} + \alpha_2 ABPA_{i(j-2)} + \alpha_3 ABPA_{i(j-3)}. \end{aligned}$$

The FEV predictions of the latter model are compared with the predictions of the conditional joint model in Table 2.7. The discrepancies in the predictions can be attributed to the fact that the estimation of the joint model is based on the full dataset, while the estimates of the time-dependent covariates model are based on data starting from the fourth time point. All three lags are significant ($\alpha_1 = 7.05$, $p < .001$; $\alpha_2 = 4.41$, $p < .001$; $\alpha_3 = 7.47, p < .001$) which indicates that the lags of ABPA have additional predictive value on top each other and the other predictors.
Next to the disadvantages already stated, the time-dependent covariates model has several drawbacks compared to the conditional joint model. First, only data starting from to fourth year can be used, since the ABPA values of the three preceding time points are utilized as predictors. Second, a decision is required about the lag and the number of lags specified in the model. Third, when values

of the time dependent covariate are highly correlated over time, multicollinearity issues can arise.

Table 2.7: *Comparison of the expected values and corresponding prediction intervals of the joint model and the time-dependent covariates model. The expected values of FEV at a certain time-point (j) are conditional on the absence of ABPA at the three preceding timepoints and the modus and medians of the time-independent covariates.*

	Joint model		Time-dependent covariates model	
j	$E[Y_{1ij}]$	PI Y_{1ij}	$E[Y_{1ij}]$	PI Y_{1ij}
3	85.1	[52.8; 137.0]	81.8	[49.8; 134.4]
4	83.6	[50.8; 137.7]	80.3	[47.7; 135.1]
5	82.2	[48.8; 138.7]	78.7	[46.0; 134.8]
6	80.8	[46.6; 140.2]	77.2	[44.0; 135.5]
7	79.4	[44.3; 142.3]	75.6	[42.1; 135.9]
8	78.1	[42.0; 145.0]	74.1	[40.0; 137.2]

Note that to ensure that lower limits of the prediction intervals are not lower than zero a log-transformation is applied to compute the prediction intervals, after which the values are transformed back to the original scale.

2.6 Concluding remarks

In many studies, multiple responses are observed and are modelled jointly. In these joint models random effects are used to capture the associations between these responses and the evolution of this association over time. Except for joint models consisting of only continuous responses, the interpretation of these random effects means examining the correlations at an underlying scale. The latter correlations are called latent correlations. A more intuitive measure are the correlations between the responses that are actually observed, i.e., the manifest correlations. In this paper, a closed-form expression is derived for the manifest correlations in the context of a probit-normal joint model.
A simpler method to obtain the manifest correlations would be the computation of the point-biserial correlations (after applying multiple imputation to handle missing values). When applied to the case study (detailed results not shown), these correlations were similar compared to the results of (13) from a joint model without covariates. However, with this method the correlations are not controlled for or dependent on characteristics of the participant. When an investigator is interested in specific subpopulations or the effect of a certain characteristic

while controlling for other characteristics, the closed-form correlations in (13) are favourable.

An alternative approach to examine the effect of one response to another response are time-dependent covariates. Disadvantages include the correct characterization of the lag relationship, the type of time-dependent covariate (endogenous versus exogenous), the treatment of missing data and the possibility that the time-dependent covariate is an intermediate variable. These drawbacks are solved by the use of a joint model; in this family of models associations between each pair of time points of the responses are modelled and can even be examined by scrutinizing the manifest correlations. Second, there is no need to discriminate between endogeneous and exogeneous covariates, since the relationship between both responses is symmetric. Hence, there is no issue if previous values of one response influence the later values of the other response. In addition, the issue of intermediate variables is resolved. Lastly, in a joint model ignorability holds for the missing response values. As a result, no additional effort has to be exerted in order to handle the missing data. This paper has outlined the conditional distributions for when interest lies in making predictions of one response given the other response.
The probit-normal model was applied to longitudinal cystic fibrosis data. The resulting model shows a heavier effect of early occurrences of ABPA on later FEV values compared to later occurrences of ABPA. This effect could be seen in both the manifest correlations and the conditional distribution of FEV given ABPA. Hence, it is in this case rewarding to examine the associations between responses at different time points. A disadvantage of the model is the impact of the random effects structure on the flexibility of the manifest correlations. In this case, the random intercept and slope for modelling of both responses allows the correlations to be flexibly estimated. However, including no random slope in one or more of the responses results in a severe restriction in the manifest correlations. Hence, the manifest correlations are equal for each time point of the response modelled without random slope. This model can be easily implemented in existing software packages such as the SAS procedure NLMIXED with limited extra coding effort.
Further research can be conducted on the bounds of the correlation function. In the context of surrogate markers, Alonso and Molenberghs (2007) note for example that their information-theoretic R^2 of a binary endpoint can generally be smaller than one. A second example is the constraints in the bounds of the correlation between dichotomous responses in the Bahadur model for clustered data Molenberghs and Verbeke (2005).

Chapter 3

A joint normal-binary (probit) model for high-dimensional longitudinal data

This chapter is based upon:

Delporte, M., Fieuws, S., Molenberghs, G., Verbeke, G, De Coninck D., & Hoorens, V. (2023). A Joint Normal-Binary (Probit) Model for High-Dimensional Longitudinal Data. *Statistical Modelling*, Accepted.

The appendix is available via `margauxdelporte.github.io/Chapter3.pdf`.

Abstract

In many biomedical studies multiple responses are collected over time, which results in high-dimensional longitudinal data. It is often of interest to model the continuous and binary responses jointly, which can be done with joint generalised mixed models in which the association is modelled through random effects. Investigating the association between the responses is often limited to scrutinizing the correlations between the latent random effects. In this paper, this approach is extended by deriving closed-form formulas for the manifest correlations (and corresponding standard errors), which reflects the correlation between the observed responses as observed. In addition, the marginal joint model is constructed, from which predictions of subvectors of one response conditional on subvectors of other response(s) and potentially a subvector of the history of the response can be derived. Corresponding prediction and confidence intervals are constructed. Two case studies are discussed, in which further pseudo-likelihood methodology is applied to reduce the computational complexity.

3.1 Introduction

In many studies, participants are followed throughout time while their characteristics are repeatedly measured. This can result in longitudinal data with many responses; this can be referred to as high-dimensional clustered data. There are several techniques applicable to model clustered data, among which random-effects models are the most popular. This group of models was introduced by Laird and Ware (1982) via the linear mixed model for continuous clustered responses. This technique models the average evolution of a specific outcome with a certain function of time. In addition, the subject-specific deviations from the latter average evolution are taken into account with random effects. Later, the generalised linear mixed model was added to the group of random-effects models in order to model non-continuous clustered data (Breslow and Clayton, 1993; Wolfinger and O'Connell, 1993; Engel and Keen, 1994). Still, the latter models can only incorporate a single response per analysis, while the data at hand regularly has multiple responses.

Joint random-effects models provide the means to model multiple responses together. This has the advantage that research questions that take several or all responses into account can be answered. For example, the effect of a covariate can be gauged on all repeated instances of the same response, but also across responses. Moreover, the association processes within a sequence as well as across sequences can be assessed. In this way, the association between the responses and the evolution of this association over time can be quantified. The random effects of univariate random-effects models are allowed then to be correlated, enabling the examination of the association between the different responses. For

example, Chakraborty et al. (2003) fitted a joint model in order to fit two continuous responses simultaneously: HIV RNA in blood and in semen. With the joint model, the correlation between the two responses could be captured and compared between two treatment groups.
The number of responses in joint models are not limited to two. Morrell et al. (2012) has, for example, applied a joint model in the context of three longitudinal continuous responses. An issue with generalised linear joint models are the computational problems that can arise when many responses or random effects are included due to dimensionality of the joint random-effects distribution. Fieuws and Verbeke (2006) provide a method where all bivariate models are separately fitted and inference for the full multivariate model follows from pseudo-likelihood arguments. They first applied this approach in a setting with 22 longitudinal continuous responses. Next, their method was extended to binary data (Fieuws et al., 2006) and the combination of various data types (Fieuws et al., 2008; Ivanova et al., 2016).
A second method to solve computational problems is the split-sample method (Molenberghs et al., 2011b). By means of sample splitting, the dataset can be subdivided into several parts and these parts can be combined for inference. It is possible to combine both methods; Ivanova et al. (2017) provide an example where sample splitting as well as pairwise fitting are combined in order to fit a multivariate joint proportional odds mixed model to three ordinal responses. In addition, it is shown that these methods provide efficient and fast inferences in high-dimensional large datasets.
The joint mixed model is able to model the density of multiple responses jointly. It fits a seperate (generalised) linear mixed model for each response, and allows the random effects to correlate in order to combine the models. The advantage of such joint models is that the responses do not necessarily have to be of the same nature. Molenberghs and Verbeke (2005) describe the hierarchical joint model for a binary and a continuous response. In addition, Delporte et al. (2022) derive the marginal probit-normal model for a binary and a continuous response. However, the setting for multiple binary and multiple continuous responses has not yet been considered. This paper handles the high-dimensional case (≥ 2 responses) where normal and binary responses are analysed. First the marginal joint binary-normal probit model is derived. This marginal model leads itself to conditional predictions and prediction intervals, where a subvector of one response can be predicted conditional on subvectors of the other responses and potentially a subvector of the same response. Hence, with joint-model methodology a time dependent covariate can be incorporated in the model as a response. The aforementioned conditional models provide an alternative to use the time-dependent covariates for prediction (Diggle, 2002). This method of prediction tackles many issues that are present in time-varying covariate methodology, such

as the possibility of endogeneity. Covariate endogeneity occurs when the time-dependent covariate can be predicted by the response at earlier timepoints (Diggle, 2002). A second issue tackled is the presence of intermediate time-dependent variables. The latter are defined as variables that are in the causal pathway from the covariate to the response. Including the intermediate variable results in the loss of the effect of the covariate on the response (Diggle, 2002). In addition, there is no need to choose or specify lags. Next, if time-dependent covariates would be used, extra steps have to be taken to take the missingness into account. In this paper, the time-dependent covariates are treated as a response by the joint-modelling methodology. As a result, missingness in the longitudinal covariates can be ignored under the assumption of MAR due to ignorability (Rubin, 1976). A direct result of using the joint model are the correlations between the random effects of the responses (latent correlations). Still, it is possible that interest lies in the correlations between the responses themselves. Therefore, closed-form formulas for the manifest correlations between the responses and their corresponding confidence intervals are derived.
The paper is organized as follows. Methodology for reducing the computational complexity and existing mixed-model methodology are reviewed in Section 3.2. In the remainder of Section 3.2, our marginal model for high-dimensional binary-normal datasets is presented. Lastly, the estimation method is specified in Section 3.2.6. The datasets are introduced in Section 3.3 and analysed in Section 3.4. In Section 3.5 concluding remarks are offered. Online Appendix A shows proofs for simplifications of several formulas. In addition, calculations for the new prediction and confidence intervals can be found in Appendix B, Appendix C and Appendix D. Next, details about the pairwise modelling approach can be found in Appendix E. Lastly, extra details of the analysis of the case studies is reported in Appendix F and Appendix G.

3.2 Methodology

3.2.1 Linear mixed models for continuous responses

Laird and Ware (1982) introduced linear mixed models for clustered continuous responses. Let y_{ij} denote the jth measurement for the ith subject, $i = 1,..,N$,$j = 1,...,n_i$, and let $\boldsymbol{y_i} = (y_{i1},..y_{in_i})$ represents the vector of all n_i measurements of subject i. The model is specified as

$$\begin{aligned} \boldsymbol{y_i}|\boldsymbol{b}_i &\sim N(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i, \boldsymbol{\Sigma}_i), \\ \boldsymbol{b}_i &\sim N(\boldsymbol{0}, \boldsymbol{D}), \end{aligned} \tag{3.1}$$

where $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are, respectively, $(n_i \times k)$ and $(n_i \times q)$ dimensional matrices of

known covariates. $\boldsymbol{\beta}$ denotes the k-dimensional vector of fixed effects and $\boldsymbol{b}_i$ is the q-dimensional vector of random effects. Next, $\boldsymbol{\Sigma}_i$ is the $(n_i \times n_i)$ dimensional covariance matrix. The subscript does not mean that the estimated variance parameters of the residuals depends on subject i. In contrast, i indicates that each subject can have a different number of measurements and hence the dimensions of the matrix can differ across individuals (Verbeke and Molenberghs, 2000). If we assume that the measurements of a subject are independent given the random effects $\boldsymbol{b}_i$, $\boldsymbol{\Sigma}_i$ simplifies to $\sigma^2 \boldsymbol{I}_{n_i}$. The latter assumption is not essential for linear mixed models, serial correlation can, for example, be included. The marginal and conditional parameters are exactly the same since $E(Y_{ij}) = E[E(Y_{ij}|\boldsymbol{b}_i)] = \boldsymbol{x}'_{ij}\boldsymbol{\beta}$. Still, one can define the marginal model as

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}^*_i, \tag{3.2}$$

where the residuals $\boldsymbol{\epsilon}^*_i \sim N(\boldsymbol{0}, \boldsymbol{V}^*_i)$ induce the intracluster correlation, and are hence correlated. As a result, the distribution of the response is

$$\boldsymbol{Y}_i \sim N(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{V}^*_i), \tag{3.3}$$

with

$$\boldsymbol{V}^*_i \;\; = \;\; \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}'_i + \boldsymbol{\Sigma}_i.$$

A more extensive overview of linear mixed models can be found in Verbeke and Molenberghs (2000).

3.2.2 Generalised linear mixed models for binary responses

The linear mixed model of Laird and Ware (1982) was later generalised for non-continuous data (Breslow and Clayton, 1993; Wolfinger and O'Connell, 1993; Engel and Keen, 1994). This model adds random effects to the generalised linear model. In this paper, only the probit link will be considered. The generalised linear mixed model makes the assumption that the elements Y_{ij} of $\boldsymbol{Y_i}$ are independent given the random effects $\boldsymbol{\delta_i}$. The model for a vector of successes is specified as

$$\Phi^{-1}[P(Y_{ij} = 1|\boldsymbol{\delta}_i)] = \boldsymbol{x}_{ij}\boldsymbol{\beta} + \boldsymbol{z}_{ij}\boldsymbol{\delta}_i, \tag{3.4}$$

where $\Phi(\cdot)$ is the cumulative normal distribution function and $\boldsymbol{\delta}_i \sim N(\boldsymbol{0}, \boldsymbol{\tau})$. In the linear mixed model, the conditional parameters are equal to the marginal parameters. In the generalised linear mixed model, this is not the case. Closed-form expressions for the marginal generalised linear mixed model have been derived by Molenberghs et al. (2010). The derivation is shown in their supplementary

materials (their Appendix D). The marginal density of the marginal linear mixed probit model is the following:

$$P(Y_{i1} = 1, Y_{i2} = 1, \ldots, Y_{ip_i} = 1) = \Phi(\boldsymbol{X}_i\boldsymbol{\beta}; \boldsymbol{L}_i^{-1}), \tag{3.5}$$

with

$$\begin{aligned} L_i &= I - \boldsymbol{Z}_i\boldsymbol{M}_i^{-1}\boldsymbol{Z}_i', \\ \boldsymbol{M}_i &= \boldsymbol{\tau}^{-1} + \boldsymbol{Z}_i'\boldsymbol{Z}_i, \end{aligned}$$

and where $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are, respectively, $(p_i \times k)$ and $(p_i \times q)$ dimensional matrices of known covariates for respectively the fixed effects and the random effects.

3.2.3 Joint mixed model

In the hierarchical joint mixed model, $\boldsymbol{b}_i$ denotes the vector of all random effects of all binary and continuous responses. This vector is distributed with a zero-mean normal distribution with covariance matrix $\boldsymbol{D}$: $\boldsymbol{b_i} \sim N(\boldsymbol{0}, \boldsymbol{D})$. The latter covariance matrix gauges both the correlation among the repeated measurements of the same response and the correlations between the vectors of measurements of separate responses. We assume that the responses are independent conditional on the random effect and hence that the random effects completely capture the correlation between the responses. As a result, the joint density of the responses conditional on the random effects equals the product of the conditional densities of the separate responses. To obtain the marginal joint density, the random effects are integrated out of the joint density. Let $\boldsymbol{Y}_{ci}$ denote the vector of all measurements of the a continuous responses: $\boldsymbol{Y}_{ci} = (Y^c_{1i1}, ...Y^c_{1in_{1i}}, Y^c_{2i1}, ..., Y^c_{2in_{2i}}, ...Y^c_{ai1}, ..., Y^c_{ain_{ai}})$. Note that the number of measurements for each response is not required to be equal. Similarly, let $\boldsymbol{Y}_{bi}$ denote a vector of all the measurements of the b binary responses: $\boldsymbol{Y}_{bi} = (Y^b_{1i1}, ...Y^b_{1ip_{1i}}, Y^b_{2i1}, ..., Y^b_{2ip_{2i}}, ...Y^b_{bi1}, ..., Y^b_{bip_{bi}})$. In addition, the matrices of covariates of the random effects of both the continuous responses and binary responses, respectively, $\boldsymbol{Z}_{ci}$ and $\boldsymbol{Z}_{bi}$ consists of the concatenation of the matrices of covariates of the separate responses $\boldsymbol{Z}_{ci} = [\boldsymbol{Z}^c_{1i}, \boldsymbol{Z}^c_{2i}, ..., \boldsymbol{Z}^c_{ai}]'$ for the continuous responses and $\boldsymbol{Z}_{bi} = [\boldsymbol{Z}^b_{1i}, \boldsymbol{Z}^b_{2i}, ..., \boldsymbol{Z}^b_{bi}]'$ for the binary responses. The matrices of covariates of fixed effects are defined similarly: for the continuous responses $\boldsymbol{X}_{ci} = [\boldsymbol{X}^c_{1i}, \boldsymbol{X}^c_{2i}, ..., \boldsymbol{X}^c_{ai}]'$ and for the binary responses $\boldsymbol{X}_{bi} = [\boldsymbol{X}^b_{1i}, \boldsymbol{X}^b_{2i}, ..., \boldsymbol{X}^b_{bi}]'$. As a consequence, the derivation of the marginal joint mixed model is almost identical as described in Delporte et al. (2022) (their Appendix A). The marginal joint density for the binary(probit)-normal model is the following:

$$f(\boldsymbol{y}_{ci}, \boldsymbol{y}_{bi} = 1) = \phi(\boldsymbol{X}_{ci}\boldsymbol{\beta}; \boldsymbol{V}_i)\Phi(\boldsymbol{X}_{bi}\boldsymbol{\beta} - \boldsymbol{\alpha}_i; \boldsymbol{B}_i), \tag{3.6}$$

with

$$\begin{aligned}
\boldsymbol{V}_i &= \boldsymbol{\Sigma}_i + \boldsymbol{Z}_{ci}\boldsymbol{D}\boldsymbol{Z}'_{ci}, \\
\boldsymbol{B}_i^{-1} &= \boldsymbol{I} - \boldsymbol{Z}_{bi}\boldsymbol{K}_i\boldsymbol{Z}'_{bi}, \\
\boldsymbol{K}_i^{-1} &= \boldsymbol{D}^{-1} + \boldsymbol{Z}'_{ci}\boldsymbol{\Sigma}_i^{-1}\boldsymbol{Z}_{ci} + \boldsymbol{Z}'_{bi}\boldsymbol{Z}_{bi}, \\
\boldsymbol{\alpha}_i &= \boldsymbol{H}_i(\boldsymbol{Y}_{ci} - \boldsymbol{X}_{ci}\boldsymbol{\beta}), \\
\boldsymbol{H}_i &= -\boldsymbol{B}_i\boldsymbol{Z}_{bi}\boldsymbol{K}_i\boldsymbol{Z}'_{ci}\boldsymbol{\Sigma}_i^{-1},
\end{aligned}$$

and where $\boldsymbol{\Sigma}_i$ is a block diagonal matrix with as blocks the variance-covariance matrices of the continuous responses

$$\boldsymbol{\Sigma}_i = \begin{bmatrix} \boldsymbol{\Sigma}_{1i} & 0 & ... & 0 \\ 0 & \boldsymbol{\Sigma}_{2i} & .. & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & \boldsymbol{\Sigma}_{ai} \end{bmatrix}.$$

3.2.4 Conditional distributions derived from the joint model

The conditional distributions of subvectors of response(s) can be derived from (3.6). Let $\boldsymbol{Y}_{ci}$ and $\boldsymbol{Y}_{bi}$ be defined identically as in (3.6): $\widetilde{\boldsymbol{Y}}_{ci}$ denotes an $\widetilde{n}_i$-dimensional subvector of the continuous response vector $\boldsymbol{Y}_{ci}$ and $\widetilde{\boldsymbol{Y}}_{bi}$ an $\widetilde{p}_i$-dimensional subvector of the binary response vector $\boldsymbol{Y}_{bi}$. Note that $\widetilde{\boldsymbol{Y}}_{ci}$ and $\widetilde{\boldsymbol{Y}}_{bi}$ can contain values from different responses. Similarly, let $\widetilde{\boldsymbol{X}}_{ci}$ and $\widetilde{\boldsymbol{X}}_{bi}$ denote the submatrices of, respectively, $\boldsymbol{X}_{ci}$ and $\boldsymbol{X}_{bi}$ corresponding to the subvectors $\widetilde{\boldsymbol{Y}}_{ci}$ and $\widetilde{\boldsymbol{Y}}_{bi}$. The expected values and prediction intervals of a subvector of one of the continuous responses conditional on a subvector of binary responses can be derived from the ratios of the marginal distributions. The conditional expected value is the following:

$$\begin{aligned}
E[\widetilde{\boldsymbol{Y}}_{\boldsymbol{ci}}|\widetilde{\boldsymbol{Y}}_{\boldsymbol{bi}} = \mathbf{1}] = \quad & \boldsymbol{E}_i(\boldsymbol{V}_i^{-1}\widetilde{\boldsymbol{X}}_{ci}\boldsymbol{\beta} + \boldsymbol{H}'_i\boldsymbol{B}_i^{-1}\boldsymbol{F}_i) \\
& + \boldsymbol{E}_i\boldsymbol{H}'_i\boldsymbol{B}_i^{-1}\boldsymbol{T}_i\big[-F_1(o_1) \quad -F_2(o_2) \quad ... \quad -F_p(o_p)\big],
\end{aligned} \tag{3.7}$$

where

$$
\begin{aligned}
\boldsymbol{E}_i^{-1} &= \boldsymbol{H}_i' \boldsymbol{B}_i^{-1} \boldsymbol{H}_i + \boldsymbol{V}_i^{-1} \\
\boldsymbol{T}_i^{-1} &= \boldsymbol{B}_i^{-1} - (\boldsymbol{H}_i' \boldsymbol{B}_i^{-1})' \boldsymbol{E}_i (\boldsymbol{H}_i' \boldsymbol{B}_i^{-1}), \\
\boldsymbol{F}_i &= \boldsymbol{T}_i \cdot (\boldsymbol{H}_i' \boldsymbol{B}_i^{-1})' \boldsymbol{E}_i (\boldsymbol{V}_i^{-1} \widetilde{\boldsymbol{X}}_{ci} \boldsymbol{\beta}), \\
o &= \widetilde{\boldsymbol{X}}_{bi} \boldsymbol{\beta} + \boldsymbol{H}_i \widetilde{\boldsymbol{X}}_{ci} \boldsymbol{\beta} - \boldsymbol{F}_i \\
F_i(x_i) &= \int_{-\infty}^{o_1} \dots \int_{-\infty}^{o_{i-1}} \int_{-\infty}^{o_{i+1}} \dots \int_{-\infty}^{o_{\widetilde{p}_i}} \varphi(x_1, ...x_{i-1}, x, x_{i+1}, ...x_{\widetilde{p}_i}) dx_{\widetilde{n}_i}, ..dx_{i+1} dx_{i-1} ... dx_1, \\
\varphi(\boldsymbol{x}) &= \begin{cases} \frac{\phi(\boldsymbol{x}, \boldsymbol{T}_i)}{\Phi(\widetilde{\boldsymbol{X}}_{bi}\boldsymbol{\beta} + \boldsymbol{H}_i \widetilde{\boldsymbol{X}}_{ci} \boldsymbol{\beta}, \boldsymbol{F}_i, \boldsymbol{T}_i)}, & \text{for } \boldsymbol{x} \leq \widetilde{\boldsymbol{X}}_{bi} \boldsymbol{\beta} + \boldsymbol{H}_i \widetilde{\boldsymbol{X}}_{ci} \boldsymbol{\beta} - \boldsymbol{F}_i, \\ 0, & \text{otherwise.} \end{cases}
\end{aligned}
$$

The derivation is shown in Online Appendix A and the prediction interval can be found in Online Appendix B.
More generally, the conditional expected value of a subvector of continuous response(s) can be specified conditional on both a subvector of continuous response(s) and binary response(s). Let $\widetilde{\boldsymbol{Y}}_{ci}^{a}$ denote an n_a-dimensional subvector of continuous response(s) for which we compute the conditional expected value. Let $\widetilde{\boldsymbol{Y}}_{ci}^{b}$ denote the subvector of length n_b of values of the continuous vector $\widetilde{\boldsymbol{Y}}_{ci}$ we will condition upon. The superscript specifies the submatrices or subvectors; superscript a and b denote, respectively, the rows a_1 until a_{n_a} and b_1 to b_{n_b}. $\widetilde{\boldsymbol{Y}}_{bi}$ still denotes a subvector of the binary response(s). The conditional expectation of $\widetilde{\boldsymbol{Y}}_{ci}^{a}$ is the following:

$$
\begin{aligned}
E[\widetilde{\boldsymbol{Y}}_{\boldsymbol{ci}}^{\boldsymbol{a}} | \widetilde{\boldsymbol{Y}}_{\boldsymbol{ci}}^{\boldsymbol{b}} = \widetilde{\boldsymbol{y}}_{\boldsymbol{ci}}^{\boldsymbol{b}}, \widetilde{\boldsymbol{y}}_{\boldsymbol{bi}} = \boldsymbol{1}] &= \Big((\boldsymbol{E}_i \boldsymbol{V}_i^{-1} \widetilde{\boldsymbol{X}}_{ci} \boldsymbol{\beta}_1)^a \\
&\quad + \boldsymbol{E}_i^{ab} (\boldsymbol{E}_i^{bb})^{-1} (\widetilde{\boldsymbol{y}}_{\boldsymbol{ci}}^{\boldsymbol{b}} - (\boldsymbol{E}_i \boldsymbol{V}_i^{-1} \widetilde{\boldsymbol{X}}_{ci} \boldsymbol{\beta}_1)^b) \Big) \\
&+ \Big((\boldsymbol{E}_i \boldsymbol{H}_i' \boldsymbol{B}_i^{-1})^a - \boldsymbol{E}_i^{ab} (\boldsymbol{E}_i^{bb})^{-1} (\boldsymbol{E}_i \boldsymbol{H}_i' \boldsymbol{B}_i^{-1})^b \Big) \\
&\times \Big(\boldsymbol{T}_i \big[-F_1(o_1) \quad -F_2(o_2) \quad ... \quad -F_p(o_p) \big] + \boldsymbol{F}_i) \Big),
\end{aligned} \tag{3.8}
$$

with

$$
\begin{aligned}
\boldsymbol{E}_i^{-1} &= \boldsymbol{H}_i' \boldsymbol{B}_i^{-1} \boldsymbol{H}_i + \boldsymbol{V}_i^{-1}, \\
\boldsymbol{T}_i^{-1} &= (\boldsymbol{E}_i \boldsymbol{H}_i' \boldsymbol{B}_i^{-1})^{b'} (\boldsymbol{E}_i^{bb})^{-1} (\boldsymbol{E}_i \boldsymbol{H}_i' \boldsymbol{B}_i^{-1})^b + \boldsymbol{B}_i^{-1} - (\boldsymbol{H}_i' \boldsymbol{B}_i^{-1})' \boldsymbol{E}_i (\boldsymbol{H}_i' \boldsymbol{B}_i^{-1}), \\
\boldsymbol{F}_i &= \boldsymbol{T}_i \cdot \Big((\boldsymbol{E}_i \boldsymbol{H}_i' \boldsymbol{B}_i^{-1})^{b'} (\boldsymbol{E}_i^{bb})^{-1} (\widetilde{\boldsymbol{y}}_{\boldsymbol{ci}}^{\boldsymbol{b}} - (\boldsymbol{E}_i \boldsymbol{V}_i^{-1} \widetilde{\boldsymbol{X}}_{ci} \boldsymbol{\beta}_1)^b) + (\boldsymbol{H}_i' \boldsymbol{B}_i^{-1})' \boldsymbol{E}_i (\boldsymbol{V}_i^{-1} \widetilde{\boldsymbol{X}}_{ci} \boldsymbol{\beta}) \Big),
\end{aligned}
$$

where the superscript bb specifies the rows b_1 until b_{n_b} and columns b_1 until b_{n_b}. The superscript ab indicates row a_1 until a_{n_a} and column b_1 until b_{n_b}. The derivation can be found in Online Appendix A and the corresponding prediction interval is shown in Online Appendix B.
Next, the conditional probability of a subvector of the the binary response(s) can be calculated conditional on a subvector of the continuous response(s). The probability of subvector of successes of the binary responses conditional on a subvector of measurements of the continuous response(s) equals the following:

$$P(\widetilde{\boldsymbol{y}}_{\boldsymbol{bi}} = \mathbf{1}|\widetilde{\boldsymbol{y}}_{\boldsymbol{ci}}) = \frac{\phi(\widetilde{\boldsymbol{X}}_{ci}\boldsymbol{\beta}; \boldsymbol{V}_i)\Phi(\widetilde{\boldsymbol{X}}_{bi}\boldsymbol{\beta} - \boldsymbol{\alpha}_i; \boldsymbol{B}_i)}{\phi(\widetilde{\boldsymbol{X}}_{ci}\boldsymbol{\beta}; \boldsymbol{V}_i)} = \Phi(\widetilde{\boldsymbol{X}}_{bi}\boldsymbol{\beta} - \boldsymbol{\alpha}_i; \boldsymbol{B}_i). \tag{3.9}$$

The corresponding confidence interval derived by Delporte et al. (2022) was constructed applying the delta method directly to the probability and therefore was not restricted to the unit interval. An alternative method to ensure correct boundaries, is computing the confidence interval on the logit scale, and then transforming the so-obtained confidence interval back using the inverse logit transformation. This can be constructed as explained in Online Appendix C.
A more general conditional probability of a subvector of successes given both a subvector of the binary responses and a subvector of the continuous responses can be derived as well. The subvector of which the probability of successes is calculated is symbolized with $\widetilde{\boldsymbol{Y}}_{\boldsymbol{bi}}^{\boldsymbol{a}}$ and we will condition on a subvector of successes $\widetilde{\boldsymbol{Y}}_{\boldsymbol{bi}}^{\boldsymbol{b}}$ and subvector of the continuous responses $\widetilde{\boldsymbol{Y}}_{\boldsymbol{ci}}$. The use of the superscripts b and bb is in analogy with 3.8.

$$P(\widetilde{\boldsymbol{y}}_{\boldsymbol{bi}}^{\boldsymbol{a}} = \mathbf{1}|\widetilde{\boldsymbol{y}}_{\boldsymbol{ci}}, \widetilde{\boldsymbol{y}}_{\boldsymbol{bi}}^{\boldsymbol{b}}) \quad = \quad \frac{\Phi(\widetilde{\boldsymbol{X}}_{bi}\boldsymbol{\beta} - \boldsymbol{H}_i(\widetilde{\boldsymbol{y}}_{\boldsymbol{ci}} - \widetilde{\boldsymbol{X}}_{ci}\boldsymbol{\beta}); \boldsymbol{B}_i)}{\Phi(\widetilde{\boldsymbol{X}}_{bi}^{b}\boldsymbol{\beta} - \boldsymbol{H}_i^b(\widetilde{\boldsymbol{y}}_{\boldsymbol{ci}} - \widetilde{\boldsymbol{X}}_{ci}\boldsymbol{\beta}); \boldsymbol{B}_i^{bb})}, \tag{3.10}$$

The corresponding confidence interval described in Delporte et al. (2022) can possibly have boundaries outside the $[0; 1]$ range. Hence, the confidence interval is constructed similarly as in (3.9). This is shown in Online Appendix C.

3.2.5 Correlation function

Delporte et al. (2022) describe the derivation of the manifest correlation function between a binary and a continuous response from the binary(probit)-normal model for two responses. The manifest correlation quantifies the correlation between the observed responses while the latent correlation is the correlation between the underlying random effects. The latter is readily available from the estimates of the joint model, while the manifest correlation has to be derived from the marginal joint model (3.6). The correlation function depends on the

covariates and the time point of the continuous response j and the time point of the binary response k. The formula of the manifest correlation function between the binary response m and the continuous response l is the following:

$$\rho_{Y_{lij},Y_{mik}} = \frac{\frac{1}{L_i^{1/2}} \boldsymbol{z}'_{lij} \boldsymbol{M}_i^{-1} \boldsymbol{z}_{mik} \phi(L_i^{1/2} \boldsymbol{x}'_{mik}\boldsymbol{\beta})}{\sqrt{\left(\boldsymbol{z}'_{lij}\boldsymbol{D}_{lm}\boldsymbol{z}_{lij} + \Sigma_{lij}\right)\Phi(L_i^{1/2}\boldsymbol{x}'_{mik}\boldsymbol{\beta})(1-\Phi(L_i^{1/2}\boldsymbol{x}'_{mik}\boldsymbol{\beta}))}}, \quad (3.11)$$

where $\boldsymbol{D}_{lm}$ denotes the submatrix of $\boldsymbol{D}$ relating to the variances and covariances of the random effects of both responses and

$$\begin{aligned} \boldsymbol{M}_i &= \boldsymbol{D_{lm}}^{-1} + \boldsymbol{z}_{mik}\boldsymbol{z}'_{mik} \\ L_i &= 1 - \boldsymbol{z}'_{mik}\boldsymbol{M}_i^{-1}\boldsymbol{z}_{mik}. \end{aligned}$$

We extended the methodology of Delporte et al. (2022) by simplifying the expression and calculating the formula of the corresponding standard errors. The proof and formulas can be found in Online Appendix D.

3.2.6 Parameter estimation

In random effects models, the marginal likelihood is maximized in order to estimate the parameters (Molenberghs and Verbeke, 2005). The likelihood function equals:

$$L(\boldsymbol{\beta}, \boldsymbol{D}) = \prod_{i=1}^{N} \int f_{ci}(\boldsymbol{Y}_{ci}|\boldsymbol{b}_i) f_{bi}(\boldsymbol{Y}_{bi} = 1|\boldsymbol{b}_i) f(\boldsymbol{b}_i|D) d\boldsymbol{b}_i, \quad (3.12)$$

where the assumption is made that the measurements within a response sequence and between response sequences are independent conditional on the random effects $\boldsymbol{b}_i$. In most cases, the integral cannot be solved analytically and numerical procedures are applied, such as Gaussian quadrature and adaptive Gaussian quadrature. A pre-specified level of accuracy can be defined and increased until numerical stability occurs (Pinheiro and Bates, 1995). This paper uses the `SAS` procedures `MIXED`, `GLIMMIX` and `NLMIXED`. In all the latter procedures it is possible to assign sample weights, to account for example for unrepresentative data. The code for the second case study be found in Online Appendix G.

3.3 Case studies

3.3.1 COVID-19 data

The COVID-19 data were collected between March 2020 and March 2021 at five key moments in the pandemic. On these time points, $1,000$ Flemish adults, representative for gender, age, educational attainment, and province, participated in an online survey about their media use and perceived vulnerability to disease in the context of the COVID-19-pandemic (De Coninck et al., 2022). Perceived vulnerability to disease can be subdivided into two continuous subconstructs. The first, perceived infectability, encompasses beliefs about the susceptibility to infectious diseases. The second, germ aversion, measures the emotional distress in situations that have a high risk for potential pathogen transmission (Duncan et al., 2009). The mean responses are plotted in Figure 3.1 on top of the individual responses of a random sample of 100 subjects. Participants were also asked how often they followed COVID-19 related news on several media sources in the week prior to the questionnaire. Answer options ranged from one (never) to five (multiple times a day). The latter variables were discretized into a binary response: one and two were transformed into zero (=low), and three through five were coded as one (=high). In this case study, the focus lies on three media sources: quality newspapers, social media of quality newspapers and the internet. The proportion of participants who rate their media consumption as high is shown in Figure 3.2. The analysis focuses on the association between the binary media consumption and the continuous perceived infectability on one hand, and between binary media consumption and the continuous germ aversion on the other hand. Each of those longitudinal variables are initially treated as a response by the joint model. In addition, it is of interest to predict the probability to consume a quality newspaper regularly in May 2022, when several COVID-19 measures were lifted. Here, the conditional model is derived from the joint model. A direct URL to the data is: https://data.mendeley.com/datasets/mhx3p7w3d6/9.

3.3.2 Vaccination data

The vaccination data are from a five-wave longitudinal study in Belgium over a six-months period (December 2020 until May 2021). A nationally representative sample of approximately 5000 participants responded to questions about their own vaccination intention and the perceived vaccination intention of others of the same age and gender (Delporte et al., 2023). There were four response options, but we discretized participants' responses. More specifically, we recoded "Try to get the vaccine as soon as possible" and "Get vaccinated, but without any haste" as a positive vaccination intention and the responses "Wait until there

is a lot of experience with it before possibly getting vaccinated" or "Certainly not get vaccinated" as a negative vaccination intention. Vaccination became available in January 2021. From then on, a vaccinated participant was coded as a participant with a positive vaccination intention. The probability of both responses at each wave can be found in Table 3.1. Besides vaccination intentions, we measured personal and comparative optimism. Personal optimism was defined as reporting that one's own probability of experiencing a negative event was low or that one's own probability of experiencing a positive event was high. In addition, individuals estimated the probability of an event for someone else of the same age and gender. Comparative optimism is defined as the difference between the perceived own probability and the perceived probability for their peers. Our research question focuses on the comparative optimism for infection (getting infected or re-infected, infecting others) and the comparative optimism for a severe outcome (being admitted to hospital, and to an intensive care unit). It is of interest to investigate whether these two continuous responses are associated with the binary vaccination intention responses. Figure 3.3 displays the evolution of the mean of both continuous responses and the corresponding standard errors.

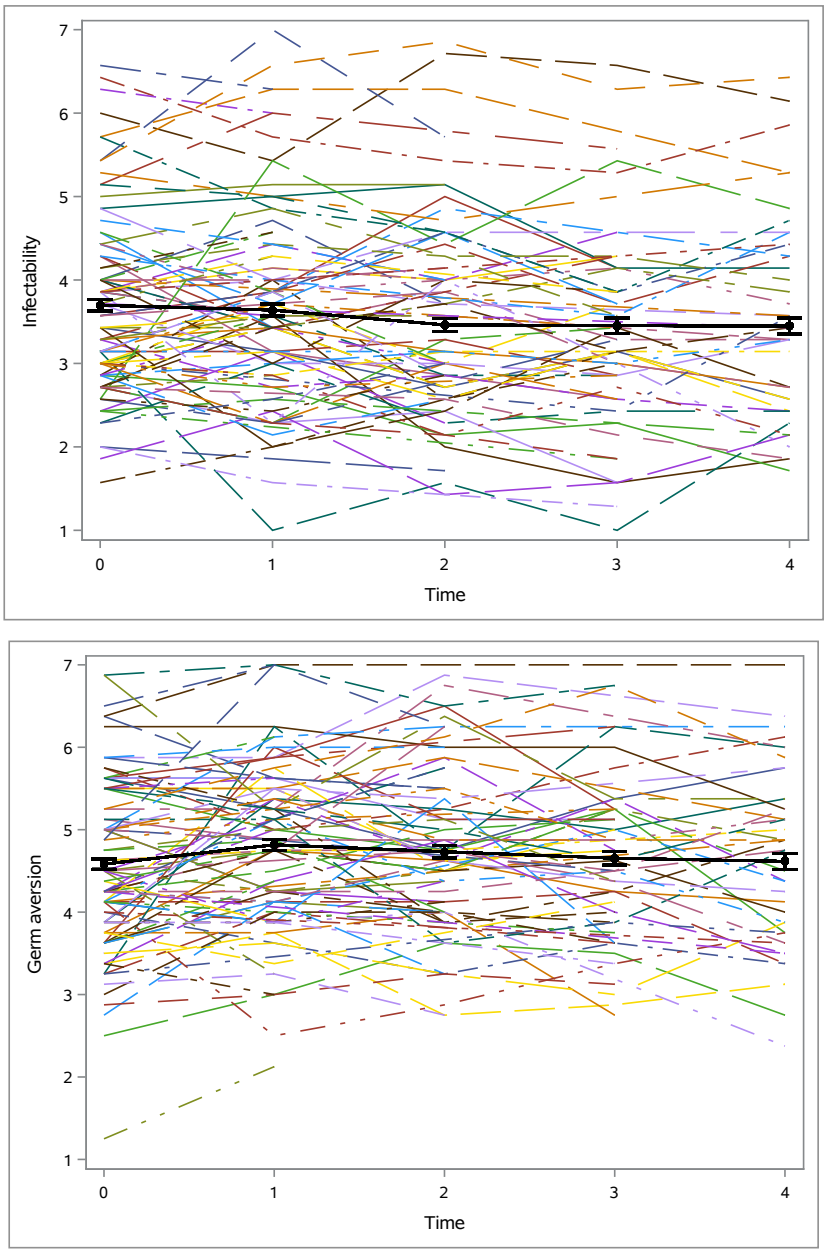

Figure 3.1: *Mean perceived infectability (upper) and germ aversion (lower) over time, with 95% confidence intervals on top of the individual responses of a random sample of* 100 *subjects in the COVID-19 study.*

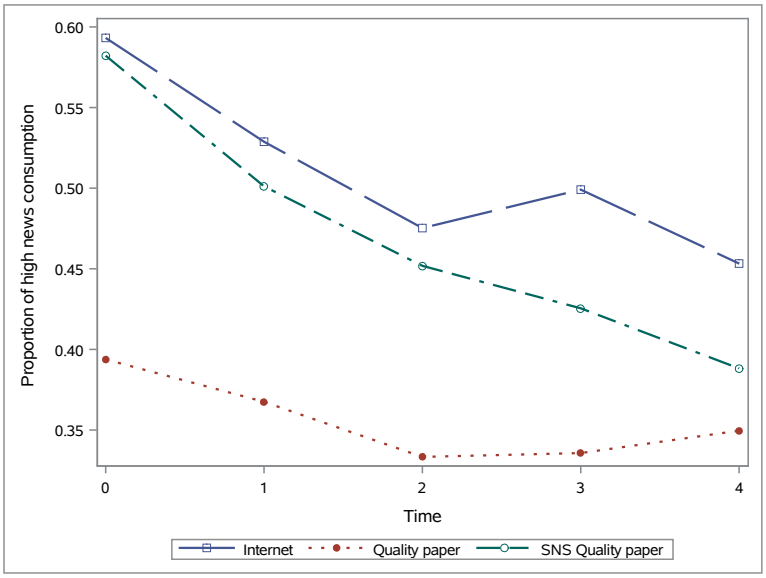

Figure 3.2: *Proportion of participants who have a high consumption of COVID-19 related news on the internet, quality newspapers or social media of quality newspapers over time in the COVID-19 study.*

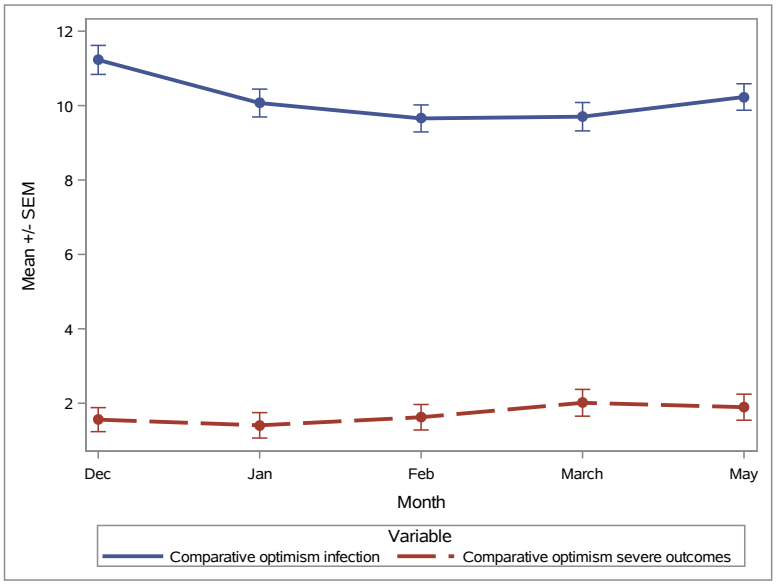

Figure 3.3: *Mean comparative optimism of infection and comparative optimism about severe outcomes over time with their standard errors in the Vaccination study.*

Table 3.1*: Prevalence of vaccination intention and perceived vaccination intention peers in the Vaccination study.*

Wave	Dec	Jan	Feb	March	May
% vaccination intention	60.11	73.68	80.06	82.08	88.65
% perceived vaccination intention peers	37.16	75.26	80.38	82.82	88.82

3.4 Analysis of the case studies

3.4.1 Analysis of the COVID-19 data

In this section, the relation between media consumption related to COVID-19 and perceived infectability and germ aversion is examined. The following joint model is constructed to model the five responses. The infectability, germ aversion, newspaper consumption, social media of newspaper consumption and internet consumption of person i at time j is denoted as, respectively, Y_{1ij},Y_{2ij},Y_{3ij}, Y_{4ij}, and Y_{5ij}. Every response is modelled by the same time-constant covariates for simplicity. It is clear from Figures 3.1 and 3.2 that the relationship between the responses and time is nonlinear. As a consequence, time is included in the model as a categorical variable with the first time point as reference category. In addition, the relationship between the responses and the six-category ordinal variable "Perceived income" was investigated. It was decided that a linear relationship was suitable, and the variable is hence treated as continuous by the model. An overview of the predictors is shown in Table 3.2.The hierarchical model is specified as:

$$\begin{aligned}
Y_{1ij} &= \boldsymbol{X}_{1ij}\boldsymbol{\beta}_1 + b_{10i} + b_{11i}t_{ij} + \epsilon_{1ij},\\
Y_{2ij} &= \boldsymbol{X}_{2ij}\boldsymbol{\beta}_2 + b_{20i} + b_{21i}t_{ij} + \epsilon_{2ij},\\
\Phi^{-1}(P(Y_{3ij}=1)) &= \boldsymbol{X}_{3ij}\boldsymbol{\beta}_3 + b_{30i} + b_{31i}t_{ij},\\
\Phi^{-1}(P(Y_{4ij}=1)) &= \boldsymbol{X}_{4ij}\boldsymbol{\beta}_4 + b_{40i} + b_{41i}t_{ij},\\
\Phi^{-1}(P(Y_{5ij}=1)) &= \boldsymbol{X}_{5ij}\boldsymbol{\beta}_5 + b_{50i} + b_{51i}t_{ij},
\end{aligned}$$

where $\boldsymbol{b}_{k0i}$ $(k = 1, .., 5)$ are the random intercepts, $\boldsymbol{b}_{k1i}$ denote the random slopes for time and ϵ_{1ij} and ϵ_{2ij} denote the usual error components. The ten random effects follow jointly a zero-mean normal distribution with a covariance matrix $\boldsymbol{D}$. The two error components follow a bivariate zero-mean normal distribution and are assumed to be uncorrelated. In addition, all fixed effects parameters are response-specific. The joint model was specified in `SAS NLMIXED`, but runtime on a TRIER 2 supercomputer exceeded 7 days. The pairwise method of Fieuws and Verbeke (2006) was implemented in order to decrease computational complexity. With this method, ten pairwise joint models had to be fitted and the

estimates combined with pseudo-likelihood methods. Sample splitting (Molenberghs et al., 2011b) was not applied, since convergence was reached within 27 hours on a regular laptop (CPU=Processor Intel(R) Core(TM) i7-8850H CPU @ 2.60GHz, 2592 Mhz, 6 Core(s), 12 Logical Processor(s), RAM=24GB). The pairwise method and sample splitting are described in Online Appendix E. The resulting parameter estimates, with as convergence criterion a relative gradient (`gconv`) equal to 10^{-8}, of the model can be found in Online Appendix F. The latent correlations between the random effects are shown in Table 3.3. A Wald test is implemented to investigate which responses are significantly correlated. The null hypothesis of the latter hypothesis tests whether all the covariances between the random effects of different responses equal 0, whereas the alternative hypothesis states that at least one covariance is not equal to 0. Table 3.4 shows the results and indicates that a significant association exists between perceived infectability and quality newspaper consumption, and between perceived infectability and internet consumption. In addition, germ aversion is significantly associated with internet consumption.

Table 3.2: *Predictors in the model.*

Predictor	Type	Options	Reference
Time	Categorical	1 to 5	
Gender	Categorical	Male or Female	Male
Age	Numerical		
Student	Categorical	Yes or No	No
Permanent disability	Categorical	Yes or No	No
Children	Categorical	Yes or No	No
Biological or adoptive (grand)parents that are 60 years or older	Categorical	Yes or No	No
Perception of income prior to the pandemic	Numerical	Very difficult to very easy	
Living situation	Categorical	Small city or large city or the suburbs of a large city or a village or in the countryside	Small city

Table 3.3*: Estimated correlation matrix of the latent random effects of the responses (the random effects b_{10i} and b_{11i} are from perceived infectability, b_{20i} and b_{21i} from germ aversion, b_{30i} and b_{31i} from newspaper consumption, b_{40i} and b_{41i} from social media consumption, and b_{50i} and b_{51i} from internet consumption).*

b_{10i}	b_{11i}	b_{20i}	b_{21i}	b_{30i}	b_{31i}	b_{40i}	b_{41i}	b_{50i}	b_{51i}
1	-.11	.18	-.05	.08	$-.03$	.05	-.02	.05	.10
-	1	$-.13$	.43	.04	0	$-.07$	.01	$-.08$	.22
-	-	1	$-.04$	$-.01$	.12	.03	$-.01$	.06	$-.10$
-	-	-	1	$-.010$	$-.08$	$-.11$	.08	$-.06$	.47
-	-	-	-	1	.05	.33	.01	.14	.24
-	-	-	-	-	1	$-.01$	.53	.06	.11
-	-	-	-	-	-	1	.09	.23	$-.13$
-	-	-	-	-	-	-	1	.06	.77
-	-	-	-	-	-	-	-	1	.05
-	-	-	-	-	-	-	-	-	1

Table 3.4*: p-values of the hypothesis tests with as null-hypothesis that there is no association between two responses.*

	Quality newspaper	Social media of quality newspaper	Internet
Infectability	0.007	0.363	0.040
Germ aversion	0.404	0.374	$<$0.001

Table 3.5*: Manifest correlations (95%CI) between perceived infectability (infect) and consumption of quality newspaper (paper). The covariates are set to specific values to maximize the correlations.*

	Wave(paper)				
Wave(infect)	0	1	2	3	4
0	.072[.071;.073]	.068[.067;.069]	.059[.057;.060]	.053[.051;.054]	.048[.046;.050]
1	.076[.075;.077]	.071[.070;.073]	.062[.061;.063]	.056[.054;.058]	.051[.049;.053]
2	.079[.078;.080]	.075[.074;.076]	.065[.063;.066]	.059[.057;.060]	.054[.052;.056]
3	.082[.080;.083]	.077[.076;.078]	.067[.066;.068]	.061[.059;.063]	.056[.054;.058]
4	.084[.082;.085]	.079[.078;.081]	.069[.067;.070]	.063[.061;.065]	.058[.056;.060]

Next, we computed the manifest correlations, using (3.11). The covariates are set to the values that result in the highest probability of a success of the binary

response to obtain the maximal correlation. The manifest correlation matrices without significant results are shown in Online Appendix F. Table 3.5 contains the manifest correlations matrices with significant results, which are between quality newspaper consumption and perceived infectability. They indicate that consumption at the first two timepoints is significantly positively linearly correlated with infectability at the other timepoints. In addition, quality newspaper consumption at the third time point is significantly linearly correlated with infectability at times 2, 3 and 4. A possible explanation is that high consumption of COVID-related news in newspapers induces anxiety in some individuals. These individuals sometimes decide to reduce their media consumption, which in turn may reduce anxiety. Hence, a negative feedback loop exists.
Next, predictions are made whether or not an individual will read newspapers at the third time point, conditional on the other responses at the first and second time point (Table 3.6). The continuous responses perceived infectability and germ aversion were either set to the mean, one standard deviation above the mean or one standard deviation below the mean. The most important effect on the predicted probability is the history of newspaper consumption, but germ aversion and infectability also have a small positive impact on the predicted probability. Note that all the predictions were conditional on no or very little consumption of COVID-19-news through the internet or social media of quality newspapers.

Table 3.6*: Predicted probability of reading the newspaper on the third time point conditional on the history of perceived infectability, germ aversion and consumption of quality newspapers, social media of quality newspapers and internet at the two preceding timepoints.*

Infectability	Germ aversion	Paper	Internet	Social media	P	CI
Mean	Mean	0	0	0	.035	[.028;.044]
Mean	-1 SD	0	0	0	.033	[.026;.043]
Mean	+1SD	0	0	0	.037	[.031;.043]
-1 SD	Mean	0	0	0	.031	[.025;.039]
-1 SD	-1 SD	0	0	0	.030	[.024;.038]
-1 SD	+1SD	0	0	0	.033	[.026;.041]
+1SD	Mean	0	0	0	.039	[.032;.046]
+1SD	-1 SD	0	0	0	.037	[.031;.044]
+1SD	+1SD	0	0	0	.040	[.034;.048]
Mean	Mean	1	0	0	.643	[.598;.686]
Mean	-1 SD	1	0	0	.635	[.591;.677]
Mean	+1SD	1	0	0	.652	[.616;.686]
-1 SD	Mean	1	0	0	.633	[.583;.680]
-1 SD	-1 SD	1	0	0	.624	[.576;.670]
-1 SD	+1SD	1	0	0	.642	[.590;.690]
+1SD	Mean	1	0	0	.655	[.620;.687]
+1SD	-1 SD	1	0	0	.647	[.612;.679]
+1SD	+1SD	1	0	0	.663	[.628;.696]

3.4.2 Analysis of the vaccination data

In this case study, we examine the relationship between comparative optimism and vaccination. Two linear mixed models are constructed for comparative optimism for infection and comparative optimism for severe outcomes of person i at time j, denoted by, respectively, Y_{1ij} and Y_{2ij}. In addition, two generalised linear mixed models with a probit link are fitted for own vaccination intention and the perceived vaccination intention of peers of person i at time j, denoted by, respectively, Y_{3ij} and Y_{4ij}. For the continuous responses, time is included as a categorical variable since there is a non-linear relationship with the response (Figure 3.3). In contrast, for the binary responses time is included as an continuous variable. Still, in the latter, an indicator variable is included for the first timepoint, since the relationship with the response only became linear on the probit-scale at the second time point. The other covariates are the categorical age group, gender and region and the interactions between these covariates and time. In each of the models, a random intercept and a random slope is included and correlations were allowed between the random effects. Hence, the full model can be denoted as follows

$$\begin{aligned}
Y_{1ij} &= \boldsymbol{X}_{1ij}\boldsymbol{\beta}_1 + b_{10i} + b_{11i}t_{ij} + \epsilon_{1ij},\\
Y_{2ij} &= \boldsymbol{X}_{2ij}\boldsymbol{\beta}_2 + b_{20i} + b_{21i}t_{ij} + \epsilon_{2ij},\\
\Phi^{-1}(P(Y_{3ij}=1)) &= \boldsymbol{X}_{3ij}\boldsymbol{\beta}_3 + b_{30i} + b_{31i}t_{ij},\\
\Phi^{-1}(P(Y_{4ij}=1)) &= \boldsymbol{X}_{4ij}\boldsymbol{\beta}_4 + b_{40i} + b_{41i}t_{ij}.
\end{aligned}$$

The random intercepts ($b_{10i}, b_{20i}, b_{30i}$ and b_{40i}) and the random slopes ($b_{11i}, b_{21i}, b_{31i}$ and b_{41i}) follow a multivariate normal distribution with mean $\mathbf{0}$ and variance-covariance $\boldsymbol{D}$. In addition, the errors of the continuous responses were assumed independent and distributed with a multivariate normal distribution around 0.
Because the dataset involved a large number of observations, computational problems arose, despite the implementation of the pairwise method of Fieuws et al. (2006). Hence the latter method was combined with the method of Molenberghs et al. (2011b) and Ivanova et al. (2017). These methods are described in Online Appendix E. Since the pairwise models on the samples could run in parallel, the computing time was the duration to fit the most computationally intensive subsample and pairwise model. Hence, the total computing time was 136 hours on a TRIER 2 supercomputer. In addition, it is feasible to perform the analysis on a regular computer as well. Model fitting was done with the `SAS` procedures `NLMIXED`, `GLIMMIX` and `MIXED`, with 5 quadrature points (if applicable) and as the convergence criterion a relative gradient (`gconv`) set to 10^{-8}. Parts of the code, resulting parameter estimates and the variance-covariance matrix of the random effects can be found in Online Appendix G. The correlation matrix between the random effects is shown in Table 3.7.

Table 3.7*: Estimated correlation matrix of the latent random effecs of the responses (the random effects b_{10i} and b_{11i} are from comparative optimism of infection, b_{20i} and b_{21i} from comparative optimism of severe outcomes, b_{30i} and b_{31i} from own vaccination intention, and b_{40i} and b_{41i} from perceived vaccination intention of peers).*

b_{10i}	b_{11i}	b_{20i}	b_{21i}	b_{30i}	b_{31i}	b_{40i}	b_{41i}
1	-0.47	0.30	-0.35	-0.01	0.01	0.05	-0.06
-	1	-0.28	0.89	-0.11	0.13	-0.03	0.05
-	-	1	-0.26	0.14	-0.14	0.03	-0.04
-	-	-	1	-0.10	0.13	-0.01	0.07
-	-	-	-	1	-0.98	0.82	-0.89
-	-	-	-	-	1	-0.51	0.75
-	-	-	-	-	-	1	-0.76
-	-	-	-	-	-	-	1

Since the correlations between the binary and continuous responses are quite small, in contrast to the correlations amongst binary responses and the correlations amongst the continuous responses, the significance of the association is tested. A Wald test is performed to test the null hypotheses that all covariances of the random effects between the responses equal zero. The results are shown in Table 3.8. It is clear that there is no association between the random effects of both responses. The manifest correlations are scrutinized as well. There are no significant linear correlations between comparative optimism and perceived vaccination intention of the peers (Online Appendix G). Still, there exists a positive association between vaccination intention at wave 1 and comparative optimism of severe outcomes at wave $1, 2$ and 3 (Table 3.9). This means that vaccination intention at the start is related to the belief that one is less likely to get severely sick from a COVID infection compared to his peers. Vaccination intention at later waves is negatively related to comparative optimism of severe outcomes at wave 1 and 2. This indicates that after the vaccine became available, individuals who perceive themselves less at risk of getting severely sick after an infection, are less motivated to get a vaccine. At the second wave, the side effects of vaccination became largely known (e.g., fever). It is possible that at wave 1, when vaccines were not yet available, individuals with vaccination intention expected to have less severe outcomes of infection. They possibly expected to be vaccinated more early than others. The vaccination campaign in Belgium had, however, a strict prioritization based on age. Once vaccines became available, they thus realised that they were not in control of when to be vaccinated. Hence, their expectation of the relative severeness of the disease could not be affected by the expectation

that they would get vaccinated sooner than their peers.
Table 3.9 suggests that vaccination intention at wave 1 is related to lower comparative optimism for infection at wave 5. Yet, vaccination intention at later waves is related to higher comparative optimism at wave 5.

Table 3.8: *p-values of the hypothesis tests with as null-hypothesis that there is no association between two responses.*

	Vaccination intention	Vaccination intention peers
Comparative optimism infection	0.73	0.95
Comparative optimism severe outcomes	0.33	0.97

Table 3.9: *Manifest correlations of own vaccine intention and comparative optimism.*

Panel A: Manifest correlations with comparative optimism of severe outcomes (CO)

	Wave(intention)				
Wave (CO)	1	2	3	4	5
1	**.058[.026;.089]**	**-.056[-.087;-.026]**	**-.057[-.085;-.028]**	**-.056[-.084;-.028]**	**-.056[-.083;-.028]**
2	**.049[.019;.078]**	**-.044[-.072;-.016]**	**-.046[-.072;-.019]**	**-.045[-.071;-.019]**	**-.045[-.071;-.019]**
3	**.039[.009;.069]**	**-.031[-.059;-.003]**	**-.033[-.060;-.007]**	**-.033[-.059;-.007]**	**-.033[-.059;-.007]**
4	.029[-.003;.060]	-.018[-.047;.012]	-.021[-.049;.008]	-.021[-.049;.007]	-.021[-.049;.007]
5	.018[-.017;.054]	-.005[-.037;.028]	-.009[-.040;.023]	-.009[-.040;.022]	-.010[-.040;.021]

Panel B: Manifest correlations with comparative optimism of infection (CO)

	Wave(intention)				
Wave (CO)	1	2	3	4	5
1	-.004[-.036;.028]	.005[-.027;.038]	.005[-.026;.036]	.005[-.026;.035]	.005[-.025;.035]
2	-.012[-.041;.017]	.016[-.012;.044]	.015[-.012;.042]	.015[-.012;.041]	.014[-.012;.041]
3	-.021[-.049;.007]	.027[.000;.053]	.025[.000;.050]	.024[-.001;.049]	.024[-.001;.049]
4	-.030[-.060;.000]	**.037[.009;.065]**	**.035[.008;.062]**	**.034[.008;.061]**	**.034[.007;.060]**
5	**-.038[-.072;-.004]**	**.047[.015;.079]**	**.045[.013;.076]**	**.043[.012;.074]**	**.043[.012;.074]**

Figure 3.4 displays the dynamic predictions for the comparative optimism of infection conditional on the history of comparative optimism and vaccination (time points before the dashed line). The predictions are conditional on a history of mean scores of comparative optimism of both infection and severe outcomes at the previous time points and either a history of consistent vaccination hesitancy (solid line) or vaccination intention (dashed line). The expected comparative optimism of infection at time 2 until 5 was lower for individuals with vaccination intention at the first time point, when vaccination was not available. However, if this vaccination intention persisted to the second time point, the predictions for comparative optimism at the next time points was higher than those of individuals without vaccination intention. This effect is also visible if the vaccination

intention persists until the third time point. If the history up to the fourth time point is taken into account, the prediction of comparative optimism of infection does not differ anymore based on the vaccination intention. This result suggests that pointing out the increased risk of infection to consistently vaccine hesitant individuals might not promote vaccination.
Figure 3.5 displays the probability of vaccination conditional on the history of both vaccination and comparative optimism. The dashed vertical line depicts up to which time point information is used for prediction. The dashed horizontal line represents individuals who are one standard deviation above the mean for both comparative optimism scores, while the solid line displays individuals with a history of 1 SD below the mean of comparative optimism. In terms of vaccination, prediction is made conditional on the fact that there is no vaccination or vaccination intention in the past. The graphs show that for individuals who have a history of vaccination hesitancy at time 2, individuals with also a history of a low comparative optimism are more inclined to vaccination at time point 3, 4, and 5. This effect is also visible at time point 3. If individuals still have no intention of vaccination at time point 4, the probability of vaccination is very low and does not seem to be affected by the history of comparative optimism. It can be hypothesized that risk perception does not play a role for individuals who are consistently vaccine hesitant up to time point 4, long after the vaccine became available.

3.5 Concluding remarks

In this paper, methodology is developed to model continuous or binary longitudinal response(s) with a set of longitudinal time-varying binary or continuous covariates. This is done by treating each of the longitudinal covariates as a response, and implementing a joint model with the response. In these joint models, univariate mixed models are applied and the random effects are allowed to be correlated. In the latter joint models, the interpretation is still conditional on the random effects. To circumvent this issue, the marginal model is derived by integrating out the random effects. Next, the conditional models are constructed to use the model for prediction. We derived conditional expected values and corresponding prediction intervals. With the latter methodology, predictions of a (sub)vector of one response type can be made given subvector(s) of the other responses and potentially a subvector of the same response. Confidence intervals for the predicted probability of the binary response were derived, where the interval is bounded between zero and one. This approach has several advantages over time dependent covariates. First, the lag does not have to be specified. In addition, no extra steps have to be taken to take missing data into account in the

responses. Next, there is no issue with endogeneity, which occurs when previous values of the response influence the later values of the time dependent covariate (Diggle, 2002). In addition, intermediate covariates do not pose a problem in the model.
Following this, manifest correlations were discussed. These are the correlations between the observed responses, in contrast to the latent correlations. Here the latent correlations are the correlations between the subject-specific underlying random effects. We extended the methodology of Delporte et al. (2022) by deriving the formulas of the standad errors of these manifest correlations.
Lastly, the methodology is illustrated in two case studies. First, a high-dimensional dataset about the relation between media consumption during COVID-19 and germ aversion and perceived infectability is analysed. In this case study, the pairwise method of Fieuws and Verbeke (2006) is implemented. The manifest correlations are examined and predictions for the binary newspaper consumption made conditional on the longitudinal covariates. Second, a case study about vaccination intention is performed. Since the data consists of four responses and a large number of measurements, the method of Ivanova et al. (2017) is implemented. Again the manifest correlations are scrutinized. In addition, dynamic predictions are made about the comparative optimism of infection, based on the history of vaccination intention and both comparative optimism of infection and severe outcomes. Next, the expected probability of vaccination is derived, conditional on the history of vaccination intention and comparative optimism about infections and severe outcomes. The results suggests that risk perception does not play a role anymore if vaccination hesitancy is persistent.
One of the main challenges in this context is the computational complexity of the high-dimensional model. A number of techniques had to be applied in order to reduce the computational complexity of the estimation of the joint models. In the first case study, the approach of Fieuws and Verbeke (2006) is implemented, where all pairwise models are fitted and are later combined into the joint model. In the second case study, the latter is combined with the partitioned sample method (Ivanova et al., 2017) with the purpose to fit the model on a subset of the data and combine the multiple subsets later on. This pairwise joint model can be easily implemented with existing software packages such as the `SAS` procedures `NLMIXED` and `GLIMMIX`. `R` was used in order to combine the subsamples and pairwise models into a single joint model, but this can also be done in `SAS`. Further research can be conducted on a marginalized joint model for other types of responses such as, for example, ordinal responses.

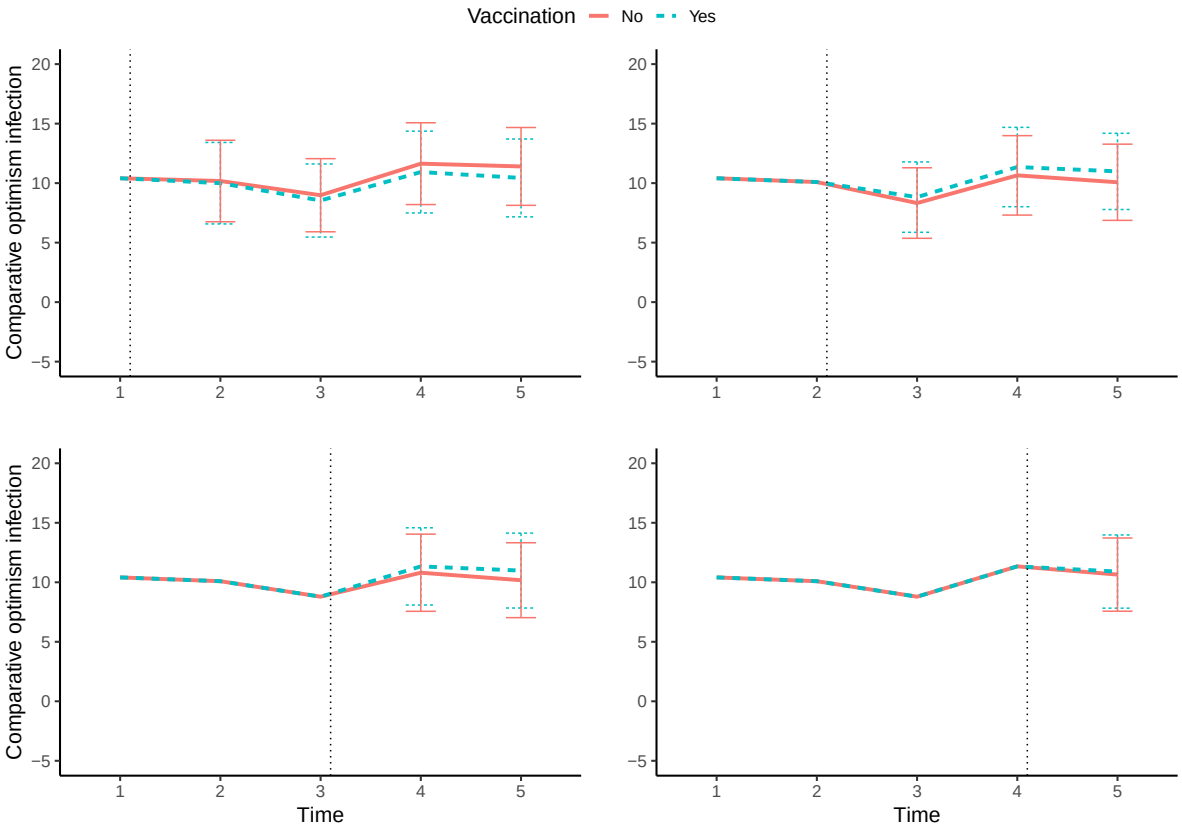

Figure 3.4: *Predictions (after the dashed line) for comparative optimism of infection based on the history (before the dashed line) of comparative optimism of infection (set to the mean) and vaccination (solid line= history of consistent vaccination hesitancy; dashed line= history of consistent vaccination intention).*

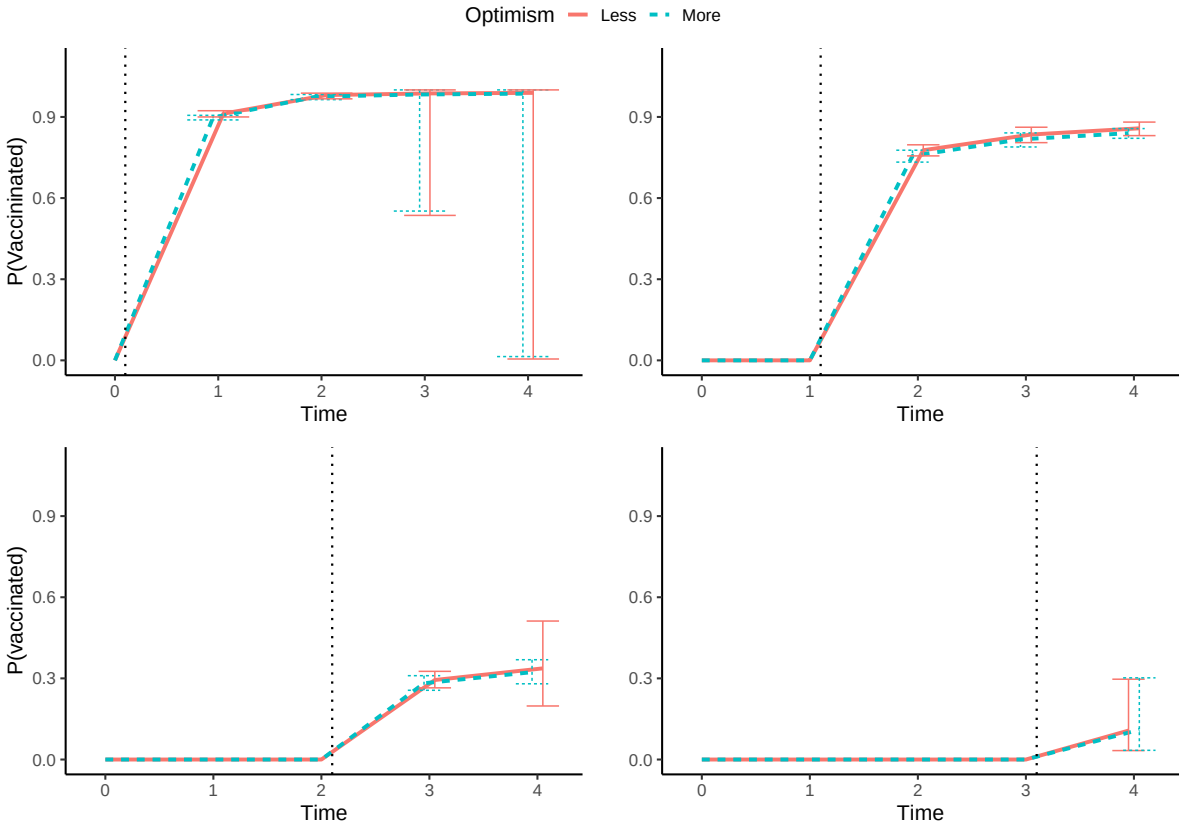

Figure 3.5: *Predictions (after the dashed line) and confidence intervals for vaccination probability based on the history (before the dashed line) of comparative optimism (red= consistently* 1 *standard deviation below the mean; blue= consistently* 1 *standard deviation above the mean) and vaccination (consistently no vaccination intention).*

Chapter 4

A joint normal-ordinal(probit) model for ordinal and continuous longitudinal data

This chapter is based upon:

Delporte, M., Molenberghs, G., Fieuws, S., & Verbeke, G. (2024). A Joint Normal-Ordinal (Probit) Model for Ordinal and Continuous Longitudinal Data. *Biostatistics*, Accepted.

The appendix is available via `margauxdelporte.github.io/Chapter4.pdf`.

Abstract

In biomedical studies, continuous and ordinal longitudinal variables are frequently encountered. In many of these studies it is of interest to estimate the effect of one of these longitudinal variables on the other. Time-dependent covariates have, however, several limitations; they can, for example, not be included when the data is not collected at fixed intervals. The issues can be circumvented by implementing joint models, where two or more longitudinal variables are treated as a response and modelled with a correlated random effect. Next, by conditioning on these response(s), we can study the effect of one or more longitudinal variables on another.We propose a normal-ordinal(probit) joint model. First, we derive closed-form formulas to estimate the model-based correlations between the responses on their original scale. In addition, we derive the marginal model, where the interpretation is no longer conditional on the random effects. As a consequence, we can make predictions for a subvector of one response conditional on the other response and potentially a subvector of the history of the response. Next, we extend the approach to a high-dimensional case with more than two ordinal and/or continuous longitudinal variables. The methodology is applied to a case study where, among others, a longitudinal ordinal response is predicted with a longitudinal continuous variable.

4.1 Introduction

In many clinical studies the same patients are repeatedly examined, which results in longitudinal data. This data can be analysed with generalised linear mixed models. This family of models encompasses linear mixed models, which was introduced by Laird and Ware (1982). Later the approach was extended to noncontinuous data by Breslow and Clayton (1993), Wolfinger and O'Connell (1993) and Engel and Keen (1994). In these models, the correlation induced by the repeated measurements is captured with random effects. These effects explicitly model the variation between the subjects.

It can be of interest to model the association between multiple longitudinal responses. A first option is to treat one of the responses as a predictor, and use it as a time-dependent covariate. However, this method has several pitfalls. First, the lag has to be correctly specified, as incorrect specification can lead to illogical results. An example given by Rizopoulos (2012) is a study where they found a positive, but insignificant, effect of smoking on the survival of patients with coronary artery disease (Cavender et al., 1992). However, there was no lagged effect in the model, and hence only the immediate effect of smoking on death was gauged. The explanation of the faulty conclusion is that most of the smokers had stopped smoking at the last time of follow up before their death. In the meantime, many patients that were still alive, were still smoking. A second

drawback is the classification of a covariate into an exogenous or endogenous time-dependent covariate. If a response at time t predicts the value of the covariate at a time $s > t$, the covariate is endogenous (Diggle, 2002). Endogenous covariates have important modelling implications. Qian et al. (2020) have, for example, found that the marginal interpretation of the parameters in a linear mixed model does not hold anymore. Third, attention has to be given to missing data, which will likely occur in patient studies with follow-up. While ignorability holds for missing data of the response values under MAR under direct likelihood (Rubin, 1976), this is not the case for missing covariate values. Fourth, the time-dependent covariate can possibly be an intermediate variable. This means that the time-dependent covariate is in the causal pathway between another covariate and the response. As a consequence, including the time-dependent covariate will make the effect of the covariate on the response disappear. Fifth, when the responses, as well as time-dependent covariates, are not collected at fixed intervals, the utilization of time-dependent covariates with lags is not possible.

Joint models provide an alternative to time-dependent covariates. Several approaches for jointly modelling responses of a mixed nature exist. An overview can be found in Molenberghs and Verbeke (2005) in their Chapter 24. They outline three approaches that are applicable in both hierarchical and non-hierarchical settings. A first approach employs a bivariate Plackett-Dale distribution and postulates the existence of an unobserved continuous response that underlies the observed binary/ordinal response. The second approach, known as the probit-normal formulation, also assumes the presence of a latent response, with the added feature of errors being correlated to the continuous response. The third approach is the generalised linear random-effects model, which we will describe here in greater detail for the longitudinal case.

In joint generalised linear random-effects models, the relation between the responses is symmetric. Here, all longitudinal variables are treated as responses and are modelled with an appropriate random effects model. The random effects of the different models are allowed to be correlated to capture the associations. One of the advantages is that the effect of a covariate can be assessed on multiple outcomes simultaneously and that the association between the responses as well as the evolution of this association can be assessed. For example, Chakraborty et al. (2003) fitted a joint model for the continuous HIV-1 RNA concentration in both blood and semen. With the joint model, he could compare the correlation between both responses between the group with and without HIV treatment. Notably, the use of joint random-effects models is not limited to continuous responses. In Delporte et al. (2022) a joint model was developed for a longitudinal continuous and a longitudinal binary response. Not only the latent correlations between the random effects were scrutinized, but also the correlations between the responses on their original scale could be gauged. They derived a closed-form

formula for the correlation function from the joint model, with the possibility to include covariates. Their case study focused on the relation between the occurrence of allergic bronchopulminary aspergilosis (ABPA) and FEV values. Based on the latent correlations, they found that a better FEV value than expected under the model resulted in higher probabilities of lung infection at baseline and that higher increase in lung function than expected under the model is positively related to a higher probability of absence of ABPA than expected at baseline on the probit scale. Still, when gauging the correlations on the original scale, the conclusions were far more clinically relevant. We found that the correlation, between the responses as observed, is slightly stronger for earlier measurements of the ABPA and later measurements of FEV. This suggests that ABPA at an early stage shows an overall frailty, which exhibits itself later in life. In addition, they proposed a prediction model where one of the responses and potentially the history of the predicted response are included as predictors in the model.

The discrepancy between the manifest and latent correlation in random-effects models is also discussed in several other papers. For example, Milanzi et al. (2015) caution against drawing misleading conclusions by using latent and manifest-based correlation reliability measures interchangeably in IRT models. They emphasize that latent correlation-based reliability measures consistently result in higher values than their manifest correlation-based counterparts. Moreover, Molenberghs and Verbeke (2005) compare in their Chapter 7 the associations found via the Bahadur, probit and Dale models. They found a strong downward bias in the marginal correlation estimates obtained from the Bahadur model in comparison to their probit model counterparts. Lastly, Fieuws and Verbeke (2006) use joint random-effects models for analysing binary questionnaire data. They stress that their interest is in the association between the (latent) concepts underlying the sets of items, in contrast to the the association between observed responses, for which they recommend other models.

Some work has been done on the joint model for a continuous and a ordinal response. Faes et al. (2004) used a Plackett-Dale approach to jointly model the birth weight (continuous) and the probabilities of degrees of malformation (ordinal) of a fetus, where they take into account the clustering induced by a common mother. Still, the model cannot be readily extended to a longitudinal setting where responses are measured at different time points. Ivanova et al. (2016) formulated a joint random-effects model in a case study of repeated measures of BMI and clinical targets of diabetes patients. "Clinical targets" was treated as an ordinal variable. The covariance between the random intercepts of the variables was examined in order to gauge the association between the responses. In this paper, we extend the approach of Ivanova et al. (2016) by deriving closed-form formulas to calculate the correlations between the responses on their original scale. In addition, a conditional model is derived in order to

construct predictions of one response conditional on the other response(s). The outline of the paper is as follows: Section 4.2 presents the case study that serves as the foundation for the subsequent analysis in Section 4.5. Section 4.3 discusses the methodology. It commences with a review of the established methods for clustered continuous and clustered ordinal responses. Following that, we introduce the normal-ordinal (probit) model and our methodology based on the joint model. In Section 4.6 concluding remarks are offered.

4.2 Case study

The dataset contains information about the occurrence and progression of cognitive impairment in 60 elderly hip fracture patients from admission to the twelfth postoperative day (Milisen et al., 1998).We will focus on the connection between cognitive abilities and functional status and how the association between both varies over time. Throughout the study, neurocognitive status and the functional performance were assessed longitudinally; neurocognitive status was measured at day 1, 3, 5, 8 and 12, while functional status was recorded at day 1, 5 and 12. Table 4.1 provides an overview of the number of measurements taken of each response at each time point. Drop-out occurred because patients were discharged from the hospital before the twelfth post-operative day. Notably, while deaths were recorded, there were no reported mortalities throughout the duration of the study.

Neurocognitive status was assessed using the mini-mental state exam (MMSE), which includes subscales for memory, linguistic ability, concentration, and psychomotor executive skills. Cognitive status was classified as no impairment (MMSE$\geq$24), moderate impairment ($18 \leq$ MMSE≤ 23), or severe impairment (MMSE≤ 17) (Milisen et al., 1998; Tombaugh and McIntyre, 1992). In addition, the functional status was measured using an adapted version of the Katz ADL-scale (ADL), which is treated as continuous. The mean ADL scores and individual profiles are presented at Figure 4.1. A higher ADL value indicates more dependence on caretakers for activities of daily living, whereas a higher category of MMSE indicates a lower level of impairment. For exploratory purposes, the point-biserial correlations between the observed responses at several time points has been calculated (see Online Appendix H). It suggests that there exists a moderately strong relation between ADL and both the event of having severe impairment and the event of having impairment. This correlation seems to slightly increase over time. However, these correlations are not corrected for covariates and are only valid when the data would be missing completely at random, which is a very strict assumption.

Table 4.1: *Number of measurements of the Mini Mental State Exam (MMSE) and Activities of Daily Living (ADL) at each time point.*

Response	Day 1	Day 3	Day 5	Day 8	Day 12
MMSE	59	58	60	52	38
ADL	60	0	60	0	40

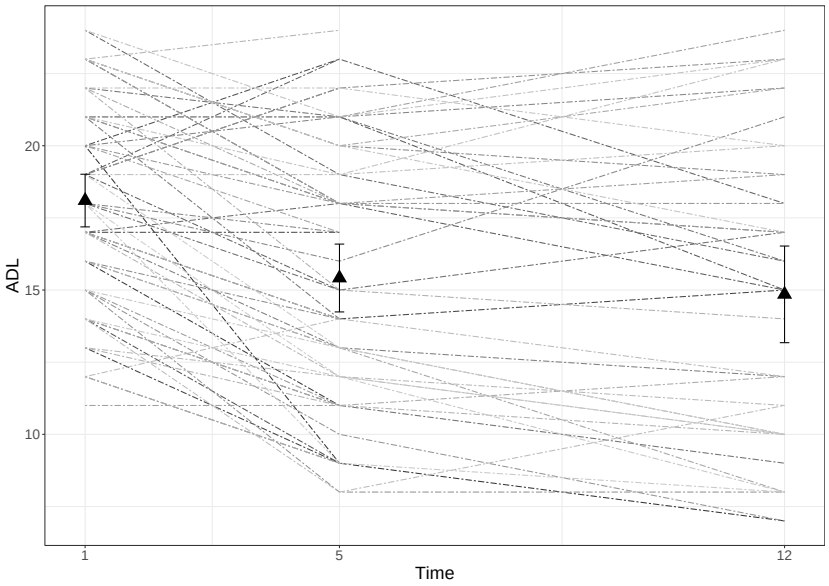

Figure 4.1: *Observed average (with 95% confidence interval) of the activities of daily living scores on day 1, 5 and 12 (solid) and individual profiles of the* 60 *subjects (dashed).*

4.3 Methodology

4.3.1 Model for a single longitudinal continuous response

One of the most popular models for longitudinal continuous variables is the linear mixed model. Suppose we have N subjects and the jth measurement for subject i is denoted by Y_{ij}. The vector $(Y_{i1}, ..Y_{in_i})$ of all n_i measurements of subject i is denoted by $\boldsymbol{Y_i}$. With this notation, we can write the model as

$$\begin{aligned} \boldsymbol{Y_i}|\boldsymbol{b}_i &\sim N(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i, \boldsymbol{\Sigma}_i), \\ \boldsymbol{b}_i &\sim N(\boldsymbol{0}, \boldsymbol{D}), \end{aligned} \tag{4.1}$$

where $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are, respectively, $(n_i \times k)$ and $(n_i \times q)$ dimensional matrices of known covariates of, respectively, the fixed effects $\boldsymbol{\beta}$ and the random effects $\boldsymbol{b}_i$. $\boldsymbol{\Sigma}_i$ denotes the $(n_i \times n_i)$ dimensional covariance matrix. Notably, i does not mean that the estimates of the variance depends on the subject. It indicates that the dimensions of the residual matrix can depend on the subject (Verbeke and Molenberghs, 2000). We can simplify $\boldsymbol{\Sigma}_i$ to $\sigma^2 \boldsymbol{I}_i$ with the assumption that the random effects fully capture the correlation between the measurements within subjects. This conditional independence assumption is however not necessary. It can be relaxed by the inclusion of, for example, serial correlation.
A property of the linear mixed model is that the parameters of the conditional model and the marginal model are exactly equal. This holds since $E[Y_{ij}] = E[E(Y_{ij}|\boldsymbol{b}_i)] = \boldsymbol{x}'_{ij}\boldsymbol{\beta}$. Still, the marginal model is defined as

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}^*_i. \tag{4.2}$$

The residuals $\boldsymbol{\epsilon}^*_i$ are here by definition correlated and are normally distributed around $\boldsymbol{0}$ with variance $\boldsymbol{V}^*_i$. As a consequence, the distribution of the response is

$$\boldsymbol{Y}_i \sim N(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{V}^*_i), \tag{4.3}$$

with $\boldsymbol{V}^*_i = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}'_i + \boldsymbol{\Sigma}_i$. More information about linear mixed models can be found in Verbeke and Molenberghs (2000).

4.3.2 Model for a single longitudinal ordinal response

A random-effects ordinal regression model can be used for clustered or repeated measures of an ordinal response. A threshold concept is applied, which assumes that the observed ordered response categories are determined by the value of an underlying continuous response. A series of threshold values $\gamma_1, \gamma_2, ..\gamma_{d-1}$ are assumed for the d categories. A response is categorized as category c if the latent response Y^*_{ik} surpasses the threshold value γ_{c-1}, but not γ_c. For the measurement at time k of this latent response of subject i, $k = 1, ..., p_i$,

$$Y^*_{ik} = \boldsymbol{x}'_{ik}\boldsymbol{\beta} + \boldsymbol{z}'_{ik}\boldsymbol{b}_i + \epsilon_{ik},$$

is the hierarchical linear mixed model, where $\boldsymbol{x}_{ik}$ is the $r \times 1$ vector that contains values for the covariates of the r-dimensional fixed effects vector $\boldsymbol{\beta}$. Next, $\boldsymbol{z}_{ik}$ is the $q \times 1$ design vector for the q random effects $\boldsymbol{b}_i$. $\boldsymbol{b}_i$ is assumed to follow a normal distribution around 0 with the covariance matrix $\boldsymbol{D}$. ϵ_{ik} are the residuals and are assumed to be independently normally distributed with mean 0 and variance σ^2.

From the latter model for Y_{ik}^*, the probabilities of the response categories can be derived. The probability that a response at time k for subject i falls into category c equals

$$P(Y_{ik} = c) = \Phi\left(\frac{\gamma_c - \zeta_{ik}}{\sigma}\right) - \Phi\left(\frac{\gamma_{c-1} - \zeta_{ik}}{\sigma}\right),$$

where $\zeta_{ik} = \boldsymbol{x}'_{ik}\boldsymbol{\beta} + \boldsymbol{z}'_{ik}\boldsymbol{b}_i$ and $\Phi(.)$ equals the cumulative normal distribution. Similarly, the probability that a response k of subject i is less than or equal to category c equals

$$P(Y_{ik} \leq c) = \Phi\left(\frac{\gamma_c - \zeta_{ik}}{\sigma}\right).$$

The choice of the unit and the origin of ζ is arbitrary (Hedeker and Gibbons, 1994). Alternatively, the logit link function can be applied (Ivanova et al., 2016), but this leads to more cumbersome calculations and less closed forms can be derived than with the probit link.

In Online Appendix A the marginal random-effects ordinal regression model is derived. In the latter model, the interpretation is no longer conditional on the random effects. Let $\boldsymbol{Z}_i$ denote the $n_i \times q$ dimensional design matrix of the random effects and $\boldsymbol{X}_i$ denote the $n_i \times r$ dimensional design matrix of the fixed effects. The marginal model is the following:

$$P(\boldsymbol{y}_i \leq \boldsymbol{c}) = \Phi(\boldsymbol{\gamma}_c - \boldsymbol{X}_i\boldsymbol{\beta}; \boldsymbol{L}_i^{-1}), \tag{4.4}$$

where

$$\boldsymbol{L}_i = \boldsymbol{I} - \boldsymbol{Z}_i(\boldsymbol{D}^{-1} + \boldsymbol{Z}'_i\boldsymbol{Z}_i)^{-1}\boldsymbol{Z}'_i.$$

4.3.3 Joint model

The joint mixed model employs a q-dimensional random effects vector $\boldsymbol{\xi}_i$ to encompass random effects linked the ordinal response as well as random effects linked with the continuous response. This vector follows a multivariate normal distribution with a mean of zero and a covariance matrix $\boldsymbol{D}$. This matrix $\boldsymbol{D}$ accounts for the correlation between repeated measurements of the same response, as well as the correlations between (the vectors of) measurements for different responses. We assume that the responses are independent given the random effects, meaning that the random effects fully capture the correlation between the responses. Consequently, the joint density of the responses, given the random effects, is equivalent to the product of the conditional densities of the individual responses.

The joint marginal density can be obtained by integrating out the random effects out of the joint density, these calculations can be found in online Online Appendix B. Note that the primary purpose of this joint marginal density is to

provide an intermediate result for future calculations. The joint marginal density is as follows:

$$f(\boldsymbol{y}_{1i}, \boldsymbol{y}_{2i} \leq \boldsymbol{c}) = \phi(\boldsymbol{X}_{1i}\boldsymbol{\beta}; \boldsymbol{V}_i)\Phi(\boldsymbol{\gamma_c} - \boldsymbol{X}_{2i}\boldsymbol{\beta} - \boldsymbol{\alpha}_i; \boldsymbol{B}_i), \tag{4.5}$$

where

$$\begin{aligned}
\boldsymbol{V}_i &= \boldsymbol{Z}_{1i}\boldsymbol{D}\boldsymbol{Z}'_{1i} + \boldsymbol{\Sigma}_i, \\
\boldsymbol{\alpha}_i &= \boldsymbol{H}_i(\boldsymbol{y}_{1i} - \boldsymbol{X}_{1i}\boldsymbol{\beta}), \\
\boldsymbol{H}_i &= \boldsymbol{B}_i\boldsymbol{Z}_{2i}\boldsymbol{K}_i\boldsymbol{Z}'_{1i}\boldsymbol{\Sigma}_i^{-1}, \\
\boldsymbol{K}_i^{-1} &= \boldsymbol{D}^{-1} + \boldsymbol{Z}'_{1i}\boldsymbol{\Sigma}_i^{-1}\boldsymbol{Z}_{1i} + \boldsymbol{Z}'_{2i}\boldsymbol{Z}_{2i}, \\
\boldsymbol{B}_i^{-1} &= \boldsymbol{I} - \boldsymbol{Z}_{2i}\boldsymbol{K}_i\boldsymbol{Z}'_{2i}.
\end{aligned}$$

It is possible to extend (4.5) to the high-dimensional case, with multiple ordinal and/or continuous responses. Let $\boldsymbol{Y}_{ci}$ represent a vector containing all the measurements of a continuous responses:
$\boldsymbol{Y}_{ci} = (Y^c_{1i1}, ..., Y^c_{1in_{1i}}, Y^c_{2i1}, ..., Y^c_{2in_{2i}}, ..., Y^c_{ai1}, ..., Y^c_{ain_{ai}})$. Notably, the number of measurements for each response does not have to be the same. Similarly, let $\boldsymbol{Y}_{bi}$ denote a vector containing all the measurements of the o ordinal responses: $\boldsymbol{Y}_{bi} = (Y^b_{1i1}, ..., Y^b_{1ip_{1i}}, Y^b_{2i1}, ..., Y^b_{2ip_{2i}}, ..., Y^b_{oi1}, ..., Y^b_{op_{oi}})$. Additionally, the matrices $\boldsymbol{Z}_{ci}$ and $\boldsymbol{Z}_{bi}$ consist of concatenated matrices of covariates for the random effects of continuous and ordinal responses, respectively. Specifically, $\boldsymbol{Z}_{ci}$ is formed by combining the matrices of covariates for the separate continuous responses: $\boldsymbol{Z}_{ci} = [\boldsymbol{Z}^c_{1i}, \boldsymbol{Z}^c_{2i}, ..., \boldsymbol{Z}^c_{ai}]'$. Similarly, $\boldsymbol{Z}_{bi}$ is formed by concatenating the matrices of covariates for the separate ordinal responses: $\boldsymbol{Z}_{bi} = [\boldsymbol{Z}^b_{1i}, \boldsymbol{Z}^b_{2i}, ..., \boldsymbol{Z}^b_{oi}]'$. The matrices of covariates for the fixed effects are defined in a similar manner: $\boldsymbol{X}_{ci}$ contains the concatenated matrices of covariates for the continuous responses: $[\boldsymbol{X}^c_{1i}, \boldsymbol{X}^c_{2i}, ..., \boldsymbol{X}^c_{ai}]'$, while $\boldsymbol{X}_{bi}$ contains the concatenated matrices of covariates for the ordinal responses: $[\boldsymbol{X}^b_{1i}, \boldsymbol{X}^b_{2i}, ..., \boldsymbol{X}^b_{oi}]'$. Further, $\boldsymbol{\Sigma}_i$ is a block diagonal matrix with as blocks the variance-covariance matrices of the continuous responses

$$\boldsymbol{\Sigma}_i = \begin{bmatrix} \boldsymbol{\Sigma}_{1i} & 0 & ... & 0 \\ 0 & \boldsymbol{\Sigma}_{2i} & .. & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & \boldsymbol{\Sigma}_{ai} \end{bmatrix}.$$

It is easy to see that the marginal hierarchical model is now

$$f(\boldsymbol{y}_{ci}, \boldsymbol{y}_{bi} \leq \boldsymbol{c}) = \phi(\boldsymbol{X}_{ci}\boldsymbol{\beta}; \boldsymbol{V}_i)\Phi(\boldsymbol{\gamma_c} - \boldsymbol{X}_{bi}\boldsymbol{\beta} - \boldsymbol{\alpha}_i; \boldsymbol{B}_i). \tag{4.6}$$

To be as general as possible, we will use the above expressions for the remainder of the paper.

Due to potential computational difficulties associated with high-dimensional models, Fieuws and Verbeke (2006) introduced a pseudo-likelihood method to simplify the model fitting. This involves fitting a bivariate model for every pair of responses and then combining the results. Kundu (2011) offers a convenient guide to implementing this method in `SAS NLMIXED`.

4.3.4 Conditional models

Conditional models offer a practical approach for making predictions of one subset of measurements, conditional on another subset. This methodology proves particularly valuable in circumventing challenges related to time-dependent covariates. In analogy to Section 4.3.3, we define $\boldsymbol{Y}_{ci}$ as a vector composed of all the measurements of a continuous responses and $\boldsymbol{Y}_{bi}$ as a vector composed of all the measurements of o ordinal responses. Next, let $\widetilde{\boldsymbol{Y}}_{ci}$ denote a $\widetilde{n}_i$-dimensional subset of the continuous response vector $\boldsymbol{Y}_{ci}$, while $\widetilde{\boldsymbol{Y}}_{bi}$ represents a $\widetilde{p}_i$-dimensional subset of the ordinal response vector $\boldsymbol{Y}_{bi}$. Notably, $\widetilde{\boldsymbol{Y}}_{ci}$ and $\widetilde{\boldsymbol{Y}}_{bi}$ can contain measurements of different, respectively, continuous and ordinal responses. By analogy, $\widetilde{\boldsymbol{X}}_{ci}$ and $\widetilde{\boldsymbol{X}}_{bi}$ represent the $\widetilde{n}_i \times q$ and $\widetilde{p}_i \times q$ submatrices of $\boldsymbol{X}_{ci}$ and $\boldsymbol{X}_{bi}$. Leveraging the ratios of the marginal distributions, it becomes feasible to derive expected values and their corresponding prediction intervals. A first conditional expected value is a subset the continuous responses, given a subset of both continuous and ordinal responses. This specific n_a-dimensional subvector of predicted continuous response(s) is denoted as $\widetilde{\boldsymbol{Y}}^{a}_{ci}$. This prediction is conditional on, on the one hand, $\widetilde{\boldsymbol{Y}}^{b}_{ci}$, the subvector of length n_b of values of the continuous response vector and, on the other hand, $\widetilde{\boldsymbol{Y}}_{bi}$, the subvector of ordinal responses. The notation will be as follows: the superscript specifies the submatrices or subvectors; superscript a and b denote, respectively, the rows a_1 until a_{n_a} and b_1 to b_{n_b}. In addition, the superscript bb specifies the rows b_1 until b_{n_b} and columns b_1 until b_{n_b}. The superscript ab indicates row a_1 until a_{n_a} and column b_1 until b_{n_b}. The conditional expected value is as follows:

$$E[\widetilde{\boldsymbol{Y}}^{\boldsymbol{a}}_{\boldsymbol{ci}}|\widetilde{\boldsymbol{Y}}^{\boldsymbol{b}}_{\boldsymbol{ci}} = \widetilde{\boldsymbol{y}}^{\boldsymbol{b}}_{\boldsymbol{ci}}, \widetilde{\boldsymbol{y}}_{bi} \leq \boldsymbol{c}] = \left((\boldsymbol{E}_i \boldsymbol{V}_i^{-1} \widetilde{\boldsymbol{X}}_{ci} \boldsymbol{\beta})^a + \boldsymbol{E}_i^{ab} (\boldsymbol{E}_i^{bb})^{-1} (\widetilde{\boldsymbol{y}}^{\boldsymbol{b}}_{\boldsymbol{ci}} - (\boldsymbol{E}_i \boldsymbol{V}_i^{-1} \widetilde{\boldsymbol{X}}_{ci} \boldsymbol{\beta})^b) \right) + \left((\boldsymbol{E}_i \boldsymbol{H}_i' \boldsymbol{B}_i^{-1})^a - \boldsymbol{E}_i^{ab} (\boldsymbol{E}_i^{bb})^{-1} (\boldsymbol{E}_i \boldsymbol{H}_i \boldsymbol{B}_i^{-1})^b \right) \kappa, \tag{4.7}$$

where κ equals the expected value of the truncated normal distribution with variance $\boldsymbol{T}_i$, mean $\boldsymbol{F}_i$ and limits$]-\infty; \boldsymbol{d}]$. This expression is implemented in standard statistical software, such as in the R package `tmvtnorm`. The analytical

expression can be found in Manjunath and Wilhelm (2021). In addition,

$$\begin{aligned}
\boldsymbol{E}_i^{-1} &= \boldsymbol{H}_i' \boldsymbol{B}_i^{-1} \boldsymbol{H}_i + \boldsymbol{V}_i^{-1}, \\
\boldsymbol{T}_i^{-1} &= (\boldsymbol{E}_i \boldsymbol{H}_i' \boldsymbol{B}_i^{-1})^{b'} (\boldsymbol{E}_i^{bb})^{-1} (\boldsymbol{E}_i \boldsymbol{H}_i' \boldsymbol{B}_i^{-1})^{b} + \boldsymbol{B}_i^{-1} - (\boldsymbol{H}_i' \boldsymbol{B}_i^{-1})' \boldsymbol{E}_i (\boldsymbol{H}_i' \boldsymbol{B}_i^{-1}), \\
\boldsymbol{F}_i &= \boldsymbol{T}_i \Bigg((\boldsymbol{E}_i \boldsymbol{H}_i' \boldsymbol{B}_i^{-1})^{b'} (\boldsymbol{E}_i^{bb})^{-1} (\widetilde{\boldsymbol{y}}_{ci}^{b} - (\boldsymbol{E}_i \boldsymbol{V}_i^{-1} \widetilde{\boldsymbol{X}}_{ci} \boldsymbol{\beta})^{b}) + (\boldsymbol{H}_i' \boldsymbol{B}_i^{-1})' \boldsymbol{E}_i (\boldsymbol{V}_i^{-1} \widetilde{\boldsymbol{X}}_{ci} \boldsymbol{\beta}) \Bigg), \\
\boldsymbol{d} &= \boldsymbol{\gamma}_c - \widetilde{\boldsymbol{X}}_{bi} \boldsymbol{\beta} + \boldsymbol{H}_i \widetilde{\boldsymbol{X}}_{ci} \boldsymbol{\beta}.
\end{aligned}$$

The corresponding prediction interval and the derivations can be retrieved in Online Appendix C.

A special case of (4.7) is when the continuous response is modelled conditional on solely the ordinal response. In this case, the expression of the expected value simplifies as follows

$$E[\widetilde{\boldsymbol{Y}}_{ci} | \widetilde{\boldsymbol{y}}_{bi} \leq \boldsymbol{c}] = \boldsymbol{E}_i \Bigg(\boldsymbol{V}_i^{-1} \widetilde{\boldsymbol{X}}_{ci} \boldsymbol{\beta} + \boldsymbol{H}_i' \boldsymbol{B}_i^{-1} \kappa \Bigg), \tag{4.8}$$

where κ is again the expected value of the truncated normal distribution with variance $\boldsymbol{T}_i^*$, mean $\boldsymbol{F}_i^*$ and limits $]-\infty; \boldsymbol{d}]$. Further,

$$\begin{aligned}
\boldsymbol{T}_i^{*-1} &= \boldsymbol{B}_i^{-1} - (\boldsymbol{H}_i' \boldsymbol{B}_i^{-1})' \boldsymbol{E}_i (\boldsymbol{H}_i' \boldsymbol{B}_i^{-1}) \\
\boldsymbol{F}_i^* &= \boldsymbol{T}_i^* \cdot (\boldsymbol{H}_i' \boldsymbol{B}_i^{-1})' \boldsymbol{E}_i (\boldsymbol{V}_i^{-1} \widetilde{\boldsymbol{X}}_{ci} \boldsymbol{\beta}).
\end{aligned}$$

The expressions for the prediction interval and the related details considering the calculations can be found in Online Appendix D.

To make predictions for the ordinal response, we can derive conditional probabilities for a subset of the ordinal response conditional on a subset of the ordinal response and the continuous response. We denote the subset for which we calculate the probability of being in category c or lower as $\widetilde{\boldsymbol{Y}}_{\boldsymbol{bi}}^{\boldsymbol{a}}$, and we calculate it conditional on a subset of the ordinal response $\widetilde{\boldsymbol{Y}}_{\boldsymbol{bi}}^{\boldsymbol{b}}$, and a subset of the continuous response, $\widetilde{\boldsymbol{Y}}_{ci}$. The use of the superscripts b, ab and bb is in analogy with (4.7). The conditional probability can be expressed as follows:

$$P(\widetilde{\boldsymbol{y}}_{\boldsymbol{bi}}^{\boldsymbol{a}} \leq \boldsymbol{c} | \widetilde{\boldsymbol{y}}_{\boldsymbol{bi}}^{\boldsymbol{b}} \leq \boldsymbol{c}, \widetilde{\boldsymbol{y}}_{ci}) = \frac{\Phi(\boldsymbol{\gamma}_c - \widetilde{\boldsymbol{X}}_{bi} \boldsymbol{\beta} - \boldsymbol{H}_i (\widetilde{\boldsymbol{y}}_{ci} - \widetilde{\boldsymbol{X}}_{ci} \boldsymbol{\beta}); \boldsymbol{B}_i)}{\Phi(\boldsymbol{\gamma}_c^b - \widetilde{\boldsymbol{X}}_{bi}^b \boldsymbol{\beta} - \boldsymbol{H}_i^b (\widetilde{\boldsymbol{y}}_{ci} - \widetilde{\boldsymbol{X}}_{ci} \boldsymbol{\beta}); \boldsymbol{B}_i^{bb})}. \tag{4.9}$$

After applying the logit transformation to ensure that the boundaries are constrained to the unit interval, the corresponding confidence interval can be calculated using the delta method. The gradients of the parameters can be found in Online Appendix E.

A special case is the conditional density of the ordinal response(s) conditional on solely a subvector of the continuous response vector. The expected probability is then simplified to

$$P(\widetilde{\boldsymbol{y}}_{bi} \leq \boldsymbol{c} | \widetilde{\boldsymbol{y}}_{ci}) = \Phi(\boldsymbol{\gamma}_c - \widetilde{\boldsymbol{X}}_{bi} \boldsymbol{\beta} - \boldsymbol{H}_i (\widetilde{\boldsymbol{y}}_{ci} - \widetilde{\boldsymbol{X}}_{ci} \boldsymbol{\beta}); \boldsymbol{B}_i). \tag{4.10}$$

Online Appendix F contains the formulas related to the standard errors.

4.3.5 Correlation function

By using the property that the responses are independent conditional on the random effects, it is feasible to deduce a correlation function from the hierarchical joint model. This correlation captures the manifest correlation, denoted as $\rho_{Y_{1ij},Y_{2ik}\leq c}$. It quantifies the relationship between the continuous response Y_{1i} at time j and the event of an ordinal response Y_{2i} below category c at time k. This model-based manifest correlation represents the correlation between the scores on the original scale, whereas the latent correlation quantifies the correlation between the underlying random effects. Although calculating the latent correlation is simpler, the scientific interest often focuses on the manifest correlation rather than the latent correlation. The formula for the manifest correlation function is as follows:

$$\rho_{Y_{1ij},Y_{2ik}\leq c} = \frac{-\frac{1}{L_i}\boldsymbol{z}'_{1ij}\boldsymbol{M}_i^{-1}\boldsymbol{z}_{2ik}\phi(\gamma_c-\boldsymbol{x}'_{2ik}\boldsymbol{\beta};L_i^{-1})}{\sqrt{\left(\boldsymbol{z}'_{1ij}\boldsymbol{D}^*\boldsymbol{z}_{1ij}+\Sigma_{1ij}\right)\Phi(\gamma_c-\boldsymbol{x}'_{2ik}\boldsymbol{\beta};L_i^{-1})(1-\Phi(\gamma_c-\boldsymbol{x}'_{2ik}\boldsymbol{\beta};L_i^{-1}))}}, \tag{4.11}$$

where $\boldsymbol{D}^*$ denotes the submatrix of $\boldsymbol{D}$ relating to the variances and covariances of the random effects of the responses Y_{1i} and Y_{2i}. In addition, $\boldsymbol{M} = (\boldsymbol{D}^*)^{-1} + \boldsymbol{z}'_{2ik}\boldsymbol{z}_{2ik}$.
The details of the derivations and the formulas regarding the standard errors can be found in Online Appendix G.

4.4 Parameter estimation

The parameters in the joint random effects model are estimated via maximum likelihood. The likelihood function of the joint random-effects model is constructed under the assumption that the responses are independent given the random effects. As a result the likelihood function for a joint model of the responses Y_{ci} and Y_{bi} equals

$$L(\theta) = \prod_{i=1}^{N}\int f_{1i}(\boldsymbol{y}_{ci}|\boldsymbol{b}_i)f_{2i}(\boldsymbol{y}_{bi}\leq c|\boldsymbol{b}_i)f(\boldsymbol{b}_i|\boldsymbol{D})d\boldsymbol{b}_i, \tag{4.12}$$

in which the vector $\boldsymbol{\theta}$ contains all parameters of the conditional distributions and the distribution of the random effects $\boldsymbol{b}_i$. In most cases, numerical approximations are needed for the integral in 4.12. In this paper, adaptive Gaussian quadrature is used for the estimation, which is implemented in the SAS procedure `NLMIXED` (Pinheiro and Bates, 1995)

4.5 Data analysis

In this section, the relationship between the continuous functioning score (ADL) and the ordinal level of impairment (MMSE) is examined. First, a joint model is implemented, as discussed in Section 4.3.3. Here, we fit a linear mixed model for ADL (as discussed in Section 4.3.1) and a generalised linear mixed model with a probit link for MMSE (as discussed in Section 4.3.2). Next, we allow the random effects of the responses to correlate to create the joint model. Since Figure 4.1 clearly indicates that the evolution of ADL is not linear, we will include time as a categorical covariate in the linear mixed model. In contrast, time since the operation is included as a continuous covariate in the generalised linear mixed model for impairment. The model can be written as

$$\begin{aligned} Y_{1ij} &= \boldsymbol{\beta}_{1,0} + \boldsymbol{\beta}_{1,1}\mathsf{I}(\mathsf{Time}_{ij} = 5) + \boldsymbol{\beta}_{1,2}\mathsf{I}(\mathsf{Time}_{ij} = 12) + \\ &\quad \boldsymbol{\beta}_{1,3}\mathsf{I}(\mathsf{Sex}_i = \mathsf{F}) + \boldsymbol{\beta}_{1,4}\mathsf{Age}_{ij} + b_{10i} + b_{11i}\tfrac{\mathsf{Time}_{ij}}{100} + \epsilon_{1ij}, \\ \Phi^{-1}(P(Y_{2ij} \leq c)) &= \gamma_c - \Bigg(\boldsymbol{\beta}_{2,1}\mathsf{Time}_{ij} + \boldsymbol{\beta}_{2,2}\mathsf{I}(\mathsf{Sex}_i = \mathsf{F}) + \boldsymbol{\beta}_{2,3}\mathsf{Age}_{ij} \\ &\quad + b_{20i} + b_{21i}\tfrac{\mathsf{Time}_{ij}}{100} \Bigg). \end{aligned}$$

The hierarchical models include several random effects: b_{10i} and b_{20i} are the random intercepts for, respectively, ADL and MMSE. Next, b_{11i} and b_{21i} are the random slopes for respectively ADL and MMSE. In order to account for the correlation between the responses, different assumptions can be made regarding the joint distribution of the random effects such as for example setting the correlations to 1 (i.e. shared random-effects model). However, we have chosen to make the distribution as flexible as possible, i.e., not to impose any restriction on the covariance matrix. We assumed that $[b_{10i}, b_{11i}, b_{20i}, b_{21i}] \sim MVN(\mathbf{0}, \boldsymbol{D})$ and $\epsilon_{1i} \sim MVN(\mathbf{0}, \sigma^2 \boldsymbol{I}_i)$. The full SAS code can be found in Online Appendix H. Convergence was reached within 10 hours and 19 minutes on a regular laptop (CPU=Processor Intel(R) Core(TM) i7-8850H CPU @ 2.60GHz, 2592 Mhz, 6 Core(s), 12 Logical Processor(s), RAM=24GB) and resulted in the parameter estimates shown in Table 4.2.

An association between the responses was indicated by a Wald test, by showing that the covariances among the random effects of the different responses significantly differ from zero ($H_0 : d_{13} = d_{14} = d_{23} = d_{24} = 0, \chi^2_{df=4} = 12.11, p = 0.02$). The latent correlations between the random effect offer a first glimpse into the relationship between the two responses (Table 4.3). These values can be interpreted in terms of the latent random effects. For instance, the correlation between the random intercepts ($r = -.69$) shows that immediately after the operation, a lower starting value of ADL (better functioning) than expected

based on the covariates, is related to a lower probability of having a more severe level of impairment than expected based on the covariates. Still, these are the correlations between the underlying (latent) random effects on the probit-scale. It can be of interest to also examine the manifest correlations, as discussed in Section 4.3.5, which are the model-based correlations between the responses on their original scale.

Table 4.2: *Parameter estimates (standard errors) of $ADLTOT$ and $MMSE$.*

Effect	ADLTOT		MMSE	
Intercept	3.42	(4.50)	-	
γ_1	-		-19.17	(5.70)
γ_2	-		-16.61	(5.50)
Time	-		0.04	(0.04)
Time 5	-2.68	(0.35)	-	
Time 12	-3.62	(0.57)	-	
Sex: Female	-1.58	(1.02)	-0.37	(1.11)
Age	0.20	(0.05)	-0.22	(0.07)
σ^2	3.02	(0.60)	-	

Table 4.3: *Latent correlations [CI] between the random effects of MMSE and ADL.*

	b_{10i}	b_{11i}	b_{20i}	b_{21i}
b_{10i}	1			
b_{11i}	.12 [-.44; .61]	1		
b_{20i}	-.70[-.89; -.31]	-.38 [-.77; .21]	1	
b_{21i}	.38 [-.80; .95]	-.07 [-.98; .98]	-.72 [-1; .95]	1

Note that to ensure that the correlation is bounded between -1 and 1, the Fisher-Z-transformation is applied to compute the confidence intervals, after which the values are transformed back to the original scale.

By the use of (4.11) the correlations between the responses on their original scale can be computed. Since these model-based correlations depend on the covariate values chosen, we computed them for a man of mean age (78). They are displayed in Table 4.4. Functional impairment is at each timepoint significantly correlated with the event of a severe cognitive impairment and the event of impairment. The correlation is quite constant over time, but better cognitive functioning is consistently related to a higher probability of having no impairment

and a lower probability of having severe impairment.
Table 4.5 presents the predicted MMSE (Mini-Mental State Examination) statuses on days 8 and 12, conditional on ADL (Activities of Daily Living) scores recorded on days 1 and 8. The latter ADL scores were set to one standard deviation below, at, or one standard deviation above the respective day's mean. Additionally, sex was set to female, and age was set to 78 years. The predictions are derived from the parameter estimates obtained from the joint model, which were then substituted into the conditional model (4.10). The results in Table 4.5 shows that a history of strong reliance on a caregiver corresponds to a high probability of cognitive impairment both in the present and in the immediate future. Furthermore, the confidence intervals emphasize that a low ADL score, indicative of low caregiver dependence, holds limited predictive value for MMSE status. This is in contrast to moderate or high ADL scores, which demonstrate stronger association with MMSE outcomes.

4.6 Concluding remarks

In this research, our primary focus has been on introducing two new methodologies that are built upon the foundation of joint models: The first methodology involves closed-form expressions to obtain the manifest correlations from the model between the responses on their original scale, providing an alternative to investigating latent correlations between the underlying random effects on the probit-scale. Our second methodology entails employing conditional joint models in lieu of time-dependent covariates to analyse the effect of a longitudinal predictor on the longitudinal response. This shift effectively sidesteps several complications associated with time-dependent covariates. Firstly, the need for specifying lags is obviated with the joint modelling approach. With manifest correlations, we can assess the effect of a predictor on the response at each time point. In addition, our conditional model allows for straightforward adjustments of lags without refitting the joint model. Secondly, due to the symmetric nature of the relationship in a joint model, challenges posed by endogeneity or intermediary variables are mitigated. Thirdly, the presence of missing data necessitates no additional steps, thanks to the principle of ignorability (Rubin, 1976). Consequently, our methodology becomes highly suitable for unbalanced data, as it operates without the requirement of lags or additional methods for handling missing data.
Moreover, our paper extends the application of our conditional model to scenarios involving multiple longitudinal predictors. By consolidating all predictors into an elongated vector and adapting the design matrices accordingly, our methodology remains seamlessly applicable.

Illustrating the practical utility of our methodology, we added a case study investigating the association of two longitudinal responses: a continuous physical functioning score and an ordinal mental functioning score. We show that predictions concerning the ordinal response can be effectively derived from the historical trajectory of the continuous response. In addition, the missingness is assumed to be at random, and hence results in no additional steps in the data analysis due to ignorability (Rubin, 1976). Implementation of the joint model can be achieved through the `NLMIXED` procedure in `SAS`. Code is provided in online Appendix H for both transforming the data in the correct format and fitting the joint model. However, it's worth noting that a limitation of our methodology is its reliance on the proportional odds assumption inherent in the ordinal regression model. Of course, extension to non-proportionality is possible by having (certain) covariate effects category-dependent. But then, as always, care needs to be taken to ensure non-negative probabilities ensue. A second drawback is the computational complexity of a joint model, especially when more than two responses are included. Various methodologies can be used, such as the pairwise fitting approach for high-dimensional data (Fieuws and Verbeke, 2006), the split-sample approach for large datasets (Molenberghs et al., 2011b), or a combination of both (Ivanova et al., 2017). A third limitation arises from the dependence of correlations and confidence intervals on the selected random effect structure. Therefore, it is recommended to model the random effects with a high degree of flexibility, potentially incorporating splines. Importantly, even within this scenario, our methodology remains applicable.
Further research can be conducted on the bounds of the manifest correlation function. Research in the context of surrogate markers (Alonso and Molenberghs, 2007) and the Bahadur model (Molenberghs and Verbeke, 2005) have shown that respectively the bounds of the R^2 or the correlation between dichotomous responses can generally be smaller than one. In addition, it can be of interest to implement the methodology in a `SAS` macro to facilitate the usability. Secondly, other approaches can be explored, such as multiple imputation models, where the value of the one longitudinal variable is imputed via a random-effects model and then included as a time-dependent covariate in the other longitudinal model. Still, some issues of time-dependent covariates would persist, such as endogeneity, possibility of intermediate variables and the definition of lags.

Table 4.4: *Correlations between ADL (higher: lower functioning) and MMSE (cognitive impairment) for a 78-year-old man.*

Panel A: Manifest correlations between ADL and the event of having severe impairment.

	Time(Impairment)				
Time (ADL)	1	3	5	8	12
1	.44 [.28 ; .58]	.44 [.29 ; .57]	.44 [.29 ; .57]	.43 [.28 ; .57]	.42 [.26 ; .57]
5	.48 [.34 ; .60]	.48 [.34 ; .60]	.48 [.34 ; .60]	.48 [.34 ; .61]	.48 [.31 ; .61]
12	.47 [.26 ; .65]	.48 [.28 ; .64]	.48 [.29 ; .64]	.49 [.29 ; .64]	.49 [.28 ; .65]

Panel B: Manifest correlations between ADL and the event of having impairment.

	Time(Impairment)				
Time (ADL)	1	3	5	8	12
1	.47 [.31 ; .60]	.47 [.32 ; .60]	.47 [.33 ; .60]	.48 [.34 ; .60]	.48 [.32 ; .61]
5	.51 [.38 ; .62]	.52 [.39 ; .62]	.52 [.41 ; .62]	.53 [.41 ; .63]	.54 [.41 ; .65]
12	.50 [.30 ; .66]	.51 [.32 ; .66]	.52 [.34 ; .66]	.54 [.37 ; .67]	.55 [.37 ; .70]

Table 4.5: *Prediction of cognitive impairment based on the history of ADL at time* 1 *and* 5 *for a female of* 78 *years.*

Timepoint prediction	History ADL (day 1 - day 5)	P(Impairment=Severe)	P(Impairment)
8	14.57 - 10.87	0.03 [0.00; 0.94]	0.22 [0.12; 0.35]
8	18.10 - 15.42	0.25 [0.17; 0.36]	0.68 [0.52; 0.80]
8	21.63 - 19.97	0.72 [0.53; 0.86]	0.96 [0.75; 0.99]
12	14.57 - 10.87	0.02 [0.00; 1.00]	0.19 [0.47; 0.84]
12	18.10 - 15.42	0.21 [0.12; 0.35]	0.66 [0.48; 0.80]
12	21.63 - 19.97	0.69 [0.47; 0.84]	0.95 [0.73; 0.99]

Chapter 5

Analysing matched continuous longitudinal data: a review

This chapter is based upon:

Delporte, M., Aerts, M., Verbeke, G.,, & Molenberghs, G. Analysing matched continuous longitudinal data: a review. Submitted.

Abstract

Longitudinal data is frequently encountered in medical research, where participants are followed throughout time. Additional structure and hence complexity occurs when there is pairing between the participants (e.g., matched case-control studies) or within the participants (e.g., analysis of participants' both eyes). We identified a large number of approaches via a systematic review for the analysis of these kind of data, where often the correlation induced by pairing was ignored. Our aim is to provide a methodological overview and guidelines for selecting the appropriate method based on the data's characteristics. The methods are compared via both a real-life case study in ophthalmology and a simulated case-control study. Key findings include the superiority of the conditional linear mixed model and multilevel models in handling paired longitudinal data in terms of precision. Moreover, the article underscores the impact of accounting for intra-pair correlations and missing data mechanisms. Focus will be on discussing the advantages and disadvantages of the approaches, rather than on the mathematical or computational details.

5.1 Introduction

Longitudinal studies are fundamental in medical research, providing valuable insights into the progression of diseases, treatment effectiveness, and patient outcomes over time. Longitudinal data, where measurements are collected on the same subjects at multiple time points, offer the opportunity to investigate within-subject changes while controlling for inter-subject variability. Clearly, it should be taken into account that the measurements from different subjects are independent, while observations within subjects are correlated. Due to this dependence, traditional techniques such as classical (generalised) linear regression models are not suitable. Still, a large number of approaches to model longitudinal data have been developed and implemented in standard statistical software (Verbeke and Molenberghs, 2000).

In practice, the subjects are not always independent, as there can be a meaningful one-to-one relationship between them. This is evident in various scenarios, such as case-control studies, where each control is carefully matched to a case based on multiple attributes to minimise confounding. Twin studies, widely utilised in both biomedical and psychological research, serve as another example, aiming to account for genetic influences. Additionally, the pairing can also be present within individuals, such as, for example, hearing thresholds measured on both ears of a set of subjects. In, for example, ophthalmology research, a case can serve as its own control when treatment is administered to only one eye.

In longitudinal studies, missing data poses additional challenges to statistical

analysis and inference. The nature of longitudinal data collection, spanning multiple time points, increases the probability of missingness due to, amongst others, participant dropout. Addressing missing data is vital for maintaining the validity of study findings. Various techniques, such as multiple imputation methods, likelihood-based or Bayesian approaches, and modelling strategies accounting for missing data mechanisms, have been developed to handle this issue (Molenberghs and Kenward, 2007). However, selecting an appropriate method requires careful consideration of the underlying missing data mechanism. When data is Missing Completely at Random (MCAR), the missingness is independent of both observed and unobserved outcomes (Rubin, 1976). When missingness is linked to observed data, it falls under the category of Missing at Random (MAR). Conversely, if it is further associated with unobserved data, it is labelled as Missing Not at Random (MNAR).

In matched longitudinal studies, the correlation induced by pairing is often ignored by researchers, in contrast to the correlation induced by repeated measurements. However, leveraging these intra-pair correlations can yield more precise estimates of the effect under investigation. Additionally, when researchers did account for the paired nature of the data, the missingness mechanism was often overlooked. Failure to properly address missing data or using inadequate techniques can lead to biased estimates and inferences. Our aim is to provide an overview and comparison of methods for the analysis of paired longitudinal data. Furthermore, we discuss the underlying assumptions of each method regarding missing data.

In Section 5.2, a motivating real-life dataset, the Ophthalmology data, is introduced. Section 5.3 delves into the modelling techniques identified through a literature review. These approaches are then contrasted using the Ophthalmology data in Section 5.4 and a simulated case-control datasets in Section 5.5. Finally, Section 5.6 presents some concluding thoughts.

5.2 Ophtomology data

The dataset at hand, provided by the DRCR Retina Network, was collected in the context of a clinical trial comparing the efficacy and safety of three treatments for central-involved Diabetic Macular Edema (DME).The study spanned 2 years with four-weekly follow-up visits. In the original study, 660 eyes were included, but our analysis was restricted to the 497 patients who had DME in one eye, but not in the other. Our research question involves comparing the evolution of the visual acuity in the eye exhibiting DME with the unaffected eye, taking into account both the correlation induced by the repeated measurements as well as the correlation due to the pairing of eyes within a subject. The mean visual acuity in our dataset over time is depicted in Figure 5.1, at this point ignoring

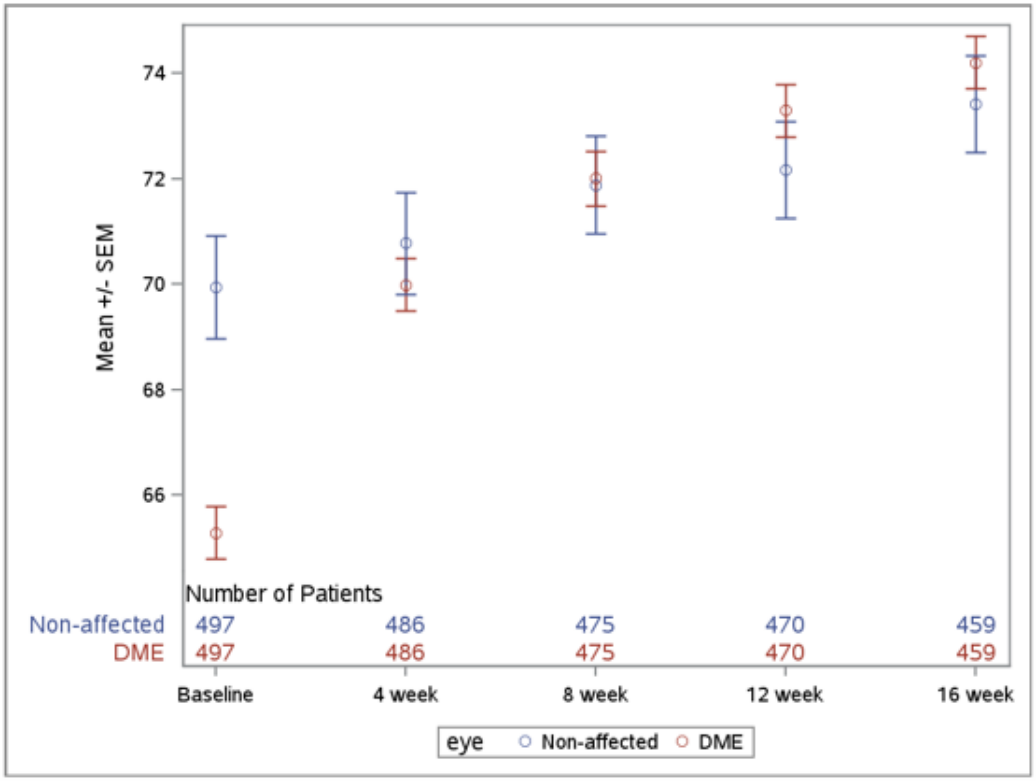

Figure 5.1: *Mean visual acuity over time, with the standard errors of the mean.*

both correlations. Figure 5.1 also displays dropout rates in both groups over the course of the study; Section 5.4 will examine and compare the methods for both the complete dataset and the complete cases.

5.3 Modelling approaches

In the remainder of the paper the notation will be as follows: the measurements of subject j of pair i at time k can be defined as Y_{ijk} with $i = 1, \ldots, N/2$, $j = 1, 2$ and $k = 1, \ldots, n_{ij}$. This notation reflects the situation of subjects nested within pairs (between subjects) and can easily be adapted in case of within-subject pairing.

5.3.1 Systematic review

To identify the methods that were used in the literature for the analysis of paired longitudinal data, a systematic review was conducted, which resulted in 56 articles that employed various methods. These methods are grouped in different categories, which are discussed in the sections below. An overview of the number of articles for each method is given in Table 5.1.

Table 5.1: *Number of studies per category of statistical method.*

Method	N
Paired t-tests at specific timepoints	8
Difference scores	3
Subject-specific slopes	1
Unpaired t-tests or non-parametric alternative	4
MANOVA	1
Linear mixed models	37
Without consideration for the paired nature of the data	(24)
With consideration for the paired nature of the data	(12)
Unclear how the researchers took into account the paired nature of the data	(1)
New methodology	2

More than half of the articles included in the systematic review ignored the paired nature of the data ($n = 29$). In addition, some researchers used sub-optimal methods such as testing differences of subject-specific slopes. The findings of this systematic review show the importance of a discussion on methodology for analysing matched longitudinal data. A more detailed discussion of the methodology of the systematic review and a one-by-one overview of the included articles can be found in Appendix S.1.

5.3.2 Paired t-tests

A first approach involves using paired t-tests or the non-parametric Wilcoxon rank sum tests to assess the pairs at particular time points. In the paired t-test, differences are used: $w_{ik} = Y_{i1k} - Y_{i2k}$, which result in a single longitudinal sequence. Next, the one-sample t-test is performed at a single time point k:

$$t = \frac{\bar{W}_k - \mu_{w_k}}{s_{w_k}},$$

which follows a Student's t-distribution with $N/2-1$ degrees of freedom when the differences w_k can be considered to be normally distributed and:

$$\bar{W}_k = \frac{\sum_i W_{ik}}{N/2}, \qquad s^2_{w_k} = \frac{\frac{\sum_i (W_{ik}-\bar{W}_k)^2}{N/2-1}}{N/2}.$$

Eight studies used this approach due to its benefits in terms of a straightforward interpretation and simplicity. The method has, however, considerable disadvantages: it does not consider 'overall' differences across all time points, and more importantly, it does not allow to study differences in evolution. In addition, corrections for multiple testing have to be applied to keep the Type I error at bay.

5.3.3 Unpaired t-tests

The independent samples t-test, or unpaired t-test can detect whether two independent groups have a different mean at a specific time point k. In analogy with the notation proposed in the introduction of this section, let Y_{i1k} denote the measurement of the first member of the pair i at time k, and let Y_{i1k} denote the measurement of the second member of the pair i at time k. $\bar{Y}_{i1k} = \sum_i^{N/2} Y_{i1k}$ equals the average of the first members of the pairs at time k (e.g. the average of the affected eyes/cases at time k), while $\bar{Y}_{i2k} = \sum_i^{N/2} Y_{i2k}$ is the average of the members of the second pair at time k (e.g. the average of the unaffected eyes/controls at time k). Under the assumption of equality of variances and normality of Y_{i1k} and Y_{i2k}, the test statistic $t = \frac{\bar{Y}_{i1k}-\bar{Y}_{i2k}}{s_p\sqrt{\frac{2}{N}}}$ follows a Student's t-distribution with $2N-2$ degrees of freedom, where $s_p = \sqrt{\frac{s^2_{Y_{i1k}}-s^2_{Y_{i2k}}}{2}}$.
Four studies performed an unpaired t-test or the non-parametric Mann-Whitney U-test to compare pairs at specific time points. Notably, in each study there was a 1:1 matching of cases and controls, as opposed to pre-existing pairs such as siblings.
However, in a paired design, the paired t-test is preferable even if the groups are only slightly correlated (Zimmerman, 1997). When the pairs are positively correlated, the paired t-test has an advantage in terms of the standard error. The correlation between the matched pairs reduces the standard error of the difference, which is apparent by the equation

$$\sigma^2_{\bar{X}-\bar{Y}} = \sigma^2_{\bar{X}} + \sigma^2_{\bar{Y}} - 2\rho_{\bar{X}\bar{Y}}\sigma_{\bar{X}}\sigma_{\bar{Y}}.$$

The size of this correlation is inversely related to the standard error, and hence positively associated with the t-ratio. Importantly, this gain is counteracted by

the loss in degrees of freedom; the one-sample t statistic of the paired t-test is evaluated at $n/2 - 1$ degrees of freedom, while the unpaired t-test is evaluated at $n - 2$ degrees of freedom. Still, Zimmerman (1997) showed that when power functions are examined, the advantages of the paired t-test are considerable even for small correlations. In addition, the unpaired t-test suffers from the same drawbacks as the paired t-test: it is not possible to test for 'overall' differences, nor differences in evolution. Similarly to the paired t-test, a correction for multiple testing should be administered.

5.3.4 Multivariate analysis of variance

Multivariate analysis of variance (MANOVA) is the multivariate extension of one-way analysis of variance (ANOVA). MANOVA is a statistical method that examines the effect of one or multiple factors on several dependent variables simultaneously. It allows for the assessment of null hypotheses regarding the effects of factor variables on the means of different groupings of dependent variables. In our scenario specifically, it can be used to test for pair (group) differences for the response at all time points simultaneously.
Beutel et al. (1996) used this methodology. While 27% of the dyads had missing data, they only included the complete cases in the MANOVA analysis. Since the resulting conclusions are only valid under the strict assumption of MCAR, extra steps such as multiple imputation should be implemented.

5.3.5 Difference scores

A fourth approach, adopted by three studies, is to calculate difference scores within the subjects between two specific time points. Let the two time points be a and b and the new difference scores $z_i = Y_{i1a} - Y_{i1b}$ and $v_i = Y_{i2a} - Y_{i2b}$. Ruhdorfer et al. (2015) administered the paired t-test on z_i and v_i to draw inference on differences between the pairs, while Goodman and Must (2011) used the non-parametric Wilcoxon test. This method allows studying differences in the evolution of the paired groups. Still, choices have to be made about which interval to use: Ruhdorfer et al. (2015) only considered the difference from baseline to the last timepoint, discarding data from intermediate time points. In contrast, Goodman and Must (2011) compared multiple pairs of difference scores between subsequent time points, necessitating correction for multiple testing.
The third study (LoCascio et al., 1998) calculated differences between baseline and each follow-up measurement for each patient, but later ignored both the longitudinal and paired nature of the data by applying ANCOVA on all difference scores simultaneously.

Closely related to the subject-specific difference scores is the calculation of a summary measure of the evolution for each subject and subsequently comparing these for the paired groups. (Schlee et al., 2021) first calculated slopes for each subject via linear regression and then used the Wilcoxon rank-sum test to test if the distributions of the slope estimates are equal in the study group and the matched control group. An advantage of this method is that each subject with more than one measurement can be included in the analysis, and it does not necessitate regular measurement intervals. In addition, no multiple testing issues arise. Still, this method treats the slopes as observed values and does not take the standard errors of the slopes into account. As a consequence, the standard errors and the corresponding p-values of this method do not reflect true uncertainty in the entire estimation process.

5.3.6 Linear mixed models

Linear mixed models were introduced by Laird and Ware (1982) for the analysis of clustered continuous responses. Using the notation introduced at the onset of this section and ignoring the existence of pairs, let Y_{ijk} denote the kth measurement of subject i of pair j. The mixed model of a longitudinal sequence is specified as:

$$\begin{aligned} Y_{ij}|\boldsymbol{b}_{ij} &\sim N(\boldsymbol{X}_{ij}\boldsymbol{\beta} + \boldsymbol{Z}_{ij}\boldsymbol{b}_{ij}, \boldsymbol{\Sigma}_{ij}), \\ \boldsymbol{b}_{ij} &\sim N(\boldsymbol{0}, \boldsymbol{D}), \end{aligned} \tag{5.1}$$

where $\boldsymbol{X}_{ij}$ and $\boldsymbol{Z}_{ij}$ denote, respectively, $(n_{ij} \times p)$ and $(n_{ij} \times q)$ dimensional matrices of known covariates. $\boldsymbol{\beta}$ is the p-dimensional vector containing fixed effects and $\boldsymbol{b}_{ij}$ denotes the q-dimensional vector of random effects. Finally, $\boldsymbol{\Sigma}_{ij}$ equals the $(n_{ij} \times n_{ij})$-dimensional residual covariance matrix and $\boldsymbol{D}$ denotes the variance-covariance matrix of the random effects. The marginal density of Y_{ij} equals

$$f(Y_{ij}) = \int f(Y_{ij}|\boldsymbol{b}_{ij}) f(\boldsymbol{b}_{ij}) d\boldsymbol{b}_{ij}, \tag{5.2}$$

which is the density function of a normal distribution with mean vector $\boldsymbol{X}_{ij}\boldsymbol{\beta}$ with covariance matrix $\boldsymbol{V}_{ij} = \boldsymbol{Z}_{ij}\boldsymbol{D}\boldsymbol{Z}'_{ij} + \boldsymbol{\Sigma}_{ij}$.
One key benefit of employing (generalised) linear mixed models lies in their fully parametric nature, enabling the use of both (restricted) maximum likelihood and Bayesian estimation. This implies ignorability when data are incomplete, as outlined in Rubin (1976), under the assumption of MAR and mild regularity conditions. Still, as we will discuss in Section 5.3.6, ignorability does not hold

when direct likelihood is not used, which is for example the case when the robust variance, or 'sandwich,' estimator is applied.
The literature review identified linear mixed models as the predominant method; the method was used in 37 of the 56 included studies. However, it should be noted that the random effects structure was sometimes unclear (e.g., Bouwmans et al. (2015)). The different strategies found in the literature review are discussed below.

Random subject effect

24 studies used exclusively random effects $\boldsymbol{b}_{ij}$ at the level of the subject. Next, a fixed effect of group membership, and an interaction between group and time was used to assess the impact of being in the case or the paired control group.
Two comments should be made. Firstly, all of these studies stated they 'matched' participants in the case and control group. However, in the literature study 'matching' was sometimes used to indicate that the distributions of age and gender were alike, instead of case-by-case matching of individual participants based on several attributes. Still, some of the studies in this category described a 1:1 matching process (e.g., Iaffaldano et al. (2021)), and subsequently did not take into account the pairing. Secondly, six of these studies indicated that they used repeated measures ANOVA. While very similar to a linear mixed model, repeated measures ANOVA assumes a common set of time points or a time schedule and time is regarded as a factor with n levels, with subjects as subplots (Krueger and Tian, 2004). As a consequence, mixed models are superior to the repeated measures ANOVA in handling multiple missing data points.
The primary disadvantage is that the method does not take into account the correlation induced by the pairing and hence assumes that the members of the pairs are independent. As a consequence, the standard errors of this method are incorrect. Still, the parameters estimates will be unbiased since the fixed effect estimates are independent of the chosen variance-covariance structure of the random effects (Lange and Ryan, 1989).

Random pair effect

In three studies, the model exclusively contained random effects $\boldsymbol{b}_i$ at the level of the pairs, omitting random effects for individual members within each pair. A study conducted by Ahmed et al. (2018) took the form of a case-control study, while in another study (Shek and Dou, 2020) an inherent link was present between members of a pair. In this particular study, a child repeatedly completed two identical questionnaires about maternal control on the one hand, and paternal

control on the other. However, this approach implicitly assumes that all the measures of the dyad are independent, given the random effect of pair.
In contrast, a case-control study conducted by Border et al. (2020) employed a random pair effect while also incorporating an autoregressive residual correlation structure to relax the conditional independence assumption. In contrast to the models with only a random effect on the pair or subject level, this methods takes into account both the correlation induced by pairs and the longitudinal aspect of the data. Still, changing the residual structure will affect the parameter estimates of the fixed effects, as will be discussed in Section 5.5.

Robust variance estimation

Another possibility found in the literature (Sibbel et al., 2016) was to use a random effect on the subject level, as described in Section 5.3.6, and combine this with a robust variance estimator to correct inferences about the fixed effects for the pairing effect. An asymptotically consistent estimator, the so-called sandwich estimator (Huber, 1967; White, 1980; Liang and Zeger, 1986) is the following:

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \left(\sum_{i=1}^{N} \boldsymbol{X}_i' \boldsymbol{X}_i \right)^{-1} \left(\sum_{i=1}^{N} \boldsymbol{X}_i' \widehat{\boldsymbol{\epsilon}}_i \widehat{\boldsymbol{\epsilon}}_i{}' \boldsymbol{X}_i \right) \left(\sum_{i=1}^{N} \boldsymbol{X}_i' \boldsymbol{X}_i \right)^{-1},$$

where $\widehat{\boldsymbol{\epsilon}}_i = \boldsymbol{y}_i - \boldsymbol{X}_i \widehat{\boldsymbol{\beta}}$. The estimator is consistent when the mean is correctly specified in the model (Verbeke and Molenberghs, 2000). Hence, when interest lies in the estimation of average longitudinal evolution and the dataset is sufficiently large, the sandwich estimator is a solution with minimal effort. But importantly, when missing data is present, very strict assumptions have to be made regarding the underlying process of missingness in order to obtain valid conclusions regarding the fixed effect based on the sandwich estimator. More specifically, it is assumed that the missing data is missing completely at random (MCAR). A second drawback is that efficiency is gained if an appropriate covariance model can be specified (Diggle, 2002).

Marginal linear mixed model

A versatile method is employing a marginal linear mixed model, as denoted in (5.2), where the random effects are integrated out of the density of the hierarchical model. Here, the population mean is modelled via the mean structure, and the dependence in the data is directly modelled via the marginal positive-definite matrix $\boldsymbol{V}_i$. Note that this model is more flexible, since it only imposes positive-definiteness on $\boldsymbol{V}_i$, in contrast to the hierarchical model that needs positive definiteness of both $\boldsymbol{\Sigma}_i$ and $\boldsymbol{D}$ (Verbeke and Molenberghs, 2000). However,

altering the residual structure will impact the parameter estimates of the fixed effects, as further discussed in Section 5.5. This method was utilised by Benestad et al. (2022), who incorporated an unstructured covariance matrix to account for the pairing and repeated nature of his matched case-control study.

Nested random effects

Six studies used nested random effects or so-called three-level multilevel models to analyse their paired longitudinal data. These models are described by (Fitzmaurice et al., 2004, Ch. 22) as follows:

$$Y_{ijk} = X_{ijk}\beta + Z^{(3)}_{ijk}b^{(3)}_i + Z^{(2)}_{ijk}b^{(2)}_{ij} + \epsilon_{ijk},$$

where $Z^{(3)}_{ijk}$ and $Z^{(2)}_{ijk}$ are the design matrices for the random effects at the level of the pair and the subject, respectively. Here, the notation of the superscript signals the levels at which the random effects vary. The model allows that the random effects are correlated within a given level, but assumes that there are no correlations between levels: $\text{Cov}(b^{(2)}_{ij}, b^{(3)}_i) = 0$. In addition, ϵ_{ijk}, the random component at the lowest level, is assumed to be independent within their level, with variance σ^2.

Conditional linear mixed model

A last possible option in the linear mixed model family, is the conditional linear mixed model. This model was employed by one study in our literature review: Gerber et al. (2016) investigated the effect of early antibiotic exposure on weight in twins during the first 8 years of life. The researchers specifically chose twins who were discordant in their early-life antibiotic exposure. Their model predicted the difference in growth trajectories in twins (weight of the exposed twin minus the weight of the unexposed twin) with a linear mixed model. As a consequence, the fixed slope of time represents the effect of antibiotic exposure on the growth evolution. It is worth noting that this method is equivalent to the conditional linear mixed model (Verbeke et al., 2001).
Verbeke and Molenberghs (2000) indicate as a main advantage of the conditional linear mixed model that inferences about the longitudinal effects can be made without making assumptions about the cross-sectional components. Ignoring the pairing, the model is as follows:

$$\boldsymbol{Y_i} = \mathbf{1}_{n_i} b^*_i + \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \epsilon_{(1)i}, \tag{5.3}$$

where b^*_i represents the cross-sectional components and is considered a nuisance. The matrices $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ and the vectors $\boldsymbol{b}_i$ and $\boldsymbol{\beta}$ are submatrices of their

original counterparts in (5.1), after the deletion of the cross-sectional effects. In a conditional linear model, the model fitting proceeds in two steps. First, there is conditioning on sufficient statistics for the nuisance parameters b_i^*. Second, the remaining parameters in the conditional density of $\boldsymbol{Y_i}$ given these sufficient statistics are estimated via (restricted) maximum likelihood.
Importantly, the conditional model can be obtained via first taking the difference between the measurements within a pair $z_{ik} = Y_{i1k} - Y_{i2k}$ and then employing a standard linear mixed model as shown in (5.1). As a result, the intercept can be interpreted as the baseline difference between the groups, while the slope of time denotes the treatment effect on the evolution.

5.3.7 Alternative methods

Two studies described new statistical methods for analysing paired longitudinal data. The study by Wilson (1979) focuses on examining individualised growth trajectories in longitudinal twin data using repeated measures ANOVA. The total variance is partitioned into various substantial sources of variance, and the magnitude of their effects are calculated. The author subsequently formulates a hypothesis test concerning twin concordance, exploring whether twins within each pair exhibit greater similarity than they do with twins from different pairs. This directly leads to the calculation of intra-class correlations, representing the concordance within pairs in the form of correlation coefficients.
A second paper by Kim (2006) is in the context of longitudinal ophthalmology data, where the paired eyes are assigned to different treatments. The authors construct methods to test the hypothesis that an interaction exists between the treatment (eye-specific factor) and race (person-specific factor). Two methods are described: a large sample-based non-parametric test statistic and a non-parametric bootstrap test analogy. Next, the results of the methods are compared with generalised estimating equations (GEE) with different working correlation structures.

5.4 Analysis of the ophthalmology data

In the previous section, several approaches to model paired continuous longitudinal data have been presented. To demonstrate how to choose the right approach for a specific scenario, we revisit the ophthalmology data presented in Section 5.2. Here, our main emphasis is on selecting the appropriate modelling approach rather than delving into the results and insights gained from the statistical analysis. The main research question is to study the impact of the medication on

the evolution of visual acuity in eyes with diabetic macular edema (DME$=0$) and without diabetic macular edema (DME$=1$).
The results of the different approaches can be found in Table 5.2 and the code is provided via `https://github.com/MargauxDelporte`. Scrutinising the results of analyses of the complete dataset, it is clear that the research question cannot be answered by the paired or the unpaired t-test. Based on these tests, we can only conclude that the acuity is different at baseline, but the differences are no longer significant at subsequent time points. As expected, the standard errors of the paired t-test are considerably smaller compared to those of the unpaired t-test. MANOVA confirms these results and shows that there is an overall difference in visual acuity in the first five time points. Note that the latter results are only valid under the assumption of MCAR.
The remaining five methods can answer the research question at hand. For instance, based on the paired t-test on the slopes of the subject-specific regression models, the conclusion can be drawn that the visual acuity has a better progression in the DME eyes. However, as the slopes are treated as 'observed' and the standard errors are not taken into account, the p-value of this analysis is incorrect. This is also true for the standard error of the 'naive' linear mixed model, where the pairing was not taken into account, and only a random intercept of the subject was included. In other words, it is assumed that all measurements of the eyes within a participant are independent, conditional on the person-specific random effect. Still, the estimated fixed effects are very similar, and in each analysis it can be concluded that when treated, there is a beneficial evolution in visual acuity in the eyes with DME. When comparing the standard errors of the three approaches that correctly take into account the pairing, they are the lowest in the conditional linear mixed model.
Next, in two columns on the right of Table 5.2, the analysis is repeated on a dataset restricted to the first five measurements of cases who have no missing data on these time points. It is apparent that the results of the MANOVA analysis are exactly equal, as by default only complete cases are considered in the analysis. All other methods take into account incomplete profiles, but differ in their assumptions with regards to the missingness. Specifically, the (un)paired t-test, MANOVA, and sandwich LMM assume missingness to be completely at random MCAR, whereas the nested and conditional LMM uphold validity under the more lenient MAR assumption. In contrast, the standard errors with regards to the slope comparisons and the naive LMM are invalid.

Table 5.2: *Analysis of the ophthalmology data.*

		Full data			Complete cases		
Method	Parameter	Estimate	SE	p-value	Estimate	SE	p-value
Paired t-test	Baseline	-4.646	0.956	<.0001	-4.618	0.999	<.0001
	4 weeks	-0.788	0.962	0.413	-0.873	0.991	0.379
	8 weeks	0.126	0.926	0.892	0.251	0.957	0.793
	12 weeks	1.128	0.911	0.216	1.230	0.941	0.192
	16 weeks	0.787	0.899	0.382	0.906	0.916	0.323
Unpaired t-test	Baseline	-4.646	1.098	<.0001	-4.618	1.161	<.0001
	4 weeks	-0.788	1.088	0.469	-0.873	1.121	0.436
	8 weeks	0.126	1.062	0.905	0.251	1.086	0.817
	12 weeks	1.128	1.050	0.283	1.230	1.084	0.257
	16 weeks	0.787	1.042	0.450	0.906	1.060	0.393
MANOVA	Wilks lambda	0.905		<.0001	0.905		<.0001
Comparison slopes		0.005	0.002	0.004	0.046	0.005	<.0001
LMM naive	DME	-0.342	0.284	0.227	-3.097	0.593	<.0001
	time*DME	0.005	0.001	<.0001	0.043	0.008	<.0001
LMM sandwich	DME	-0.342	0.858	0.690	-3.097	0.993	0.002
	time*DME	0.005	0.001	<.0001	0.043	0.006	<.0001
LMM nested	DME	-0.363	0.982	0.656	-3.255	0.939	0.001
	time*DME	0.004	0.001	<.0001	0.046	0.004	<.0001
Conditional LMM	DME	-0.366	0.813	0.653	-3.256	0.935	0.001
	time*DME	0.004	<0.001	<.0001	0.046	0.004	<.0001

5.5 Simulation study

To compare the performance of various methods across a wide array of settings, we simulated data from case-control studies. Specifically, 100 datasets were simulated, each comprising five measurements from 200 pairs of subjects. While the treatment effects remained fixed across datasets, the other parameters were varying. The exact specifications of the parameters in the simulation study can be found in Table 5.3. In addition, the code can be found via `https://github.com/MargauxDelporte`.

Table 5.3: *Specifications of the simulation study.*

Parameter	Specification in simulation study
Treatment effect baseline	-0.560
Treatment effect slope	-0.230
Intercept	$\sim N(0, 1)$
Slope	$\sim N(10, 100)$
Pair-specific effect	$\sim N(0, 4)$
Variance-Covariance matrix $\mathbf{D}$	$\sim \text{Wishart}(df = 3, \mathbf{\Sigma} = \mathcal{I}_2)$
Subject-specific intercept and slope	$\sim N(\mathbf{0}, \mathbf{D})$
Residual ϵ_{ijk}	$\sim N(0, 4)$

The resulting estimates and standard errors from the different categories of linear mixed models are averaged and presented in Table 5.4. Comparing the parameter estimates, it is clear that some are slightly different from the others. This is the case for both the marginal model and a combination method, where a random effect of the pair is combined with an autoregressive correlation structure in the residual variance-covariance matrix. Previous studies (Lange and Ryan, 1989) showed that in the absence of missing data, fixed effects do not depend on the chosen variance-covariance structure of the random effects, but the same does not hold true for the variance-covariance structure of the residuals. While the naive, sandwich, nested, and conditional models assume that the residuals are uncorrelated given the random effects, this is not the case for the marginal and the combination model. In the marginal model, the residual variance-covariance matrix is unstructured, and in the combination model, auto-regressive correlation is assumed.

Next, the differences in average standard errors of the treatment effect on the evolution are negligible, while larger differences exist in the average standard errors of the baseline effects. The smallest standard errors are found in the conditional LMM and the nested random effects model. Comparing the averaged

Table 5.4: *Average parameter estimates, average standard errors and standard deviation of the estimates of the treatment effect at baseline and the treatment effect on the evolution of* 100 *simulated datasets.*

Method	Parameter	Avg. Estimate	Avg. SE	SD estimates
LMM naive	treated	-0.5372	0.4637	0.2067
	time*treated	-0.2411	0.1028	0.1711
LMM sandwich	treated	-0.5372	0.2967	0.2067
	time*treated	-0.2411	0.1710	0.1711
LMM marginal	treated	-0.5377	0.2956	0.2050
	time*treated	-0.2414	0.1708	0.1698
LMM nested	treated	-0.5372	0.2243	0.2067
	time*treated	-0.2411	0.1714	0.1711
LMM combination	treated	-0.5303	0.3380	0.2271
	time*treated	-0.2423	0.1404	0.1761
Conditional LMM	treated	-0.5372	0.2243	0.2067
	time*treated	-0.2411	0.1711	0.1711

SE to the standard deviation of the estimates, it is apparent that these estimates are also more accurate compared to the other models. Notably, the standard errors of the naive model are proven to be incorrect since they do not take the intra-pair correlation into account.

5.6 Concluding remarks

In this research, we focused on the analysis of paired longitudinal data, where the pairing is either within the participant, or between the participants. First, a systematic review has been conducted to identify the methods that are used in the literature. Next to showing the broad range of methods that are used for this kind of data, the systematic review demonstrated that most studies ignored the pairing, while it could be used in favour of getting more precise estimates and correct inferences.

We presented the various methods that emerged from the systematic review and discussed the possible research questions they can answer, along with their respective advantages and limitations. For instance, while (un)paired t-tests are suitable for comparing pairs at different time-points, they fall short in assessing differences in progression over time. The questions can be answered by linear mixed models, or alternatively, by deriving summary measures (like slopes) for comparison among groups. However, it is crucial to account for standard errors of the summary statistics in the analysis. Furthermore, in linear mixed models with only a random effect at the subject level, standard errors can be misleading since pairs are assumed to be independent. These nuances underscore the importance of selecting the appropriate modelling approach.

In addition, special attention has been given to missing data, which is frequently encountered in longitudinal studies. Some methods are only valid under the very restrictive assumption of Missing Completely at Random (MCAR), which assumes that the missingness does not depend under observed nor unobserved data. This is the case for MANOVA, as well as linear mixed models with the robust sandwich estimator. In contrast, in linear mixed models that do not employ the sandwich estimator, ignorability holds under the less restrictive Missing at Random assumption.

Following the methodological exploration, we applied these techniques to a real-life ophthalmology dataset, where both eyes of participants were examined concurrently. Interestingly, neither the (un)paired t-test nor MANOVA could effectively address the specific research question concerning the differences in evolution between eyes with and without diabetic macular edema. Moreover, we concluded that the standard errors in analyses comparing slopes via the paired t-test and the linear mixed model with only a random subject effect were flawed. Our analysis demonstrated that the most precise estimates are obtained via the conditional linear mixed model.

Next, we compared the linear-mixed model based techniques on 100 simulated datasets of a case-control study with identical treatment effects. Our analysis revealed slight disparities in parameter estimates attributable to variations in residual covariance structures. However, these differences proved inconsequen-

tial, as did variations in standard errors regarding the estimation of treatment effects on evolution. Notably, the standard errors of the treatment effect at baseline were the most accurately estimated when employing either the multilevel (nested) linear mixed model or the conditional linear mixed model.

Chapter 6

Accelerating computation: a faster pairwise fitting technique for multivariate probit models

This chapter is based upon:

Delporte, M., Verbeke, G., Fieuws, S., & Molenberghs, G. Accelerating computation: a faster pairwise fitting technique for multivariate probit models. Submitted.

Abstract

When fitting multivariate probit models via maximum likelihood, except for very small numbers of responses, computational challenges, in terms of computation time and/or difficulty to reach convergence, are inevitable; these are compounded for ordinal data. This article will introduce an efficient computational approach based on a pairwise fitting technique within a pseudo-likelihood framework. The proposed methodology is applied to medical case studies, specifically utilising a trivariate probit model. Further, the correlation structure among outcomes is allowed to depend on covariates, enhancing model flexibility and interpretability. Through simulation and real data application, we demonstrate the superiority of our approach in terms of computational efficiency when the dimension of the outcome vector increases. The flexibility in capturing covariate-dependent correlations makes our method particularly appealing in medical research, where understanding complex associations among health outcomes is often of scientific interest.

6.1 Introduction

When researchers encounter a single binary or ordinal response variable, logistic and probit regression are often methods of choice. Although probit regression is occasionally considered, its advantages over logistic regression are often perceived as negligible. However, the situation changes when researchers face multiple correlated binary or ordinal response variables. In such cases, employing separate logistic regressions is not able to seize the relationship between these variables. Moreover, research questions regarding the association between these variables remain unanswered. In these cases, the polychoric correlation, which is the correlation between the underlying latent continuous variables, induced by multivariate probit models can be of interest.

Of course, there are different possible models for multivariate binary data, which result in different ways to quantify the association between the responses. Molenberghs and Verbeke (2005) compared the correlations resulting from the trivariate Bahadur, probit and Dale model and found a strong downward bias of the correlations of the Bahadur model in comparison with the correlations of the probit model. This is due to the strong parameter space restrictions in the Bahadur case, especially when higher-order correlations are omitted. A second factor contributing to the difference is the nature of the correlations in the two models. In the Bahadur model, the correlations are manifest, meaning they are on the scale of the observed responses in the data (Aerts et al., 2002). In contrast, the correlations in the probit model are latent; they represent polychoric correlations between underlying continuous variables. While the concept of polychoric correlation appears straightforward in theory, the process of model fitting is of

high computational complexity. Particularly as the dimensionality increases, the calculations quickly become cumbersome.
In the bivariate scenario, there are some model fitting options available within standard statistical software packages. For instance, in `SAS`, the likelihood can be manually programmed and optimised using `PROC NLMIXED`. In `R`, there are also solutions for bivariate probit models with multiple predictors, such as the `VGAM` or the `GJRM` packages, both for binary responses. On the other hand, the software implementation of multivariate probit regression is more limited. Currently, the `mvProbit` function from the `mvProbit` package is able of such analysis for binary responses. However, the correlation cannot depend on the parameters and the documentation comes with a caveat: "WARNING: this function is experimental and extremely (perhaps even unusably) slow!" A second possible package is the `mvord` package for ordinal data, where it is possible to let the correlation depend on the parameters (Hirk et al., 2020). The latter function uses the composite likelihood approach, where the full likelihood is approximated by lower dimensional marginal distributions in a pseudo-likelihood framework. Hence, the sum of the bivariate likelihoods is maximised instead of the full multivariate likelihood.
In the context of multivariate longitudinal data, Fieuws and Verbeke (2006) used another pseudo-likelihood approach to avoid computational problems. They proposed to first maximise the likelihood of each pair of responses separately and afterwards combine the results. It poses considerable advantages since the bivariate likelihoods can be optimised in parallel via, for example, high-performance computing, which is not the case in classical composite likelihood. Moreover, in the latter approach, all parameters are still estimated simultaneously, resulting in high computational complexity, in contrast to the pairwise fitting approach. In this paper, we will expand the pairwise fitting approach from random-effects models to the multivariate probit model and compare the results with the classical pseudo-likelihood approach.
A drawback of pseudo-likelihood is that some useful properties of maximum-likelihood estimation are lost. Most importantly, in the maximum likelihood framework ignorability holds (Rubin, 1976) under the assumption of Missingness at Random (MAR). MAR occurs when the missingness depends on the observed data, but not further on unobserved data. In contrast, pseudo-likelihood inference is only valid under Missingness Completely at Random (MCAR), when the missingness does not depend on observed nor unobserved data. Still, when MAR is assumed, pseudo-likelihood can be performed after multiple imputation or certain weighting methods (Molenberghs et al., 2011a). Of course, both maximum likelihood and pseudo-likelihood are not valid under Missingness Not at Random (MNAR), which means that missingness depends on unobserved outcomes, in addition to possible dependence on observed outcomes.

The paper is organised as follows: Section 6.2 presents the case studies that serve as the foundation for the subsequent analysis in Section 6.6. Section 6.3 discusses the methodology. It commences with a review of the established methods for probit models. Following that, we introduce the pseudo-likelihood estimation with the pairwise fitting approach applied to multivariate probit models. A simulation study comparing the different approaches is presented in Section 6.5. In Section 6.7, concluding remarks are offered.

6.2 Motivating case studies

6.2.1 The Belgian Interuniversity Research on Nutrition and Health (BIRNH) study

Between 1980 and 1984, 11302 Belgian individuals were enrolled in a study that aimed to investigate the impact of diet on health. The BIRNH study involved assessing 156 food items, as well as measuring blood pressure, serum lipids, psychosocial factors, and drinking and smoking habits. Focus was on alcohol and smoking habits and their correlation with specific demographic factors.

Alcohol consumption was categorised into four groups based on daily intake, while smoking was divided into three categories: never smoked, ex-smoker, and current smoker. The proportion of each combination of categories is shown in Figure 6.1, for individuals with and without elevated cholesterol levels (cholesterol> 235). As predictors we used the following variables: SEX (1: female, 0: male), AGE, BMI, SITE (1: Flanders, 0: rest of Belgium), SOC1 (1: Employed, 0: unemployed); SOC2 (1: working at home, 0: working outside).

Several questions motivated the study: the relationship between drinking and smoking and elevated cholesterol, the association between these variables and demographics and whether the correlation between smoking, cholesterol and drinking varied across demographic subgroups.

The dataset was analysed by Lesaffre and Molenberghs (1991), utilising a bivariate probit model on the categorical smoking and drinking variables. However, the analysis was initially performed on a random sample comprising only 10% of the total study population. Given the advancements with our computational approach, we will now repeat the analysis using the entire study population and including cholesterol as an additional response.

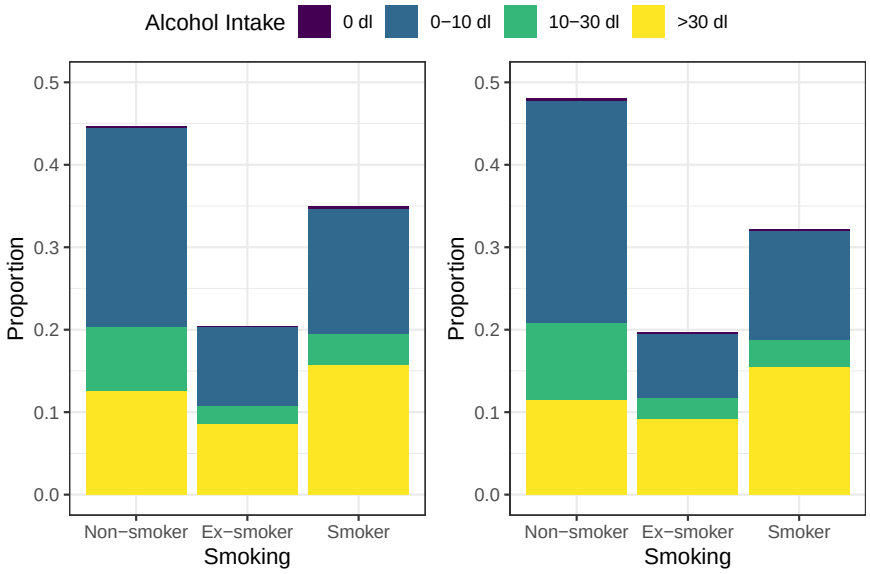

Figure 6.1: *Proportions of smoking and daily alcohol intake categories, according to cholesterol level (left: not elevated, right: elevated).*

6.2.2 The Project On Preterm and Small-for-gestational age infants (POPS) study

The Project On Preterm and Small-for-gestational age infants (POPS) conducted a study involving 1338 infants born in the Netherlands in 1983, with a gestational age of less than 32 weeks and/or a birth weight less than $1,5$ kg (see Verloove-Vanhorick et al. (1986) for more details). A total of 133 clinics participated in the study, representing 94% of the births in that year with similar gestational age and birth weight characteristics. The study collected prenatal, perinatal, and postnatal information, as well as two-year follow-up data. The main research question was to investigate the relationship between three ability scores measured at the age of two years and risk factors measured at birth: neonatal seizures (NSZ), congenital malformation (CGM), and the highest bilirubin value since birth (BIL). These ability scores, recorded dichotomously, assessed the child's physical and mental abilities.

Due to the two-year gap between enrolment and assessment, the data showed considerable missingness, as shown in Appendix S.3. When the dataset was introduced and analysed in Molenberghs and Lesaffre (1994), only the complete cases (comprising 60% of the data) were considered. In our analysis, we will compare the results under different assumptions: independence between responses, MCAR, and MAR.

6.3 Probit models

6.3.1 Univariate probit models

Probit models are appropriate for the analysis of categorical responses. Suppose that the response Y is binary, taking values 0 and 1, then the probit model assumes an underlying continuous latent variable Y^* that follows a normal distribution:

$$Y_i^* = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \varepsilon_i = \boldsymbol{X}_i\boldsymbol{\beta} + \varepsilon_i \qquad (6.1)$$

where $\beta_0, \beta_1, \ldots, \beta_k$ are the coefficients to be estimated, $x_1, x_2, \ldots, x_k$ are the predictor variables and ε is the error term, assumed to be normally distributed with mean 0 and variance σ^2.
The observed binary outcome Y is determined by:

$$Y_i = \begin{cases} 1 & \text{if } Y_i^* > 0 \\ 0 & \text{if } Y_i^* \leq 0 \end{cases}$$

The probabilities are calculated using the cumulative standard normal distribution function $\Phi(\cdot)$:

$$P(Y_i = 1) = \Phi\left(\frac{\boldsymbol{X}_i\boldsymbol{\beta}}{\sigma}\right).$$

For ordinal data with J ordered categories ($Y = 1, 2, ..., J$), the model assumes a latent variable Y^*, defined in (6.1) with threshold parameters τ_j separating the categories. The observed ordinal outcome y is determined by the latent variable via a threshold concept:

$$Y_i = j \text{ if } \tau_{j-1} \leq Y_i^* < \tau_j,$$

where $\tau_0 = -\infty$ and $\tau_J = \infty$.
The probabilities for each category j are given by:

$$P(Y_i = j) = \Phi\left(\frac{\tau_j - \boldsymbol{X}_i\boldsymbol{\beta}}{\sigma}\right) - \Phi\left(\frac{\tau_{j-1} - \boldsymbol{X}_i\boldsymbol{\beta}}{\sigma}\right).$$

The unit and origin of $\boldsymbol{X}_i\boldsymbol{\beta}$ can be chosen arbitrarily (Hedeker and Gibbons, 1994) and is typically set to 1 and 0, respectively. The same choice will be made in the remainder of this paper.

6.3.2 Bivariate probit models

In instances where there is heterogeneity within the study group, relying solely on a single two-by-two table can distort the true correlation between the two responses. This distortion occurs because the correlation might be affected by other variables, either measured or unmeasured, that introduces heterogeneity. The bivariate probit model can address this issue by incorporating such effects when computing the correlation coefficient.
In analogy with the univariate probit model of the previous section, we assume that there is threshold concept involving an underlying latent variable for each of the categorical responses. The latent variables are here denoted with Y_1^* for the first categorical variable and Y_2^* for the second categorical variable. We will now assume that $\boldsymbol{Y}^* = (Y_1^*, Y_2^*)$ follows a bivariate normal density with mean $(\boldsymbol{X}_{1i}\boldsymbol{\beta}_1, \boldsymbol{X}_{2i}\boldsymbol{\beta}_2)$ and correlation ρ. The variances are set to 1. Hence, the association between both categorical responses is modelled via the polychoric correlation between the underlying latent variables. It is possible to let this correlation depend on the covariates, such that $\rho_i = \boldsymbol{X}_{3i}\boldsymbol{\beta}_3$ (Morimune, 1979). Often, a Fisher's z transformation is applied in this case to avoid that estimates jump out of the $[-1, +1]$ interval. Naturally, when the polychoric correlation equals 0, the joint probabilities equal the product of the marginal probabilities.
For the binary case, the probit model assumes the following joint probabilities

$$\begin{aligned}
P(Y_{1i} = 1, Y_{2i} = 1) &= \Phi\left(\boldsymbol{X}_{1i}\boldsymbol{\beta}_1, \boldsymbol{X}_{2i}\boldsymbol{\beta}_2, \rho\right) \\
P(Y_{1i} = 0, Y_{2i} = 1) &= \Phi\left(\boldsymbol{X}_{1i}\boldsymbol{\beta}_1\right) - \Phi\left(\boldsymbol{X}_{1i}\boldsymbol{\beta}_1, \boldsymbol{X}_{2i}\boldsymbol{\beta}_2, \rho\right) \\
P(Y_{1i} = 1, Y_{2i} = 0) &= \Phi\left(\boldsymbol{X}_{2i}\boldsymbol{\beta}_2\right) - \Phi\left(\boldsymbol{X}_{1i}\boldsymbol{\beta}_1, \boldsymbol{X}_{2i}\boldsymbol{\beta}_2, \rho\right) \\
P(Y_{1i} = 0, Y_{2i} = 0) &= 1 - \Phi\left(\boldsymbol{X}_{1i}\boldsymbol{\beta}_1\right) - \Phi\left(\boldsymbol{X}_{2i}\boldsymbol{\beta}_2\right) - \Phi\left(\boldsymbol{X}_{1i}\boldsymbol{\beta}_1, \boldsymbol{X}_{2i}\boldsymbol{\beta}_2, \rho\right)
\end{aligned}$$

This can easily be generalised to the ordinal case, where more than one threshold is implemented.

6.3.3 Multivariate probit models

The probit model can be generalised to the multivariate case, where L ordinal or binary responses are analysed. Now the L-dimensional vector of underlying latent variables $\boldsymbol{Y}^* = (Y_1^*, Y_2^*, \ldots, Y_L^*)$ follows an L-dimensional normal distribution. Thus, a multivariate probit model of dimension L comprises L marginal probability distributions, each corresponding to a specific characteristic, and $\frac{L(L-1)}{2}$ polychoric correlations indicating the relationship between the occurrences of the L characteristics. If the correlations are zero, the marginal probability distributions alone can generate the probabilities of all response combinations. If not, multivariate probability distributions are required. A detailed discussion on the

multivariate probit model can be found in Molenberghs and Lesaffre (1994) and Molenberghs and Verbeke (2005).

6.4 Pseudo-likelihood estimation

6.4.1 Introduction

When using maximum likelihood estimation, the primary approach involves maximising the (log-) likelihood to estimate the unknown parameters. For univariate continuous models, this process poses no significant computational hurdles and has been widely integrated into statistical software like SAS. However, when dealing with non-normal or high-dimensional data, specifying the full likelihood can be computationally demanding, particularly for large datasets. Instead of specifying the full likelihood, the concept of pseudo-likelihood, or composite likelihood, is often employed. The idea of pseudo-likelihood is to replace a computationally challenging joint density by a simpler function consisting of appropriate factors (Molenberghs et al., 2011a).

6.4.2 Pairwise likelihood estimation

An important special case of the pseudo-likelihood framework involves specifying all bivariate likelihoods among the complete set of possible response pairs instead of the full likelihood. In the case of bivariate likelihood, each subject's likelihood contribution to the full likelihood is substituted with a product of bivariate likelihoods. For instance, when there are four measurements per subject, the likelihood contribution of subject i is replaced by the product of the six bivariate densities, and the corresponding log-likelihood is adjusted accordingly. In the general scenario with L responses per subject i, the subject's contribution to the log pseudo-likelihood is computed as the sum of $L(L-1)/2$ bivariate log-likelihoods, and the marginal log-pseudo-likelihood is derived by summing these contributions across all subjects and two-by-two combinations. Let $\boldsymbol{\theta}$ be the vector containing the parameters of al $L(L-1)/2$ bivariate models. The pseudo-likelihood then takes the following form:

$$pl(\theta) = l_{12}(\boldsymbol{\theta}_{1,2}|\boldsymbol{Y}_1, \boldsymbol{Y}_2) + l_{13}(\boldsymbol{\theta}_{1,3}|\boldsymbol{Y}_1, \boldsymbol{Y}_3) + ... + l_{(L-1)L}(\boldsymbol{\theta}_{(L-1),L}|\boldsymbol{Y}_{L-1}, \boldsymbol{Y}_L). \tag{6.2}$$

Within the context of pseudo-likelihood, as demonstrated for the case of vector-valued parameters by Arnold and Strauss (1991), as well as by Geys et al. (1999), $\hat{\boldsymbol{\theta}}$ follows the following asymptotic multivariate distribution:

$$\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \sim N(\mathbf{0}, \boldsymbol{J}^{-1}\boldsymbol{K}\boldsymbol{J}^{-1}), \tag{6.3}$$

where $\boldsymbol{J}^{-1}\boldsymbol{K}\boldsymbol{J}^{-1}$ is a robust 'sandwich' estimator. More specifically, $\boldsymbol{J}$ is a block-diagonal matrix with blocks $\boldsymbol{J}_{pp}$ and $\boldsymbol{K}$ is a symmetric matrix containing blocks $\boldsymbol{K}_{pq}$

$$\begin{aligned}\boldsymbol{J}_{pp} &= -\frac{1}{N}\sum_{i=1}^{N} E\left(\frac{\partial^2 l_{pi}}{\partial\boldsymbol{\theta}_p\partial\boldsymbol{\theta}'_p}\right)\\ \boldsymbol{K}_{pq} &= \frac{1}{N}\sum_{i=1}^{N} E\left(\frac{\partial l_{pi}\partial l_{qi}}{\partial\boldsymbol{\theta}_p\partial\boldsymbol{\theta}'_q}\right),\end{aligned}$$

with $p, q = 1 \ldots L(L-1)/2$.

6.4.3 Pairwise fitting approach in multivariate probit models

A very useful variation to the theme of pseudo-likelihood estimation is the pairwise fitting approach, introduced by Fieuws and Verbeke (2006). Here, each bivariate model is fitted separately, independent from the others, rather than optimising the likelihood of the entire multivariate model simultaneously as described earlier. This is achieved by maximising likelihoods of the form $l_{rs}(\boldsymbol{\theta}_{r,s}|\boldsymbol{Y}_r, \boldsymbol{Y}_s)$ individually for each pair (r, s), where $\boldsymbol{\theta}_{r,s}$ represents the parameter vector for the specific pair. In this way, the computational burden is reduced considerably, as not all the parameters are estimated simultaneously. Moreover, the bivariate models can be fitted independently in parallel, which reduces the computation time even more. However, this means that some elements in the parameter vector $\boldsymbol{\theta}^*$ are estimated more than once. Therefore, these estimates are averaged in a subsequent step to obtain a single estimate for each parameter in $\boldsymbol{\theta}^*$ of the full joint model.
Let $\boldsymbol{\theta}$ then be the stacked vector combining all pair-specific parameter vectors $\boldsymbol{\theta}_{r,s}$. Note that the vectors $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^*$ are not identical. While some parameters in $\boldsymbol{\theta}^*$ correspond one-to-one with those in $\boldsymbol{\theta}$ (e.g., polychoric correlations between two responses), others have multiple counterparts (e.g., estimated thresholds for a response). In such cases, a single estimate is obtained by averaging the corresponding pair-specific maximum likelihood estimates in $\boldsymbol{\theta}^*$. This linear combination of maximum likelihood estimates inherits the asymptotic properties of its constituents. The standard errors of the resulting estimates cannot simply be calculated by averaging standard errors, since the variability among the estimates of each pair needs to be considered. Additionally, when two estimates correspond to pairwise models sharing a common outcome, they are based on overlapping information and are therefore correlated. This correlation needs to be taken into consideration when assessing the sampling variability of the combined estimates

in $\hat{\boldsymbol{\theta}}^*$. This can be constructed by the use of (6.3). To summarise, the parameter estimates are obtained via averaging over the estimates resulting from all response pairs: $\hat{\boldsymbol{\theta}}^* = \boldsymbol{A}\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}^* \sim \text{MVN}(\boldsymbol{\theta}^*, \boldsymbol{A}\Sigma(\hat{\boldsymbol{\theta}})\boldsymbol{A})$, where the matrix $\boldsymbol{A}$ contains the coefficients to calculate the averages and $\Sigma(\hat{\boldsymbol{\theta}})$ is obtained via (6.3).

6.5 Simulation study

To compare the performance of regular pseudo-likelihood (pairwise likelihood) and the pairwise fitting approach across a wide array of standardised datasets, simulated data with correlated ordinal responses $(n = 2, 3, 4, 5)$ are employed. Specifically, we simulated 100 datasets, each comprising 2000 subjects. The responses were all of ordinal nature with four categories. The responses could be predicted with age, sex, and region, while the polychoric correlation between the responses varied across sexes. We analysed the data with a regular laptop (CPU=Processor Intel(R) Core(TM) i7-8850H CPU @ 2.60GHz, 2592 Mhz, 6 Core(s), 12 LogicalProcessor(s), RAM=24GB) and averaged the computation time across the 100 datasets. The code of both the simulation and the analyses of the simulated datasets can be found via `https://github.com/MargauxDelporte`. Figure 6.2 shows the resulting computation times for pseudo-likelihood and the pairwise fitting approach for varying numbers of responses. For the pseudo-likelihood approach, the computation time increases exponentially with an increasing number of responses. The computation time for the pairwise fitting approach stays constant, since the bivariate models can be run independently in parallel. Suppose that the bivariate models would be run sequentially, the computation time would increase quadratically, instead of exponentially. Hence, even when applied sequentially, the pairwise fitting method still results in considerable time savings. Since the classical maximum-likelihood approach is not feasible for more than three responses, we did not examine the computation time of this method. But of course, this is a very important consideration in its own right.

6.6 Analysis of the case studies

6.6.1 BIRNH

Here, the BIRNH dataset is analysed, introduced in Section 6.2. In order to reproduce the results of Lesaffre and Molenberghs (1991), we kept the same predictors of the probit model for the ordinal responses *Alcohol* and *Smoking* and the association between the responses. For the new binary variable *Cholesterol*, we performed backwards variable selection, where only SEX was dropped from

the model. The code, both for pseudo-likelihood estimation and the pairwise fitting approach, can be found on `https://github.com/MargauxDelporte`. The table with resulting parameter estimates, standard errors and p-values, identical for both methods, is displayed in Appendix S.2. The results are in line with Lesaffre and Molenberghs (1991): women drink and smoke less than men and on average the people in Flanders drink less alcohol. Being unemployed results in more drinking, but less smoking and a lower probability of high cholesterol. While the estimates in Table S.2 are on the Fisher's z-scale, the polychoric correlation for the different combinations of classes is shown in Figure 6.3. It shows for example that while the polychoric correlation between smoking and alcohol is negative for women who work outside, it is positive for men who work at home.

6.6.2 POPS

As a second illustration, we will analyse the POPS dataset, introduced in Section 6.2. In this study, it is of main interest to estimate the polychoric correlation between the three ability scores, while controlling for confounding variables: neonatal seizures, congenital malformation and maximum bilirubin since birth. The code can be found on `https://github.com/MargauxDelporte` and the detailed results are shown in Appendix S.3. The results under MCAR indicate that the presence of neonatal seizures and congenital malformation significantly decreases the probability of successfully performing the any of the three tests. In addition, a similar significant effect of bilirubin is observed for the first and second ability score. Still, the parameter estimates and conclusions also depend greatly on the assumptions regarding the the missingness mechanism, since 40% of the cases is incomplete and 18% has missing covariate values. More importantly, pseudo-likelihood estimation is only valid under the assumption MCAR, while multiple imputation is valid under the less restrictive assumption of MAR. We imputed the dataset ten times using fully conditional specification (van Buuren, 2007). Each variable was imputed with an appropriate model for its response type, based on all the variables included in the analysis. Specifically, we used the three ability scores, congenital malformation, neonatal seizures, and the highest bilirubin value since birth. Interaction terms or polynomials were not included in the imputation model. Performing the analysis after multiple imputation, results in a strong decrease of the effect of neonatal seizures and the polychoric correlations between the ability scores.
A notable aspect of the multivariate approach is its ability to compute joint probabilities. For instance, it enables the calculation of the combined probability of failing all three tests. This may vary considerably from the joint probability derived under the assumption of independent responses. This probability is depicted for the different models, for varying bilirubin values and given that both

congenital malformation and neonatal seizures are present, in Figure 6.4. Notably, the complete case analysis makes the assumption of MCAR, while the model after multiple imputation makes the less restrictive assumption of MAR. As indicated in the figure, the probabilities depend considerably on whether or not independence, MCAR or MAR assumptions are made. Note that the models assuming independence or MAR use all available cases, resulting in smaller confidence intervals compared to the complete case approach.

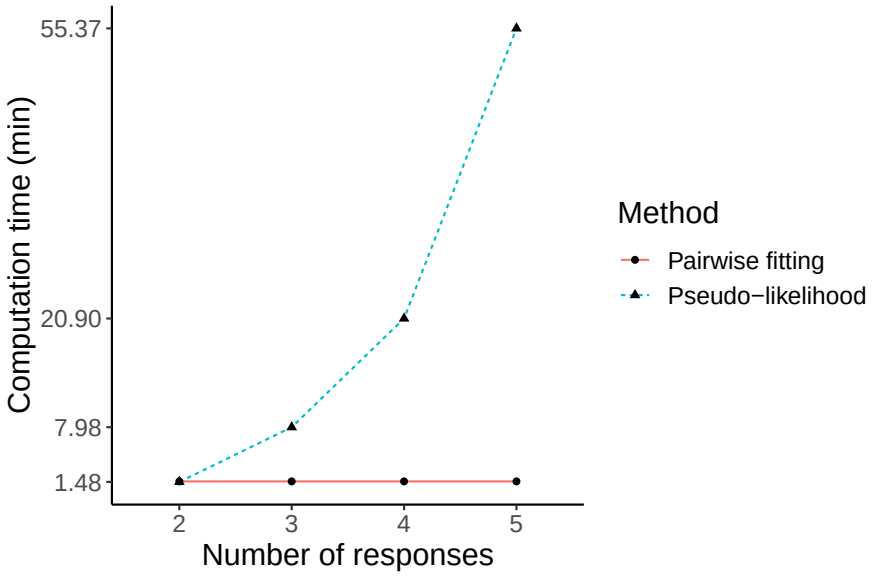

Figure 6.2: *Comparison of computation times for regular pseudo-likelihood (pairwise likelihood) and pairwise fitting approaches across varying numbers of responses.*

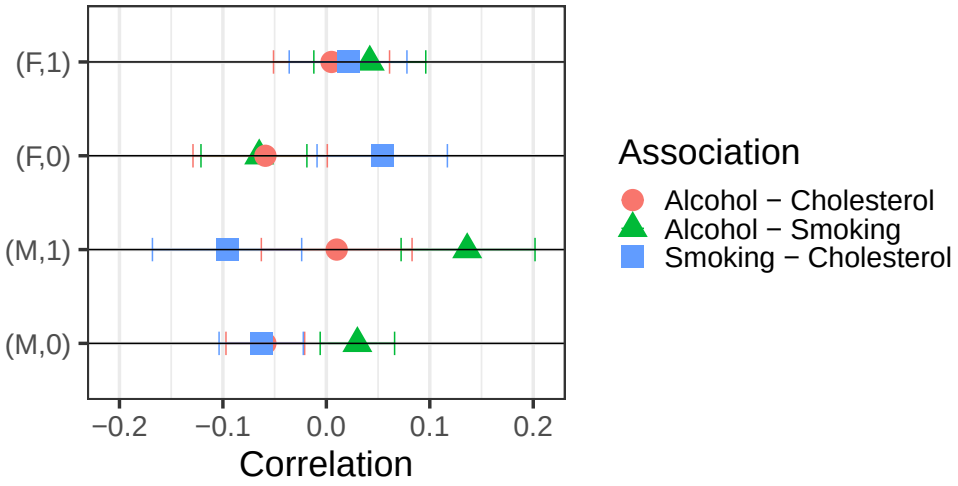

Figure 6.3: *Polychoric correlations with 95% confidence intervals, according to sex and SOC2 (1: working at home, 0: working outside)*

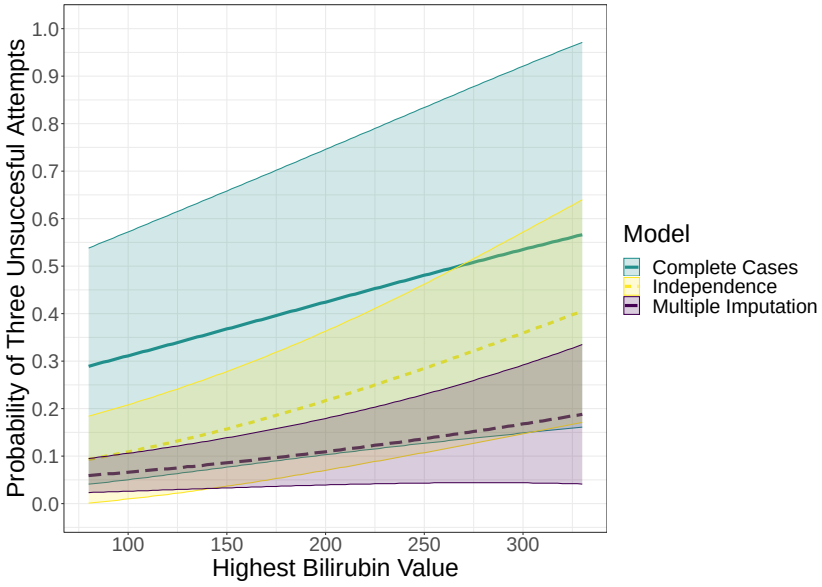

Figure 6.4: *Probability that a child fails on all three ability scores for a range of bilirubin values, evaluated under three fitted multivariate probit models, with 95% confidence intervals.*

6.7 Discussion

In this paper, we explore the application of the pairwise fitting technique as a means to drastically reduce the computational burden when fitting multivariate probit models. The multivariate probit model is relevant in medical research due to its ability to account for and to estimate correlations between multiple responses. However, its primary limitation is the substantial computation time required when fitting the model with maximum-likelihood, which becomes unfeasible when dealing with more than three responses.

A viable alternative to maximum likelihood estimation is the pseudo-likelihood method, which uses an appropriate approximation of the full likelihood function. For example, the full high-dimensional likelihood can be adequately approximated by the sum of the bivariate likelihoods, and the latter is optimised. For instance, Hirk et al. (2020) developed an R package that employs this method. Nonetheless, our findings demonstrate that the computational burden can be further mitigated by fitting bivariate pairs instead of optimising the full sum of bivariate likelihoods, as evidenced by our simulation study. The associated SAS code is available on `https://github.com/MargauxDelporte`, allowing for easy implementation on personal datasets or for replication of our results.

Despite its computational efficiency, the pseudo-likelihood method does have drawbacks, particularly the loss of some properties inherent to maximum likelihood estimation. Notably, under the assumption of Missing at Random (MAR), ignorability is no longer valid. Consequently, the pseudo-likelihood approach necessitates additional steps when the missingness mechanism is not Missing Completely at Random (MCAR). These steps may include multiple imputation or weighting mechanisms, as discussed by Molenberghs et al. (2011a). In other words, when multiple imputation is applied, inferences with the pseudo-likelihood approach are valid under MAR. When the assumption of MCAR holds, analysing complete pairs or available cases can lead to small differences between the pairwise fitting approach and the pseudo-likelihood method due to variations in weighting mechanisms. For an in-depth discussion on this topic, we refer to the comments section on the paper of Fitzmaurice et al. (1997).

In our case studies, we demonstrate the practical importance of our multivariate probit approach in analysing medical datasets. By re-analysing older data that previously faced computational constraints, we were able to leverage our method effectively. For example, in the BIRNH case study, we analysed the complete dataset ($n = 11302$) rather than a 10% random sample to calculate the polychoric correlations between latent variables underlying health risks across different demographic groups. For example, the analysis revealed that the association between smoking and cholesterol is negative for men but positive for women. In the POPS case study, we examined the effects of assuming independence and

the MAR versus the MCAR assumption. Our findings indicated that assumptions regarding independence and the missingness mechanism strongly impacted the estimates of correlations between responses. Again, the code of the analyses of both case studies is shared on `https://github.com/MargauxDelporte` to facilitate the straightforward implementation of our method.

Chapter 7

Concluding remarks

7.1 General discussion

This dissertation introduced and applied new statistical techniques aimed at addressing specific research questions. Their use was demonstrated within the context of medical research, though these methods can also be adapted for use with similarly structured data from other areas. The main focus of this dissertation was on the analysis of multivariate longitudinal data, since little alternatives existed for TDC's. These alternative approaches were very important, since TDC's are not applicable in every scenario. For example, TDC's with a lag cannot be used when the responses, and/or the time-dependent covariates, are not collected at fixed intervals.

In Chapter 2, we discussed a joint model for two responses, one of binary nature and one of continuous nature. The approach was based on joint generalised linear mixed models with correlated random effects. Traditionally, examining the association between the responses was limited to scrutinising the correlations between the random effects, which are on an underlying (latent) scale. Our approach extended the existing methodology by deriving closed-form formulas for the computation of the manifest correlations between the responses, which are on the observed scale. Next, we integrated the random effects out of the joint density to derive the marginal model. From the marginal model, we derived conditional expected values, conditional probabilities and the corresponding prediction/confidence intervals for a subvector of one of the responses conditional on another and potentially a subvector of the response treated as outcome.

In Chapter 3 the latter approach was extended to the high-dimensional case, where we analysed more than two binary and/or more than two continuous responses. By rearranging matrices and vectors and altering the design matrices accordingly, the calculations for the bivariate case were still valid in the high-dimensional approach. Further, we extended the methodology by deriving

closed-form formulas for the confidence intervals of the manifest correlations between the responses on the observed scale. Additionally, we applied the pairwise fitting approach to avoid computational problems when fitting our complex joint model on a large high-dimensional dataset.
Chapter 4 extended our approach from the combination of continuous and binary responses to the combination of continuous and ordinal responses, where we continued to use a probit link to analyse the categorical response. Also in this case, joint modelling was applied, where we allowed correlations between the random effects of the generalised linear mixed models. Closed-form formulas were deduced in order to calculate the correlations between the responses on the observed scale at two given time points, together with corresponding confidence intervals. Additionally, we derived the marginal model, where the random effects were integrated out of the joint density. From the marginal model, we derived closed-form formulas to make predictions of one response conditional on the other response(s), potentially the history of the predicted response, and the covariates. As a consequence, we could make predictions for a subvector of one response conditional on the other response and potentially a subvector of the history of the predicted response. Again, by rearranging matrices and altering design matrices, the approach could be seamlessly extended to the high-dimensional case.
A new data structure was examined in Chapter 5. More specifically, we investigated possible methods to analyse paired longitudinal data. Hence, this data structure had additional complexity since there was pairing between the participants (e.g., matched case-control studies) or within the participants (e.g., analysis of participants' both eyes). In other words, there were correlations between members of a pair, next to the correlations of measurements within members of a pair. Various modelling approaches, identified through a systematic review, were explored, including (un)paired t-tests, multivariate analysis of variance (MANOVA), difference scores, linear mixed models (LMM), and newer statistical methods. Emphasis was placed on choosing suitable models based on data characteristics. The methods were compared in both a simulation study, and in the analysis of a real-life medical dataset. We found that the standard errors of the treatment effect at baseline were the most accurately estimated when employing either the multilevel (nested) linear mixed model or the conditional linear mixed model.
The last chapter, Chapter 6, was about multivariate probit models. In this chapter, we examined the application of the pairwise fitting technique to reduce the computational complexity of fitting multivariate probit models, which were important in medical research for estimating correlations between multiple binary and/or ordinal responses. The primary challenge of these models was the substantial computation time required for maximum-likelihood estimation, which became impractical when dealing with more than three responses. Our proposed

alternative approach was based on the pseudo-likelihood method, which approximates the full likelihood function, often using the sum of bivariate likelihoods. For instance, an R package developed by Hirk et al. (2020) implemented this method. Our study demonstrated that further computational efficiency, compared to Hirk et al. (2020), could be achieved by fitting bivariate pairs instead of optimising the full sum of bivariate likelihoods.

7.2 Clinical relevance

The joint modelling approach, developed for, amongst others, analysing longitudinal data with irregularly measured responses, holds substantial clinical relevance. This is especially the case in complex and chronic conditions where patient data was not always collected at regular intervals. This method's flexibility and robustness made it particularly valuable for a variety of medical research applications, including studies with practical constraints on data collection schedules.
One famous example where our methods could hold considerable benefits over traditional TDC's is in the analysis of data from the Baltimore Longitudinal Study of Aging (Ferrucci, 2008), an observational study in which patient data are collected during occasional hospital visits, which do not occur at uniform time intervals. Traditional methods (TDC's) would require regularly spaced data points for the use of lags. However, our proposed method could effectively handle such irregular intervals, providing more accurate insights into disease progression and patient outcomes. By accommodating the real-world scenario of irregular hospital visits, our method could potentially enhance the study's relevance and applicability, offering more insights for developing personalised treatment plans and monitoring strategies. We demonstrated in Chapter 3 the potential of our method for dynamic predictions, where we adjusted our predictions when new data became available at later time-points.
The clinical relevance of this method was further underscored by its application in cystic fibrosis research. In a study exploring the relationship between Forced Expiratory Volume (FEV) and Allergic Bronchopulmonary Aspergillosis (ABPA), the focus was on the evolution of the association between both longitudinal responses. The ability of this method to discover the increase in correlations with increasing time intervals between the responses was of particular clinical relevance. It raised the hypothesis that the occurrence of ABPA might be an early signal for frailty.
Another interesting example of the method's clinical relevance was its use in a study investigating the link between cognitive impairment and physical dependency in hospitalised patients after a hip fracture. Cognitive and physical assessments were conducted with different frequency; the cognitive assessment

was taken more regularly compared to the physical assessment. Traditional methods could not make use of all the data or incorporate lags, while this did not pose issues for joint modelling. The method addressed the research question of how cognitive decline impacts physical recovery and dependency over time. It showed that high dependency in activities of daily living during the first five post-operative days had predictive value for mental impairment at a later stage. This insight could be of potential use for developing targeted rehabilitation programs and support systems for patients recovering from hip fractures, ultimately improving their quality of life and reducing the burden on healthcare systems.
The clinical relevance of Chapter 5 was evidenced by the systematic review. Paired longitudinal data is frequently encountered in medical research, for instance in a longitudinal analysis of matched participants. In more than half of the articles included in our systematic review the paired nature of the data was ignored. However, accounting for these intra-pair correlations could lead to more precise estimates of the studied effect. Even more, when researchers accounted for the paired nature of the data, they often did not elaborate on the impact of missing data on the method. This chapter aimed to offer a thorough overview and discussion of the available methodologies for matched longitudinal data, enabling researchers to make informed decisions regarding the methodology based on their specific data and research questions.
In the case studies of Chapter 6, we highlighted the practical significance of our multivariate probit approach in analysing medical datasets. We identified two medical case studies where the multivariate probit model was used, but computational limitations hindered fully leveraging the richness of the data. In the first case study, we analysed the full dataset instead of a random subset of 10% of the original sample, as was done previously in order to tackle computational issues. Additionally, alongside smoking and drinking, we included a third response in our analysis: elevated cholesterol. In this study, polychoric correlations between latent variables associated with health risks across various demographic groups were calculated. This analysis revealed, for example, that the correlation between smoking and cholesterol is negative for men but positive for women.In the original analysis of the second case study, only the complete cases were analysed, which comprised only 60% of the entire data set. As a consequence, a lot of information was lost. With our method, we were able to analyse 10 imputed datasets and make less stringent assumptions about the missing data mechanism.

7.3 Limitations

Like many other methods in the literature, our proposals and implementations have potential for enhancement.

First and foremost, the joint modelling approach was much more involved in comparison with the TDC approach. The complexity of the code needed to implement joint models is substantial, often necessitating advanced statistical programming skills. To improve the feasibility of our method for non-statisticians, developing `SAS` macros and `R` functions is essential. Additionally, datasets must frequently be reorganised to fit the model's requirements, adding another layer of complexity to the analysis process. In addition, when working with large datasets or high-dimensional data, model fitting can take a long time and computational problems can arise, such as convergence issues. While there are some situations where TDC's cannot be applied (e.g., endogeneity or irregular time intervals), time dependent covariates remain a relative straightforward way to study the effect of one longitudinal response on another. The results of models with TDC's are very intuitive and readily available, in contrast to our method where complex formulas need to be applied before obtaining predictions.

Another drawback of the joint modelling approach was its dependence on the random effects structure for the flexibility of the manifest correlation matrix. When both responses included random intercepts and slopes, the estimated correlations could vary more over time. However, if a random slope was not included for one or more responses, this led to a significant limitation in the time-dependent changes in the correlations on the observed scale.

Furthermore, it was worth noting that a limitation of our joint modelling methodology with ordinal responses was its reliance on the proportional odds assumption inherent in the ordinal regression model. Of course, extension to non-proportionality was possible by having (certain) covariate effects category-dependent. But then, as always, care needs to be taken to ensure non-negative probabilities.

Despite its computational efficiency, the pseudo-likelihood method (pairwise fitting approach) for multivariate probit models had some drawbacks, especially the loss of certain properties intrinsic to maximum likelihood estimation. Particularly, ignorability under the assumption of Missing at Random (MAR) was no longer applicable. Therefore, the pseudo-likelihood method required additional steps when the missingness mechanism is not Missing Completely at Random (MCAR). These steps might include methods such as multiple imputation or weighting mechanisms, as noted by Molenberghs et al. (2011a). Essentially, when multiple imputation was used, inferences with the pseudo-likelihood method were valid under MAR.

If the MCAR assumption holds, analysing complete pairs or available cases may

result in minor differences between the pairwise fitting approach and the pseudo-likelihood method due to differences in weighting mechanisms. For a comprehensive discussion on this topic, we referred to the comments section of the paper by Fitzmaurice et al. (1997).

7.4 Future research

In the joint modelling of longitudinal data, different directions for future research are possible. One potential direction is to compare the results of our conditional expected value approach with those obtained using models with TDC's. Specifically, it would be valuable to examine the differences in (the length of) confidence intervals and prediction intervals between the two methods. Although we have already assessed our approach against models with TDC's using a real-life dataset, conducting a simulation study could provide further insights by evaluating these differences across various scenarios. For example, it could be interesting to assess the results with missing data (e.g., in combination with the pairwise fitting approach) or under the scenario of small sample sizes. The latter would be particularly relevant in rare disease research. Another research direction involves extending our approach beyond the current combination of continuous-binary and continuous-ordinal data. Extending its application to combinations of continuous data with other response types, such as count data or multinomial data, could be particularly useful. For multinomial data, we anticipate that the calculations would closely resemble those used for binary data. A third option would be to explore the implementation of our methods in surrogate marker research and (observational) studies with time-varying treatments.
In the analysis of paired longitudinal data, we only considered continuous responses. In practice, paired longitudinal responses of binary nature, multinomial nature or other non-normal types are also frequently encountered. Future research could involve identifying possible methods for the analysis of paired non-continuous longitudinal data, and comparing these methods. In addition, when simulating paired longitudinal data to compare the different methods, we generated only complete datasets. Future research could explore the potential impact of missing data under different missing data mechanisms. This consideration also applies to the pairwise fitting approach in multivariate probit models. Throughout the methodological section of our paper, the MCAR assumption was maintained. However, it is well-known that using pseudo-likelihood instead of maximum likelihood results in the loss of several desirable properties, such as ignorability under MAR. We investigated the impact of multiple imputation compared to complete case analysis, but the approach could be further extended. For instance, incorporating weighting mechanisms in the pairwise fitting approach could be further

explored. Nevertheless, we do not anticipate that weighting mechanisms would offer any advantages over multiple imputation.

To facilitate the broader adoption and implementation of advanced statistical methods, there is a need to develop and refine R packages and SAS macros. These tools should include the closed-form formulas derived from the joint models (conditional expected values, probabilities and manifest correlations) and functionalities for fitting multivariate probit models. Creating user-friendly and well-documented software solutions will allow researchers and practitioners to apply these complex methodologies more easily.

Summary

In biomedical research, hierarchical binary, ordinal and continuous responses are often jointly modelled using generalised linear mixed models with correlated random effects. This allows for examining the association structure between various responses and the effect of covariates on all outcomes. However, the investigation is usually limited to the correlations between the underlying subject-specific random effects. This thesis extends the methodology by providing ways to compute manifest correlations, i.e., correlations between the responses on the observed scale, and deriving a marginal model where we integrated the random effects out of the joint density. Expected values and probabilities are derived for one subvector of responses conditional on another. Pseudo-likelihood methodology is employed to handle computational complexity in high-dimensional cases. The methods are applied in medical case studies, including lung function and allergic bronchopulmonary aspergillosis in cystic fibrosis patients.

Further, longitudinal data in medical research often involves complex structures, such as the combination of dyads and repeated measures (within members) of the dyad. Various modelling approaches, including t-tests, MANOVA, difference scores, and linear mixed models, are reviewed. These methods are compared in both a real-life case study in ophthalmology and simulated case-control studies, which led to a favourable result for nested linear mixed models and conditional linear mixed models in handling paired longitudinal data. Moreover, an efficient computational approach is introduced for multivariate probit models using pairwise fitting within a pseudo-likelihood framework. Our approach improves computational efficiency in capturing covariate-dependent correlations in medical research.

Bibliography

Aerts, M., Geys, H., Molenberghs, G., and Ryan, L. (2002). *Topics in Modelling of Clustered Data*. Number 2. Chapman & Hall, New York, 1 edition.

Ahmed, B., King, W., Gourash, W., Belle, S., Hinerman, A., Pomp, A., Dakin, G., and Courcoulas, A. (2018). Long-term weight change and health outcomes for sleeve gastrectomy (sg) and matched roux-en-y gastric bypass (rygb) participants in the longitudinal assessment of bariatric surgery (labs) study. *Surgery (United States)*, 164(4):774–783.

Alonso, A. and Molenberghs, G. (2007). Surrogate marker evaluation from an information theory perspective. *Biometrics*, 63(1).

Arnold, B. C. and Strauss, D. (1991). Pseudolikelihood Estimation: Some Examples. *The Indian Journal of Statistics, Series B*, 53(2):233–243.

Benestad, M., Drageset, J., Eide, G., Vollsæter, M., Halvorsen, T., and Vederhus, B. (2022). Development of health-related quality of life and subjective health complaints in adults born extremely preterm: a longitudinal cohort study. *Health and Quality of Life Outcomes*, 20(1).

Beutel, M., Willner, H., Deckardt, R., Von Rad, M., and Weiner, H. (1996). Similarities and differences in couples' grief reactions following a miscarriage: Results from a longitudinal study. *Journal of Psychosomatic Research*, 40(3):245–253.

Border, W., Sachdeva, R., Stratton, K., Armenian, S., Bhat, A., Cox, D., Leger, K., Leisenring, W., Meacham, L., Sadak, K., Sivanandam, S., Nathan, P., and Chow, E. (2020). Longitudinal changes in echocardiographic parameters of cardiac function in pediatric cancer survivors. *JACC: Cardiooncology*, 2(1):26–37.

Bouwmans, M., Bos, E., Booij, S., van Faassen, M., Oldehinkel, A., and de Jonge, P. (2015). Intra- and inter-individual variability of longitudinal daytime melatonin secretion patterns in depressed and non-depressed individuals. *Chronobiology International*, 32(3):441–446.

Breslow, N. E. and Clayton, D. G. (1993). Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association*, 88(421).

Cavender, J. B., Rogers, W. J., Fisher, L. D., Gersh, B. J., Coggin, C. J., and Myers, W. O. (1992). Effect of smoking on survival and morbidity in patients randomized to medical or surgical therapy in the coronary artery surgery study (CASS): 10-Year follow-up. *Journal of the American College of Cardiology*, 20(2).

Chakraborty, H., Helms, R. W., Sen, P. K., and Cohen, M. S. (2003). Estimating correlation by using a general linear mixed model: Evaluation of the relationship between the concentration of HIV-1 RNA in blood and semen. *Statistics in Medicine*, 22(9).

De Coninck, D., d'Haenens, L., Molenberghs, G., Declercq, A., Delecluse, C., Van Roie, E., and Matthijs, K. (2022). Updating 'Perceptions and opinions on the COVID-19 pandemic in Flanders, Belgium' with data of two additional waves of a longitudinal study. *Data in Brief*, 42.

Delporte, M., Fieuws, S., Molenberghs, G., Verbeke, G., Situma Wanyama, S., Hatziagorou, E., and De Boeck, C. (2022). A joint normal-binary (probit) model. *International Statistical Review*.

Delporte, M., Luyts, M., Molenberghs, G., Verbeke, G., Demarest, S., and Hoorens, V. (2023). Do optimism and moralization predict vaccination? A five-wave longitudinal study. *Health Psychology*.

Diggle, P. (2002). *Analysis of longitudinal data*, volume 25. Oxford statistical science series, New York, 2nd ed edition.

Duncan, L. A., Schaller, M., and Park, J. H. (2009). Perceived vulnerability to disease: Development and validation of a 15-item self-report instrument. *Personality and Individual Differences*, 47(6):541–546.

Efendi, A., Molenberghs, G., Njagi, E. N., and Dendale, P. (2013). A joint model for longitudinal continuous and time-to-event outcomes with direct marginal interpretation. *Biometrical Journal*, 55(4):572–588.

Engel, B. and Keen, A. (1994). A simple approach for the analysis of generalizea linear mixed models. *Statistica Neerlandica*, 48(1).

Faes, C., Geys, H., Aerts, M., Molenberghs, G., and Catalano, P. J. (2004). Modeling combined continuous and ordinal outcomes in a clustered setting. *Journal of Agricultural, Biological, and Environmental Statistics*, 9(4):515–530.

Ferrucci, L. (2008). The baltimore longitudinal study of aging (blsa): A 50-year-long journey and plans for the future. *The journals of gerontology. Series A, Biological sciences and medical sciences*, 63(12):1416–1419.

Fieuws, S. and Verbeke, G. (2006). Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal profiles. *Biometrics*, 62(2).

Fieuws, S., Verbeke, G., Boen, F., and Delecluse, C. (2006). High dimensional multivariate mixed models for binary questionnaire data. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, 55(4).

Fieuws, S., Verbeke, G., Maes, B., and Vanrenterghem, Y. (2008). Predicting renal graft failure using multivariate longitudinal profiles. *Biostatistics*, 9(3).

Fitzmaurice, G., Davidian, M., Verbeke, G., and Molenberghs, G. (2008). *Longitudinal data analysis*, volume 1. Chapman and Hall/CRC, New York, 1 edition.

Fitzmaurice, G. M., Heath, A. F., and Cox, D. R. (1997). Detecting Overdispersion in Large Scale Surveys: Application to a Study of Education and Social Class in Britain. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 46(4):415–432.

Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. (2004). *Applied longitudinal analysis.* Wiley series in probability and statistics. Wiley, Hoboken.

Gerber, J., Bryan, M., Ross, R., Daymont, C., Parks, E., Localio, A., Grundmeier, R., Stallings, V., and Zaoutis, T. (2016). Antibiotic exposure during the first 6 months of life and weight gain during childhood. *Journal of the American Medical Association*, 315(12):1258–1265.

Geys, H., Molenberghs, G., and Ryan, L. M. (1999). Pseudolikelihood Modeling of Multivariate Outcomes in Developmental Toxicology. *Journal of the American Statistical Association*, 94(447):734–745.

Goodman, E. and Must, A. (2011). Depressive symptoms in severely obese compared with normal weight adolescents: Results from a community-based longitudinal study. *Journal of Adolescent Health*, 49(1):64–69.

Gothe, F., Kappler, M., and Griese, M. (2017). Increasing Total Serum IgE, Allergic Bronchopulmonary Aspergillosis, and Lung Function in Cystic Fibrosis. *Journal of Allergy and Clinical Immunology: In Practice*, 5(6).

Hedeker, D. and Gibbons, R. D. (1994). A Random-Effects Ordinal Regression Model for Multilevel Analysis. *Biometrics*, 50(4):933–944.

Hirk, R., Hornik, K., and Vana, L. (2020). Mvord: An R package for fitting multivariate ordinal regression models. *Journal of Statistical Software*, 93.

Huber, P. (1967). The behavior of maximum likelihood estimates under non-standard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, University of California Press*, 1:221–233.

Iaffaldano, P., Lucisano, G., Caputo, F., Paolicelli, D., Patti, F., Zaffaroni, M., Morra, V., Pozzilli, C., De Luca, G., Inglese, M., Salemi, G., Maniscalco, G., Cocco, E., Sola, P., Lus, G., Conte, A., Amato, M., Granella, F., Gasperini, C., Bellantonio, P., Totaro, R., Rovaris, M., Salvetti, M., Clerici, V., Bergamaschi, R., Maimone, D., Scarpini, E., Capobianco, M., Comi, G., Filippi, M., and Trojano, M. (2021). Long-term disability trajectories in relapsing multiple sclerosis patients treated with early intensive or escalation treatment strategies. *Therapeutic Advances in Neurological Disorders*, 14.

Iddi, S. and Molenberghs, G. (2012). A joint marginalized multilevel model for longitudinal outcomes. *Journal of Applied Statistics*, 39(11).

Ivanova, A., Molenberghs, G., and Verbeke, G. (2016). Mixed models approaches for joint modeling of different types of responses. *Journal of Biopharmaceutical Statistics*, 26(4).

Ivanova, A., Molenberghs, G., and Verbeke, G. (2017). Fast and highly efficient pseudo-likelihood methodology for large and complex ordinal data. *Statistical Methods in Medical Research*, 26(6).

Kim, J. (2006). Hypothesis testing problems in an unbalanced longitudinal ophthalmology study. *Communications in Statistics - Theory and Methods*, 35(3):461–476.

Krueger, C. and Tian, L. (2004). A comparison of the general linear mixed model and repeated measures anova using a dataset with multiple missing data points. *Biological Research For Nursing*, 6(2):151–157. PMID: 15388912.

Kundu, M. G. (2011). Implementation of Pairwise Fitting Technique for Analyzing Multivariate Longitudinal Data in SAS.

Laird, N. M. and Ware, J. H. (1982). Random-Effects Models for Longitudinal Data. *Biometrics*, 38(4).

Lange, N. and Ryan, L. (1989). Assessing normality in random effects models. *The Annals of Statistics*, 17(2):624–642.

Lesaffre, E. and Molenberghs, G. (1991). Multivariate probit analysis: a neglected procedure in medical statistics. *Statistics in Medicine*, 10(9):1391–1403.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22.

LoCascio, V., Ballanti, P., Milani, S., Bertoldo, F., LoCascio, C., Zanolin, E., and Bonucci, E. (1998). A histomorphometric long-term longitudinal study of trabecular bone loss in glucocorticoid-treated patients: Prednisone versus deflazacort. *Calcified Tissue International*, 62(3):199–204.

Manjunath, B. G. and Wilhelm, S. (2021). Moments Calculation for the Doubly Truncated Multivariate Normal Density. *Journal of Behavioral Data Science*, 1(1):13–33.

Milanzi, E., Molenberghs, G., Alonso, A., Verbeke, G., and De Boeck, P. (2015). Reliability measures in item response theory: Manifest versus latent correlation functions. *British Journal of Mathematical and Statistical Psychology*, 68(1):43–64.

Milisen, K., Abraham, I. L., and Broos, P. L. (1998). Postoperative variation in neurocognitive and functional status in elderly hip fracture patients. *Journal of Advanced Nursing*, 27(1).

Molenberghs, G. and Kenward, M. (2007). *Missing Data in Clinical Studies.* Wiley series in probability and statistics. Wiley, Hoboken.

Molenberghs, G., Kenward, M. G., Verbeke, G., and Birhanu, T. (2011a). Pseudo-likelihood estimation for incomplete data. *Statistica Sinica*, 20(1):187–206.

Molenberghs, G. and Lesaffre, E. (1994). Marginal Modeling of Correlated Ordinal Data Using a Multivariate Plackett Distribution. *Journal of the American Statistical Association*, 89(426):633–644.

Molenberghs, G. and Verbeke, G. (2005). *Models for discrete longitudinal data. New York*. Springer, New York, NY.

Molenberghs, G., Verbeke, G., Demétrio, C. G., and Vieira, A. M. (2010). A Family of generalized linear models for repeated measures with normal and conjugate random effects. *Statistical Science*, 25(3).

Molenberghs, G., Verbeke, G., and Iddi, S. (2011b). Pseudo-likelihood methodology for partitioned large and complex samples. *Statistics and Probability Letters*, 81(7).

Morimune, K. (1979). Comparisons of Normal and Logistic Models in the Bivariate Dichotomous Analysis. *Econometrica*, 47(4):957–975.

Morrell, C. H., Brant, L. J., Sheng, S., and Metter, E. J. (2012). Screening for prostate cancer using multivariate mixed-effects models. *Journal of Applied Statistics*, 39(6).

Pepe, M. S. and Anderson, G. L. (1994). A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Communications in Statistics - Simulation and Computation*, 23(4).

Pepe, M. S. and Couper, D. (1997). Modeling partly conditional means with longitudinal data. *Journal of the American Statistical Association*, 92(439):991–998.

Pinheiro, J. C. and Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics*, 4(1).

Qian, T., Klasnja, P., and Murphy, S. A. (2020). Linear Mixed Models with Endogenous Covariates: Modeling Sequential Treatment Effects with Application to a Mobile Health Study. *Statistical Science*, 35(3).

Rizopoulos, D. (2012). *Joint models for longitudinal and time-to-event data: With applications in R*. CRC Press, Boca Ranton, FL.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3).

Ruhdorfer, A., Wirth, W., Dannhauer, T., and Eckstein, F. (2015). Longitudinal (4 year) change of thigh muscle and adipose tissue distribution in chronically painful vs painless knees - data from the osteoarthritis initiative. *Osteoarthritis and Cartilage*, 23(8):1348–1356.

Schlee, W., Simoes, J., and Pryss, R. (2021). Auricular acupressure combined with self-help intervention for treating chronic tinnitus: A longitudinal observational study. *Journal of Clinical Medicine*, 10(18).

Shek, D. and Dou, D. (2020). Perceived parenting and parent-child relational qualities in fathers and mothers: Longitudinal findings based on hong kong adolescents. *International Journal of Environmental Research and Public Health*, 17(11):1–20.

Sibbel, S., Hunt, A., Laplante, S., Beck, W., Gellens, M., and Brunelli, S. (2016). Comparative effectiveness of dialyzers: A longitudinal, propensity score-matched study of incident hemodialysis patients. *ASAIO Journal*, 62(5):613–622.

Tombaugh, T. N. and McIntyre, N. J. (1992). The Mini-Mental State Examination: A Comprehensive Review. *Journal of the American Geriatrics Society*, 40:922–935.

van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, 16(3):219–242.

Verbeke, G., Fieuws, S., Molenberghs, G., and Davidian, M. (2014). The analysis of multivariate longitudinal data: A review. *Statistical Methods in Medical Research*, 23(1):42–49.

Verbeke, G. and Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. Springer Series in Statistics, New-York.

Verbeke, G. and Molenberghs, G. (2003). The Use of Score Tests for Inference on Variance Components. Technical report.

Verbeke, G., Spiessens, B., and Lesaffre, E. (2001). Conditional linear mixed models. *American Statistician*, 55.

Verloove-Vanhorick, P., Verwey, R. A., Brand, R., Bennebroek Gravenhorst, J., Keirse, M. J. C., and Ruys, J. H. (1986). Neonatal Mortality Risk in Relation to Gestational Age and Birthweigh. *The Lancet*, 1:55–57.

White, H. (1980). A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica*, 48(4):817–838.

Wilson, R. (1979). Analysis of longitudinal twin data. basic model and applications to physical growth measures. *Acta Geneticae Medicae et Gemellologiae*, 28(2):93–105.

Wolfinger, R. and O'Connell, M. (1993). Generalized linear mixed models: a pseudo-likelihood approach. *Journal of Statistical Computation and Simulation*, 48.

Zeger, S. L., Liang, K.-Y., and Albert, P. S. (1988). Models for Longitudinal Data: A Generalized Estimating Equation Approach. Technical Report 4.

Zimmerman, D. W. (1997). A note on interpretation of the paired-samples t test. *Journal of Educational and Behavioral Statistics*, 22:349–360.

Acknowledgment

At this point, trying to describe the gratitude I feel only seems to diminish it, as words can only convey so much. Thank you, Geert and Geert, for providing me with the opportunity to pursue a PhD and for mentoring me along the way. Thank you Steffen for your expertise and invaluable encouragement. Throughout my PhD trajectory, I felt truly blessed to be supervised by such an amazing team.

I am deeply thankful to my internal and external jury members, for their assistance and expertise. Their constructive feedback greatly enriched the quality of this work.

My sincere gratitude is extended to my colleagues and friends of L-BioStat. I greatly enjoyed the stimulating work environment and our lunches, birthday parties and trips together. Special thanks to Evert, Maxime and Dries V who have been by my side since the first day of the Master of Statistics and continue to be a great source of support.

My heartfelt appreciation goes to my parents, parents-in-law, sister, and friends for their unwavering love, encouragement, and understanding throughout this journey. Their willingness to lend an empathetic ear has been invaluable. Special thanks to my partner for his undying support and for being the backbone of our household.

Lastly, I extend my gratitude to all those who have directly or indirectly contributed to this work, whether through discussions, feedback, or moral support. Your contributions have been immensely valuable.

Scientific Acknowledgement, Conflict of Interest and Personal Contribution

Scientific Acknowledgments

Conflict of Interest

There is no conflict of interest.

Personal Contribution

Chapter 2: Calculation of the formulas for the manifest correlations, conditional distribution, conditional expected values and conditional probabilities. Proof of W=V in the calculation of the marginal joint density. Analysis of the Cystic Fibrosis data and interpretation of the results.
Chapter 3: Formulation of the high-dimensional models in analogy with the bivariate model in Delporte et al. (2022). Analysis of the COVID-19 and Vaccination data and interpretation of the results.
Chapter 4: Calculation of the formulas for the marginal joint model, manifest correlations, conditional distribution, conditional expected values and conditional probabilities. Analysis of the Hip Fracture data and interpretation of the results.
Chapter 5: Performing the systematic review on matched continuous longitudinal data. Analysis of the Ophthalmology data and interpretation of the results. Conducting the simulation study and interpretation of the results.
Chapter 6: Implementation of the pairwise fitting method on the multivariate probit model. Application of pairwise fitting method and pseudo-likelihood

method on the case study and in the simulation study. Interpretation of the results.

Curriculum vitae

Margaux Delporte
Groot Begijnhof 91/202
3000 Leuven, Belgium

Phone: +32 16 32 16 78

Email: margauxdelporte@gmail.com
Website: `https://margauxdelporte.github.io/`

Born: August 18, 1996—Kortrijk, Belgium
Nationality: Belgian

Education

2017-2019 M.Sc. in Statistics, KU Leuven
2014-2017 B.sc. in Psychology, KU Leuven

Experience

2019-... Teaching Assistant – KU Leuven
Summer 2019 Data Science Intern – De Persgroep-Medialaan
2018-2019 Junior Statistician – Leuven Statistics Research Centre
Summer 2018 Data Science Intern – Vente-Exclusive via Exellys

Skills

Languages
Dutch (Native)
English (Fluent)
French (Intermediate)
Programming
R
SAS (SAS Certified Base Programmer for SAS 9)
SPSS
Python
SQL

Supplementary materials for Chapter 5

S.1 Literature review

S.1.1 Data sources and searches

Computerized bibliographic databases Web of science, Pubmed, and Scopus were used to identify studies. These databases were searched on October 18, 2023 and without limitations regarding to the year of publication. The search criteria were restricted to containing '*longitudinal*' in the title (or abstract in Web of Science) and '*matched*' and/or '*paired*' in the abstract.

S.1.2 Study selection

Publications were included in this systematic review if the following inclusion criteria were met: 1) the study is longitudinal and subjects have measurements on more than two time points. 2) the data is paired, which means that there is an obvious and meaningful one-to-one correspondence between subjects. This can also be the case when there is one-to-one (propensity score) matching. 3) the response is continuous. The yield of the database search were first screened based on title and abstract and next the full text of the selected articles was screened for relevance.

S.1.3 Data extraction

The studies included were grouped based on the various statistical methodologies employed. These categories were:

1. Difference scores
2. Comparison of subject-specific slopes
3. Paired t-tests or non-parametric alternative
4. Unpaired t-tests or non-parametric alternative

5. Linear mixed models
   - Without consideration for the paired nature of the data
   - With consideration for the paired nature of the data

6. New methodology

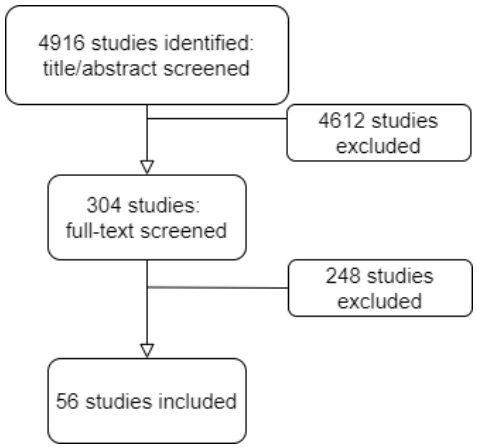

Figure 1: *Flowchart of the literature search.*

S.1.4 Results

The search strategy in the database identified 4916 potentially relevant studies (953 in Web of Science, 5 in Pubmed, and 4143 in Scopus). Based on the title and abstract, 304 articles appeared to meet the selection criteria, but after reading the full text of these studies, only 56 fulfilled the inclusion criteria (see flowchart in Fig. 1). The studies are grouped in different categories based on the used methods in Table 5.1. An article-by-article overview can be found in Table S.1.

Table S.1*: Overview of the studies included in the systematic review.*

Authors (year)	Journal	Category
Aguilar-Mediavilla et al. (2019)	Frontiers in Psychology	LMM: Ignore pairing
Ahmed et al. (2018)	Surgery (United States)	LMM: Random effect pair
Alfieri et al. (2022)	Genes	LMM: Ignore pairing
Ancoli-Israel et al. (2014)	Supportive Care in Cancer	LMM: Ignore pairing
Andreas et al. (2006)	Journal of Consulting and Clinical Psychology	LMM: Ignore pairing
Benestad et al. (2022)	Health and Quality of Life Outcomes	LMM: Marginal
Bergstrom et al. (2016)	Language Learning	LMM: Ignore pairing
Beutel et al. (1996)	Journal of Psychosomatic Research	MANOVA
Border et al. (2020)	Jacc: Cardiooncology	LMM: Random effect pair
Bouwmans et al. (2015)	Chronobiology International	LMM: Unclear structure
Büttner et al. (2021)	Physical Therapy in Sport	LMM: Ignore pairing
Ceroni et al. (2012)	BMC Musculoskeletal Disorders	Paired t-test
Dall et al. (2017)	BMC Public Health	LMM: Nested
Dayal et al. (2017)	Pediatria Polska	Unpaired t-test
De Paul and Domenech (2000)	Child Abuse and Neglect	LMM: Ignore pairing
Gardner and Boellaard (2007)	Family Relations	LMM: Ignore pairing
Gerber et al. (2016)	Journal Of The American Medical Association	LMM: Conditional
Goodman and Must (2011)	Journal Of Adolescent Health	Difference score
Gurucharri et al. (1984)	Journal of Consulting and Clinical Psychology	LMM: Ignore pairing
Hancock et al. (1993)	Paraplegia	LMM: Ignore pairing
Hands (2008)	Journal of Science and Medicine in Sport	LMM: Ignore pairing
Hull et al. (2020)	Journal of Applied Developmental Psychology	LMM: Ignore pairing
Iaffaldano et al. (2021)	Therapeutic Advances In Neurological Disorders	LMM: Ignore pairing
Isberg et al. (1993)	Scandinavian Journal of Plastic and Reconstructive Surgery and Hand Surgery	Paired t-test
Keresztes et al. (2003)	Heart & Lung	LMM: Ignore pairing
Kim (2006)	Communications in Statistics-Theory and Methods	New methodology
Kleinbub et al. (2015)	Frontiers in Psychology	LMM: Ignore pairing
Kretzschmar et al. (2016)	Osteoarthritis And Cartilage	LMM: Nested
Langer et al. (2003)	Psycho-Oncology	Paired t-test
Liu et al. (2022)	Neurology	LMM: Ignore pairing
LoCascio et al. (1998)	Calcified Tissue International	Difference score
Loher et al. (2014)	Applied Neuropsychology: Child	LMM: Ignore pairing
Luo et al. (2019)	Scientific Reports	LMM: Ignore pairing
Magnusson and Nauclér (1990)	Clinical Linguistics and Phonetics	Paired t-test
Metallinou et al. (2021)	Brain Sciences	Paired t-test
Mok et al. (2023)	Eye	Unpaired t-test
Moyle et al. (2007)	Journal of Speech Language and Hearing Research	LMM: Nested
Nickols-Richardson et al. (1999)	Journal of Bone and Mineral Research	LMM: Ignore pairing
Oertel et al. (2019)	Journal of Neuroinflammation	LMM: Nested
Oshima et al. (2020)	ERJ Open Research	LMM: Ignore pairing
Peetsma et al. (2001)	Educational Review	Paired t-test
Rebibo et al. (2013)	Surgical Endoscopy	Unpaired t-test
Roberts et al. (2018)	Obesity Surgery	Paired t-test
Ruhdorfer et al. (2015)	Osteoarthritis and Cartilage	Difference score
Schlee et al. (2021)	Journal of Clinical Medicine	Subject-specific slopes
Scholten-Peeters et al. (2020)	Journal of Headache and Pain	LMM: Ignore pairing
Shek and Dou (2020)	International Journal of Environmental Research and Public Health	LMM: Random effect pair
Shih et al. (2019)	Current Medical Research And Opinion	Unpaired t-test
Sibbel et al. (2016)	ASAIO Journal	LMM: Sandwich
Torgalsbøen et al. (2023)	Schizophrenia Research	LMM: Ignore pairing
Vasunilashorn et al. (2015)	Journals Of Gerontology Series A	Paired t-test
Weiler et al. (1997)	Early Human Development	LMM: Ignore pairing
Wilson (1979)	Acta Geneticae Medicae et Gemellologiae	New methodology
Yang et al. (2022)	Journal of Behavioral Medicine	LMM: Nested
Zhao et al. (2014)	Investigative Ophthalmology and Visual Science	LMM: Nested
Zorrilla-Vaca et al. (2022)	Canadian Journal Of Anesthesia	LMM: Ignore pairing

Supplementary Materials for Chapter 6

S.2 Analysis of the BIRNH data

Table S.2: *Parameter estimates, standard errors and p-values of the multivariate probit model for the BIRNH data.*

Parameter	Estimate	SE	t-value	p-value
Alcohol				
Threshold 1	-2.709	0.054	-50.10	<.001
Threshold 2	-0.230	0.028	-8.28	<.001
Threshold 3	0.158	0.028	5.70	<.001
SEX	-0.546	0.025	-22.11	<.001
REGIO	-0.095	0.024	-3.94	<.001
SOC1	0.225	0.024	9.20	<.001
Smoking				
Threshold 1	-2.286	0.102	-22.37	<.001
Threshold 2	-1.651	0.101	-16.33	<.001
SEX	-1.239	0.030	-40.68	<.001
BMI	-0.029	0.003	-9.51	<.001
AGE	-0.014	0.001	-12.32	<.001
SOC1	-0.157	0.032	-4.93	<.001
SOC2	-0.253	0.029	-8.70	<.001
Cholesterol				
Threshold 1	-2.422	0.099	-24.43	<.001
BMI	-0.035	0.003	-11.42	<.001
AGE	-0.028	0.001	-23.59	<.001
REGIO	0.127	0.026	4.85	<.001
SOC1	-0.078	0.031	-2.52	0.012
SOC2	-0.169	0.028	-6.13	<.001
Association Alcohol-Smoking				
Constant	0.030	0.018	1.63	0.104
SEX	-0.095	0.031	-3.03	0.002
SOC2	0.107	0.034	3.18	0.001
Assocation Alcohol-Cholesterol				
Constant	-0.059	0.019	-3.04	0.002
SEX	-0.005	0.035	-0.13	0.899
SOC2	0.069	0.037	1.86	0.063
Association Smoking-Cholesterol				
Constant	-0.063	0.021	-3.02	0.003
SEX	0.117	0.035	3.33	0.001
SOC2	-0.033	0.037	-0.90	0.370

S.3 Analysis of the POPS data

The analysis is performed via the pairwise fitting approach, as well as via the regular pseudo-likelihood method. While the computation time of the former is about eight times longer that the latter, negligible differences exists between the results of both methods when incomplete datasets, being complete pairs or available cases, are analysed. The results from the pairwise fitting approach are shown in Table S.4 and those of classical pseudo-likelihood in Table S.5. The reason is that the weights differ between the methods in the case of incomplete data. In the case of the pairwise likelihood, one is forced to divide in (6.3) between the number of 'independent clusters', here the number of subjects. One can argue to weight the contributions, according to the size of the cluster (number of measurements). Hence, when the data is incomplete, this can differ from the total sample size. A discussion about effects of clustering and cluster weights can be found on the comment section of Fitzmaurice et al. (1997), but we argue that the results will be asymptotically equal. Still, it is important to note that the analysis of available cases, as well as complete pairs, are only valid under the restrictive assumption of MCAR. In the analysis of the imputed datasets under the less restrictive assumptions of MAR, as well as in the complete case analysis, the methods yield the same estimates.

Table S.3*: Missing data patterns in the POPS dataset.*

ABIL1	ABIL2	ABIL3	NSZ	CGM	BIL	Freq	Percent
X	X	X	X	X	X	793	59.27
.	.	.	X	X	X	304	22.72
.	.	.	X	X	.	167	12.48
X	X	X	X	X	.	25	1.87
X	X	.	X	X	X	22	1.64
.	X	.	X	X	X	16	1.20
.	X	X	X	X	X	7	0.52
X	.	X	X	X	X	1	0.07
X	.	.	X	X	X	1	0.07
.	X	.	X	X	.	1	0.07
.	.	X	X	X	X	1	0.07

Table S.4: *Trivariate probit model for ability scores: results from pairwise fitting approach.*

	Complete cases	Complete pairs	Available cases	Imputation
First ability score				
Intercept	2.02 (0.26)	2.03 (0.26)	1.98 (0.26)	1.82 (0.23)
Neonat seiz	-1.14 (0.28)	-1.14 (0.28)	-1.11 (0.27)	-0.50 (0.21)
Cong malf	-0.61 (0.19)	-0.59 (0.19)	-0.59 (0.18)	-0.32 (0.17)
100 x bilirubin	-0.32 (0.14)	-0.33 (0.14)	-0.30 (0.14)	-0.24 (0.12)
Second ability score				
Intercept	2.20 (0.26)	2.19 (0.26)	2.14 (0.26)	2.00 (0.22)
Neonat seiz	-1.28 (0.28)	-1.25 (0.28)	-1.18 (0.26)	-0.53 (0.23)
Cong malf	-0.56 (0.19)	-0.55 (0.19)	-0.52 (0.18)	-0.32 (0.19)
100 x bilirubin	-0.42 (0.14)	-0.41 (0.14)	-0.40 (0.13)	-0.33 (0.12)
Third ability score				
Intercept	1.85 (0.28)	1.83 (0.28)	1.79 (0.27)	1.73 (0.24)
Neonat seiz	-0.94 (0.28)	-0.92 (0.28)	-0.90 (0.27)	-0.42 (0.21)
Cong malf	-0.48 (0.19)	-0.50 (0.19)	-0.48 (0.18)	-0.32 (0.17)
100 x bilirubin	-0.21 (0.15)	-0.20 (0.15)	-0.18 (0.15)	-0.17 (0.14)
Assocation parameters				
(1,2): ρ	0.73 (0.05)	0.72 (0.05)	0.73 (0.05)	0.56 (0.06)
(1,3): ρ	0.82 (0.04)	0.82 (0.04)	0.82 (0.04)	0.61 (0.06)
(2,3): ρ	0.73 (0.04)	0.74 (0.05)	0.75 (0.05)	0.56 (0.06)

Table S.5: *Trivariate probit model for ability scores: results from pseudo-likelihood estimation.*

	Complete cases	Complete pairs	Available cases	Imputation
First ability score				
Intercept	2.02 (0.26)	2.01 (0.26)	1.97 (0.26)	1.82 (0.23)
Neonat seiz	-1.14 (0.28)	-1.12 (0.28)	-1.13 (0.27)	-0.50 (0.21)
Cong malf	-0.61 (0.19)	-0.59 (0.18)	-0.58 (0.18)	-0.32 (0.17)
100 x bilirubin	-0.32 (0.14)	-0.32 (0.14)	-0.30 (0.14)	-0.24 (0.12)
Second ability score				
Intercept	2.20 (0.26)	2.19 (0.26)	2.10 (0.25)	2.00 (0.22)
Neonat seiz	-1.28 (0.28)	-1.25 (0.27)	-1.19 (0.27)	-0.54 (0.23)
Cong malf	-0.56 (0.19)	-0.55 (0.19)	-0.51 (0.18)	-0.32 (0.19)
100 x bilirubin	-0.41 (0.14)	-0.41 (0.13)	-0.38 (0.13)	-0.33 (0.12)
Third ability score				
Intercept	1.85 (0.28)	1.83 (0.28)	1.77 (0.27)	1.73 (0.24)
Neonat seiz	-0.94 (0.28)	-0.92 (0.28)	-0.90 (0.27)	-0.42 (0.21)
Cong malf	-0.48 (0.19)	-0.48 (0.19)	-0.48 (0.19)	-0.32 (0.17)
100 x bilirubin	-0.21 (0.15)	-0.20 (0.15)	-0.17 (0.14)	-0.17 (0.14)
Assocation parameters				
(1,2): ρ	0.73 (0.05)	0.72 (0.05)	0.73 (0.05)	0.56 (0.06)
(1,3): ρ	0.82 (0.04)	0.82 (0.04)	0.82 (0.04)	0.61 (0.06)
(2,3): ρ	0.73 (0.05)	0.74 (0.05)	0.74 (0.05)	0.56 (0.06)