Bayesian Imputation of Missing Covariates

Nicole S. Erler

Bayesian Imputation of Missing Covariates

Nicole S. Erler

ISBN: 978-94-6323-623-2
Cover design: Marijke van Zuilen (https://marijkeengerjannevanzuilen.nl)
Printed by: Gildeprint

The work presented in this thesis was conducted at the Department of Biostatistics and the Department of Epidemiology, Erasmus Medical Center, Rotterdam, the Netherlands.

Bayesian Imputation of Missing Covariates

Bayesiaanse imputatie van ontbrekende covariabelen

Thesis

to obtain the degree of Doctor from the
Erasmus University Rotterdam
by command of the
rector magnificus

Prof.dr. R.C.M.E. Engels

and in accordance with the decision of the Doctorate Board.

The public defence shall be held on

Wednesday, 12 June 2019 at 15:30 hrs

by

Nicole Stephanie Erler

born in Weilheim i. OB, Germany

Erasmus University Rotterdam

Doctoral committee

Promotors:	Prof.dr. D. Rizopoulos
	Prof.dr. E.M.E.H. Lesaffre
Other members:	Prof.dr. S. van Buuren
	Prof.dr. E. Boersma
	Prof.dr. A. Burdorf

Contents

1 General Introduction

It can be assumed that missing values exist since we started to collect data. The need to find ways to deal with them emerged as the amount of data that was collected grew, but the willingness to provide information to the parties collecting it (often the government for the purpose of the census) abated. The inefficiency due to the loss of information resulting from only using the complete cases for analysis was no longer considered acceptable (Scheuren 2005).

The methods that were used to handle missing values in the 50s and 60s replaced the missing values by a single value, such as "hot-deck" imputation, where a missing value is replaced by an observed value of a case that is similar to the case with the missing value with regards to other, observed characteristics (Andridge and Little 2010; Behrmann 1954; Nordbotten 1963; Gorinson 1969). The idea of replacing a missing datum with multiple values (and arguably the beginning of the development of more sophisticated missing data methods) traces back to Donald B. Rubin, who first described his idea in 1977 (Rubin 2004). He states that

> "[...] of course (1) imputing *one* value for missing datum can't be correct in general, and (2) in order to insert sensible values for a missing datum we must rely more or less on some model relating unobserved values to observed values."

The Bayesian framework naturally lends itself to missing data settings, treating missing values as unobserved random variables that have a distribution which depends on the observed data (Rubin 1978a; Rubin 1978b). Nevertheless, initially it was not used in these settings since calculations can quickly become very involved when parts of the data are unobserved, and the computational procedures nowadays used to overcome these difficulties were not yet available. In this thesis, we focus on inference on missing data under the Bayesian paradigm but compare it to other, commonly used approaches.

1.1 Mechanisms and Patterns of Missing Data

1.1.1 Missing Data Mechanism

In the context of missing data, strictly speaking, the term "data" does not only refer to the values of those variables that were intended to be measured, but also includes the missing data indicator, a binary variable, that describes if a value was observed or not.

Using this missingness indicator, a model for the missing data mechanism can be postulated. This model describes how the probability of a value being missing is

related to other characteristics of the same unit. It can be written as

$$p(\mathbf{R} \mid \mathbf{X}_{obs}, \mathbf{X}_{mis}, \boldsymbol{\psi}),$$

where $\mathbf{R}$ is the missing data indicator matrix and $\mathbf{X}_{obs}$ and $\mathbf{X}_{mis}$ denote the part of the data that is completely observed and the part of the data that has missing values for some subjects, respectively.

In general, the model for the missing data mechanism must be taken into account when analysing incomplete data, however, under certain conditions, it may be ignored. Conditions for *ignorability* were introduced by Rubin (1976) and are closely connected to the well-established classification of missing data mechanisms described below (Little and Rubin 2002).

Missing Completely at Random

The data is said to be *missing completely at random* (MCAR) when the probability of a value being missing does not depend on any data, that is, when $p(\mathbf{R} \mid \mathbf{X}_{obs}, \mathbf{X}_{mis}, \boldsymbol{\psi}) = p(\mathbf{R} \mid \boldsymbol{\psi})$. This type of missingness may occur when samples are lost, participants miss a visit or do not fill in a question in a questionnaire, because of reasons that are unrelated with the study for which the data was supposed to be collected.

Missing at Random

In settings in which the probability of a value being missing does depend on factors associated with the study, but those factors have been observed, i.e., are part of $\mathbf{X}_{obs}$, the missing data mechanism is called *missing at random* (MAR) and can be formally described as $p(\mathbf{R} \mid \mathbf{X}_{obs}, \mathbf{X}_{mis}, \boldsymbol{\psi}) = p(\mathbf{R} \mid \mathbf{X}_{obs}, \boldsymbol{\psi})$. Missing values of this type may occur for instance when in a survey about the consumption of sweets overweight participants are less likely to respond to the question how much chocolate they eat, and the weight of all participants is known. Another example is a longitudinal study where subjects are excluded from future measurements once a critical value is exceeded. MAR includes MCAR as a special case.

Missing Not at Random

When the assumption of MAR does not hold, and the probability of a value being missing does depend on unobserved data, which may either be the missing value itself, unobserved values of other variables, or variables that have not been recorded at all, the missing data mechanism is called *missing not at random* (MNAR) and can be written as $p(\mathbf{R} \mid \mathbf{X}_{obs}, \mathbf{X}_{mis}, \boldsymbol{\psi}) \neq p(\mathbf{R} \mid \mathbf{X}_{obs}, \boldsymbol{\psi})$. Examples for data

missing not at random would be if, in the above survey, participants who are overweight are more likely not to report their weight, or, if in the longitudinal study the value that exceeded the threshold would not be recorded. Since $\mathbf{X}_{mis}$ is not observed, it is not possible to test whether the assumption of MAR holds. Good knowledge of how the data was obtained and why values are missing is necessary to make appropriate assumptions about the missing data mechanism.

Ignorability

In settings where the missing data mechanism is MAR and the parameters of the missingness model, $\boldsymbol{\psi}$, are a priori independent / distinct of the parameters of the data model, $\boldsymbol{\theta}$, i.e., $p(\boldsymbol{\psi}, \boldsymbol{\theta}) = p(\boldsymbol{\psi})\, p(\boldsymbol{\theta})$, the missingness process does not need to be explicitly modelled when Bayesian or likelihood inference for $\boldsymbol{\theta}$ is performed. Then, the missingness is called *ignorable*. Throughout this thesis, we assume that this is the case.

1.1.2 Missing Data Pattern

When values are missing in multiple variables, different patterns of missingness can arise. If variables can be ordered such that if a value is missing, the values of all variables following in the sequence are also always missing, the missing data pattern is called *monotone* (left panel of Figure 1.1). If such an order does not exist, the missing data pattern is *non-monotone* (right panel of Figure 1.1).

Monotone missing data patterns typically arise in longitudinal studies with dropout, where once a patient has left the study, no further measurements are available. In studies where multiple variables measuring different aspects are obtained, either at the same time point or over time, non-monotone missing data patterns occur more frequently, since study participants do not return a particular questionnaire or miss a single visit to the research centre. In this thesis, we consider general, non-monotone missing data patterns.

1.2 Multiple Imputed Datasets

When the concept of multiple imputation was developed in the 1970s, a requirement for a practical way to deal with missing data was that it allowed many researchers to analyse the incomplete data. Moreover, it was essential that these analyses could be done using only standard techniques and software tools, which required complete, balanced data, and without the need of in-depth knowledge of missing data methods (Scheuren 2005). Especially the second part of this requirement is still relevant today since analyses are often performed by researchers

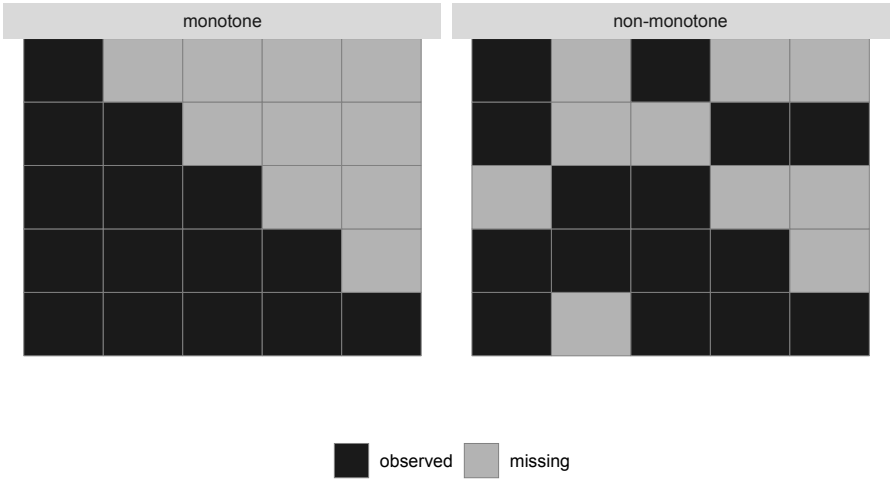


Figure 1.1: Visualization of a monotone and a non-monotone missing data pattern. Each column represents a variable, while rows represent different patterns of missing values.

without expert knowledge in statistics, who are usually only familiar with standard complete data methods and are not versed in Bayesian methodology.

Rubin's solution to this practicality issue was to perform imputation once and to distribute the imputed data to various researchers. In this imputation procedure, multiply imputed, i.e., completed, versions of the original data were produced, which differed only in the values that had been filled in for the missing observations. This allowed retaining some information on the uncertainty about the missing values, while each of the datasets could be analysed using standard methods. Since the imputed datasets are not identical, the estimates obtained from the analysis of each dataset will vary. This variation in the obtained estimates allows assessing the additional uncertainty in the effect estimates that is caused by the missing values (Rubin 2004).

In the following, we first give an overview of some popular (Bayesian and non-Bayesian) methods that have been proposed for creating imputed values and then describe the most commonly used procedure to pool the results from the analyses of multiple imputed datasets.

1.2.1 Methods for Imputation

The task to create the imputed values requires to sample from the (posterior) predictive distribution of the unobserved data, given the observed data. Especially in larger datasets, with missing values in multiple variables, possibly of different measurement levels, and a non-monotone missing data pattern, this distribution is multivariate and not of any standard form.

Joint Model Imputation

Rubin's suggestion in his initial paper on multiple imputation from 1977 (Rubin 2004) was to approximate the posterior predictive distribution with a multivariate normal distribution if variables are continuous, or with a multinomial distribution if data is categorical. Since sampling from either of these distributions is fast and readily available in software, this approach, especially using the multivariate normal distribution, is still used today (Carpenter and Kenward 2013). To allow for covariates of mixed type, the assumption is made that categorical variables have an underlying normal distribution and that different categories are observed depending on the value of that latent distribution. Due to its ability to impute incomplete baseline as well as time-varying covariates in multi-levels settings, we apply and investigate this approach in Chapter 5 of this thesis.

Expectation Maximization

A general approach that allows performing likelihood inference when parts of the data are unobserved, is the *Expectation Maximization* (EM) algorithm, introduced by Dempster et al. (1977). The algorithm alternates between the expectation (E) step, in which the expected value of missing data, conditional on observed data and current estimates of the parameters, is obtained, and the maximisation (M) step, in which parameter values are estimated by maximizing the likelihood of the parameters given the current values of the missing data. Even though the approach was not specifically developed to create multiple imputations, it could be applied in this way, if, once the algorithm has converged, not the expectation of missing values is determined, but instead multiple values are drawn from the estimated distribution. In settings where incomplete variables have non-linear associations with other variables, however, the distribution in the M-step may not have a closed form and updating the parameters becomes difficult.

Data Augmentation

Data augmentation, an approach similar to the EM algorithm, was proposed by Tanner and Wong (1987). It can be thought of as a Bayesian version of the EM

algorithm since it has the same two-step structure, but the E step is replaced by imputation of missing values and the M step by estimation of the posterior distribution instead of maximization of the likelihood. The motivation behind augmenting the data is that sampling from $p(\boldsymbol{\theta} \mid \mathbf{X})$ is often easier than sampling from $p(\boldsymbol{\theta} \mid \mathbf{X}_{obs})$ and works well if sampling of $p(\mathbf{X}_{mis} \mid \mathbf{X}_{obs}, \boldsymbol{\theta})$ is possible. Nonetheless, even when Monte Carlo integration is used, sampling from distributions that are not members of the exponential family may be difficult, diminishing the attractiveness of the approach. While Tanner and Wong (1987) aimed to estimate the posterior distribution, Li (1988) focused on imputing values but used essentially the same algorithm.

The approach for analysis of incomplete data and imputation of missing values followed throughout this thesis uses data augmentation in conjunction with the Gibbs sampler. The joint distribution of all data, observed and unobserved, as well as the parameters, is specified in a sequence of univariate distributions. Using Gibbs sampling, missing values are imputed by draws from the full conditional distributions arising from this joint distribution. Once the observed data is augmented by the imputed values, posterior inference for the parameters of interest can be obtained as if the data had been complete. The approach is described in detail in Chapter 2.

Multiple Imputation using Chained Equations

The nowadays most popular approach for creating multiple imputations, introduced by Van Buuren, Boshuizen, et al. (1999), also uses the idea of the Gibbs sampler but is a (mostly) frequentist approach. In *multiple imputation using chained equations* (MICE), also called multiple imputation using a *fully conditional specification* (FCS), a full conditional distribution is specified for each incomplete variable and imputed values are sampled from these distributions. If a multivariate distribution exists that has the specified distributions as its full conditionals, the algorithm is a Gibbs sampler.

Specifying the full conditional distributions directly has the advantage that it allows for a flexible algorithm, in which distributions can be tailored to the measurement level of each variable, and sampling is performed on a variable by variable basis using samplers that are easy to implement. The MICE algorithm (Van Buuren 2012) starts by randomly drawing starting values from the set of observed values. Then, in each iteration $t = 1, \ldots, T$, it cycles once through all incomplete variables $\mathbf{x}_j,\ j = 1, \ldots, p$. For each incomplete variable j, the currently completed data except x_j is defined as $\dot{\mathbf{X}}^t_{-j} = \left(\dot{\mathbf{x}}^t_1, \ldots, \dot{\mathbf{x}}^t_{j-1}, \dot{\mathbf{x}}^{t-1}_{j+1}, \ldots, \dot{\mathbf{x}}^{t-1}_p\right)$. The parameters of the model for $\mathbf{x}_j$, $\dot{\boldsymbol{\theta}}^t_j$, are sampled from their distribution conditional on

the observed part of $\mathbf{x}_j$ and the currently completed data of the other variables from subjects that have $\mathbf{x}_j$ observed:

$$p(\boldsymbol{\theta}_j^t \mid \mathbf{x}_j^{obs}, \dot{\mathbf{X}}_{-j}, \mathbf{R}).$$

Imputed values $\dot{\mathbf{x}}_j^t$ are drawn from the predictive distribution of the missing values $\mathbf{x}_j^{mis}$ given the other variables and parameters $\dot{\boldsymbol{\theta}}_j^t$,

$$p(\mathbf{x}_j^{mis} \mid \mathbf{X}_{-j}^t, \mathbf{R}, \boldsymbol{\theta}_j^t).$$

By filling in the imputed values of the last iteration, i.e., $\left(\dot{\mathbf{x}}_1^\top, \ldots, \dot{\mathbf{x}}_p^\top\right)$ into the original, incomplete, data, one imputed dataset is created. The algorithm is run multiple times with different starting values to create a set of multiply imputed datasets.

A drawback of the MICE algorithm is that there is no guarantee that a joint distribution exists that has the specified conditional distributions as its full conditionals. If no such distribution exists, the algorithm may not converge to the correct distribution. Despite this theoretical limitation, it has been shown to work well in practice as long as the conditional distributions fit the data well enough (Zhu and Raghunathan 2015). In settings where incomplete covariates are involved in non-linear functional forms or interactions, or with complex outcomes, such as survival or longitudinal outcomes, specification of correct imputation models is often not feasible (Bartlett et al. 2015; Carpenter and Kenward 2013). Even specification of models that adequately include all information necessary to obtain valid imputations is not straightforward, and, in practice, often not even attempted when imputation is performed by researchers who are unaware that naive use of imputation software will lead to violation of important assumptions and thereby faulty imputations and biased inference. The performance of MICE when used naively for imputation of covariates in longitudinal data is the topic of Chapter 2 of this thesis.

1.2.2 Pooling Results from Multiple Imputed Datasets

Irrespective of the method used for producing imputations, the results from the analyses of the multiple imputed datasets need to be combined in a manner that takes into account the added uncertainty due to the missing values. The formulas for pooling of such results proposed by Rubin and Schenker (1986) (see also Rubin (1987)), usually referred to as *Rubin's Rules*, have gained wide acceptance and are outlined in the following.

For a parameter vector $\mathbf{Q}$ the overall estimate, pooled over the analyses of m imputed datasets, can be calculated as the mean over the estimates from these

analyses,

$$\overline{\mathbf{Q}} = \frac{1}{m}\sum_{\ell=1}^{m}\widehat{\mathbf{Q}}_\ell,$$

where $\widehat{\mathbf{Q}}_\ell$ denotes the estimate obtained from the ℓ-th imputed set. The overall variance of $\mathbf{Q}$ consists of the within imputation variances $\overline{\mathbf{W}}$, which can be calculated by averaging over the estimated variances of the $\mathbf{Q}_\ell$ from each imputed dataset, $\widehat{\mathbf{W}}_\ell$,

$$\overline{\mathbf{W}} = \frac{1}{m}\sum_{\ell=1}^{m}\widehat{\mathbf{W}}_\ell,$$

and the between imputation variance, $\mathbf{B}$, calculated as

$$\mathbf{B} = \frac{1}{m-1}\sum_{\ell=1}^{m}\left(\widehat{\mathbf{Q}}_\ell - \overline{\mathbf{Q}}\right)\left(\widehat{\mathbf{Q}}_\ell - \overline{\mathbf{Q}}\right)^\top.$$

Following Rubin and Schenker (1986), the total variance, $\mathbf{T}$ is the sum of within and between imputation variance, plus an additional term $\mathbf{B}/m$, correcting for the finite number of imputations, i.e., $\mathbf{T} = \overline{\mathbf{W}}+\mathbf{B}+\mathbf{B}/m$. The relative increase in variance that is due to the missing values can be estimated as $r_m = (\mathbf{B}+\mathbf{B}/m)/\overline{\mathbf{W}}$.

The $(1-\alpha)$ 100% confidence interval for scalar Q can be calculated as $\overline{Q} \pm t_\nu(\alpha/2)\sqrt{T}$, where t_ν is the $\alpha/2$ quantile of the t-distribution with $\nu = (m-1)\left(1+r_m^{-1}\right)^2$ degrees of freedom.

The corresponding p-value is the probability $Pr\{F_{1,\nu} > \left(Q_0-\overline{Q}\right)^2/T\}$, where $F_{1,\nu}$ is a random variable that has an F distribution with 1 and ν degrees of freedom, and Q_0 is the null hypothesis value (typically zero).

In the above calculation, the degrees of freedom ν are derived under the assumption that there are infinite degrees of freedom in the complete data, denoted ν_{com} (Barnard and Rubin 1999). Since this is not a reasonable assumption for small datasets, Barnard and Rubin (1999) proposed a different calculation of the degrees of freedom for the t-distribution

$$\tilde{\nu} = \left(\frac{1}{\nu}+\frac{1}{\hat{\nu}_{obs}}\right)^{-1} = \nu_m\left(1+\frac{\nu}{\hat{\nu}_{obs}}\right)^{-1} = \nu_{com}\left[\{\lambda(\nu_{com})(1-\hat{\gamma}_m)\}^{-1}+\frac{\nu_{com}}{\nu}\right]$$

where the observed-data degrees of freedom, ν_{obs}, are estimated as $\hat{\nu}_{obs} = \lambda(\nu_{com})\nu_{com}(1-\hat{\gamma}_m)$, $\hat{\gamma}_m = r_m/(1+r_m)$ and $\lambda(\nu) = (\nu+1)(\nu+3)$. This small-sample version of Rubin's Rules is implemented in the R package **mice** and used in this thesis.

1.3 The Bayesian Framework

Since the focus of this thesis is on inference for missing data under the Bayesian paradigm, in this section we will briefly introduce the Bayesian framework and some relevant concepts.

The idea behind the Bayesian paradigm is that inference about an unknown parameter can be obtained by updating an initial guess or prior belief about that parameter with data (Bayes et al. 1763; Laplace 1774).

1.3.1 Bayes Theorem

The *posterior distribution*, i.e., the distribution of a parameter $\boldsymbol{\theta}$ conditional on the data $\mathbf{X}$, can be expressed as

$$p(\boldsymbol{\theta} \mid X) = \frac{p(\mathbf{X} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})}{\int_{-\infty}^{\infty} p(\mathbf{X} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta}}.$$

In the above equations, $p(\boldsymbol{\theta})$ denotes the *prior distribution* of $\boldsymbol{\theta}$, i.e., the distribution of $\boldsymbol{\theta}$ that is assumed without taking into account the collected data, $p(\mathbf{X} \mid \boldsymbol{\theta})$ is the *likelihood* of the data given the parameter, and the denominator constitutes the marginal distribution of the data, i.e., the distribution of data under all possible values of $\boldsymbol{\theta}$, and is often called the *normalizing constant*. Since this normalizing constant does not depend on $\boldsymbol{\theta}$, Bayes theorem is often simplified to

$$p(\boldsymbol{\theta} \mid \mathbf{X}) \propto p(\mathbf{X} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta}),$$

i.e., the posterior distribution is proportional to the product of the likelihood and the prior distribution.

In the Bayesian framework, the distribution relating unobserved values, $\mathbf{X}_{mis}$, to observed values, $\mathbf{X}_{obs}$, referred to by Rubin in his statement given above, is called the *posterior predictive distribution* of the missing data given the observed data,

$$p(\mathbf{X}_{mis} \mid \mathbf{X}_{obs}) = \int p(\mathbf{X}_{mis} \mid \mathbf{X}_{obs}, \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \mathbf{X}_{obs})\, d\boldsymbol{\theta}.$$

In practice, an analytic calculation of the posterior distribution is often not feasible. In that case, it has to be approximated or determined numerically. A numeric method that is frequently used and works in complex settings is the *Monte Carlo* method.

1.3.2 Monte Carlo Methods

Instead of determining the posterior distribution analytically, the Monte Carlo method (Metropolis and Ulam 1949) draws random samples from it and uses these samples to calculate summary measures of the distribution, to approximate the corresponding measures of the posterior distribution.

Using the central limit theorem, the precision of this approximation to the sample mean, $\bar{\theta}$, can be determined as

$$\bar{\theta'} \pm 1.96 \, \mathrm{sd}(\theta')/\sqrt{K},$$

where K is the number of independently sampled values θ', and $\mathrm{sd}(\theta')/\sqrt{K}$ is called the *Monte Carlo error*.

In high-dimensional settings, however, even "just" sampling from the posterior distribution may not be feasible, since no sampler is available for direct sampling from a multivariate distribution of unknown form, and factorization of the distribution would require solving multiple integrals (Lesaffre and Lawson 2012). The development of *Markov Chain Monte Carlo* methods (Metropolis, Rosenbluth, et al. 1953; Hastings 1970) was crucial in overcoming this difficulty.

1.3.3 Markov Chain Monte Carlo

The idea behind Markov Chain Monte Carlo (MCMC) sampling is to construct a Markov chain that has the distribution to be sampled from as its stationary distribution. It is an iterative procedure in which a sequence of random variables is created by repeatedly drawing values from a distribution that depends on the previously drawn value. MCMC methods, hence, perform *dependent sampling*. By creating a Markov chain that has the posterior distribution as its stationary distribution, samples from that stationary distribution can, thus, be regarded as a sample from the posterior distribution.

1.3.4 Gibbs Sampler

The Gibbs sampler, introduced by Geman and Geman (1984), facilitates sampling from high-dimensional distributions by splitting a multivariate problem into a set of univariate problems. It uses the property that a joint distribution is fully specified by its corresponding set of full conditional distributions. Iteratively sampling from these, often univariate, distributions is usually relatively easy. Using Gibbs sampling to obtain draws in an MCMC chain, hence, allows sampling from high-dimensional posterior distributions.

1.3.5 Convergence, Mixing and Thinning

As previously mentioned, samples from an MCMC chain are only samples from the posterior distribution once the chain has *converged*, i.e., when the distribution of the values remains stable throughout further iterations. In order to obtain valid inferences, convergence of the chains must be checked, and, if necessary, samples from before the chain has converged need to be discarded. The iterations before the chain has converged are called *burn-in* period.

Convergence may be checked visually, by plotting the drawn values against the iteration number in a so-called *traceplot*, which should show a horizontal band with no apparent trends or patterns. In this thesis, we additionally evaluated convergence using a statistical criterion developed by Gelman and Rubin (1992). It uses multiple, independent chains for the same parameter, and compares within and between chain variances. When this criterion, which we refer to as the *Gelman-Rubin criterion* in the rest of this thesis, is close enough to one, say, no more than 1.2, the MCMC chains can be assumed to have converged.

Another potential issue when working with MCMC methods is that they perform dependent sampling. When this dependence is strong it can take many iterations before the MCMC chain converges and, moreover, many iterations until the chain has sufficiently explored the whole range of the posterior distribution. In order to provide enough information about the posterior distribution, a chain with high auto-correlation may have to be continued for more iterations than can practically be handled. To reduce the number of samples that have to be stored, a chain may be thinned out so that only a reduced number of samples is saved.

1.4 Motivating Datasets

The research presented in this thesis is motivated by several large cohort studies. Such studies are especially prone to missing values since typically a large number of variables are measured, and many variables are self-reported (i.e., by questionnaire) or participants have to visit a research centre for measurements to be taken. Moreover, participants are from the general population and may not always see a direct personal benefit in complying with the study protocol.

Two datasets that were used in this thesis for demonstration of the statistical methods under investigation, as well as the real world application of the proposed approach, are briefly introduced in the following sections.

The Generation R Study

The Generation R Study (Kooijman et al. 2016) is an ongoing longitudinal population-based prospective cohort study from fetal life until young adulthood, investigating growth development and health of children, conducted in Rotterdam, the Netherlands. Approximately 10000 pregnant women from the Rotterdam area with an expected delivery date between 2002 and 2006 were included and are followed up, together with their children, until the offspring is 18 years of age. Data is collected at scheduled visits to the research centre as well as by questionnaire, phone interviews or home visits, and augmented by registry information. Among other things, information on maternal diet and lifestyle, health and complications during pregnancy, child growth and health outcomes (e.g., asthma or infectious diseases), behaviour and cognition, body composition and obesity, and eye and tooth development, is collected at different time points.

The National Health and Nutrition Examination Survey

The National Health and Nutrition Examination Survey (NHANES), conducted by the National Center for Health Statistics in the United States, is a large cohort of children and adults and investigates a broad range of health and nutrition related issues (National Center for Health Statistics (NCHS) 2011). Since 1999, a new cohort of approximately 5000 participants, chosen representatively of the US population, is included every year. Data on demographic, socioeconomic, dietary and health-related topics is obtained by interviews and physiological, dental and medical examinations, and laboratory tests are performed. The study is designed to, among other things, investigate risk factors for and prevalence of diseases, get insights into how nutrition (advice) may be used for disease prevention and to inform public health policies.

1.5 Outline of this Thesis

This section provides a brief overview of the content of the subsequent chapters.

In Chapter 2 a fully Bayesian approach to analysis and imputation of data with incomplete covariate information is described in detail for the setting with a continuous longitudinal outcome and incomplete baseline covariates.

Moreover, different approaches to the handling of a longitudinal outcome in multiple imputation using MICE are investigated, both the naive but commonly used, and other, more sophisticated techniques. MICE and the fully Bayesian approach are compared using a simulation study and a real data example from the Generation R Study, in which the association between maternal diet during pregnancy

and gestational weight throughout pregnancy is analysed. This application is motivated by the study conducted in Chapter 3.

Chapters 3 and 4 contain applications of the presented Bayesian approach using data from the Generation R Study. In Chapter 3 the association between guideline-based (*a priori*) and data-driven (*a posteriori*) dietary patterns with gestational weight gain and trajectories of gestational weight are investigated. Using the approach presented in Chapter 2, simultaneous analysis of the trajectories of gestational weight and imputation of incomplete baseline covariates is performed. The imputed data is then used in the secondary analysis to investigate the association of diet with weight gain.

Chapter 4 examines the effect of sugar containing beverage consumption by pregnant women on body composition of their offspring. Measures of body composition are the body mass index (BMI), which was measured repeatedly, and the fat mass index and fat-free mass index, both measured when children where approximately six years of age. All three outcomes are modelled jointly, and imputation of incomplete baseline covariates is performed using the Bayesian approach presented in Chapter 2.

In Chapter 5, the approach of Chapter 2 is extended to time-varying covariates. Additional issues that arise with such covariates, specifically the potential endogeneity and non-linear shape of the association with the outcome are considered. Advantages and disadvantages as well as the performance of the proposed approach are compared to multiple imputation using a multivariate normal model in a simulation study and two research questions from the Generation R Study: the association between blood pressure and weight of mothers during pregnancy, and the association of maternal gestational weight and child BMI from birth until five years of age.

The implementation of our fully Bayesian approach to jointly analyse and impute incomplete data into the R package **JointAI** is described in Chapter 6. Functionality of the package to analyse incomplete data using generalized linear mixed models or generalized linear regression models, which may include non-linear forms or interaction terms, is demonstrated in detail. Data from the NHANES study as well as data simulated to mimic data from a longitudinal cohort study, such as the Generation R Study, is used to illustrate this functionality.

The thesis concludes in Chapter 7 with a general discussion of limitations of the proposed approach which arise from the model assumptions, and propositions as to how the approach and its implementation in software should be further improved and extended to facilitate valid inference with incomplete data in a wider range of applications and settings.

References

Andridge, R. R. and R. J. Little (2010). "A review of hot deck imputation for survey non-response". *International Statistical Review*, **78**(1):40–64. DOI: 10.1111/j.1751-5823.2010.00103.x.

Barnard, J. and D. B. Rubin (1999). "Miscellanea. Small-sample degrees of freedom with multiple imputation". *Biometrika*, **86**(4):948–955. DOI: 10.1093/biomet/86.4.948.

Bartlett, J. W. et al. (2015). "Multiple imputation of covariates by fully conditional specification: accommodating the substantive model". *Statistical Methods in Medical Research*, **24**(4):462–487. DOI: 10.1177/0962280214521348.

Bayes, T., R. Price, and J. Canton (1763). "An Essay towards solving a Problem in the Doctrine of Chances". *Philosophical Transactions*, **53**:370–418. DOI: 10.1098/rstl.1763.0053.

Behrmann, H. I. (1954). "Sampling Technique in an Economic Survey of Sugar Cane Production". *South African Journal of Economics*, **22**(3):326–336. ISSN: 1813-6982. DOI: 10.1111/j.1813-6982.1954.tb01646.x.

Carpenter, J. R. and M. G. Kenward (2013). *Multiple Imputation and its Application*. John Wiley & Sons, Ltd. DOI: 10.1002/9781119942283.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). "Maximum likelihood from incomplete data via the EM algorithm". *Journal of the Royal Statistical Society. Series B (methodological)*, (1):1–38. URL: http://www.jstor.org/stable/2984875.

Gelman, A. and D. B. Rubin (1992). "Inference from Iterative Simulation Using Multiple Sequences". *Statistical Science*, **7**(4):457–472. DOI: 10.1214/ss/1177011136.

Geman, S. and D. Geman (1984). "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6):721–741. DOI: 10.1109/TPAMI.1984.4767596.

Gorinson, M. (1969). "How the census data will be processed". *Monthly Labor Review (pre-1986); Washington*, **92**(12):42. URL: http://www.jstor.org/stable/41837533.

Hastings, W. K. (1970). "Monte Carlo Sampling Methods Using Markov Chains and Their Applications". *Biometrika*, **57**(1):97–109. DOI: doi:10.2307/2334940.

Kooijman, M. N. et al. (2016). "The Generation R Study: design and cohort update 2017". *European Journal of Epidemiology*, **31**(12):1243–1264. DOI: 10.1007/s10654-016-0224-9.

Laplace, P.-S. (1774). "Mémoire sur la probabilité des causes par les évènemens". *Mémoires de Mathématque et Physique, Presentés à l'Académie Royale des*

Sciences, par divers Savans & lûs dans ses Assemblées, Tome Sixiéme, **66**:621–56.

Lesaffre, E. M. and A. B. Lawson (2012). *Bayesian Biostatistics*. John Wiley & Sons. DOI: 10.1002/9781119942412.

Li, K.-H. (1988). "Imputation using Markov chains". *Journal of Statistical Computation and Simulation*, **30**(1):57–79.

Little, R. J. and D. B. Rubin (2002). *Statistical Analysis with Missing Data*. Hoboken, New Jersey: John Wiley & Sons, Inc. DOI: 10.1002/9781119013563.

Metropolis, N., A. W. Rosenbluth, et al. (1953). "Equation of State Calculations by Fast Computing Machines". *The Journal of Chemical Physics*, **21**(6):1087–1092. DOI: 10.1063/1.1699114.

Metropolis, N. and S. Ulam (1949). "The Monte Carlo Method". *Journal of the American Statistical Association*, **44**(247):335–341. DOI: 10.2307/2280232.

National Center for Health Statistics (NCHS) (2011). *National Health and Nutrition Examination Survey Data*. URL: https://www.cdc.gov/nchs/nhanes/.

Nordbotten, S. (1963). "Automatic editing of individual statistical observations". Statistical Standards and Studies (3). United Nations Statistical Commission and Economic Commission for Europe. Conference of European Statisticians. URL: http://hdl.handle.net/11250/178265.

Rubin, D. B. (1976). "Inference and Missing Data". *Biometrika*, **63**(3):581–592. DOI: 10.2307/2335739.

— (1978a). "Bayesian Inference for Causal Effects: The Role of Randomization". *The Annals of Statistics*, **6**(1):34–58. DOI: 10.1214/aos/1176344064.

— (1978b). "Multiple Imputations in Sample Surveys - A Phenomenological Bayesian Approach to Nonresponse". *Proceedings of the Survey Research Methods Section of the American Statistical Association*. Vol. 1. American Statistical Association, p. 20–34.

— (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley Series in Probability and Statistics. Wiley. DOI: 10.1002/9780470316696.

— (2004). "The Design of a General and Flexible System for Handling Nonresponse in Sample Surveys". *The American Statistician*, **58**(4):298–302. DOI: 10.1198/000313004X6355.

Rubin, D. B. and N. Schenker (1986). "Multiple Imputation for Interval Estimation From Simple Random Samples With Ignorable Nonresponse". *Journal of the American Statistical Association*, **81**(394):366–374. DOI: 10.2307/2289225.

Scheuren, F. (2005). "Multiple Imputation: How It Began and Continues". *The American Statistician*, **59**(4):315–319. DOI: 10.1198/000313005X74016.

Tanner, M. A. and W. H. Wong (1987). "The Calculation of Posterior Distributions by Data Augmentation". *Journal of the American Statistical Association*, **82**(398):528–540. DOI: 10.2307/2289457.

Van Buuren, S. (2012). *Flexible Imputation of Missing Data*. Chapman & Hall/CRC Interdisciplinary Statistics. Taylor & Francis.

Van Buuren, S., H. C. Boshuizen, and D. L. Knook (1999). "Multiple imputation of missing blood pressure covariates in survival analysis". *Statistics in Medicine*, **18**(6):681–694. DOI: 10.1002/(SICI)1097-0258(19990330)18:6<681::AID-SIM71>3.0.CO;2-R.

Zhu, J. and T. E. Raghunathan (2015). "Convergence Properties of a Sequential Regression Multiple Imputation Algorithm". *Journal of the American Statistical Association*, **110**(511):1112–1124. DOI: 10.1080/01621459.2014.948117.

2 Dealing with Missing Covariates in Epidemiologic Studies

This chapter is based on

> Nicole S. Erler, Dimitris Rizopoulos, Joost van Rosmalen, Vincent W. V. Jaddoe, Oscar H. Franco and Emmanuel M. E. H. Lesaffre. Dealing with missing covariates in epidemiologic studies: a comparison between multiple imputation and a full Bayesian approach. *Statistics in Medicine*, 2016; **35**(17), 2955 – 2974. doi:10.1002/sim.6944

Abstract

Incomplete data are generally a challenge to the analysis of most large studies. The current gold standard to account for missing data is multiple imputation (MI), and more specifically multiple imputation with chained equations (MICE). Numerous studies have been conducted to illustrate the performance of MICE for missing covariate data. The results show that the method works well in various situations. However, less is known about its performance in more complex models, specifically when the outcome is multivariate as in longitudinal studies. In current practice, the multivariate nature of the longitudinal outcome is often neglected in the imputation procedure or only the baseline outcome is used to impute missing covariates. In this chapter, we evaluate the performance of MICE using different strategies to include a longitudinal outcome into the imputation models and compare it with a fully Bayesian approach that jointly imputes missing values and estimates the parameters of the longitudinal model. Results from simulation and a real data example show that MICE requires the analyst to correctly specify which components of the longitudinal process need to be included in the imputation models in order to obtain unbiased results. The fully Bayesian approach, on the other hand, does not require the analyst to explicitly specify how the longitudinal outcome enters the imputation models. It performed well under different scenarios.

2.1 Introduction

Missing values are a common complication in the analysis of cohort studies. Since epidemiologic studies are often adjusted for a large number of possible confounders, the treatment of missing covariate values is of special interest and our focus in this chapter. There have been numerous publications that show that naive ways to handle missing data, like complete case analysis, often lead to biased estimates and considerable loss of power (Molenberghs and Kenward 2007; Donders et al. 2006; Janssen et al. 2010; Knol et al. 2010; Sterne et al. 2009; Van Buuren 2012).

One standard approach to tackle this problem is to perform multiple imputation (MI) (Rubin 1987). Among the different flavours of MI, the Multiple Imputation with Chained Equations (MICE) (Van Buuren 2012) approach has gained the widest acceptance due to its good performance and ease of use. MI, and hence also MICE, works in three steps: First, a small number (often 5 or 10) of datasets are created by imputing each missing value multiple times. Each of the completed datasets can then be analysed using standard complete data methods. To obtain the overall result and to take additional uncertainty created by the missing values into account, the derived estimates are being pooled in the last step.

As a consequence of the separation of imputation and analysis steps, the relations between the outcome and the covariates, which are modelled in the analysis step, must be included explicitly in these imputation models. This means that the imputation models should not only contain other covariates in the predictor but also the outcome (Moons et al. 2006). When the outcome is univariate (and the model does not contain non-linear effects or interactions that involve incomplete covariates) this can be easily done because the outcome is just one of the variables in the dataset. However, when the outcome has a multivariate nature, such as the longitudinal outcome in our motivating study, it is not directly evident how this can be achieved, since longitudinal outcomes are often unbalanced with different subjects providing a different number of measurements at different time points. To overcome this problem one could consider simple or complex summaries of the long trajectories, such as including only a single value (e.g., baseline or the last available one) or the area under the trajectory. However, it is not clear which of these representations is the most adequate one for a specific analysis model and moreover, except in very simple situations, these summaries discard relevant information. As will be shown later, inclusion of inadequate summary measures of the subjects' trajectories can lead to bias.

To prevent the problem of having to specify the appropriate summary measure in the MICE approach we propose here a fully Bayesian approach which combines the analysis model with the imputation models. The essential difference to

MICE is that by combining the imputation and analysis in one estimation procedure, the Bayesian approach obtains inferences on the posterior distribution of the parameters and missing covariates directly and jointly. Thereby, the whole trajectory of the longitudinal outcome is implicitly taken into account in the imputation of missing covariate values and no summary representation is necessary. A common approach to specify the joint distribution is to assume (latent) normal distributions for all variables and model it as multivariate normal (Carpenter and Kenward 2013; Goldstein et al. 2009). In the present work, we chose a different approach and follow Ibrahim et al. (2002), who propose a decomposition of the likelihood into a series of univariate conditional distributions. This produces the sequential fully Bayesian (SFB) approach which is a flexible and easy to implement alternative to MICE. Furthermore, since missing values are continuously imputed in each iteration of the estimation procedure, the uncertainty about the missing values is automatically taken into account and no pooling is necessary.

Besides approaches using multiple imputation or the fully Bayesian framework, other methods that can handle missing covariates in longitudinal settings have been investigated in the literature. Stubbendick and Ibrahim (2003), for instance, approach the missing data problem using a likelihood-based approach that factorized the joint likelihood into a sequence of conditional distributions, analogue the SFB approach. Other authors (Chen, Grace, et al. 2010; Chen and Zhou 2011) have shown how to apply weighted estimating equations for inference in settings with incomplete data.

In the present study, we describe different strategies to include a longitudinal outcome in MICE and compare them with the SFB approach. Both methods were evaluated using simulation and a motivating real data example that required the analysis of a large dataset with missing values in several variables of different types. The rest of this chapter is organized as follows: Section 2.2 briefly describes the data motivating this study. In Section 2.3 we introduce the problem of missing data, and describe and compare the two methods of interest, MICE and the SFB approach. Both methods are applied to a real data example in Section 2.4 and evaluated in a simulation study, which is described in Section 2.5. Section 2.6 concludes the chapter with a discussion.

2.2 Generation R Data

We have taken a subset of variables measured within the Generation R Study, a population-based prospective cohort study from early fetal life onwards in Rotterdam, the Netherlands (Jaddoe et al. 2012), to illustrate both approaches. It was extracted with the aim to analyse the effect of diet, represented by three

principal components, on gestational weight gain and contains a number of incomplete covariates of mixed type. The relationship between diet and weight gain during pregnancy is of special interest to epidemiologists because diet may influence the amount of weight gained during pregnancy and subsequently affect the risk of adverse pregnancy outcomes. Furthermore, the Body Mass Index (BMI) before pregnancy is an important related factor on which general guidelines and recommendations for gestational weight gain are based.

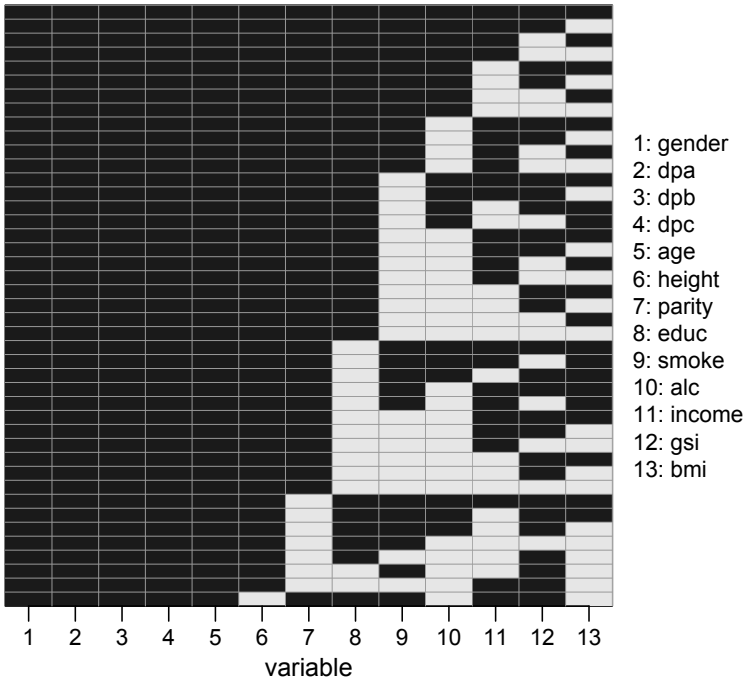


Figure 2.1: Missing data pattern of the Generation R data.

In the present study, data of 3374 pregnant women for whom dietary information was available, were analysed. Each woman was asked for her pre-pregnancy weight (baseline) and had up to three weight measurements during pregnancy, one in each trimester. There were 2297 women for whom all four weight measurements were recorded, 917 for whom three weight measurements were observed, 146 women had two measurements, and 14 women had only one measurement of weight. The gestational age at each measurement was recorded and the time point of the baseline measurement was set to be zero for all women. Table 2.1 in Appendix 2.A gives an overview of the available data. All covariates are cross-sectional. The variable `bmi` was calculated from baseline weight ($\texttt{weight}_0$) and `height`. Except

for `gender`, `age` and the dietary pattern variables, all variables had missing values, in proportions ranging between 0.03% and 14.17%. Reasons for missing covariate values in this study are usually (item) non-response in the questionnaires used. Missing baseline weight or weight measurement in the first trimester occurs when women are included in the study at a later gestational age. The missing pattern of the covariates is visualized in Figure 2.1. Variables 2 – 6, 12 and 13 are continuous, variables 1, 7 and 11 are binary and variables 8 – 10 are categorical with three categories each. 2236 individuals had complete covariate data.

2.3 Dealing with Missing Data

A standard modelling framework for studying the relationship between a longitudinal outcome $\mathbf{y}$ and predictor variables $\mathbf{X}$ is a linear mixed model:

$$y_{ij} = \mathbf{x}_{ij}^{\top}\boldsymbol{\beta} + \mathbf{z}_{ij}^{\top}\mathbf{b}_i + \varepsilon_{ij},$$

where y_{ij} is the j-th observation of individual i, measured at time t_{ij}, $\boldsymbol{\beta}$ denotes the vector of regression coefficients of the design matrix of the fixed effects $\mathbf{X}_i$, where $\mathbf{x}_{ij}$ is a column vector that contains the j-th row of that matrix, and, analogously, z_{ij} denotes a row of the design matrix $\mathbf{Z}_i$ of the random effects $\mathbf{b}_i$ and contains a subset of the variables in x_{ij}. Furthermore, the vector $\mathbf{b}_i$ follows a multivariate normal distribution with mean zero and covariance matrix $\mathbf{D}$, and ε_{ij} is an error term that is normally distributed with mean zero and variance σ_y^2.

In a complete data setting, the probability density function of interest is $p(y_{ij} \mid \mathbf{x}_{ij}, \mathbf{z}_{ij}, \boldsymbol{\theta}_{Y|X})$, where $\boldsymbol{\theta}_{Y|X}$ denotes the vector of parameters of the model. When some of the covariates are incomplete, $\mathbf{X}$ consists of two parts, the completely observed variables $\mathbf{X}_{obs}$ and those variables that containing missing values, $\mathbf{X}_{mis}$. The measurement model $p(y_{ij} \mid \mathbf{x}_{ij,obs}, \mathbf{x}_{ij,mis}, \mathbf{z}_{ij}, \boldsymbol{\theta}_{Y|X})$ then depends on unobserved data and standard complete data methods cannot be used any more.

In this chapter, we restrict our attention to the imputation of cross-sectional covariates. The missing data mechanism of the outcome is assumed to be ignorable, i.e., Missing At Random (MAR) or Missing Completely at Random (MCAR) (Little and Rubin 2002; Seaman, Galati, et al. 2013), and hence the missing outcome values do not require special treatment when using mixed effects models.

2.3.1 Multiple Imputation using Chained Equations

The underlying principle of MI is to divide the analysis of incomplete data into three steps: imputation, analysis and pooling. MICE is a popular implementation of the imputation step since it allows for multivariate missing data and does not

require a specific missingness pattern. The idea behind MICE is that, under certain regularity conditions, the multivariate distribution

$$p(\mathbf{x}_{i,mis} \mid \mathbf{y}_i, \mathbf{x}_{i,obs}, \boldsymbol{\theta}_X), \tag{2.1}$$

with $\mathbf{x}_{i,mis} = (x_{i,mis_1}, \ldots, x_{i,mis_p})^\top$ and $\mathbf{x}_{i,obs} = (x_{i,obs_1}, \ldots, x_{i,obs_q})^\top$, can be uniquely determined by its full conditional distributions and hence Gibbs sampling of the conditional distributions can be used to produce a sample from (2.1) (Van Buuren 2012). However, the MICE procedure does not actually start from a specification of (2.1), but it directly defines a series of conditional, predictive models of the form

$$p(x_{i,mis_\ell} \mid \mathbf{x}_{i,mis_{-\ell}}, \mathbf{x}_{i,obs}, \mathbf{y}_i, \boldsymbol{\theta}_{X_\ell}), \tag{2.2}$$

that link each incomplete predictor variable x_{i,mis_ℓ}, $\ell = 1, \ldots, p$, with other incomplete and complete predictors, $\mathbf{x}_{i,mis_{-\ell}}$ and $\mathbf{x}_{i,obs}$, respectively, and importantly with the outcome. These predictive distributions are typically members of the extended exponential family (extended with models for ordinal data) with linear predictor

$$g_\ell \left\{ E\left(x_{i,mis_\ell} \mid x_{i,obs}, x_{i,mis_{-\ell}}, \mathbf{y}_i, \boldsymbol{\gamma}_\ell, \boldsymbol{\xi}_\ell, \boldsymbol{\alpha}_\ell\right) \right\} = \boldsymbol{\gamma}_\ell^\top \mathbf{x}_{i,obs} + \boldsymbol{\xi}_\ell^\top \mathbf{x}_{i,mis_{-\ell}} + \boldsymbol{\alpha}_\ell^\top f(\mathbf{y}_i),$$

where $g_\ell(\cdot)$ is the one-to-one monotonic link function for the ℓ-th covariate and $\boldsymbol{\gamma}_\ell$, $\boldsymbol{\xi}_\ell$ and $\boldsymbol{\alpha}_\ell$ are vectors of parameters relating the complete and missing covariates and the outcome to x_{i,mis_ℓ}.

The function $f(\cdot)$ specifies how the outcome enters the linear predictor. In the univariate case, the default choice for $f(y_i)$ is simply the identity function. However, when we have a multivariate $\mathbf{y}_i$, such as a longitudinal outcome, we cannot always simply specify $\boldsymbol{\alpha}_\ell^\top \mathbf{y}_i$ because $\mathbf{y}_i$ may have different length than $\mathbf{y}_k$ for $i \neq k$, and the time points t_{ij} and t_{kj} of the observations y_{ij} and y_{kj} may be very different. Hence, it is not meaningful to use the same regression coefficient $\alpha_{\ell j}$ to connect outcomes of different individuals with x_{i,mis_ℓ} and a representation needs to be found that summarizes $\mathbf{y}_i$ and that has the same number of elements which also have the same interpretation for all individuals.

Some examples for $f(\mathbf{y}_i)$ could be

$$f(\mathbf{y}_i) = 0, \tag{2.3}$$

$$f(\mathbf{y}_i) = y_{ij}, \tag{2.4}$$

$$f(\mathbf{y}_i) = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}, \tag{2.5}$$

$$f(\mathbf{y}_i) = \sum_{j=1}^{n_i-1}(t_{ij+1} - t_{ij})\frac{y_{ij} + y_{ij+1}}{2}, \tag{2.6}$$

$$f(\mathbf{y}_i) = \hat{\mathbf{b}}_i = \widehat{\mathbf{D}}\widetilde{\mathbf{Z}}_i^\top \left(\widetilde{\mathbf{Z}}_i\widehat{\mathbf{D}}\widetilde{\mathbf{Z}}_i^\top + \widehat{\mathbf{\Sigma}}_i\right)^{-1}\left(\mathbf{y}_i - \widetilde{\mathbf{X}}_i\hat{\boldsymbol{\beta}}\right), \tag{2.7}$$

$$f(\mathbf{y}_i) = \int \widetilde{\mathbf{X}}_i(t)\hat{\boldsymbol{\beta}} + \widetilde{\mathbf{Z}}_i(t)\hat{\mathbf{b}}_i \, dt. \tag{2.8}$$

Here, $f(\mathbf{y}_i)$ from (2.3) results in omitting the outcome completely from the imputation procedure. Equations (2.4) – (2.6) are examples of representations that directly use the observed outcome, where (2.4) uses only one observation, e.g., the first/baseline outcome if $j = 1$, (2.5) uses the mean of the observed outcome and (2.6) uses the area under the observed trajectory to summarize $\mathbf{y}_i$. Functions (2.7) and (2.8) are examples of representations that are based on the fit of a preliminary model. Such a preliminary model could, for instance, include the time variable and possibly completely observed covariates. In (2.7) we use as a summary of the trajectory the empirical Bayes estimates of the random effects, $\hat{\mathbf{b}}_i$, where $\widetilde{\mathbf{X}}$ and $\widetilde{\mathbf{Z}}$ are the design matrices of the preliminary model and subsets of $\mathbf{X}$ and $\mathbf{Z}$, respectively, $\widehat{\mathbf{\Sigma}}_i$ is the estimated covariance matrix of the error terms and $\hat{\boldsymbol{\beta}}$ are the restricted maximum likelihood estimates of the regression coefficients from that model. Equation (2.8) describes the area under the estimated individual trajectory from (2.7).

Naturally, this list of possible summary functions is not exhaustive. In addition, combinations of these could be considered as well. Only in very simple settings is it possible to determine which function of $\mathbf{y}_i$ is appropriate. Carpenter and Kenward (2013) show that under a random intercept model for $\mathbf{y}_i$, (2.5) is the appropriate summary function in the imputation model for a normal cross-sectional covariate. For more complex analysis models or discrete covariates it is not straightforward to derive the appropriate summary functions.

In settings where the outcome is balanced or close to balanced and does not have a large number of repeated measurements, another approach could be to impute missing outcome values so that all individuals have the same number of

measurements at approximately the same time points, and to include all outcome variables as separate variables in the linear predictor of the imputation models.

Two important requirements that are necessary to obtain valid imputations using the MICE procedure as described above are that the imputation models need to be specified correctly and that the missing data mechanism needs to be ignorable, i.e., Missing Completely at Random (MCAR) or Missing At Random (MAR) (Little and Rubin 2002; Seaman, Galati, et al. 2013). In this case, the missing data mechanism does not need to be modelled specifically. A common assumption is that the missing values are MAR given all the observed values. This implies that also the values of the time variable of a mixed model should be included in the imputation models. Since the time variable is not constant over time, a summary representation has to be specified for this variable as well.

2.3.2 Fully Bayesian Imputation

The choice of a summary representation of a multivariate outcome can be avoided by using a fully Bayesian approach. In the Bayesian setting, the complete data likelihood is combined with prior information to compute the complete data posterior, which can be written as

$$\begin{aligned} p(\boldsymbol{\theta}_{Y|X}, \boldsymbol{\theta}_X, \mathbf{x}_{i,mis} \mid y_{ij}, \mathbf{x}_{i,obs}) \quad &\propto \quad p(y_{ij} \mid \mathbf{x}_{i,obs}, \mathbf{x}_{i,mis}, \boldsymbol{\theta}_{Y|X}) \\ &\quad\; p(\mathbf{x}_{i,mis} \mid \mathbf{x}_{i,obs}, \boldsymbol{\theta}_X) \pi(\boldsymbol{\theta}_{Y|X}) \pi(\boldsymbol{\theta}_X), \end{aligned}$$

where $\boldsymbol{\theta}_X$ is a vector containing parameters that are associated with the likelihood of the partially observed covariates $\mathbf{X}_{mis}$ and $\pi(\boldsymbol{\theta}_{Y|X})$ and $\pi(\boldsymbol{\theta}_X)$ are prior distributions.

A convenient way to specify the joint likelihood of the missing covariates $p(\mathbf{x}_{i,mis} \mid \mathbf{x}_{i,obs}, \boldsymbol{\theta}_X)$ is to use a sequence of conditional univariate distributions (Ibrahim et al. 2002)

$$\begin{aligned} &p(x_{i,mis_1}, \ldots, x_{i,mis_p} \mid \mathbf{x}_{i,obs}, \boldsymbol{\theta}_X) = \\ &\quad p(x_{i,mis_1} \mid \mathbf{x}_{i,obs}, \boldsymbol{\theta}_{X_1}) \prod_{\ell=2}^{p} p(x_{i,mis_\ell} \mid \mathbf{x}_{i,obs}, x_{i,mis_1}, \ldots, x_{i,mis_{\ell-1}}, \boldsymbol{\theta}_{X_\ell}), \end{aligned} \tag{2.9}$$

with $\boldsymbol{\theta}_X = (\boldsymbol{\theta}_{X_1}, \ldots, \boldsymbol{\theta}_{X_p})^\top$. Each of these distributions is again chosen from the extended exponential family, according to the type of the respective variable. Writing the joint distribution of the covariates in such a sequence provides a straightforward way to specify the joint distribution even when the covariates are of mixed type.

After specifying the prior distributions $\pi(\boldsymbol{\theta}_{Y|X})$ and $\pi(\boldsymbol{\theta}_X)$, Markov chain Monte Carlo (MCMC) methods, such as Gibbs sampling, can be used to draw samples from the joint posterior distribution of all parameters and missing values. Since all missing values are imputed in each iteration of the Gibbs sampler, the additional uncertainty created by the missing values is automatically taken into account and no pooling is necessary.

The advantage of working with the full likelihood instead of the series of predictive models (2.2) is that we can choose how to factorize this full likelihood. More specifically, by factorizing the joint distribution $p(y_{ij}, \mathbf{x}_{i,mis}, \mathbf{x}_{i,obs} \mid \boldsymbol{\theta}_{Y|X}, \boldsymbol{\theta}_X)$ into the conditional distribution $p(y_{ij} \mid \mathbf{x}_{i,obs}, \mathbf{x}_{i,mis}, \boldsymbol{\theta}_{Y|X})$ and the marginal distribution $p(\mathbf{x}_{i,mis} \mid \mathbf{x}_{i,obs}, \boldsymbol{\theta}_X)$, the joint posterior distribution can be specified without having to include the outcome into any predictor and no summary representation $f(\mathbf{y}_i)$ is needed. This becomes clear when writing out the full conditional distribution of the incomplete covariates, used by the Gibbs sampler:

$$
\begin{aligned}
p(x_{i,mis_\ell} \mid \mathbf{y}_i, \mathbf{x}_{i,obs}, \mathbf{x}_{i,mis_{-\ell}}, \mathbf{b}_i, \boldsymbol{\theta}) &\propto \left\{ \prod_{j=1}^{n_i} p\left(y_{ij} \mid \mathbf{x}_{i,obs}, \mathbf{x}_{i,mis}, \mathbf{b}_i, \boldsymbol{\theta}_{Y|X}\right) \right\} \\
&\quad p(\mathbf{x}_{i,mis} \mid \mathbf{x}_{i,obs}, \boldsymbol{\theta}_X) \pi(\mathbf{b}_i) \pi(\boldsymbol{\theta}_{Y|X}) \pi(\boldsymbol{\theta}_X) \\
&\propto \left\{ \prod_{j=1}^{n_i} p\left(y_{ij} \mid \mathbf{x}_{i,obs}, \mathbf{x}_{i,mis}, \mathbf{b}_i, \boldsymbol{\theta}_{Y|X}\right) \right\} \\
&\quad p(x_{i,mis_\ell} \mid \mathbf{x}_{i,obs}, \mathbf{x}_{i,mis_{<\ell}}, \boldsymbol{\theta}_{X_\ell}) \\
&\quad \left\{ \prod_{k=\ell+1}^{p} p(x_{i,mis_k} \mid \mathbf{x}_{i,obs}, \mathbf{x}_{i,mis_{<k}}, \boldsymbol{\theta}_{X_k}) \right\} \\
&\quad \pi(\mathbf{b}_i) \pi(\boldsymbol{\theta}_{Y|X}) \pi(\boldsymbol{\theta}_{X_\ell}) \prod_{k=\ell+1}^{p} \pi(\boldsymbol{\theta}_{X_k}),
\end{aligned}
$$

where n_i is the number of repeated measurements of individual i, $\boldsymbol{\theta}_{Y|X}^\top = (\boldsymbol{\gamma}_y^\top, \boldsymbol{\xi}_y^\top)$, $\boldsymbol{\theta}_{X_\ell}^\top = (\boldsymbol{\gamma}_\ell^\top, \boldsymbol{\xi}_\ell^\top)$ and $\boldsymbol{\theta}_{X_k}^\top = (\boldsymbol{\gamma}_k^\top, \boldsymbol{\xi}_k^\top)$, and $\mathbf{x}_{i,mis_{<\ell}} = (x_{i,mis_1}, \ldots, x_{i,mis_{\ell-1}})^\top$ and $\mathbf{x}_{i,mis_{<k}} = (x_{i,mis_1}, \ldots, x_{i,mis_{k-1}})^\top$ denote the subset of variables in the sequence before $\mathbf{x}_{i,mis_\ell}$ and $\mathbf{x}_{i,mis_k}$, respectively.

The densities $p(y_{ij} \mid \mathbf{x}_{i,obs}, \mathbf{x}_{i,mis}, \mathbf{b}_i, \boldsymbol{\theta}_{Y|X})$, $p(x_{i,mis_\ell} \mid \mathbf{x}_{i,obs}, \mathbf{x}_{i,mis_{<\ell}}, \boldsymbol{\theta}_{X_\ell})$ and $p(x_{i,mis_k} \mid \mathbf{x}_{i,obs}, \mathbf{x}_{i,mis_{<k}}, \boldsymbol{\theta}_{X_k})$ are members of the extended exponential family with linear predictors

$$E\left(y_{ij} \mid \mathbf{x}_{i,obs}, \mathbf{x}_{i,mis}, \mathbf{b}_i, \boldsymbol{\theta}_{Y|X}\right) = \boldsymbol{\gamma}_y^\top \mathbf{x}_{i,obs} + \xi_y^\top \mathbf{x}_{i,mis} + \mathbf{z}_{ij}^\top \mathbf{b}_i, \quad (2.10)$$

$$g_\ell\left\{E\left(x_{i,mis_\ell} \mid \mathbf{x}_{i,obs}, \mathbf{x}_{i,mis_{<\ell}}, \boldsymbol{\theta}_{X_\ell}\right)\right\} = \boldsymbol{\gamma}_\ell^\top \mathbf{x}_{i,obs} + \sum_{s=1}^{\ell-1} \xi_{\ell_s} x_{i,mis_s}, \quad (2.11)$$

$$g_k\left\{E\left(x_{i,mis_k} \mid \mathbf{x}_{i,obs}, \mathbf{x}_{i,mis_{<k}}, \boldsymbol{\theta}_{X_k}\right)\right\} = \boldsymbol{\gamma}_k^\top \mathbf{x}_{i,obs} + \sum_{s=1}^{k-1} \xi_{k_s} x_{i,mis_s}, \quad (2.12)$$
$$k = \ell + 1, \ldots, p.$$

Equation (2.10) represents the predictor of the linear mixed model, Equation (2.11) the predictor of the imputation model of x_{mis_ℓ} with link function g_ℓ from the extended exponential family, and Equation (2.12) represents the predictors of the covariates that have x_{mis_ℓ} as a predictive variable, with $g_k(\cdot)$ being the corresponding link function. As can easily be seen, none of the Equations (2.10) – (2.12) contains the outcome on its right-hand side, whereby the SFB approach avoids the need for a summary representation.

It has been mentioned before that it is not obvious how the imputation models in the sequence should be ordered (Bartlett et al. 2015) and, from a theoretical point of view, different orderings may result in different joint distributions, leading to different results. Chen and Ibrahim (2001) suggest to condition the categorical imputation models on the continuous covariates. In the context of MI it has been suggested to impute variables in a sequence so that the missing pattern is close to monotone (Van Buuren 2012; Schafer 1997; Bartlett et al. 2015). Our convention is to order the imputation models in (2.9) according to the number of missing values, starting with the variable with the least missing values. It has been shown, however, that sequential specifications, as used in the Bayesian approach, are quite robust against changes in the ordering (Chen and Ibrahim 2001; Zhu and Raghunathan 2015) and results may be unbiased even when the order of the sequence is misspecified as long as the imputation models fit the data well enough. Preliminary results of our own work (not shown here) indicated that our convention may lead to shorter computational times.

Like MICE, the SFB approach, in the form described above, is valid only under ignorable missing data mechanisms and when the analysis model, as well as the conditional distributions of the covariates, are correctly specified.

2.4 Data Analysis

2.4.1 Design

The data described in Section 2.2 was imputed and analysed using MICE and a frequentist linear mixed model as well as with the SFB approach.

The analysis model of interest was the linear mixed model

$$\begin{aligned}\texttt{weight}_{ij} =& \beta_0 + \beta_1 \texttt{gender}_i + \beta_2 \texttt{dpa}_i + \beta_3 \texttt{dpb}_i + \beta_4 \texttt{dpc}_i + \beta_5 \texttt{age}_i + \\ & \beta_6 \texttt{parity}_i + \beta_7 \texttt{educ}_{2i} + \beta_8 \texttt{educ}_{3i} + \beta_9 \texttt{smoke}_{2i} + \\ & \beta_{10} \texttt{smoke}_{3i} + \beta_{11} \texttt{alc}_{2i} + \beta_{12} \texttt{alc}_{3i} + \beta_{13} \texttt{income}_i + \\ & \beta_{14} \texttt{gsi}_i + \beta_{15} \texttt{bmi}_i + \beta_{16} \texttt{time}_{ij} + \beta_{17} \texttt{time}_{ij} \times \texttt{dpa}_i + \\ & \beta_{18} \texttt{time}_{ij} \times \texttt{dpb}_i + \beta_{19} \texttt{time}_{ij} \times \texttt{dpc}_i + \beta_{20} \texttt{time}_{ij}^2 + \\ & b_{i0} + b_{i1} \texttt{time}_{ij} + \varepsilon_{ij},\end{aligned}$$

where `time` is a centred version of the gestational age (in weeks) and b_{i0} and b_{i1} are correlated random effects. Interaction terms between the dietary pattern variables and time were included in the model to allow the slope of the weight trajectories to be associated with diet. The variable names used here are explained in Table 2.1 in Appendix 2.A.

For the imputation with MICE, we chose five different strategies to represent $\mathbf{y}_i$, three simple and commonly used strategies and two more sophisticated strategies:

- *S1:* exclude the outcome from the imputation models, i.e., $f(\mathbf{y}_i)$ as in (2.3),
- *S2:* include only the baseline outcome, $\texttt{weight}_0$, i.e., $f(\mathbf{y}_i)$ as in (2.4) with $j = 1$,
- *S3:* include the mean of the observed weight measurements, i.e., $f(\mathbf{y}_i)$ as in (2.5),
- *S4:* obtain empirical Bayes estimates of the random effects $\hat{\mathbf{b}}_i$ from a preliminary analysis model with `time` as only explanatory variable and a simplified random effects structure that only contained a random intercept, and include those estimates in the imputation, i.e., $f(\mathbf{y}_i)$ from (2.7) with $\hat{\mathbf{b}}_i = (\hat{b}_{i,0})$,
- *S5:* fit a preliminary model with `time` as only explanatory variable, using the random effects structure used in the analysis model, and include the empirical Bayes estimates from that model in the imputation step, i.e. $f(\mathbf{y}_i)$ from (2.7) with $\hat{\mathbf{b}}_i = (\hat{b}_{i,0}, \hat{b}_{i,1})$.

For each of the five strategies, we calculated $f(\texttt{weight}_i)$ and $f(\texttt{time}_i)$, using the same function $f(\cdot)$, and appended both as two or more new variables to the dataset.

Since in our data the baseline value of `time` is zero for all individuals it does not help the imputation in *S2* and was hence not included in that strategy. Note that strategy *S3* is implemented in the current version (2.22) of the R-package **mice** (Van Buuren and Groothuis-Oudshoorn 2011) for the imputation of continuous cross-sectional covariates in a 2-level model but not for variables of other type. It is also important to note that strategies *S1* and *S2* may generally not lead to valid imputations since they do not include the entire outcome. They are included here to demonstrate the bias that may be introduced by a naive use of MICE.

Ten datasets were created for each strategy by taking the imputed values of the 20th iteration from the MICE algorithm using the R-package **mice**. Covariates were imputed according to their measurement level using (Bayesian) normal regression, logistic regression and a proportional odds model for ordered factors. For details on these models see Van Buuren (2012) and Van Buuren and Groothuis-Oudshoorn (2011). Since the variable `gsi` is restricted to positive values, we imputed the square root of `gsi`, but used `gsi` on the original scale as predictor in the other imputation models. For the analysis, categorical variables were re-coded into dummy variables. The analysis model described above was then fitted for each of the completed datasets using `lmer()` from the R-package **lme4** (Bates et al. 2015). Finally, the results were pooled using Rubin's rules (Rubin 1987), as implemented in **mice**.

The SFB approach was implemented in R (R Core Team 2013) and JAGS (Plummer 2003) using a normal hierarchical model

$$
\begin{aligned}
\texttt{weight}_{ij} &\sim N(\mu_{ij}, \sigma^2_{\texttt{weight}}) \\
\mu_{ij} &= \beta_0 + \beta_1 \texttt{gender}_i + \ldots + \beta_{15} \texttt{bmi}_i + \beta_{16} \texttt{time}_{ij} + \\
&\quad \beta_{17} \texttt{time}_{ij} \times \texttt{dpa}_i + \beta_{18} \texttt{time}_{ij} \times \texttt{dpb}_i + \beta_{19} \texttt{time}_{ij} \times \texttt{dpc}_i + \\
&\quad \beta_{20} \texttt{time}^2_{ij} + b_{i0} + b_{i1} \texttt{time}_{ij} \\
\\
(b_{i,0}, b_{i,1})^\top &\sim N(\mathbf{0}, \mathbf{D}) \\
\mathbf{D} &\sim \text{inv-Wishart}(\mathbf{R}, 2) \\
\mathbf{R} &= \text{diag}(r_1, r_2) \\
r_1, r_2 &\overset{iid}{\sim} \text{Ga}(0.1, 0.01) \\
\sigma^2_{\texttt{weight}} &\sim \text{inv-Ga}(0.001, 0.001),
\end{aligned}
$$

where μ_{ij} has the same structure as the frequentist mixed model described above.

The conditional distributions for the missing covariates from Equation (2.9) were specified as linear, logistic and cumulative logistic regression models. Uninformative priors were used for all parameters. Following the advice of Garrett and Zeger

(2000), we assumed independent normal distributions with mean zero and variance 9/4 for the regression coefficients in categorical models (logistic and cumulative logistic), since that choice leads to priors that are approximately uniform on the probability scale obtained from the "expit" transformation. The first 100 iterations of the MCMC sample were discarded as burn-in. Three chains with 5000 iterations each were used to compute the posterior summary measures. Convergence of the MCMC chains was checked using the Gelman-Rubin criterion (Gelman, Meng, et al. 1996). The posterior estimates were considered precise enough if the MCMC error was less than five per cent of the parameter's standard deviation (Lesaffre and Lawson 2012).

2.4.2 Results

Figure 2.2 shows parameter estimates and 95% confidence intervals (CIs; 2.5% and 97.5% quantiles for the SFB approach) for the regression coefficients from the Generation R example.

Several coefficients demonstrate substantial differences in the estimates and CIs between different imputation strategies. Overall, MICE strategies *S1* and *S2* resulted in similar estimates and CIs, which however differed considerably from the estimates and/or CIs of the other three MICE strategies and the SFB approach, for most parameters. The naive MICE approaches *S1* and *S2* estimated non-significant effects for `dpa`, `dpc` and `educ`, whereas the other approaches estimated effects that were significantly different from zero. The largest difference between the strategies was observed for the parameter of `bmi`. The posterior mean and CIs for all regression coefficients are displayed in Table 2.2 in Appendix 2.B. These results indicate that, in the present example, a considerable amount of information is lost when the outcome is not represented by a summary of all its observations but only the first measurement or excluded completely from the imputation models in MICE.

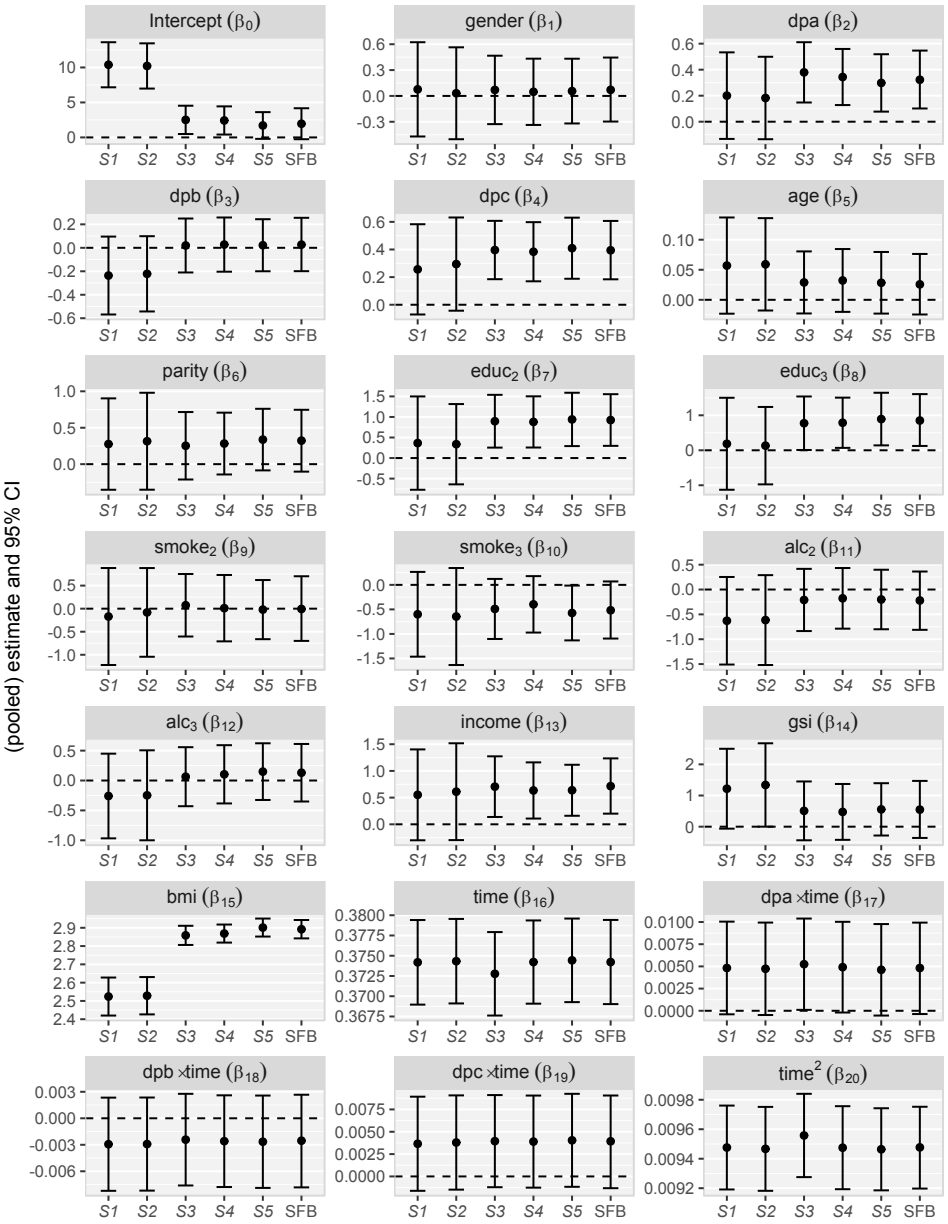


Figure 2.2: Parameter estimates and 95% CIs for the regression coefficients from the Generation R example for the five MICE strategies (*S1* – *S5*) and the sequential fully Bayesian approach (SFB). The displayed CIs for SFB are the 2.5% and 97.5% quantiles of the MCMC sample.

2.5 Simulation Study

2.5.1 Design

To compare the performance of MICE and the SFB approach under different scenarios of missing data mechanisms, we performed a simulation study in which the missingness of the covariates depended on the outcome in different ways.

We simulated data for 500 subjects from a linear mixed-effects model

$$\begin{aligned} y_{ij} =& \beta_0 + \beta_1 \texttt{time}_{ij} + \beta_2 \texttt{time}_{ij}^2 + \beta_3 \texttt{B}_i + \beta_4 \texttt{C}_i + \beta_5 \texttt{C}_i \times \texttt{time}_{ij} + \\ & b_{i0} + \texttt{time}_{ij} b_{i1} + \texttt{time}_{ij}^2 b_{i2} + \varepsilon_{ij}, \end{aligned}$$

with $j = 1, \ldots, 6$, a `time` variable that was uniformly distributed between 0 and 3.5, a binary cross-sectional covariate $\texttt{B}_i$, creating two groups of equal size, a continuous cross-sectional covariate $\texttt{C}_i$ that follows a standard normal distribution and is independent from $\texttt{B}_i$, and error terms ε_{ij} from a normal distribution with mean zero and variance $\sigma_y^2 = 0.25$. The random effects were assumed to be multivariate normal with mean $\mathbf{0}$ and covariance matrix $\mathbf{D}$. The values of $\boldsymbol{\beta}$ and $\mathbf{D}$ that were used in each of the scenarios can be found in Appendix 2.C. To create an unbalanced design, repeated measurements were removed with probability 0.2 and under the restriction that each individual had to have at least three repeated measurements. The remaining time points were sorted for each subject.

In order to create missingness that depends on the longitudinal structure of the outcome, in *Scenario 1* we obtained ordinary least squares estimates $\hat{\boldsymbol{\lambda}}_i$ of the regression coefficients from preliminary linear models

$$\texttt{y}_{ij} = \lambda_{0i} + \lambda_{1i} \texttt{time}_{ij} + \lambda_{2i} \texttt{time}_{ij}^2 + \tilde{\varepsilon}_{ij}, \quad \tilde{\varepsilon}_{ij} \sim N(0, \sigma_{\tilde{\varepsilon}}^2),$$

for each individual i and $\texttt{C}_i$ was put to missing with probability

$$\text{logit}(p_i) = \gamma_0 + \gamma_1 \hat{\lambda}_{0i} + \gamma_2 \hat{\lambda}_{1i} + \gamma_3 \hat{\lambda}_{2i},$$

where $\text{logit(x)} = \log\{x/(1-x)\}$. The values of $\boldsymbol{\gamma}$ used for the simulation can be found in Table 2.4 in Appendix 2.C.

In *Scenario 2*, missing values were created in $\texttt{B}_i$ as well as $\texttt{C}_i$, depending on the area under the curve (AUC) described by the observed outcome $\texttt{y}_{ij}$. The AUC for individual i was computed as

$$\text{AUC}_i = \sum_{j=1}^{n_i - 1} (\texttt{time}_{ij+1} - \texttt{time}_{ij}) \frac{\texttt{y}_{ij} + \texttt{y}_{ij+1}}{2},$$

where n_i is the number of repeated measurements of individual i, and scaled to be centred around zero and have standard deviation equal to 1, for computational reasons.

The probability p_{1i} of the continuous variable $\texttt{C}_i$ to be missing was then computed as

$$\text{logit}(p_{1i}) = \gamma_{10} + \gamma_{11}\text{AUC}_i + \gamma_{12}\text{AUC}_i^2.$$

Observations in the binary variable $\texttt{B}_i$ were deleted depending on the AUC and the missingness of $\texttt{C}_i$ with probability p_{2i}:

$$\text{logit}(p_{2i}) = \gamma_{20} + \gamma_{21}\mathbb{1}(\texttt{C}_i = \text{NA}) + \gamma_{22}\text{AUC}_i,$$

where $\mathbb{1}$ is the indicator function which is 1 if $\texttt{C}_i$ is missing and 0 otherwise.

In both scenarios, we performed the analysis based on the complete data as well as datasets with 20% and 50% missings. In Scenario 2, 60% of that desired missing percentage was missing in $\texttt{C}_i$, the remaining 40% in $\texttt{B}_i$. The coefficients $\gamma_1, \gamma_2, \gamma_3, \gamma_{11}, \gamma_{12}, \gamma_{21}$ and γ_{22} were specified a priori and the intercepts γ_0, γ_{10} and γ_{20} chosen so that (approximately) the desired proportion of data was missing. All parameter values used for the simulation can be found in Appendix 2.C.

Under both scenarios, the simulated data were again imputed and fitted, using the correct analysis model, with the five strategies using MICE described in Section 2.4, as well as the SFB approach. Here, $\texttt{time}_{i1}$ was included in strategy *S2*, strategy *S4* used empirical Bayes estimates of the random intercept and slope and strategy *S5* used estimates of the random intercept and the random effects for `time` and $\texttt{time}^2$. We created 10 imputed datasets with each MICE strategy. In the SFB approach, we simulated three chains. The number of iterations was determined based on two criteria: the Gelman-Rubin criterion (Gelman, Meng, et al. 1996) for convergence had to be smaller than 1.2 and the MCMC error less than five per cent of the parameter's standard deviation (Lesaffre and Lawson 2012). An example of the JAGS syntax used for *Scenario 2* can be found in Appendix 2.E.

It is important to note that *S1* and *S2* are not correct models under the MAR assumption in neither of the two scenarios since not all available information of $\mathbf{y}_i$ and $\texttt{time}_i$ is included.

2.5.2 Results

Under *Scenario 1*, the most severely biased estimates were observed for β_1 (related to `time`) and β_5 (related to the interaction between `time` and `C`). Figures 2.3 and

2.4 compare the relative bias (estimated $\hat{\beta}$ divided by true β) and coverage rate of the 95% CIs, respectively, for all imputation strategies and missing percentages in *Scenario 1*.

Table 2.6 in Appendix 2.D summarizes the results from both simulation studies. For each covariate, the first line gives the average relative bias of the estimated parameter over all 500 simulations, the second line gives the mean squared error (MSE) of the estimates multiplied by ten and the third line gives the proportion of simulations in which the estimated 95% CI, or the 95% credible interval for the Bayesian analysis, covered the true β.

MICE models *S1* and *S2* performed the worst, with a relative bias of 0.10 and 0.25 under 20% missing values, and a relative bias of -0.39 and -0.61 under 50% missing values. The coverage rate of the 95% CIs for both models was 20% and 30% respectively, under 20% missing values, and 5% and 0%, respectively, when only half of the individuals had `C` observed.

Summarizing $\mathbf{y}_i$ by its mean (*S3*) resulted in less biased results (relative bias 0.82, 92% coverage under 20% missing values). For 50% missing values the relative bias worsened to 0.39 and the coverage rate dropped to 54%. *S4* gave slightly biased estimates for the effect of `time` and the continuous covariate `C` and the CI of the parameter for `C` covered the true parameter only in 85.8% of the simulations under 50% missing values. *S5* and the SFB approach were unbiased for all missing percentages and had only minor deviations from the desired coverage rate of 95%. The same difference between the imputation strategies was observed for the interaction between `time` and `C`, the intercept and the parameter of the continuous variable `C`, although less severe in the latter case. Again, *S1* and *S2* led to the most biased parameters and had (very) poor coverage rate, *S3* was less biased and had insufficient coverage rate for the parameter of the interaction term. Also for these parameters, the SFB approach provided unbiased estimates with the desired coverage rate. The SFB approach had slightly smaller MSE than *S4* and *S5* for most parameters.

In *Scenario 2*, *S1* and *S2* gave biased estimates and produced CIs with insufficient coverage rate for most parameters (see Figures 2.5 and 2.6 in Appendix 2.D). Estimates from *S1* were more biased than those from *S2* except for the parameter of the continuous covariate and $\texttt{time}^2$. *S3* performed better than *S1* and *S2* for most parameters but still gave biased estimates and had coverage rate below 85% under 50% missing for the intercept and the effects for `B` and the interaction between `B` and `time`. The two more sophisticated MICE strategies *S4* and *S5*, as well as the SFB approach, provided unbiased estimates and CIs with the desired coverage rate for all parameters.

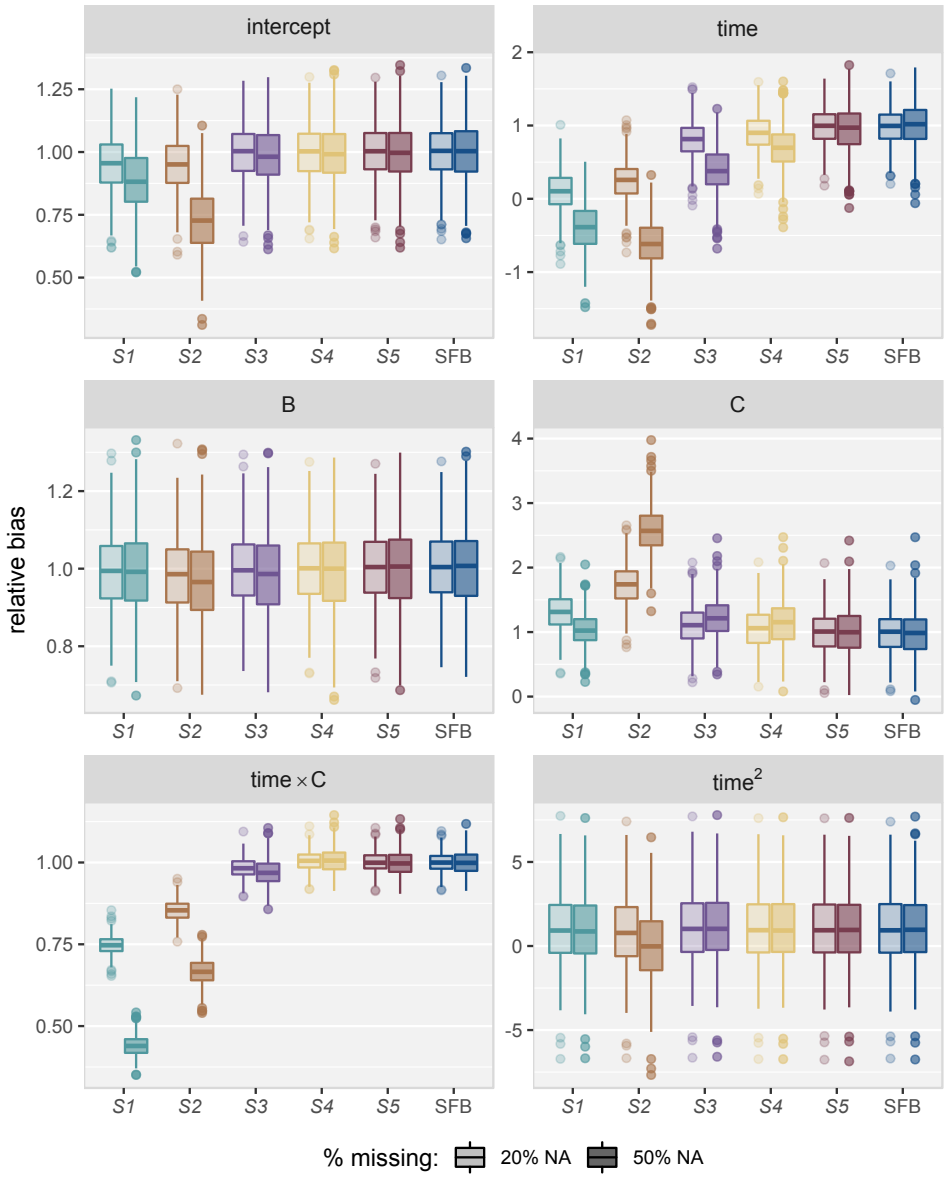


Figure 2.3: Relative bias in simulation *Scenario 1*, for the five imputation strategies using MICE (*S1*–*S5*) and the sequential fully Bayesian approach (SFB).

The average computation time per simulation for the complete data was less than one second for `lmer()` and 15 seconds for the Bayesian model. For incomplete

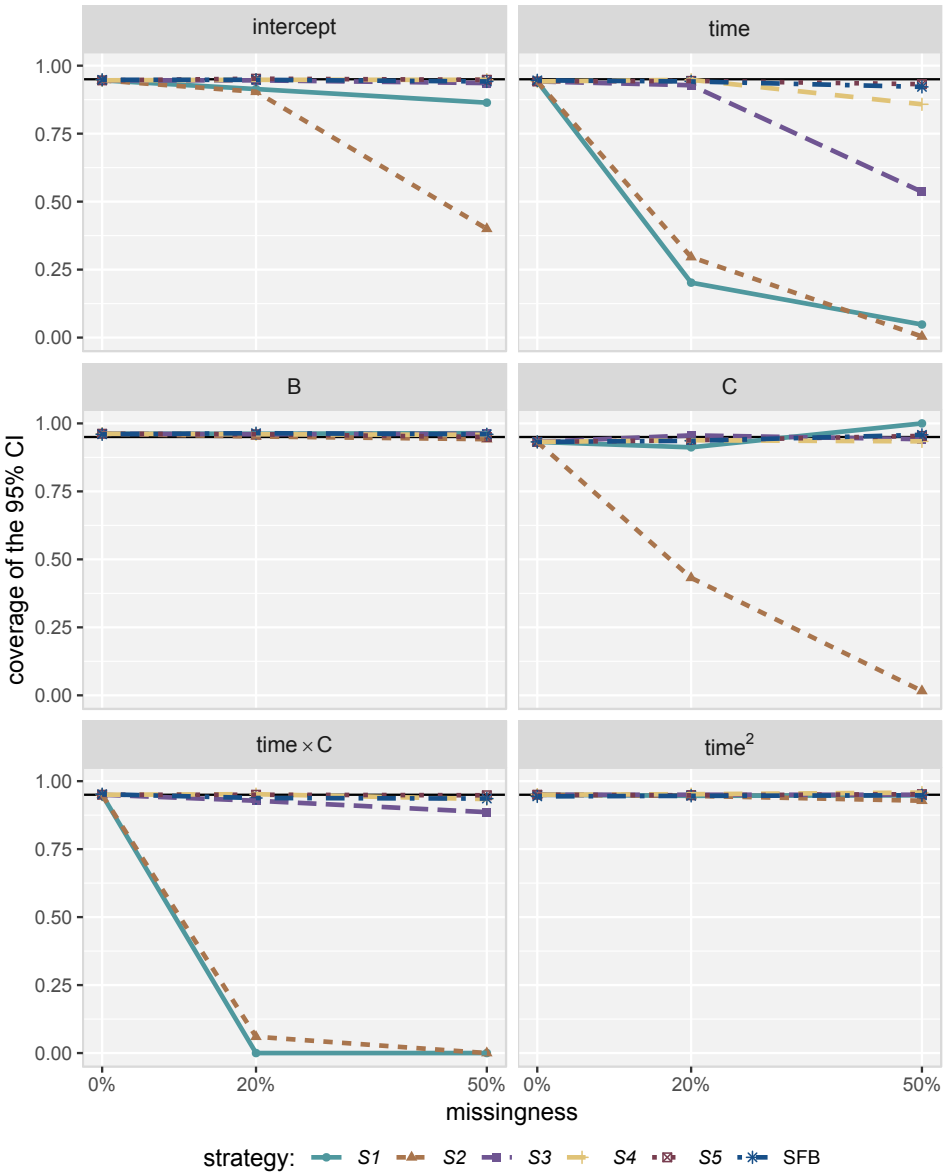


Figure 2.4: Coverage rate in simulation *Scenario 1*, for the five imputation strategies using MICE (*S1*–*S5*) and the sequential fully Bayesian approach (SFB).

data, the computational time increased to 7 – 10 seconds for imputation with **mice** and subsequent analysis with `lmer()` and to 16 – 38 seconds for the SFB

approach. The exact numbers can be found in Table 2.5 in Appendix 2.D.

2.6 Discussion

In this chapter, we have studied the statistical analysis of longitudinal outcomes with incomplete covariates. We have contrasted the popular MICE approach with a full specification of the joint distribution of the outcome and the covariates using the SFB approach. Our theoretical and simulation results suggest that the key in obtaining unbiased estimates is to sample the missing covariates from conditional distributions which are compatible with the other imputation models and with the analysis model of interest.

In MICE the full conditionals are specified directly and the outcome has to be included explicitly in the imputation models. When the outcome is multivariate and the covariates are of mixed type it is not straightforward to derive a functional form of the outcome to include in the corresponding linear predictors that would lead to compatible imputation models.

One way to ensure compatibility is to specify the joint distribution of all variables and derive the full conditionals from there. Bartlett et al. (2015) use a similar idea to extend MICE by proposing imputation using a "substantive model compatible full conditional specification". Another approach (Carpenter and Kenward 2013; Goldstein et al. 2009) is to specify the joint distribution as (latent) multivariate normal and to draw multiple imputations from the resulting conditional distributions. The SFB approach uses a sequence of univariate conditional distributions to specify the joint distribution. This has the advantage that, since the analysis model is part of the conditional distributions, the parameters of interest are estimated simultaneously with the imputation of the missing values and no additional analysis (and pooling) is necessary. Furthermore, the specification of the univariate conditional imputation models allows for a straightforward and very flexible implementation.

The separation of the imputation and analysis step in MI is often considered an advantage because the set of completed datasets can be used for several analyses. The same is possible when data are imputed with the SFB approach: random draws from the posterior samples of the imputed values may be used to form multiple imputed datasets. However, the compatibility of the imputation models with the new analysis models cannot be guaranteed. If the analysis model changes considerably, the imputation has to be re-done. When using the SFB approach this may be time-consuming.

The possibility to include auxiliary variables, which are not part of the analysis model but are used to aid the imputation, is another attractive feature of MI. The SFB approach can be extended to include such auxiliary variables by including them in predictors of the imputation models but omitting them from the predictor of the analysis model or fixing the corresponding regression coefficients in the analysis model to zero. The implied assumption is therefore that the outcome is conditionally independent of the auxiliary variables.

A major reason why MICE is widely used is its availability in many software packages (Sterne et al. 2009; White et al. 2011; Yucel 2011). There are several packages in R that make use of the Bayesian framework for imputation (Quartagno and Carpenter 2015; Zhao and Schafer 2015; Van Buuren and Groothuis-Oudshoorn 2011; Bartlett 2015) but to our knowledge, none of them can currently handle missing baseline covariates of mixed types in a longitudinal setting. With the use of programs like WinBUGS (D. J. Lunn et al. 2000) or JAGS (Plummer 2003) however, the implementation and estimation of the SFB approach is straightforward (see example syntax in Appendix 2.E).

Even though it has been shown that MICE performs well in most standard situations and that the issue of incompatible imputation models is mainly a theoretical issue (Zhu and Raghunathan 2015), our results demonstrate the advantage of working with the full likelihood in more complex situations. This may as well be relevant when using other multivariate models, such as for instance latent mixture models, where imputation with MICE without taking into account the actual structure of the measurement model may lead to poorly imputed values which will distort the analysis. The SFB approach we presented in this chapter can be extended to such other multivariate outcomes in a straightforward and natural way and even allows for settings with multiple outcomes of possibly different types.

Appendix

2.A Variable Description

Table 2.1: Description of the covariates in the Generation R data.

name	description	level	missings (%)
$\texttt{weight}_0$	maternal pre-pregnancy weight in kg	continuous	478 (14.2 %)
$\texttt{weight}_1$	maternal weight in first trimester in kg	continuous	550 (16.3 %)
$\texttt{weight}_2$	maternal weight in second trimester in kg	continuous	124 (3.7 %)
$\texttt{weight}_3$	maternal weight in third trimester in kg	continuous	99 (2.9 %)
`dpa`, `dpb`, `dpc`	dietary patterns (first three components from PCA)	continuous	–
`gender`	gender of the child (boys, girls)	binary	–
`age`	mother's age at intake in years	continuous	–
`height`	mother's height at intake in cm	continuous	1 (<0.1 %)
`parity`	previous births (none vs one or more)	binary	8 (0.2 %)
`educ`	educational level of the mother (low and midlow, midhigh, high)	categorical	45 (1.3 %)
`smoke`	maternal smoking during pregnancy (never, until pregnancy was known, continued during pregnancy)	categorical	250 (7.4 %)
`alc`	maternal alcohol intake during pregnancy (never, until pregnancy was known, continued during pregnancy)	categorical	272 (8.1 %)
`income`	monthly net household income ($\leq$ 2200 €, $>$ 2200 €)	binary	347 (10.28 %)
`gsi`	Global Severity Index (stress measure)	continuous	406 (12.0 %)
`bmi`	maternal BMI at intake $\left(\text{computed as } \frac{\texttt{weight}_0}{(\texttt{height}/100)^2}\right)$	continuous	478 (14.2 %)

2.B Results from the Generation R Data

Table 2.2: Pooled parameter estimates and 95% CIs or posterior means and 2.5% and 97.5% quantiles (SFB) of the analysis of the Generation R data. (*S1*: no outcome, *S2*: baseline outcome, *S3*: mean outcome, *S4*: simple random effects, *S5*: full random effects, SFB: sequential fully Bayesian; continued on page 45)

	S1		*S2*		*S3*	
intercept	10.39	[7.16, 13.61]	10.21	[6.98, 13.45]	2.51	[0.49, 4.53]
gender	0.08	[-0.47, 0.62]	0.03	[-0.50, 0.56]	0.07	[-0.33, 0.47]
dpa	0.20	[-0.13, 0.53]	0.18	[-0.13, 0.50]	0.38	[0.15, 0.61]
dpb	-0.24	[-0.57, 0.10]	-0.22	[-0.54, 0.10]	0.02	[-0.21, 0.25]
dpc	0.26	[-0.07, 0.58]	0.29	[-0.04, 0.63]	0.40	[0.18, 0.61]
time	0.37	[0.37, 0.38]	0.37	[0.37, 0.38]	0.37	[0.37, 0.38]
time^2	0.01	[0.01, 0.01]	0.01	[0.01, 0.01]	0.01	[0.01, 0.01]
age	0.06	[-0.02, 0.14]	0.06	[-0.02, 0.14]	0.03	[-0.02, 0.08]
educ_2	0.36	[-0.77, 1.50]	0.34	[-0.64, 1.31]	0.90	[0.25, 1.54]
educ_3	0.19	[-1.13, 1.50]	0.13	[-0.97, 1.24]	0.77	[0.01, 1.54]
parity	0.28	[-0.35, 0.90]	0.31	[-0.35, 0.98]	0.25	[-0.21, 0.72]
BMI	2.52	[2.42, 2.63]	2.53	[2.43, 2.63]	2.86	[2.81, 2.91]
smoke_2	-0.17	[-1.22, 0.88]	-0.08	[-1.04, 0.88]	0.07	[-0.60, 0.75]
smoke_3	-0.60	[-1.46, 0.26]	-0.65	[-1.63, 0.34]	-0.49	[-1.10, 0.12]
alc_2	-0.63	[-1.51, 0.25]	-0.61	[-1.52, 0.29]	-0.21	[-0.84, 0.42]
alc_3	-0.26	[-0.97, 0.45]	-0.25	[-1.00, 0.50]	0.06	[-0.43, 0.56]
gsi	1.22	[-0.06, 2.50]	1.34	[0.00, 2.68]	0.51	[-0.44, 1.45]
income	0.55	[-0.30, 1.40]	0.61	[-0.30, 1.52]	0.71	[0.14, 1.27]
dpa×time	0.00	[-0.00, 0.01]	0.00	[-0.00, 0.01]	0.01	[0.00, 0.01]
dpb×time	-0.00	[-0.01, 0.00]	-0.00	[-0.01, 0.00]	-0.00	[-0.01, 0.00]
dpc×time	0.00	[-0.00, 0.01]	0.00	[-0.00, 0.01]	0.00	[-0.00, 0.01]

Table 2.2: (Continued from page 44) Posterior means and 95% CIs or 2.5% and 97.5% quantiles (SFB) of the analysis of the Generation R data. (*S1*: no outcome, *S2*: baseline outcome, *S3*: mean outcome, *S4*: simple random effects, *S5*: full random effects, SFB: sequential fully Bayesian)

	S4		*S5*		SFB	
`intercept`	2.42	[0.41, 4.44]	1.71	[-0.20, 3.61]	1.96	[4.17, -0.26]
`gender`	0.05	[-0.34, 0.43]	0.06	[-0.32, 0.43]	0.07	[-0.30, 0.45]
`dpa`	0.34	[0.13, 0.56]	0.30	[0.08, 0.52]	0.32	[0.10, 0.55]
`dpb`	0.03	[-0.20, 0.26]	0.02	[-0.20, 0.24]	0.03	[-0.20, 0.26]
`dpc`	0.38	[0.17, 0.60]	0.41	[0.19, 0.63]	0.39	[0.18, 0.61]
`time`	0.37	[0.37, 0.38]	0.37	[0.37, 0.38]	0.37	[0.37, 0.38]
$\texttt{time}^2$	0.01	[0.01, 0.01]	0.01	[0.01, 0.01]	0.01	[0.01, 0.01]
`age`	0.03	[-0.02, 0.08]	0.03	[-0.02, 0.08]	0.03	[-0.02, 0.08]
$\texttt{educ}_2$	0.88	[0.26, 1.50]	0.94	[0.29, 1.59]	0.92	[0.30, 1.55]
$\texttt{educ}_3$	0.79	[0.07, 1.50]	0.89	[0.14, 1.64]	0.85	[0.12, 1.61]
`parity`	0.28	[-0.14, 0.71]	0.34	[-0.09, 0.76]	0.32	[-0.10, 0.75]
`BMI`	2.87	[2.82, 2.92]	2.90	[2.85, 2.95]	2.89	[2.84, 2.94]
$\texttt{smoke}_2$	0.01	[-0.71, 0.73]	-0.02	[-0.66, 0.62]	-0.01	[-0.70, 0.70]
$\texttt{smoke}_3$	-0.40	[-0.97, 0.18]	-0.57	[-1.13, -0.01]	-0.52	[-1.09, 0.07]
$\texttt{alc}_2$	-0.18	[-0.79, 0.43]	-0.20	[-0.80, 0.40]	-0.22	[-0.81, 0.36]
$\texttt{alc}_3$	0.10	[-0.38, 0.59]	0.15	[-0.33, 0.62]	0.13	[-0.35, 0.61]
`gsi`	0.47	[-0.42, 1.37]	0.56	[-0.28, 1.40]	0.55	[-0.36, 1.47]
`income`	0.63	[0.11, 1.16]	0.64	[0.16, 1.12]	0.71	[0.20, 1.23]
$\texttt{dpa}\times\texttt{time}$	0.00	[-0.00, 0.01]	0.00	[-0.00, 0.01]	0.00	[-0.00, 0.01]
$\texttt{dpb}\times\texttt{time}$	-0.00	[-0.01, 0.00]	-0.00	[-0.01, 0.00]	-0.00	[-0.01, 0.00]
$\texttt{dpc}\times\texttt{time}$	0.00	[-0.00, 0.01]	0.00	[-0.00, 0.01]	0.00	[-0.00, 0.01]

2.C Simulation Settings

Table 2.3: Parameters $\boldsymbol{\beta}$ used for simulation.

	`Intercept`	`time`	`B`	`C`	$\texttt{time}\times\texttt{C}$	$\texttt{time}^2$
Scenario 1	0.97	0.36	1.4	0.25	2.46	0.01
Scenario 2	0.97	0.36	1.4	0.25	2.46	0.35

$$D = \begin{pmatrix} 2.300 & -0.750 & -0.060 \\ -0.750 & 1.550 & 0.004 \\ -0.060 & 0.004 & 0.058 \end{pmatrix}$$

Table 2.4: Parameters $\boldsymbol{\gamma}$ used for simulation.

20% NA	50% NA	20% NA & 50% NA		
$\gamma_0 \approx$ -13.2	$\gamma_0 \approx$ -6.9	$\gamma_1 = 1.0$	$\gamma_2 =$ -3.5	$\gamma_3 = 2.5$
$\gamma_{10} \approx$ -2.8	$\gamma_{10} \approx$ -1.2	$\gamma_{11} = 2.5$	$\gamma_{12} =$ -0.5	
$\gamma_{20} \approx$ -3.0	$\gamma_{20} \approx$ -2.1	$\gamma_{21} = 1.5$	$\gamma_{22} =$ 0.5	

2.D Simulation Results

Table 2.5: Computational times in seconds for simulation in both scenarios.

	Scenario 1			*Scenario 2*		
	0% NA	20% NA	50% NA	0% NA	20% NA	50% NA
S1	0.59	7.77	8.14	0.59	8.67	9.50
S2	0.59	6.89	7.44	0.59	8.62	8.85
S3	0.59	7.70	7.87	0.59	9.58	9.64
S4	0.59	6.82	6.87	0.59	8.72	8.72
S5	0.59	7.29	7.31	0.59	9.30	9.27
SFB	14.99	16.01	23.03	14.71	26.58	37.66

Table 2.6: Simulation results from both scenarios: relative bias, 10 times the mean squared error (MSE) and coverage rate of the 95% CI (*S1*: no outcome, *S2*: baseline outcome, *S3*: mean outcome, *S4*: simple random effects, *S5*: full random effects, SFB: sequential fully Bayesian; continued on page 48)

		full data		20% missing values						50% missing values					
		`lmer()`	SFB	*S1*	*S2*	*S3*	*S4*	*S5*	SFB	*S1*	*S2*	*S3*	*S4*	*S5*	SFB
Scenario 1															
	rel. bias	1.00	1.00	0.95	0.95	1.00	1.00	1.00	1.00	0.88	0.73	0.98	0.99	1.00	1.00
`Intercept`	10 · MSE	0.11	0.11	0.15	0.15	0.12	0.11	0.11	0.12	0.28	0.86	0.13	0.13	0.13	0.13
	coverage	95%	95%	91%	90%	95%	95%	95%	95%	86%	40%	94%	95%	95%	94%
	rel. bias	0.99	0.99	0.10	0.25	0.82	0.91	0.99	0.99	-0.39	-0.61	0.39	0.70	0.96	1.01
`time`	10 · MSE	0.07	0.07	1.14	0.81	0.12	0.09	0.07	0.07	2.65	3.49	0.61	0.24	0.15	0.13
	coverage	94%	95%	20%	30%	93%	95%	94%	94%	5%	0%	54%	86%	93%	92%
	rel. bias	1.00	1.00	0.99	0.98	1.00	1.00	1.00	1.00	0.99	0.97	0.99	0.99	1.00	1.00
`B`	10 · MSE	0.16	0.16	0.20	0.20	0.18	0.17	0.17	0.17	0.22	0.25	0.23	0.23	0.23	0.20
	coverage	96%	96%	96%	95%	96%	96%	96%	96%	96%	94%	96%	96%	95%	96%
	rel. bias	0.99	0.99	1.31	1.74	1.10	1.05	1.00	0.99	1.04	2.59	1.22	1.14	1.01	0.97
`C`	10 · MSE	0.06	0.06	0.12	0.41	0.07	0.07	0.07	0.07	0.04	1.65	0.09	0.09	0.08	0.08
	coverage	93%	93%	91%	43%	96%	94%	94%	94%	100%	2%	94%	93%	95%	96%
	rel. bias	1.00	1.00	0.75	0.85	0.98	1.00	1.00	1.00	0.44	0.67	0.97	1.01	1.00	1.00
`time`×`C`	10 · MSE	0.04	0.04	3.92	1.37	0.07	0.06	0.05	0.05	19.06	6.84	0.15	0.09	0.08	0.08
	coverage	95%	95%	0%	6%	93%	95%	95%	94%	0%	0%	89%	93%	95%	94%
	rel. bias	1.04	1.04	1.01	0.84	1.10	1.03	1.04	1.04	0.97	0.02	1.12	1.02	1.02	1.02
`time`^{2}	10 · MSE	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00
	coverage	95%	94%	95%	95%	95%	95%	95%	95%	95%	93%	95%	96%	95%	95%

Table 2.6: (Continued from page 47) Simulation results from both scenarios: relative bias, 10 times the mean squared error (MSE) and coverage rate of the 95% CI (*S1*: no outcome, *S2*: baseline outcome, *S3*: mean outcome, *S4*: simple random effects, *S5*: full random effects, SFB: sequential fully Bayesian)

		full data		20% missing values						50% missing values					
		`lmer()`	SFB	*S1*	*S2*	*S3*	*S4*	*S5*	SFB	*S1*	*S2*	*S3*	*S4*	*S5*	SFB
Scenario 2															
	rel. bias	1.00	1.00	1.11	1.04	1.05	1.00	1.00	1.00	1.28	1.12	1.14	1.01	1.00	1.00
`Intercept`	10 · MSE	0.11	0.11	0.24	0.14	0.14	0.12	0.12	0.12	0.85	0.28	0.31	0.14	0.14	0.14
	coverage	95%	94%	87%	94%	94%	96%	95%	95%	42%	84%	82%	95%	94%	95%
	rel. bias	0.99	0.99	1.33	1.17	1.00	0.99	1.00	0.99	1.72	1.41	1.01	0.98	0.99	0.99
`time`	10 · MSE	0.07	0.07	0.22	0.12	0.07	0.07	0.07	0.07	0.77	0.30	0.08	0.07	0.07	0.07
	coverage	94%	94%	82%	91%	95%	94%	95%	94%	35%	75%	95%	94%	94%	94%
	rel. bias	1.00	1.00	0.88	0.96	0.93	1.00	1.00	1.00	0.69	0.89	0.81	0.98	1.00	1.00
`B`	10 · MSE	0.16	0.16	0.46	0.21	0.27	0.18	0.18	0.18	2.05	0.47	0.91	0.25	0.25	0.23
	coverage	96%	96%	87%	95%	93%	96%	95%	96%	32%	88%	69%	95%	94%	96%
	rel. bias	0.99	0.99	0.92	1.16	1.00	1.00	0.99	1.00	0.80	1.41	1.05	1.03	1.00	1.00
`C`	10 · MSE	0.06	0.06	0.06	0.08	0.06	0.06	0.06	0.06	0.08	0.18	0.07	0.07	0.07	0.07
	coverage	93%	93%	97%	92%	96%	93%	95%	94%	95%	82%	96%	94%	94%	94%
	rel. bias	1.00	1.00	0.91	0.95	0.99	1.00	1.00	1.00	0.78	0.87	0.98	0.99	1.00	1.00
`time`×`C`	10 · MSE	0.04	0.04	0.53	0.19	0.05	0.05	0.05	0.05	2.93	1.00	0.08	0.05	0.05	0.05
	coverage	95%	95%	45%	78%	94%	95%	95%	95%	0%	13%	90%	95%	95%	95%
	rel. bias	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.01	1.00	1.00	1.00	1.00
`time`^{2}	10 · MSE	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	coverage	95%	95%	96%	96%	95%	95%	95%	95%	95%	95%	95%	95%	95%	94%

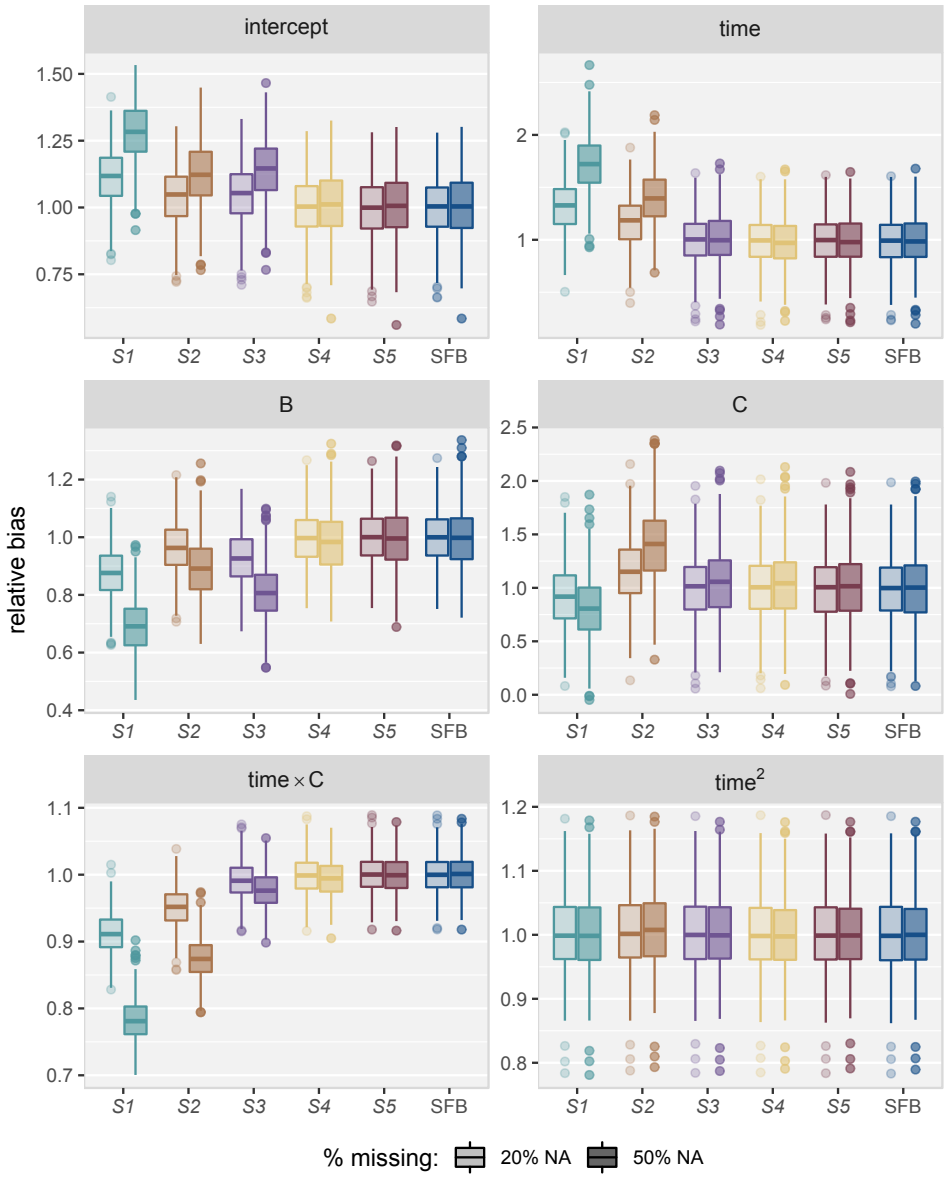


Figure 2.5: Relative bias in simulation *Scenario 2*, for the five imputation strategies using MICE (*S1*–*S5*) and the sequential fully Bayesian approach (SFB).

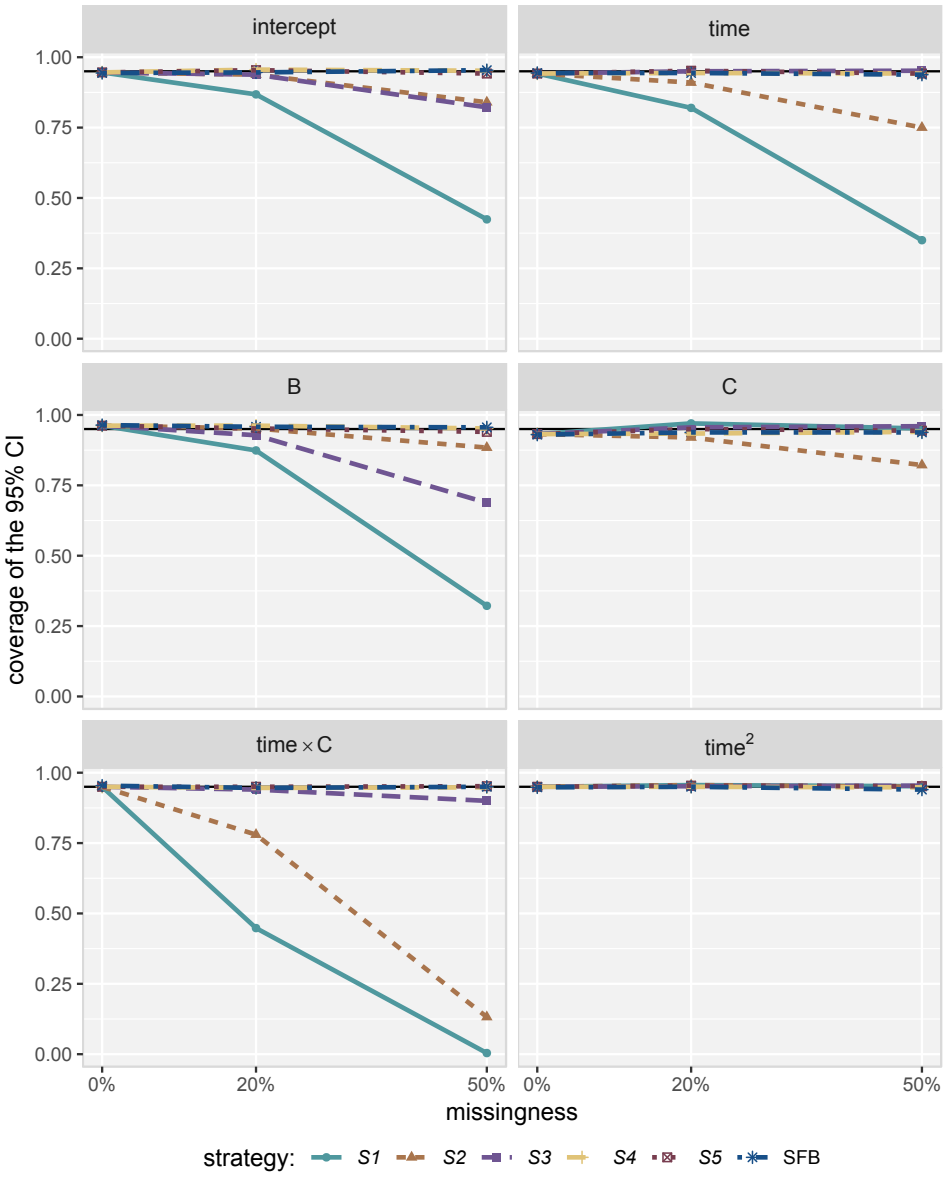


Figure 2.6: Coverage rate in simulation *Scenario 2*, for the five imputation strategies using MICE (*S1–S5*) and the sequential fully Bayesian approach (SFB).

2.E JAGS Syntax for Simulation Scenario 2

It is convenient to implement the SFB approach using hierarchical centring. This means that the fixed effects enter the linear predictor through the random effects, i.e., the random effects are not centred around zero but around the fixed effects.

Data / Notation:

- `TN`: number of observations in the dataset
- `N`: number of individuals
- `priorR`: 3×3 diagonal matrix of NA's

```
model {
  for(j in 1:TN){
    # Linear mixed effects model for y
    y[j] ~ dnorm(mu.y[j], tau.y)
    mu.y[j] <- inprod(b[subj[j], ], Z[j, ]) # lin. predictor with
                                            # hierarchic. centring
                                            # specification
  }

  for(i in 1:N){
    b[i, 1:3] ~ dmnorm(mu.b[i, ], inv.D[ , ])
    mu.b[i, 1] <- beta[1] + beta[2] * B[i] + beta[3] * C[i]
    mu.b[i, 2] <- beta[4] + beta[5] * C[i]
    mu.b[i, 3] <- beta[6]
  }

  # Priors for analysis model: fixed effects
  for(k in 1:6){
    beta[k] ~ dnorm(0, 0.1)
  }
  tau.y ~ dgamma(0.001, 0.001)
  sigma.y <- sqrt(1/tau.y)

  # Priors for analysis model: random effects
  for(k in 1:3){
    priorR.invD[k,k] ~ dgamma(0.1, 0.01)
  }
  inv.D[1:3, 1:3] ~ dwish(priorR[, ], 3)
  D[1:3, 1:3] <- inverse(inv.D[, ])
```

```
  # imputation models
  for(i in 1:N){
    # normal regression for C
    C[i] ~ dnorm(mu.C[i], tau.C)
    mu.C[i] <- alpha[1]

    # binary regression for B
    B[i] ~ dbern(p.B[i])
    logit(p.B[i]) <- alpha[2] + alpha[3] * C[i]
  }

  # Priors for imputation of C
  for(k in 1:1){
    alpha[k] ~ dnorm(0, 0.001)
  }
  tau.C ~ dgamma(0.01, 0.01)

  # Priors for imputation of B
  for(k in 2:3){
    alpha[k] ~ dnorm(0, 4/9)
  }
}
```

References

Bartlett, J. W. (2015). *smcfcs: Multiple Imputation of Covariates by Substantive Model Compatible Fully Conditional Specification*. R package version 1.0.0. URL: http://CRAN.R-project.org/package=smcfcs.

Bartlett, J. W. et al. (2015). "Multiple imputation of covariates by fully conditional specification: accommodating the substantive model". *Statistical Methods in Medical Research*, **24**(4):462–487. DOI: 10.1177/0962280214521348.

Bates, D. et al. (2015). "Fitting Linear Mixed-Effects Models Using lme4". *Journal of Statistical Software*, **67**(1):1–48. DOI: 10.18637/jss.v067.i01.

Carpenter, J. R. and M. G. Kenward (2013). *Multiple Imputation and its Application*. John Wiley & Sons, Ltd. DOI: 10.1002/9781119942283.

Chen, B., Y. Y. Grace, and R. J. Cook (2010). "Weighted generalized estimating functions for longitudinal response and covariate data that are missing at random". *Journal of the American Statistical Association*, **105**(489):336–353. DOI: 10.1198/jasa.2010.tm08551.

Chen, B. and X.-H. Zhou (2011). "Doubly robust estimates for binary longitudinal data analysis with missing response and missing covariates". *Biometrics*, **67**(3):830–842. DOI: 10.1111/j.1541-0420.2010.01541.x.

Chen, M.-H. and J. G. Ibrahim (2001). "Maximum likelihood methods for cure rate models with missing covariates". *Biometrics*, **57**(1):43–52. DOI: 10.1111/j.0006-341X.2001.00043.x.

Donders, A. R. T. et al. (2006). "Review: A gentle introduction to imputation of missing values". *Journal of Clinical Epidemiology*, **59**(10):1087–1091. DOI: 10.1016/j.jclinepi.2006.01.014.

Garrett, E. S. and S. L. Zeger (2000). "Latent Class Model Diagnosis". *Biometrics*, **56**(4):1055–1067. DOI: 10.1111/j.0006-341X.2000.01055.x.

Gelman, A., X.-L. Meng, and H. Stern (1996). "Posterior predictive assessment of model fitness via realized discrepancies". *Statistica Sinica*, **6**(4):733–760.

Goldstein, H. et al. (2009). "Multilevel models with multivariate mixed response types". *Statistical Modelling*, **9**(3):173–197. DOI: 10.1177/1471082X0800900301.

Ibrahim, J. G., M.-H. Chen, and S. R. Lipsitz (2002). "Bayesian methods for generalized linear models with covariates missing at random". *Canadian Journal of Statistics*, **30**(1):55–78. DOI: 10.2307/3315865.

Jaddoe, V. W. et al. (2012). "The Generation R Study: design and cohort update 2012". *European Journal of Epidemiology*, **27**(9):739–756. DOI: 10.1007/s10654-012-9735-1.

Janssen, K. J. et al. (2010). "Missing covariate data in medical research: To impute is better than to ignore". *Journal of Clinical Epidemiology*, **63**(7):721–727. DOI: 10.1016/j.jclinepi.2009.12.008.

Knol, M. J. et al. (2010). "Unpredictable bias when using the missing indicator method or complete case analysis for missing confounder values: an empirical example". *Journal of Clinical Epidemiology*, **63**(7):728–736. DOI: 10.1016/j.jclinepi.2009.08.028.

Lesaffre, E. M. and A. B. Lawson (2012). *Bayesian Biostatistics*. John Wiley & Sons. DOI: 10.1002/9781119942412.

Little, R. J. and D. B. Rubin (2002). *Statistical Analysis with Missing Data*. Hoboken, New Jersey: John Wiley & Sons, Inc. DOI: 10.1002/9781119013563.

Lunn, D. J. et al. (2000). "WinBUGS – a Bayesian modelling framework: Concepts, structure, and extensibility". *Statistics and Computing*, **10**(4):325–337. DOI: 10.1023/A:1008929526011.

Molenberghs, G. and M. G. Kenward (2007). *Missing Data in Clinical Studies*. Statistics in Practice. Wiley. DOI: 10.1002/9780470510445.

Moons, K. G. et al. (2006). "Using the outcome for imputation of missing predictor values was preferred". *Journal of Clinical Epidemiology*, **59**(10):1092–1101. DOI: 10.1016/j.jclinepi.2006.01.009.

Plummer, M. (2003). "JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling". *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*. Ed. by K. Hornik, F. Leisch, and A. Zeileis. ISSN: 1609-395X.

Quartagno, M. and J. R. Carpenter (2015). *jomo: A package for Multilevel Joint Modelling Multiple Imputation*. URL: http://CRAN.R-project.org/package=jomo.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: http://www.R-project.org/.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley Series in Probability and Statistics. Wiley. DOI: 10.1002/9780470316696.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis. ISBN: 9780412040610.

Seaman, S. R., J. Galati, et al. (2013). "What Is Meant by "Missing at Random"?" *Statistical Science*, **28**(2):257–268. DOI: 10.1214/13-STS415.

Sterne, J. A. et al. (2009). "Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls". *BMJ: British Medical Journal*, **338**:DOI: 10.1136/bmj.b2393.

Stubbendick, A. L. and J. G. Ibrahim (2003). "Maximum Likelihood Methods for Nonignorable Missing Responses and Covariates in Random Effects Models". *Biometrics*, **59**(4):1140–1150. DOI: 10.1111/j.0006-341X.2003.00131.x.

Van Buuren, S. (2012). *Flexible Imputation of Missing Data*. Chapman & Hall/CRC Interdisciplinary Statistics. Taylor & Francis.

Van Buuren, S. and K. Groothuis-Oudshoorn (2011). "mice: Multivariate Imputation by Chained Equations in R". *Journal of Statistical Software*, **45**(3):1–67. DOI: 10.18637/jss.v045.i03.

White, I. R., P. Royston, and A. M. Wood (2011). "Multiple imputation using chained equations: Issues and guidance for practice". *Statistics in Medicine*, **30**(4):377–399. DOI: 10.1002/sim.4067.

Yucel, R. M. (2011). "State of the multiple imputation software". *Journal of Statistical Software*, **45**(1):DOI: 10.18637/jss.v045.i01.

Zhao, J. H. and J. L. Schafer (2015). *pan: Multiple imputation for multivariate panel or clustered data*. R package version 1.3. URL: https://CRAN.R-project.org/package=pan.

Zhu, J. and T. E. Raghunathan (2015). "Convergence Properties of a Sequential Regression Multiple Imputation Algorithm". *Journal of the American Statistical Association*, **110**(511):1112–1124. DOI: 10.1080/01621459.2014.948117.

3 Dietary Patterns and Gestational Weight Gain

This chapter is based on

Myrte J. Tielemans, Nicole S. Erler, Elisabeth T.M. Leermakers, Marion van den Broek, Vincent W.V. Jaddoe, Eric A.P. Steegers, Jessica C. Kiefte-de Jong and Oscar H. Franco. A priori and a posteriori dietary patterns during pregnancy and gestational weight gain: The Generation R Study. *Nutrients*, 2015; **7**(11), 9383–9399. doi:10.3390/nu7115476

Abstract

Abnormal gestational weight gain (GWG) is associated with adverse pregnancy outcomes. We examined whether dietary patterns are associated with GWG in 3374 pregnant women participating in a population-based cohort study in the Netherlands. Dietary intake during pregnancy was assessed with food-frequency questionnaires. Three *a posteriori*-derived dietary patterns were identified using factor analysis: a "Vegetable, oil and fish", a "Nuts, high-fibre cereals and soy", and a "Margarine, sugar and snacks" pattern. An *a priori*-defined dietary pattern was based on national dietary recommendations. Weight was repeatedly measured around 13, 20 and 30 weeks of pregnancy; pre-pregnancy and maximum weight were self-reported. Normal-weight women with 1 SDS higher score on the "Nuts, high-fibre cereals and soy" pattern gained weight slightly slower, while normal-weight women with higher adherence to the "Margarine, sugar and snacks" pattern had higher weight (0.3kg; 95% CI [0.07, 0.52]) throughout pregnancy. Normal-weight women with higher adherence to the "Vegetable, oil and fish" pattern had 16 g/week (95% CI [6.5, 25.6]) higher early-pregnancy GWG. The *a priori*-defined pattern was not associated with GWG. In conclusion, specific dietary patterns may play a role in early pregnancy but were not consistently associated with GWG in our data.

3.1 Introduction

Abnormal maternal weight gain during pregnancy (i.e., too little or too much) has been associated with unfavourable pregnancy outcomes in both mother and child. Insufficient gestational weight gain (GWG) is associated with both preterm birth and low birthweight (Han et al. 2011), and excessive GWG increases the risk of giving birth to large-for-gestational-age infants (Kim et al. 2014). Excessive GWG is also associated with maternal pregnancy complications, including hypertensive disorders (Johnson et al. 2013; Gaillard et al. 2013) and gestational diabetes (Hedderson et al. 2010), which can increase the risk of the mother developing cardiometabolic diseases after pregnancy (Brown et al. 2013; Bellamy et al. 2009).

Energy intake during pregnancy is associated with GWG (Gaillard et al. 2013; Streuling et al. 2011), but literature is scarce on whether GWG could be influenced by dietary composition. Some studies have examined the influence of food groups on GWG (Stuebe et al. 2009; Olafsdottir et al. 2006; Martins and Benicio 2011). These studies found no association of fruit or vegetable intake with GWG (Stuebe et al. 2009; Martins and Benicio 2011) but unhealthier foods (e.g., sweets and processed foods) were associated with higher prevalence of excessive GWG (Stuebe et al. 2009; Olafsdottir et al. 2006; Martins and Benicio 2011). Weight gain during pregnancy involves both maternal components (e.g., blood volume increase, fat accretion) and fetal components (e.g., weight of the fetus, amniotic fluid) (Pitkin 1976). Therefore, the effect of diet on weight gain may differ between pregnant and non-pregnant women.

Assessing overall diet in relation to GWG has several advantages over studying individual foods or nutrients. First, the intakes of different nutrients are often highly correlated, which complicates the assessment of individual nutrients (Hu 2002). Second, possible associations between nutrient intake and GWG might be affected by biological interactions between nutrients (Hu 2002). For these reasons, evaluating diet using a dietary pattern approach may improve our understanding of which patterns of diet are most beneficial during pregnancy. Moreover, this approach can facilitate future food-based dietary guidelines (World Health Organization and Food and Agriculture Organization of the United Nations 1998).

Only a few studies have focused on the relationship between dietary patterns and GWG (Uusitalo et al. 2009; Shin et al. 2014; Rifas-Shiman et al. 2009; Hillesund et al. 2014). However, no study evaluated dietary patterns and longitudinal development of weight during pregnancy. We hypothesized that specific dietary patterns may influence the development of maternal weight during pregnancy. In addition, dietary patterns are likely to differ between countries and populations

(Hu 2002), hence, it is important to identify country-specific dietary patterns that may be associated with GWG.

Therefore, the purpose of our study was to determine whether *a priori*-defined and *a posteriori*-derived dietary patterns are associated with weight development during pregnancy and GWG during different phases in pregnancy in Dutch women participating in a population-based cohort.

3.2 Experimental Section

3.2.1 Study Design

This study was embedded in the Generation R Study, a population-based prospective cohort from fetal life onwards in Rotterdam, the Netherlands. Details of this study have been described previously (Kruithof et al. 2014). Briefly, pregnant women with an expected delivery date between April 2002 and January 2006, living in the urban area around Rotterdam, were approached to participate. All participants provided written informed consent. The study was conducted according to the World Medical Association Declaration of Helsinki and was approved by the Medical Ethics Committee, Erasmus Medical Center Rotterdam (the Netherlands, MEC 198.782.2001.31).

3.2.2 Population of Analysis

For the current analysis, women of Dutch ancestry who entered the Generation R Study during pregnancy (n = 4097) were included. Women of non-Dutch ancestry were excluded from the analysis since the dietary assessment method that was used was designed to evaluate a Dutch diet. Furthermore, women with missing dietary information (n = 538) were excluded and only women with singleton live births (n = 3479) were considered. For five women weight during pregnancy was not measured and women who were underweight before pregnancy (body mass index (BMI) < 18.5 kg/m^2; n = 100) were excluded, leaving 3374 women for the current analysis (Figure 3.1).

3.2.3 Dietary Assessment

Dietary intake in early pregnancy was assessed at enrolment (median: 13.4 weeks of gestation, Q3 – Q1: 12.2 – 15.5) using a 293-item semi-quantitative food-frequency questionnaire (FFQ) that covered dietary intake over the preceding three months. The FFQ contained questions regarding foods that are frequently consumed in a traditional Dutch diet, their consumption frequency, portion size

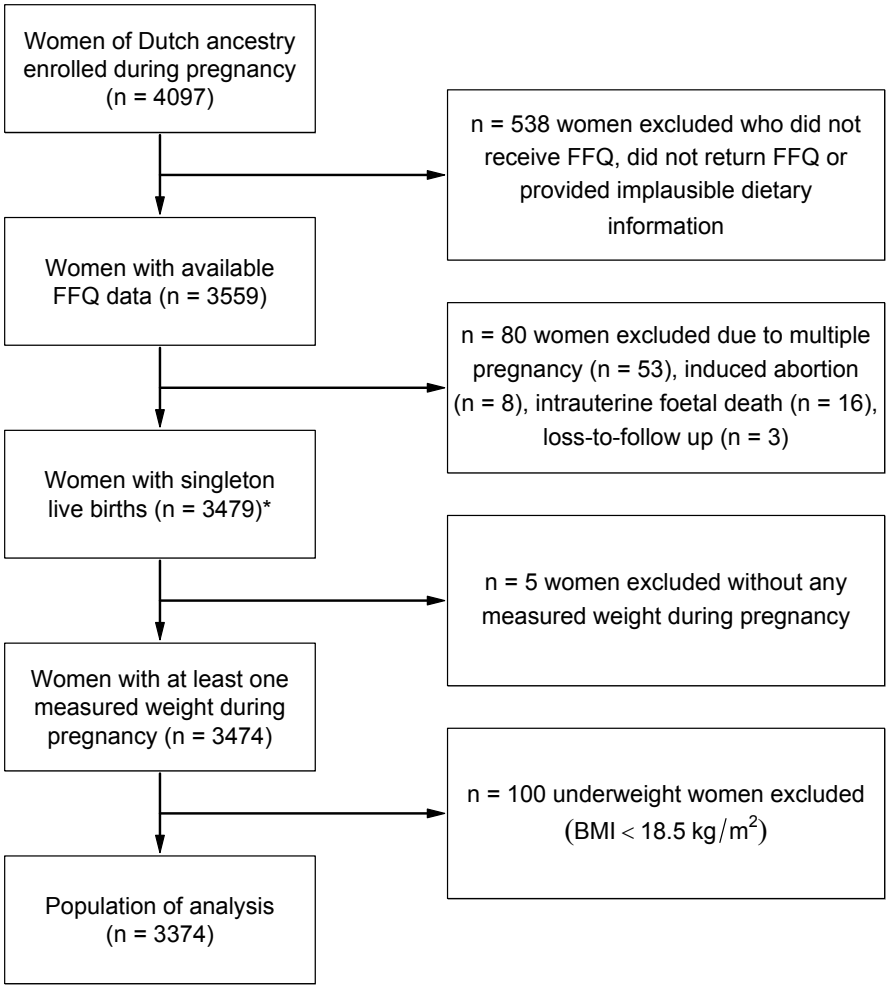


Figure 3.1: Flow chart of the study population: the Generation R Study (2002–2006). * Population in which the *a posteriori*-derived dietary patterns were determined. (BMI: body mass index; FFQ: food-frequency questionnaire)

(Donders-Engelen and Heijden 2003), preparation methods, and additions to foods. The average daily intake of energy and nutrients was calculated using the Dutch food-composition table (Netherlands Nutrition Center 2006). The FFQ was designed for and validated in an elderly population (Klipstein-Grobusch et al. 1998), and has additionally been validated against three 24-h dietary recalls in 71 Dutch pregnant women who visited a midwifery in Rotterdam. The intra-class correlation coefficients for energy-adjusted macronutrients ranged between 0.48 and 0.68.

A Posteriori-Derived Dietary Patterns

We used three *a posteriori*-derived dietary patterns, which had been identified previously by factor analysis using the principal component method and Varimax rotation (Hu 2002; Kaiser 1958), and are described in detail elsewhere (Van den Broek et al. 2015). Briefly, the 293 individual food items from the FFQ were aggregated into 23 food groups. Subsequently, factors that had an eigenvalue of ≥ 1.5 were extracted and are regarded as dietary patterns (Jolliffe 2011). The factor loadings, which describe the contribution of each food group to each of the patterns, are presented in Table 3.1. Finally, factor scores (i.e., adherence scores) for each participant and each pattern were determined by calculating the individual sum of the intake of the food groups, weighting them with their factor loadings and standardizing those weighted sums to have mean zero and standard deviation one (standard deviation score). A higher factor score indicates that a woman's diet was closer to that dietary pattern.

Three *a posteriori*-derived dietary patterns were thereby identified, namely a "Vegetable, oil and fish" pattern, a "Nuts, high-fibre cereals and soy" pattern and a "Margarine, sugar and snacks" pattern, together explaining 25.8% of the variance in maternal dietary intake.

A Priori-Defined Dietary Pattern

The *a priori*-defined dietary pattern was based on the Dutch Healthy Diet Index (Van Lee et al. 2012). This index was developed to measure adherence to the Dutch guidelines for a healthy diet (Health Council of the Netherlands 2006) and consisted of ten components: physical activity, vegetable, fruit, dietary fibre, fish, saturated fatty acids, trans-fatty acids, consumption of acidic drinks and foods, sodium, and alcohol. We omitted the components physical activity, trans-fatty acids, and the consumption of acidic drinks and foods because this information had not been collected. Furthermore, we did not include the alcohol component as alcohol abstinence is recommended during pregnancy. The score of each com-

ponent ranged between 0 and 10 points, resulting in a total score ranging from 0 to 60 points. A higher score on the Dutch Healthy Diet Index corresponds with a higher adherence to the 2006 Dutch healthy diet guidelines and thus reflects a healthier diet. Finally, to facilitate comparison between all dietary patterns, we standardized the "Dutch Healthy Diet Index" pattern to a standard deviation score.

3.2.4 Maternal Weight Gain

Information on pre-pregnancy weight was collected at enrolment using a questionnaire and was used to calculate pre-pregnancy BMI (kg/m^2). Women visited our research centre three times at median gestational ages of 12.9 (Q3 – Q1: 12.1 – 14.4) weeks (first visit), 20.4 (Q3 – Q1: 19.9 — 21.1) weeks (second visit), and 30.2 (Q3 – Q1: 29.9 – 30.8) weeks (third visit). During each visit, maternal height and weight were measured without shoes and heavy clothing. Six weeks after childbirth, women were asked to report their highest weight during pregnancy using a questionnaire, which we used as maximum weight in pregnancy.

Gestational Weight Gain during Different Phases in Pregnancy

GWG in different phases of pregnancy was calculated for three consecutive periods, namely early-pregnancy GWG (calculated as weight at the first visit minus pre-pregnancy weight, divided by follow-up duration (g/week), n = 2425), mid-pregnancy GWG (calculated as weight at the second visit minus weight the first visit, divided by follow-up duration (g/week), n = 2748), and late-pregnancy GWG (calculated as weight at the third visit minus weight at the second visit, divided by follow-up duration (g/week), n = 3158). GWG until early-third trimester was calculated as weight at the third visit minus pre-pregnancy weight, divided by follow up duration (g/week, n = 2815).

3.2.5 Covariates

Several maternal sociodemographic and lifestyle characteristics were considered as potential confounders. We obtained information from prenatal questionnaires that were sent in different trimesters regarding maternal age, educational level (Van Rossem et al. 2009), household income ($\leq$ 2200 vs >2200 euro/month), parity (no child vs $\geq$ 1 child), pre-pregnancy weight, pre-existing comorbidities, vomiting, smoking or alcohol consumption (both categorized as never during pregnancy, stopped when pregnancy was known, or continued throughout pregnancy), folic acid supplementation (started periconceptionally, started first 10 weeks, or no supplementation), energy intake, and stress during pregnancy (using the Global

Severity Index (Derogatis and Spencer 1993)). To calculate pre-pregnancy BMI, height was measured at enrolment. Gestational age was determined based on ultrasound examination, and during the third visit an ultrasound was performed to estimate fetal weight. Information on fetal sex was obtained from delivery reports.

3.2.6 Statistical Analyses

In order to adequately estimate the relationship between *a posteriori*-derived dietary patterns and trajectories of gestational weight in the presence of incomplete covariates, we performed a longitudinal analysis using linear mixed modelling in the Bayesian framework. This method has been described in detail previously (Erler, Rizopoulos, Rosmalen, et al. 2016) and allows for simultaneous analysis and imputation of missing values. Briefly, by modelling the joint distribution of exposure, outcome and covariates, all available information is used to impute the missing values, and parameter estimates are obtained that take into account the added uncertainty due to the missing values.

We considered two sets of possible confounders in the analysis. Model 1 was adjusted for gestational age at the mid-point of the follow-up interval and pre-pregnancy BMI. Model 2 was further adjusted for age, educational level, household income, parity, smoking during pregnancy, alcohol consumption during pregnancy, stress during pregnancy, and fetal sex. The selection of potential confounders was based on factors found in the literature and on a change of at least 10% in effect estimate in a preliminary analysis assessing the association of dietary patterns with GWG until early-third trimester. As GWG is related to BMI (Institute of Medicine and National Research Council 2009) and preliminary analyses showed significant interaction terms for the "Vegetable, oil and fish" pattern and the "Nuts, high-fibre cereals and soy" pattern with pre-pregnancy BMI, we stratified all analyses (including the Bayesian linear mixed model) on the basis of weight status (normal-weight: BMI < 25 kg/m^2 and overweight: BMI ≥ 25 kg/m^2).

In the Bayesian linear mixed model, all main effects from Model 2 (except gestational age at the mid-point of the follow-up interval), interaction terms between the dietary pattern variables and a linear and quadratic effect for gestational age were included in the fixed effects structure. Correlation between weight measurements within each individual was modelled by including random effects for the intercept and slope (for gestational age) into the model. For this analysis, the reported parameter estimates and 95% credible intervals were obtained by taking the mean and 2.5% and 97.5% quantiles of the posterior samples.

To analyse the association of the *a priori*-defined and *a posteriori*-derived dietary

patterns with GWG during different phases in pregnancy, specifically GWG during each trimester, until early third trimester and maximal GWG, we performed multivariable linear regression analysis on imputed datasets. For these analyses, ten imputed values were drawn from the posterior samples of each of the incomplete covariates derived in the Bayesian analysis described above. Missing observations of gestational weight were not imputed. The reported results from the cross-sectional models were pooled over all ten completed datasets using Rubin's rules (Rubin 1987).

Sensitivity Analyses

To test the stability of the results, four sensitivity analyses for the association between dietary patterns and GWG until early-third trimester were performed in Model 2. First, because energy intake may be an intermediate factor in the association of maternal diet with GWG, we further adjusted for energy intake (kcal/day). Second, we excluded women with pre-existing comorbidities (n = 182) and women with hypertensive complications in pregnancy (Coolman et al. 2010) or gestational diabetes (n = 272) since these conditions may influence both dietary intake and GWG. Third, we excluded women who reported vomiting more than once per week during the three months prior to enrolment (n = 421), since this might alter dietary intake and GWG. Moreover, we explored effect modification of the association between dietary patterns and GWG with educational level and household income.

Additionally, we evaluated whether the associations of dietary patterns with GWG would markedly change when using self-reported maximum weight during pregnancy instead of measured weight at the third visit (n = 1917). Furthermore, we explored long-term maternal weight gain and evaluated whether this long-term weight gain differed in women with inadequate, adequate or excessive GWG using Analysis of Variance (ANOVA). The cut-off values of adequate weekly GWG were 0.35 — 0.50 kg/week for normal weight women, 0.23 — 0.33 kg/week for overweight women, and 0.17 — 0.27 kg/week for obese women (Institute of Medicine and National Research Council 2009). Finally, we calculated the correlation between weight at the third visit and weight 6 years after childbirth.

All statistical analyses were performed in SPSS version 21.0 (IBM Corp., Armonk, NY, USA) or R version 3.2.1 (R Foundation for Statistical Computing, Vienna, Austria) and JAGS version 3.4.0 (Plummer 2003).

Table 3.1: Factor loadings food groups in the *a posteriori*-derived dietary patterns (Van den Broek et al. 2015). Food groups that are considered to have a strong association with a dietary pattern (factor loading ≥ 0.2 or ≤ -0.2) are shown in bold. The three factor loadings with the highest positive factor loading are used to name the dietary pattern and have grey background.

	Dietary Pattern		
Food Group	Vegetable, Oil and Fish	Nuts, High-Fibre Cereals and Soy	Margarine, Sugar and Snacks
potatoes and other tubers	0.05	**-0.53**	**0.21**
vegetables	**0.78**	0.17	-0.03
fruits	0.13	**0.37**	0.02
dairy products—high fat	**0.26**	**-0.26**	**0.29**
dairy products—low fat	-0.15	**0.29**	0.16
cereals—high fiber	**0.24**	**0.43**	**0.36**
cereals—low fiber	**0.23**	-0.16	**0.25**
meat and meat products	0.08	**-0.54**	**0.33**
fish and shellfish	**0.45**	**0.24**	-0.11
eggs and egg products	**0.27**	0.05	0.19
vegetable oils	**0.74**	0.08	-0.12
margarine and butter	-0.06	-0.03	**0.61**
sugar and confectionary and cakes	-0.11	0.13	**0.56**
snacks	0.05	0.08	**0.4**
coffee and tea	**0.28**	**0.34**	0.1
sugar-containing beverages	-0.14	**-0.28**	**0.29**
light soft drinks	0.13	**0.28**	0.02
alcoholic beverages	**0.35**	0	-0.04
condiments and sauces	0.05	-0.09	**0.39**
soups and bouillon	0.19	-0.02	0.15
nuts, seeds and olives	0.03	**0.64**	**0.3**
soy products	0.01	**0.39**	-0.1
legumes	**0.44**	-0.02	0.07

3.3 Results

3.3.1 Study Population

Baseline characteristics for normal-weight women (n = 2544; 75%) and overweight women (n = 830; 25%) are presented in Table 3.2. The mean score ± SD on the Dutch Healthy Diet Index was 32 ± 8 and ranged from 8 to 59. Overall, 43% of women had excessive GWG (n = 826); excessive GWG was found in 37% of the normal-weight women (n = 557) and in 63% of the overweight women (n = 269).

3.3.2 Dietary Patterns and Trajectories of Gestational Weight

Results from the Bayesian linear mixed models for normal-weight and overweight women are summarized in Table 3.3 and visualized in Figure 3.2. The figure shows the expected gestational weight over time with corresponding 95% credible intervals under six different scenarios. In each scenario the expected weight, either for normal-weight or overweight women, is shown for a -0.5 SDS and 0.5 SDS adherence to one of the *a posteriori*-derived dietary patterns, while the adherence to the other patterns and confounders are held constant at reference values (0 SDS for adherence to other dietary patterns, and population median and reference category for continuous and categorical confounders, respectively).

In both normal-weight and overweight women, there was no evidence of either main effects of diet or interaction with gestational age for most dietary patterns (Table 3.3). Only adherence to the "Margarine, sugar and snacks" pattern was associated with higher weight in normal-weight women (0.30; 95% CI [0.07, 0.52]) and the "Nuts, high-fibre cereals and soy" pattern was associated with slightly slower weight gain in normal-weight women (-0.01; 95% CI [-0.02, -0.00]).

3.3.3 Dietary Patterns and Gestational Weight Gain in Different Phases in Pregnancy

For normal-weight women a 1 SDS higher adherence to the "Vegetable, oil and fish" pattern was associated with a 16 g/week (95% CI [6.5, 25.6]) greater early-pregnancy GWG when adjusting for lifestyle and sociodemographic variables (Table 3.4). We observed no such association in overweight women (Table 3.5). The "Nuts, high-fibre cereals and soy" pattern was associated with lower early-pregnancy GWG in Model 1 in both normal-weight and overweight women. However, after additional adjustment (Model 2) this pattern was no longer significantly associated with early-pregnancy GWG. Neither the "Margarine, sugar and snacks"

pattern nor the "Dutch Healthy Diet Index" pattern was associated with early-pregnancy GWG.

No significant associations were found for any of the dietary patterns with mid-pregnancy GWG in normal-weight or overweight women (Tables 3.6 and 3.7). Tables 3.8 and 3.9 show that in normal-weight women, only the "Nuts, high-fibre cereals and soy" pattern was inversely associated with late-pregnancy GWG in Model 1, but these results largely attenuated after adjustment for sociodemographic and lifestyle factors (Model 2). In overweight women, none of the dietary patterns was significantly associated with late-pregnancy GWG.

3.3.4 Sensitivity Analyses

The results of the sensitivity analyses are presented in Tables 3.10 – 3.14. In line with the results from early-pregnancy GWG, normal-weight women with a 1 SDS higher adherence to the "Vegetable, oil and fish" pattern had 6.7 g/week (95% CI [0.9, 12.5]) higher GWG until the early-third trimester, whereas no association was found in overweight women. The other dietary patterns were not associated with GWG until the early-third trimester (Table 3.10).

Additional adjustment for energy intake resulted in attenuation of the effect estimate of the "Vegetable, oil and fish" pattern with GWG until early-third trimester. The results did not alter greatly after exclusion of women who vomited more than once per week (Table 3.12) or exclusion of women with pre-existing comorbidities or pregnancy complications (Table 3.13). The evaluation of maximum GWG showed results that were in line with the results of the previous analyses (Table 3.14).

The association between dietary patterns and GWG was not modified by educational level or household income.

Six years after childbirth, women had gained on average 3.4 kg (Q1, Q3: 0.4, 7.0) compared to their pre-pregnancy weight (n = 2247). The median (Q1, Q3) long-term weight gain was significantly different between the categories of GWG adequacy: women with inadequate GWG gained 2.2 kg (-0.6, 5.2), those with adequate GWG gained on average 2.6 kg (0.2, 5.2), and women with excessive GWG were 4.6 kg (1.4, 8.8) heavier (F-test 27.5, p-value < 0.001). The weight 6 years after childbirth was highly correlated with the weight at the third visit in pregnancy (R = 0.85; p-value < 0.001).

Table 3.2: Subject characteristics. Values represent n (%) for categorical variables, and mean (± SD) or median (1^{st} quartile, 3^{rd} quartile) for continuous variables. Missing data: educational level (1.3%), household income (10.3%), parity (0.2%), pre-pregnancy BMI (14.2%), smoking during pregnancy (7.4%), alcohol consumption during pregnancy (8.1%), stress during pregnancy (12.0%), gestational weight gain (43.2%), adequacy of gestational weight gain (43.2%). Numbers may not add up to total due to rounding after imputation.

Subject Characteristics	**normal-weight** (n = 2544)	**overweight** (n = 830)
age (years)	31.6 (±4.3)	31.0 (±4.4)
educational level		
low and midlow	307 (12.1%)	201 (24.2%)
midhigh	1283 (50.4%)	436 (52.5%)
high	954 (37.5%)	193 (23.3%)
household income		
<2200 euro/month	620 (24.4%)	266 (32.1%)
≥2200 euro/month	1924 (75.6%)	564 (67.9%)
parity		
0	1554 (61.1%)	465 (56.0%)
≥ 1	990 (38.9%)	365 (44.0%)
pre-pregnancy BMI (kg/m^2)	21.6 (20.4, 23.0)	27.7 (26.0, 30.5)
smoking during pregnancy		
never during pregnancy	1911 (75.1%)	612 (73.7%)
until pregnancy was known	233 (9.2%)	61 (7.3%)
continued throughout	400 (15.7%)	157 (19.0%)
alcohol consumption during pregnancy		
never during pregnancy	764 (30.0%)	359 (43.2%)
until pregnancy was known	416 (16.4%)	138 (16.6%)
continued throughout	1364 (53.6%)	334 (40.2%)
stress during pregnancy (score 0–4)	0.12 (0.06, 0.24)	0.13 (0.06, 0.26)
energy intake (kcal/day)	2162 (507)	2090 (514)
Dutch Healthy Diet Index (score 0–60)	32 (±8)	30 (±8)
fetal sex		
male	1287 (50.6%)	415 (50.0%)
female	1257 (49.4%)	415 (50.0%)
gestational weight gain (kg)	14.7 (±7.3)	12.9 (±7.7)
adequacy of gestational weight gain		
inadequate	370 (24.8%)	89 (20.9%)
adequate	565 (37.9%)	67 (15.8%)
excessive	557 (37.3%)	269 (63.3%)

Table 3.3: Posterior means and 95% credible intervals of the longitudinal analysis of *a posteriori*-derived dietary patterns and weight development in pregnancy in normal-weight and overweight women using Bayesian linear mixed models. Effects reflect a difference in weight (kg). ($\text{pattern}_{\text{VOF}}$: "Vegetable, oil and fish"; $\text{pattern}_{\text{NCS}}$: "Nuts, high-fibre cereals and soy"; $\text{pattern}_{\text{MSS}}$: "Margarine, sugar and snacks")

	normal-weight	**overweight**
$\text{pattern}_{\text{VOF}}$	0.22 [-0.01, 0.44]	0.26 [-0.18, 0.72]
$\text{pattern}_{\text{NCS}}$	0.15 [-0.08, 0.39]	-0.31 [-0.77, 0.14]
$\text{pattern}_{\text{MSS}}$	0.30 [0.07, 0.52]	0.32 [-0.07, 0.71]
gest. age	0.13 [0.12, 0.14]	0.10 [0.08, 0.12]
gest. age^2	0.01 [0.01, 0.01]	0.01 [0.01, 0.01]
gest. age $\times$ $\text{pattern}_{\text{VOF}}$	-0.00 [-0.01, 0.01]	0.01 [-0.01, 0.02]
gest. age $\times$ $\text{pattern}_{\text{NCS}}$	-0.01 [-0.02, -0.00]	0.01 [-0.00, 0.02]
gest. age $\times$ $\text{pattern}_{\text{MSS}}$	0.01 [-0.00, 0.01]	0.00 [-0.01, 0.02]

Table 3.4: Association of dietary patterns with gestational weight gain during early pregnancy in normal-weight women (n = 1849). Shown are regression coefficients with 95% confidence intervals from multivariable linear regression on imputed data, reflecting the difference in weight gain (g/week) for a 1 SD increase in dietary pattern score.

	Model 1	**Model 2**
***a posteriori* dietary patterns**		
Vegetable, Oil and Fish	14.3 [5.2, 23.4]	16.0 [6.5, 25.6]
Nuts, High-Fibre Cereals and Soy	-14.5 [-23.7, -5.3]	-6.5 [-16.4, 3.5]
Margarine, Sugar and Snacks	7.6 [-1.7, 17.0]	4.4 [-4.9, 13.7]
***a priori* dietary pattern**		
Dutch Healthy Diet Index	0.8 [-8.3, 9.9]	-6.2 [-15.5, 3.1]

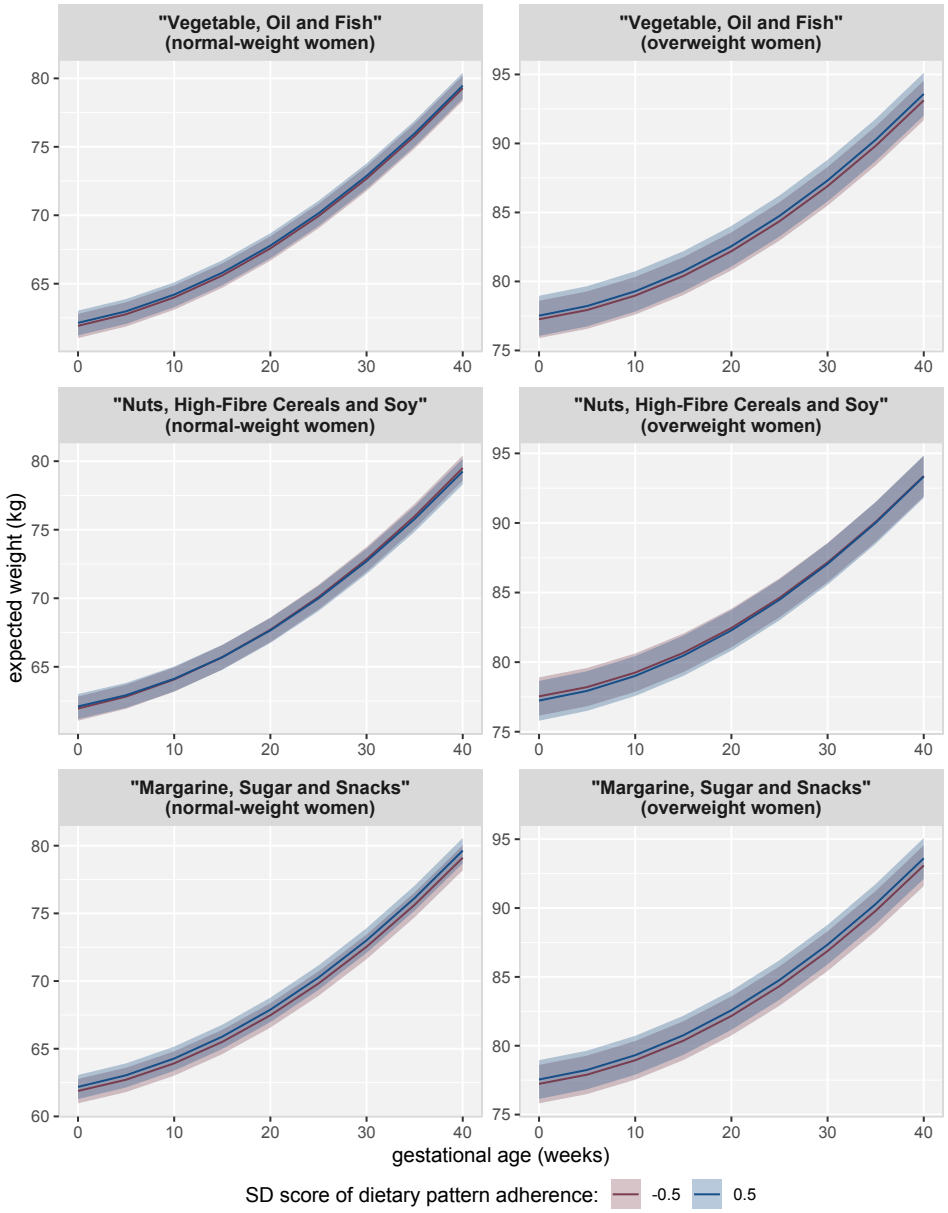


Figure 3.2: Visualization of the effect of diet on gestational weight over time. Shown are the expected value and corresponding 95% CIs for hypothetical cases that have either -0.5 or 0.5 SDS for adherence to the respective dietary pattern and reference values for the other patterns and confounders.

Table 3.5: Association of dietary patterns with gestational weight gain during early pregnancy in overweight women (n = 576). Shown are regression coefficients with 95% confidence intervals from multivariable linear regression on imputed data, reflecting the difference in weight gain (g/week) for a 1 SD increase in dietary pattern score.

	Model 1	**Model 2**
***a posteriori* dietary patterns**		
Vegetable, Oil and Fish	5.8 [-17.8, 29.5]	14.8 [-11.2, 40.7]
Nuts, High-Fibre Cereals and Soy	-28.8 [-52.2, -5.4]	-24.0 [-48.8, 0.7]
Margarine, Sugar and Snacks	14.1 [-7.3, 35.5]	12.4 [-9.0, 33.9]
***a priori* dietary pattern**		
Dutch Healthy Diet Index	10.8 [-10.3, 31.9]	2.4 [-19.5, 24.2]

Table 3.6: Association of dietary patterns with gestational weight gain during mid pregnancy in normal-weight women (n = 2086). Shown are regression coefficients with 95% confidence intervals from multivariable linear regression on imputed data, reflecting the difference in weight gain (g/week) for a 1 SD increase in dietary pattern score.

	Model 1	**Model 2**
***a posteriori* dietary patterns**		
Vegetable, Oil and Fish	-8.5 [-22.0, 5.0]	-7.1 [-21.4, 7.2]
Nuts, High-Fibre Cereals and Soy	9.5 [-4.1, 23.2]	11.3 [-3.6, 26.2]
Margarine, Sugar and Snacks	3.9 [-9.7, 17.5]	4.2 [-9.5, 18.0]
***a priori* dietary pattern**		
Dutch Healthy Diet Index	3.2 [-10.2, 16.6]	2.8 [-11.2, 16.7]

Table 3.7: Association of dietary patterns with gestational weight gain during mid pregnancy in overweight women (n = 662). Shown are regression coefficients with 95% confidence intervals from multivariable linear regression on imputed data, reflecting the difference in weight gain (g/week) for a 1 SD increase in dietary pattern score.

	Model 1	**Model 2**
***a posteriori* dietary patterns**		
Vegetable, Oil and Fish	35.6 [6.1, 65.1]	19.1 [-13.8, 52.0]
Nuts, High-Fibre Cereals and Soy	23.5 [-5.7, 52.7]	3.5 [-28.0, 34.9]
Margarine, Sugar and Snacks	8.3 [-18.3, 34.8]	14.5 [-12.3, 41.4]
***a priori* dietary pattern**		
Dutch Healthy Diet Index	-17.9 [-44.7, 8.9]	-2.2 [-30.0, 25.7]

Table 3.8: Association of dietary patterns with gestational weight gain during late pregnancy in normal-weight women (n = 2393). Shown are regression coefficients with 95% confidence intervals from multivariable linear regression on imputed data, reflecting the difference in weight gain (g/week) for a 1 SD increase in dietary pattern score.

	Model 1	**Model 2**
***a posteriori* dietary patterns**		
Vegetable, Oil and Fish	-8.3 [-18.3, 1.6]	-3.9 [-14.4, 6.6]
Nuts, High-Fibre Cereals and Soy	-13.2 [-23.4, -3.0]	-4.6 [-15.6, 6.3]
Margarine, Sugar and Snacks	0.3 [-9.8, 10.5]	0.4 [-9.8, 10.5]
***a priori* dietary pattern**		
Dutch Healthy Diet Index	3.2 [-6.7, 13.2]	-1.5 [-11.8, 8.8]

Table 3.9: Association of dietary patterns with gestational weight gain during late pregnancy in overweight women (n = 765). Shown are regression coefficients with 95% confidence intervals from multivariable linear regression on imputed data, reflecting the difference in weight gain (g/week) for a 1 SD increase in dietary pattern score.

	Model 1	**Model 2**
***a posteriori* dietary patterns**		
Vegetable, Oil and Fish	-3.8 [-24.5, 16.9]	7.4 [-15.5, 30.2]
Nuts, High-Fibre Cereals and Soy	4.4 [-17.0, 25.8]	9.5 [-13.4, 32.3]
Margarine, Sugar and Snacks	-8.5 [-28.0, 11.1]	-9.8 [-29.5, 10.0]
***a priori* dietary pattern**		
Dutch Healthy Diet Index	2.1 [-17.6, 21.7]	0.2 [-20.1, 20.5]

3.4 Discussion

3.4.1 Summary of the Main Findings

The results from our analysis of a population-based Dutch cohort suggest that specific *a posteriori*-derived dietary patterns have limited influence on early-pregnancy GWG, the prevalence of excessive GWG, and weight development in pregnancy. We neither found consistent associations of any dietary pattern with the prevalence of inadequate GWG nor was the *a priori*-defined dietary pattern associated with GWG.

3.4.2 Interpretation and Comparison with other Studies

The association of dietary patterns during pregnancy with GWG has been evaluated previously in a few studies (Uusitalo et al. 2009; Shin et al. 2014; Rifas-Shiman et al. 2009; Hillesund et al. 2014), but these studies did not evaluate longitudinal development of gestational weight and were conducted in different populations. Uusitalo et al. (2009) found that higher adherence to an *a posteriori*-derived dietary pattern characterized by high intake of sweets, fast food and snacks was associated with higher weekly GWG. In line with these results, we found that higher adherence to the unhealthy "Margarine, sugar and snacks" pattern was associated with higher prevalence of excessive GWG. Additionally, Uusitalo et al. (2009) reported that a pattern that was high in vegetables, fish and fruits was not associated with GWG. In contrast, we found that the "Vegetable, oil and fish" pattern, a relatively healthy pattern, was associated with higher GWG in early pregnancy.

In our study, the *a priori*-defined "Dutch Healthy Diet Index" pattern was not consistently associated with any measure of GWG. This result was in accordance with two studies showing no relationship between the *a priori*-defined "US healthy eating index of 2005" (HEI-2005) and the "Alternate Healthy Eating Index, slightly modified for pregnancy" (AHEI-P) with the prevalence of inadequate or excessive GWG (Shin et al. 2014; Rifas-Shiman et al. 2009). Nevertheless, a large population-based cohort study of over 66,000 participants found that high adherence to the *a priori*-defined "New Nordic Diet score" was associated with a 7% lower prevalence of excessive GWG in normal weight women, compared with low adherence (Hillesund et al. 2014). The inconsistency in findings regarding the associations between the *a priori*-defined dietary patterns may be due to different items that were included in the diet scores. The "New Nordic Diet score" contained items on meal patterns and the type of beverages consumed, among others (Hillesund et al. 2014); these items were not evaluated in our Dutch Healthy Diet Index,

nor in other *a priori*-defined dietary patterns (Shin et al. 2014; Rifas-Shiman et al. 2009).

The association of *a posteriori*-derived dietary patterns with weight trajectories over pregnancy has not been evaluated previously, to our knowledge. Studying this association longitudinally has the advantage that all available weight measurements can be used while taking into account the correlation between these measurements. In addition, weekly GWG is not constant over pregnancy and differs considerably by individual (Institute of Medicine and National Research Council 2009; Carmichael et al. 1997), which complicates cross-sectional comparisons of GWG. Our longitudinal analysis showed that women with higher adherence to the "Nuts, high-fibre and soy" pattern had a more moderate increase in weight during pregnancy than did women with low adherence to this dietary pattern, although absolute differences were small.

Results from both observational and interventional studies indicated that women with higher energy intake had higher GWG compared with women who have lower energy intake (Streuling et al. 2011), results that were also found in our cohort (Gaillard et al. 2013). In our analyses, the association of the "Vegetable, oil and fish" pattern remained significantly associated with GWG after additional adjustment for energy intake. This may indicate that dietary patterns are associated with GWG beyond energy intake.

Evaluating weight gain in pregnancy is important because GWG has been associated with many adverse pregnancy and birth outcomes. Gaining excessive weight during pregnancy can have short-term consequences such as delivery complications, and giving birth to a child that is large for its gestational age (Johnson et al. 2013; Gaillard et al. 2013; Institute of Medicine and National Research Council 2009). Additionally, it has been associated with long-term health consequences including post-partum obesity of the mother (Nehring et al. 2011) due to retaining their excess fat mass, and childhood obesity (Gaillard et al. 2013). Indeed, in our population, six years after childbirth women had gained on average 3.4 kg compared to their pre-pregnancy weight.

Weight gain during pregnancy consists of several maternal and fetal components that contribute differently to GWG over time (Pitkin 1976). For example, during the first half of pregnancy, maternal fat gain is a major contributor of GWG (Clapp et al. 1988; Kopp-Hoolihan et al. 1999), and most of the fat gain that takes place during pregnancy is in that period (Forsum et al. 1988). In our study, the higher GWG in women with high adherence to the "Vegetable, oil and fish" pattern could not be explained by fetal growth and was mainly found in early

pregnancy, meaning this higher GWG is likely due to maternal components, e.g., fat mass.

Our results for normal-weight women differed from those for overweight women, particularly for the "Vegetable, oil and fish" pattern and for the "Nuts, high-fibre cereals and soy" pattern. Similarly, Hillesund et al. (2014) reported differential associations for women below and above a BMI of 25 kg/m^2. These difference in findings may be explained by differences in the reporting of dietary intake (Freisling et al. 2012) or by differing contribution of the individual components of GWG for normal-weight and overweight women (Butte et al. 2003). In addition, our longitudinal analyses showed that over the whole course of pregnancy, normal-weight women with higher adherence to the "Margarine, sugar and snacks" pattern tend to be heavier than women with lower adherence.

3.4.3 Strengths and Limitations

A strength of our study is that we used a comprehensive approach to analyse the relation between diet and GWG by evaluating the associations of dietary patterns with trajectories of gestational weight and GWG during different phases in pregnancy. Another strength is the use of two distinct methods to define dietary patterns, which enabled us to evaluate the effects of dietary patterns derived by a data-driven and by a hypothesis-driven approach. Dietary patterns represent the combined effects of all foods consumed (Hu 2002), which may lead to a more powerful effect than the effects of the individual components, although it may also have led to a dilution of the effects of individual components that are associated with GWG (Willett 2012). For example, the food groups of vegetables and high-fat dairy products were strongly associated with the "Vegetable, oil and fish" dietary pattern. Yet, higher intake of fruits and vegetables has been associated with lower GWG (Olson and Strawderman 2003), whereas dairy products were associated with higher GWG (Stuebe et al. 2009; Olafsdottir et al. 2006). Consequently, this may result in an overall null effect of the dietary pattern. Furthermore, imputing the missing covariate values in the Bayesian framework allowed us to use all available information in the imputation. Especially in settings with a longitudinal outcome, imputation methods that are available in standard software and, hence, are more commonly used, often fail to appropriately include the outcome into the imputation procedure which may lead to severely biased results (Erler, Rizopoulos, Rosmalen, et al. 2016). Other strengths of our study are its population-based design, the collection of numerous covariates, and that the population was restricted to women of Dutch ancestry. We excluded women with other ethnicities to minimize measurement error since the FFQ was designed to evaluate a Dutch

diet. However, this restriction may have reduced the generalizability of our results to other ethnicities.

Our study also has some limitations. First, maternal weight before pregnancy, as well as maximum weight, were obtained using questionnaires, which may have resulted in a larger measurement error. Although we found no indication of systematic measurement error, random error may have resulted in loss of precision in GWG assessment. Furthermore, we were not able to calculate GWG per trimester because we did not have weight measurements at the required time points and the available data was insufficient for imputing those values. Another limitation is the lack of information on the separate components of GWG, in particular, maternal fat mass, and the lack of information on postpartum maternal weight. Future studies should collect detailed information on maternal body composition during pregnancy or measure the participants' weight a few weeks postpartum to evaluate associations with the different components of GWG. Furthermore, we could not use information on absolute dietary intake because dietary information collected using an FFQ does not provide this information. However, FFQs have been shown to be accurate in ranking participants according to their dietary intake (Kipnis et al. 2003a). Furthermore, we assessed maternal diet only once during pregnancy and were therefore not able to account for changes in dietary intake. Nevertheless, dietary patterns and macronutrient composition may not change largely during pregnancy despite an increased energy intake (Crozier et al. 2009; Rad et al. 2011). Additionally, we found that our results did not change after excluding women who may have altered their dietary intake due to illness or vomiting. Finally, the numerous statistical analyses performed may have resulted in chance findings (type I error). However, our results for weight trajectories and early-pregnancy GWG remained statistically significant when a more stringent significance level of $1 - 0.05/4 = 0.9875$ was used.

3.4.4 Conclusions and Implications

In conclusion, our results suggest that dietary composition during pregnancy may play a role in GWG in early pregnancy but has limited influence on total GWG in a population of Dutch women. The strength of the associations between dietary patterns and GWG differs for different definitions of dietary patterns and GWG. This suggests that the relationship between dietary patterns and GWG may be complex and may need further elucidation in order to facilitate the development of dietary guidelines during pregnancy and to adequately advise pregnant women on their diet.

Appendix

3.A Supplementary Materials

Table 3.10: Association of dietary patterns with gestational weight gain until early third trimester in normal-weight and overweight women (n = 2815). Results from multivariable linear regression on imputed data. Values (regression coefficients with 95% confidence interval) reflect the difference in gestational weight gain until early-third trimester (g/week) for a 1 SD increase in dietary pattern score, pooled over 10 imputed datasets.

	normal-weight (n = 2141)	**overweight** (n = 674)
***a posteriori* dietary patterns**		
Vegetable, Oil and Fish	6.7 [0.9, 12.5]	10.0 [-5.0, 25.1]
Nuts, High-Fibre Cereals and Soy	-2.0 [-8.1, 4.0]	-5.6 [-20.7, 9.4]
Margarine, Sugar and Snacks	2.5 [-3.2, 8.2]	6.0 [-7.2, 19.1]
***a priori* dietary pattern**		
Dutch Healthy Diet Index	-4.3 [-10.0, 1.4]	1.6 [-11.7, 14.9]

Table 3.11: Association of dietary patterns with gestational weight gain until early third trimester in normal-weight and overweight women (n = 2815); additionally adjusted for energy intake. Results from multivariable linear regression on imputed data. Values (regression coefficients with 95% confidence interval) reflect the difference in gestational weight gain until early-third trimester (g/week) for a 1 SD increase in dietary pattern score, pooled over 10 imputed datasets.

	normal-weight (n = 2141)	**overweight** (n = 674)
***a posteriori* dietary patterns**		
Vegetable, Oil and Fish	3.4 [-3.0, 9.8]	8.9 [-7.3, 25.2]
Nuts, High-Fibre Cereals and Soy	-4.8 [-11.2, 1.7]	-6.5 [-22.3, 9.3]
Margarine, Sugar and Snacks	-11.8 [-24.6, 1.0]	1.0 [-29.8, 31.8]
***a priori* dietary pattern**		
Dutch Healthy Diet Index	-4.3 [-10.0, 1.4]	1.6 [-11.7, 14.9]

Table 3.12: Association of dietary patterns with gestational weight gain during early third trimester in normal-weight and overweight women, excluding women with comorbidities or pregnancy complications (n = 2469). Results from multivariable linear regression on imputed data. Values (regression coefficients with 95% confidence interval) reflect the difference in gestational weight gain until early-third trimester (g/week) for a 1 SD increase in dietary pattern score, pooled over 10 imputed datasets.

	normal-weight (n = 1937)	**overweight** (n = 532)
***a posteriori* dietary patterns**		
Vegetable, Oil and Fish	8.0 [2.0, 14.0]	9.5 [-6.0, 24.9]
Nuts, High-Fibre Cereals and Soy	-1.4 [-7.6, 4.9]	-6.0 [-21.6, 9.7]
Margarine, Sugar and Snacks	1.9 [-4.0, 7.8]	9.5 [-4.8, 23.8]
***a priori* dietary pattern**		
Dutch Healthy Diet Index	-5.7 [-11.6, 0.2]	1.3 [-12.6, 15.1]

Table 3.13: Association of dietary patterns with gestational weight gain until early third trimester in normal-weight and overweight women, excluding women who vomited more than once per week in the previous three months (n = 2450). Results from multivariable linear regression on imputed data. Values (regression coefficients with 95% confidence interval) reflect the difference in gestational weight gain until early-third trimester (g/week) for a 1 SD increase in dietary pattern score, pooled over 10 imputed datasets.

	normal-weight (n = 1890)	**overweight** (n = 560)
***a posteriori* dietary patterns**		
Vegetable, Oil and Fish	4.8 [-1.3, 10.8]	14.2 [-2.1, 30.5]
Nuts, High-Fibre Cereals and Soy	-2.3 [-8.6, 4.0]	-7.7 [-23.8, 8.5]
Margarine, Sugar and Snacks	0.5 [-5.3, 6.4]	5.2 [-9.5, 19.8]
***a priori* dietary pattern**		
Dutch Healthy Diet Index	-5.2 [-11.0, 0.7]	-4.8 [-19.3, 9.7]

Table 3.14: Association of dietary patterns with maximal gestational weight gain in normal-weight and overweight women (n = 1703). Results from multivariable linear regression on imputed data. Values (regression coefficients with 95% confidence interval) reflect the difference in gestational weight gain until early-third trimester (g/week) for a 1 SD increase in dietary pattern score, pooled over 10 imputed datasets.

	normal-weight (n = 1343)	**overweight** (n = 360)
***a posteriori* dietary patterns**		
Vegetable, Oil and Fish	4.4 [-4.9, 13.8]	-9.3 [-34.0, 15.4]
Nuts, High-Fibre Cereals and Soy	1.4 [-8.3, 11.2]	5.2 [-19.5, 29.8]
Margarine, Sugar and Snacks	8.1 [-1.2, 17.5]	17.8 [-6.0, 41.6]
***a priori* dietary pattern**		
Dutch Healthy Diet Index	-0.8 [-10.3, 8.7]	14.4 [-8.3, 37.2]

References

Bellamy, L. et al. (2009). "Type 2 diabetes mellitus after gestational diabetes: a systematic review and meta-analysis". *The Lancet*, **373**(9677):1773–1779. DOI: 10.1016/S0140-6736(09)60731-5.

Brown, M. C. et al. (2013). "Cardiovascular disease risk in women with pre-eclampsia: systematic review and meta-analysis". *European Journal of Epidemiology*, **28**(1):1–19. DOI: 10.1007/s10654-013-9762-6.

Butte, N. F. et al. (2003). "Composition of gestational weight gain impacts maternal fat retention and infant birth weight". *American Journal of Obstetrics & Gynecology*, **189**(5):1423–1432. DOI: 10.1067/S0002-9378(03)00596-9.

Carmichael, S., B. Abrams, and S. Selvin (1997). "The pattern of maternal weight gain in women with good pregnancy outcomes". *American Journal of Public Health*, **87**(12):1984–1988. DOI: 10.2105/AJPH.87.12.1984.

Clapp, J. F. et al. (1988). "Maternal physiologic adaptations to early human pregnancy". *American Journal of Obstetrics & Gynecology*, **159**(6):1456–1460. DOI: 10.1016/0002-9378(88)90574-1.

Coolman, M. et al. (2010). "Medical record validation of maternally reported history of preeclampsia". *Journal of Clinical Epidemiology*, **63**(8):932–937. DOI: 10.1016/j.jclinepi.2009.10.010.

Crozier, S. R. et al. (2009). "Women's Dietary Patterns Change Little from Before to During Pregnancy". *The Journal of Nutrition*, **139**(10):1956–1963. DOI: 10.3945/jn.109.109579.

Derogatis, L. R. and P. Spencer (1993). *Brief symptom inventory: BSI*. Pearson Upper Saddle River, NJ.

Donders-Engelen, M. and L. van der Heijden (2003). "Maten, Gewichten en Codenummers 2003". Zeist, The Netherlands: Wageningen UR, Vakgroep Humane Voeding Wageningen and TNO Voeding.

Erler, N. S., D. Rizopoulos, J. van Rosmalen, et al. (2016). "Dealing with missing covariates in epidemiologic studies: a comparison between multiple imputation and a full Bayesian approach". *Statistics in Medicine*, **35**(17):2955–2974. DOI: 10.1002/sim.6944.

Forsum, E., A. Sadurskis, and J. Wager (1988). "Resting metabolic rate and body composition of healthy Swedish women during pregnancy". *The American Journal of Clinical Nutrition*, **47**(6):942–947.

Freisling, H. et al. (2012). "Dietary reporting errors on 24 h recalls and dietary questionnaires are associated with BMI across six European countries as evaluated with recovery biomarkers for protein and potassium intake". *British Journal of Nutrition*, **107**(6):910–920. DOI: 10.1017/S0007114511003564.

Gaillard, R. et al. (2013). "Risk factors and outcomes of maternal obesity and excessive weight gain during pregnancy". *Obesity*, **21**(5):1046–1055. DOI: 10.1002/oby.20088.

Han, Z. et al. (2011). "Low gestational weight gain and the risk of preterm birth and low birthweight: a systematic review and meta-analyses". *Acta Obstetricia et Gynecologica Scandinavica*, **90**(9):935–954. DOI: 10.1111/j.1600-0412.2011.01185.x.

Health Council of the Netherlands (2006). *Guidelines for a healthy diet 2006*.

Hedderson, M. M., E. P. Gunderson, and A. Ferrara (2010). "Gestational Weight Gain and Risk of Gestational Diabetes Mellitus". *Obstetrics & Gynecology*, **115**(3):597. DOI: 10.1097/AOG.0b013e3181cfce4f.

Hillesund, E. R. et al. (2014). "Development of a New Nordic Diet score and its association with gestational weight gain and fetal growth – a study performed in the Norwegian Mother and Child Cohort Study (MoBa)". *Public Health Nutrition*, **17**(9):1909–1918. DOI: 10.1017/S1368980014000421.

Hu, F. B. (2002). "Dietary pattern analysis: a new direction in nutritional epidemiology". *Current Opinion in Lipidology*, **13**(1):3–9.

Institute of Medicine and National Research Council (2009). *Weight Gain During Pregnancy: Reexamining the Guidelines*. Ed. by K. M. Rasmussen and A. L. Yaktine. Washington, DC: The National Academies Press. DOI: 10.17226/12584.

Johnson, J. et al. (2013). "Pregnancy Outcomes With Weight Gain Above or Below the 2009 Institute of Medicine Guidelines". *Obstetrics & Gynecology*, **121**(5):969. DOI: 10.1097/AOG.0b013e31828aea03.

Jolliffe, I. (2011). "Principal Component Analysis". *International Encyclopedia of Statistical Science*. Ed. by M. Lovric. Berlin, Heidelberg: Springer, p. 1094–1096. DOI: 10.1007/978-3-642-04898-2_455.

Kaiser, H. F. (1958). "The varimax criterion for analytic rotation in factor analysis". *Psychometrika*, **23**(3):187–200. DOI: 10.1007/BF02289233.

Kim, S. Y. et al. (2014). "Association of Maternal Body Mass Index, Excessive Weight Gain, and Gestational Diabetes Mellitus With Large-for-Gestational-Age Births". *Obstetrics & Gynecology*, **123**(4):737. DOI: 10.1097/AOG.0000000000000177.

Kipnis, V. et al. (2003a). "Structure of Dietary Measurement Error: Results of the OPEN Biomarker Study". *American Journal of Epidemiology*, **158**(1):14–21. DOI: 10.1093/aje/kwg091.

Klipstein-Grobusch, K. et al. (1998). "Dietary assessment in the elderly: validation of a semiquantitative food frequency questionnaire". *European Journal of Clinical Nutrition*, **52**(8):588–596. DOI: 10.1038/sj.ejcn.1600611.

Kopp-Hoolihan, L. et al. (1999). "Fat mass deposition during pregnancy using a four-component model". *Journal of Applied Physiology*, **87**(1):196–202. DOI: 10.1152/jappl.1999.87.1.196.

Kruithof, C. J. et al. (2014). "The Generation R Study: Biobank update 2015". *European Journal of Epidemiology*, **29**(12):911–927. DOI: 10.1007/s10654-014-9980-6.

Martins, A. P. B. and M. H. D. Benicio (2011). "Influence of dietary intake during gestation on postpartum weight retention". *Revista de Saúde Pública*, **45**(5):870–877. DOI: 10.1590/S0034-89102011005000056.

Nehring, I. et al. (2011). "Gestational weight gain and long-term postpartum weight retention: a meta-analysis". *The American Journal of Clinical Nutrition*, **94**(5):1225–1231. DOI: 10.3945/ajcn.111.015289.

Netherlands Nutrition Center (2006). "Nevo: Dutch food composition database 2006". *The Hague, The Netherlands: Netherlands Nutrition Centre*.

Olafsdottir, A. S. et al. (2006). "Maternal diet in early and late pregnancy in relation to weight gain". *International Journal of Obesity*, **30**(3):492. DOI: 10.1038/sj.ijo.0803184.

Olson, C. M. and M. S. Strawderman (2003). "Modifiable behavioral factors in a biopsychosocial model predict inadequate and excessive gestational weight gain". *Journal of the Academy of Nutrition and Dietetics*, **103**(1):48–54. DOI: 10.1053/jada.2003.50001.

Pitkin, R. M. (1976). "Nutritional support in obstetrics and gynecology". *Clinical Obstetrics and Gynecology*, **19**(3):489–513.

Plummer, M. (2003). "JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling". *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*. Ed. by K. Hornik, F. Leisch, and A. Zeileis. ISSN: 1609-395X.

Rad, N. T. et al. (2011). "Longitudinal analysis of changes in energy intake and macronutrient composition during pregnancy and 6 weeks post-partum". *Archives of Gynecology and Obstetrics*, **283**(2):185–190. DOI: 10.1007/s00404-009-1328-1.

Rifas-Shiman, S. L. et al. (2009). "Dietary Quality during Pregnancy Varies by Maternal Characteristics in Project Viva: A US Cohort". *Journal of the Academy of Nutrition and Dietetics*, **109**(6):1004–1011. DOI: 10.1016/j.jada.2009.03.001.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley Series in Probability and Statistics. Wiley. DOI: 10.1002/9780470316696.

Shin, D. et al. (2014). "Is Gestational Weight Gain Associated with Diet Quality During Pregnancy?" *Maternal and Child Health Journal*, **18**(6):1433–1443. DOI: 10.1007/s10995-013-1383-x.

Streuling, I. et al. (2011). "Weight gain and dietary intake during pregnancy in industrialized countries – a systematic review of observational studies". *Journal of Perinatal Medicine*, **39**(2):123–129. DOI: 10.1515/jpm.2010.127.

Stuebe, A. M., E. Oken, and M. W. Gillman (2009). "Associations of diet and physical activity during pregnancy with risk for excessive gestational weight gain". *American Journal of Obstetrics & Gynecology*, **201**(1):58–e1. DOI: 10.1016/j.ajog.2009.02.025.

Uusitalo, U. et al. (2009). "Unhealthy dietary patterns are associated with weight gain during pregnancy among Finnish women". *Public Health Nutrition*, **12**(12):2392–2399. DOI: 10.1017/S136898000900528X.

Van den Broek, M. et al. (2015). "Maternal dietary patterns during pregnancy and body composition of the child at age 6 y: the Generation R Study". *The American Journal of Clinical Nutrition*, **102**(4):873–880. DOI: 10.3945/ajcn.114.102905.

Van Lee, L. et al. (2012). "The Dutch Healthy Diet index (DHD-index): an instrument to measure adherence to the Dutch Guidelines for a Healthy Diet". *Nutrition Journal*, **11**(1):49. DOI: 10.1186/1475-2891-11-49.

Van Rossem, L. et al. (2009). "Are Starting and Continuing Breastfeeding Related to Educational Background? The Generation R Study". *Pediatrics*, **123**(6):e1017–e1027. DOI: 10.1542/peds.2008-2663.

Willett, W. (2012). *Nutritional Epidemiology*. Oxford University Press. ISBN: 9780199754038.

World Health Organization and Food and Agriculture Organization of the United Nations (1998). *Preparation and use of food-based dietary guidelines: report of a joint FAO/WHO consultation*.

4 Mothers' SCB Intake and Body Composition of their Children

This chapter is based on

Vincent Jen, Nicole S. Erler, Myrte J. Tielemans, Kim V.E. Braun, Vincent W.V. Jaddoe, Oscar H. Franco and Trudy Voortman. Mothers' intake of sugar-containing beverages during pregnancy and body composition of their children during childhood: the Generation R Study. *The American Journal of Clinical Nutrition*, 2017; **105**(4), 834–841. doi:10.3945/ajcn.116.147934

Abstract

Background: High intake of sugar-containing beverages (SCBs) has been linked to increased risk of obesity. However, associations of SCB intake during pregnancy with child body composition have been unclear.

Objectives: We explored whether SCB intake during pregnancy was associated with children's body mass index (BMI) and detailed measures of body composition. In addition, we examined different types of SCBs (i.e., fruit juice, soda, and concentrate).

Design: We included 3312 mother-child pairs of the Generation R Study, a prospective cohort from fetal life onward in the Netherlands. Energy-adjusted SCB intake was assessed in the first trimester with a food-frequency questionnaire. Anthropometric data of the children were collected repeatedly until six years of age, and BMI was calculated. At six years of age, we further measured the fat mass index (FMI) and fat-free mass index with dual-energy X-ray absorptiometry. All outcomes were sex- and age-standardized. Associations of SCB intake with children's body composition and BMI trajectories were analysed with multivariable linear and multivariable linear mixed models.

Results: Results from linear mixed models showed that, after adjustment for confounders including the SCB intake of the child itself, mothers' total SCB intake was positively associated with children's BMI until six years of age (per serving per day: 0.04 SD score (SDS); 95% CI: [0.00, 0.07]). In addition, intakes of total SCBs and fruit juice, but not of soda or concentrate, were associated with a higher FMI (total SCBs: 0.05 SDS, 95% CI: [0.01, 0.08]; fruit juice: 0.04 SDS, 95% CI: [0.01, 0.06]) of the six-year-old children. These associations remained significant after additional adjustment for gestational weight gain, birth weight, and children's insulin concentrations.

Conclusion: Our study suggests that maternal SCB intake during pregnancy is positively associated with children's BMI during early childhood and particularly with higher fat mass.

4.1 Introduction

The rapidly increasing prevalence of childhood overweight and obesity is of great concern because excess weight during childhood is associated with health problems over the child's life course, including cardiometabolic disturbances (Batch and Baur 2005). Intake of sugar-containing beverages (SCBs) has increased substantially during the past decades (Malik, Popkin, et al. 2010). Several studies have revealed that high intakes of SCBs by adults and children are related to overweight and obesity (Malik, Popkin, et al. 2010; Krebs and Jacobson 2003; Leermakers et al. 2015). Although this relation has been well established in the general population, the association of pregnant women's SCB intake with their offspring's growth and body composition has been unclear. Intake of SCBs during pregnancy may influence the intrauterine programming of the child toward obesity (Phelan et al. 2011). Potential effects of SCB intake during pregnancy on child body composition may be hypothesized by several mechanisms including changes in the insulin response.

To our knowledge, only one study has prospectively investigated the association between intake of SCBs during pregnancy and child body weight (Phelan et al. 2011). In a US cohort of 285 mothers and their infants, Phelan et al. (2011) observed that intakes of soft drinks and fruit juice during pregnancy were not associated with child birth weight or weight at 6 months of age. However, the later onset of childhood overweight and obesity was not explored. Also, more detailed measurements of body composition (i.e., fat mass and lean mass) were not explored in the study. Previous studies in the general population have suggested that higher intake of SCBs may be associated with a higher fat mass but not lean mass (Zheng et al. 2015; Raben et al. 2002; Tordoff and Alleva 1990). Therefore, we hypothesized that mothers' SCB intakes during pregnancy would be associated with higher BMI (in kg/m^2) and, in particular, with higher fat mass in their children.

The aim of this study was to explore the associations of mothers' SCB intakes during early pregnancy with BMI trajectories of their children until six years of age and with the children's fat masses and fat-free masses at six years of age. In addition, we analysed whether these associations differed by types of SCBs (fruit juice, concentrate, and soda).

4.2 Methods

4.2.1 Study Design

This study was embedded in the Generation R Study, which is an ongoing population-based prospective cohort from fetal life onwards in Rotterdam, the Netherlands. Details of the study design and procedures have been described previously (Kooijman et al. 2016). The study was approved by the Medical Ethics Committee at the Erasmus Medical Center. Pregnant women with an expected delivery date between 2002 and 2006 were included. Written informed consent was obtained for all participants. No measurements were performed when a child was not willing to participate.

4.2.2 Study Population

The selection process of the population for analysis is shown in Figure 4.1. We restricted our analyses to women who were of Dutch origin (n = 4545). Of these women, 3558 individuals provided valid dietary data of whom 3478 had singleton live births. Of these children, 3312 had available information on BMI for at least one time point. More detailed body-composition outcomes were available for 2660 children at six years of age.

4.2.3 Dietary Assessment of the Mother

We assessed dietary intake in pregnant women at enrolment (median: 13.4 weeks of gestation; 2.5% and 97.5% quantiles: 9.9 and 22.8 weeks of gestation) with the use of a self-administered semi-quantitative food-frequency questionnaire (FFQ) (Klipstein-Grobusch et al. 1998). The FFQ covered intakes of 293 food items that were consumed in the preceding three months and included questions about their consumption frequencies and portion sizes. We obtained information on the following three types of SCBs: soda (soft drinks, sports drinks, and energy drinks), fruit juice (fresh and boxed; 100% fruit juice only), and concentrate (juice and lemonades concentrate with added sugars). In this study, we defined total intake of SCBs as the sum of soda, fruit juice, and concentrate intakes, which were expressed as servings per day. We chose not to include sugar-containing milk products because of their different macronutrient composition (Malik, Schulze, et al. 2006). Daily energy and macronutrient intakes of the pregnant women were calculated with the use of the Dutch food-composition table (Netherlands Nutrition Center 2006).

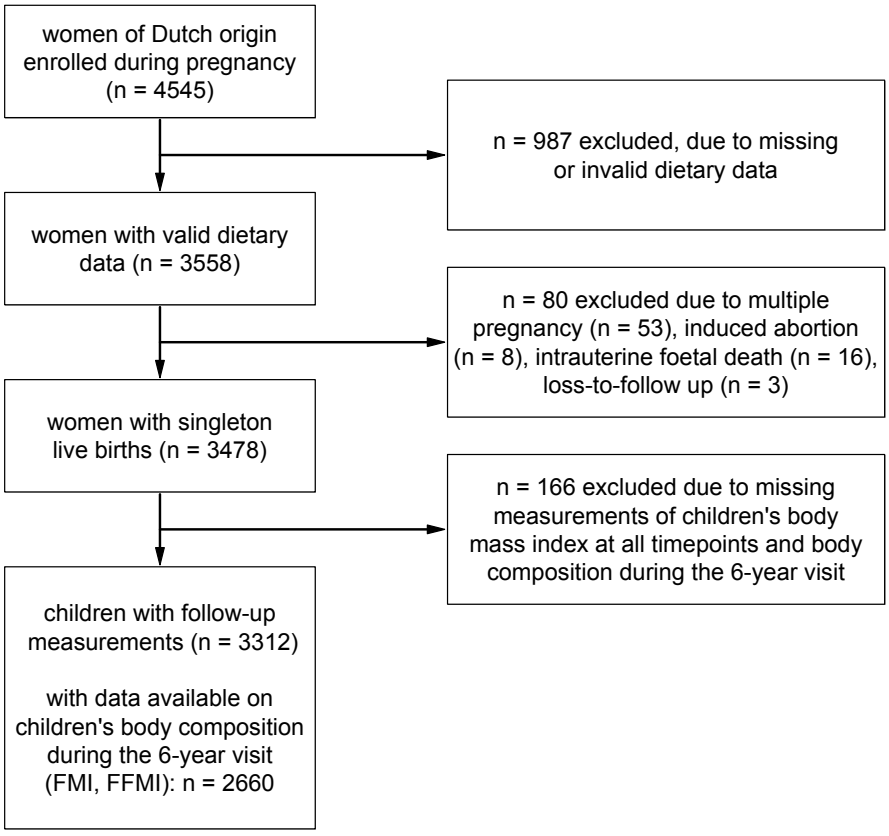


Figure 4.1: Flowchart of the population for analysis.

4.2.4 Child Anthropometric Measures and Body Composition

Children visited Child Health Centres at the median (2.5%, 97.5% quantiles) ages of 1.1 months (0.8, 1.8), 2.2 months (2.0, 2.9), 3.3 months (3.0, 3.9), 4.4 months (4.0, 4.9), 6.2 months (5.3, 8.2), 11.1 months (10.1, 12.5), 14.3 months (13.5, 16.0), 18.3 months (17.3, 21.4), 24.9 months (23.4, 28.0), 30.6 months (29.2, 33.8), 36.7 months (35.4, 40.6) and 45.8 months (44.4, 48.5). During all visits, heights and weights of the children were measured with the children not wearing shoes or heavy clothing. When children were a median age of six years (2.5% and 97.5% quantiles: 5.7, 7.4), well-trained staff measured children's anthropometric variables and body compositions in a dedicated research centre of Sophia Children's Hospital in Rotterdam (Kooijman et al. 2016). We measured height in a standing position

to the nearest millimetre with the use of a Harpenden stadiometer (Holtain Ltd.). Weight was measured with the use of a mechanical personal scale (SECA). BMI was calculated for all time points.

In addition, to measure the body composition of children at six years of age, we used a dual-energy X-ray absorptiometry (DXA) scanner (iDXA; Ge-Lunar, 2008) (Kooijman et al. 2016). The DXA scanner measured fat mass, lean mass, and bone mass of the total body with the use of enCORE software (v.13.6; GE Healthcare). We calculated the fat mass index (FMI) as fat mass (kilograms) divided by the square of height (in meters) and the fat-free mass index (FFMI) as lean mass plus bone mass (both in kilograms) divided by the square of height (in meters). All outcomes were standardized for age and sex on the basis of the Generation R Study population. We used cut-offs that were recommended by Cole et al. (2000) to determine the weight status of each child at six years of age as normal, or as overweight or obese.

4.2.5 Covariates

At enrolment, we used self-administered questionnaires to collect information on maternal age, marital status (no partner, married / living together), education (low or high) (Statistics Netherlands 2003), use of folic acid supplements (never, started during the first 10 weeks, or started periconceptional), parity (nulliparous or multiparous), and net household income ($<$ € 2400/month or $\geq$ € 2400/month). Maternal height and weight were measured at enrolment to calculate BMI, and a fetal ultrasound was performed to determine the gestational age. The maternal diet quality (range: 0 to 14) was scored on the basis of adherence to Dutch dietary guidelines with the use of the dietary intake data that were obtained from the FFQ at enrolment. This score included items on intakes of vegetables, fruit, legumes, whole grains, nut, fish, dairy, tea, soft fat and oils, red meat, alcohol, salt, and folic acid supplements (Nguyen et al. 2017). Psychiatric symptoms during pregnancy (Global Severity Index; range: 0 to 4) (Derogatis and Melisaratos 1983) and vomiting ($\geq$ 1 or $<$ 1 time/week) were also assessed with the use of questionnaires at enrolment, but also during pregnancy.

With the use of questionnaires, we assessed smoking (never, until the pregnancy was known, or continued during pregnancy) and alcohol consumption (never, until the pregnancy was known, occasionally during pregnancy, or frequently during pregnancy). Body weight was measured in each trimester, and the total gestational weight gain was calculated by subtracting weight that was assessed at enrolment from weight that was measured in the third trimester (Kooijman et al. 2016).

After pregnancy, we obtained information on pregnancy-related diseases including

pregnancy-induced hypertension, pre-eclampsia (Coolman et al. 2010), and gestational diabetes from midwives and obstetricians. We collected information on child sex, gestational age, and birth weight from hospital medical records, and z-scores were calculated for birth weight with the use of reference data (Niklasson et al. 1991). When a child was six years old, information on screen time (watching television and the use of the computer in hours per day), sports participation (yes or no), and intake of SCBs (servings per day) was obtained with the use of a questionnaire. In addition, nonfasting blood was collected at the research centre, and insulin concentrations were analysed with the use of enzymatic methods (Cobas 8000; Roche) (Voortman, van den Hooven, et al. 2016).

4.3 Statistical Analyses

To reduce bias due to missing values ($\leq$ 22.8% per covariate), we used a fully Bayesian approach that allowed for simultaneous imputation of missing covariates and analysis of associations with multiple cross-sectional and longitudinal outcomes. A detailed description of this approach can be found elsewhere (Erler, Rizopoulos, Rosmalen, et al. 2016). Results of the main analyses were obtained directly from the Bayesian approach and are presented as posterior means and 95% credible intervals (CIs). To perform subsequent sensitivity analyses, 10 imputed datasets were created by random selection of imputed values from the Bayesian approach. These analyses were performed separately in each of the imputed datasets, and pooled effect estimates and 95% confidence intervals (CIs) are presented. All decisions with regards to the structure of the final models were based on preliminary analyses conducted in the unimputed data.

We specified linear mixed models to analyse associations of SCB intake during pregnancy with trajectories of children's BMI between one month and six years of age. To explore these associations, we constructed three models with a fixed-effects structure that included maternal SCB intake and possible confounders, and a random-effects structure that included a random intercept and slope (for age). In Model 1 (crude model), we adjusted for child sex and further included energy intake in the model. Model 2 (confounder model) was additionally adjusted for the following covariates: maternal parity, age, gestational age at enrolment, marital status, income, education, psychiatric symptoms during pregnancy, protein intake (percentage of energy intake), diet quality, smoking during pregnancy, alcohol intake during pregnancy, use of folic acid supplements, and child sports and screen times. The confounders were selected on the basis of previous literature and a $\geq$ 10% change in effect estimates (Mickey and Greenland 1989). In Model 3 (SCB child model), we adjusted for child SCB intake at age six years in addition

to the confounders used in Model 2 (Leermakers et al. 2015). We investigated non-linearity of trajectories of children's BMI over time and potential effect modification of the trajectories by SCB intake using natural cubic splines for age and an interaction term between the spline and SCB intake. Based on this preliminary analysis there was no evidence for the need of either non-linear trajectories or the interaction term and the simpler model assuming linear trajectories and associations was used in the final analysis.

In addition, we specified linear regression models to examine the association between intake of SCBs during pregnancy and children's FMI and FFMI at age six years. We used the previously mentioned Models 2 and 3 to which we further added child age at the six-year visit. The analyses for BMI trajectories, FMI, and FFMI were modelled jointly in the Bayesian analysis.

Intake of SCBs was examined both with and without energy adjustment. Adjustment for energy intake was performed using the residual method (Willett et al. 1997). We report results from energy-adjusted SCBs as the main results in this study. To study whether the associations differed by the type of SCB, the analyses were repeated for intakes of fruit juice, concentrate, and soda instead of total SCB intake.

For relevant associations (i.e., when the 95% CI excluded zero in Model 3), we also examined the role of the potential mediators gestational weight gain, child birth weight, maternal BMI at enrolment (Kalk et al. 2009), and child serum insulin concentrations at six years of age (Reichetzeder et al. 2016) by including them in the model. These analyses were performed in the multiply imputed data obtained from the Bayesian model. Moreover, we tested whether associations were modified by child sex (Reichetzeder et al. 2016) or by maternal BMI at enrolment by adding the product term of the potential effect modifier and maternal SCB intake to Models 1 and 2 and re-evaluating them in the imputed data. Main analyses were stratified if the pooled estimate of the interaction term was significantly different from zero.

To test the robustness of our findings, we performed two sensitivity analyses based on the imputed data. First, we repeated our analyses by restricting them to women who vomited less than one time per week because vomiting in pregnancy may alter dietary intake (Chortatos et al. 2013). Second, we restricted analyses to women who did not experience comorbidities during pregnancy (i.e., diabetes mellitus, gestational diabetes, or hypertensive disorders) because these comorbidities could alter the maternal diet during pregnancy and may affect child birth weight (Hutcheon et al. 2011; Boerschmann et al. 2010). All statistical analyses were performed with the use of SPSS 21.0 (IBM Corp.), R version 3.3.1

(R Core Team 2016), and JAGS (version 4.2.0; http://mcmc-jags.sourceforge.net)(Plummer 2003).

4.4 Results

4.4.1 Subject Characteristics

Characteristics of the 3312 mothers and their children included in this study are presented in Table 4.1. Median intake of total SCBs during pregnancy was 1.9 servings/day (2.5% and 97.5% quantiles: 0.1 and 7.4 servings/day). Most intake came from fruit juices, with a median intake of 1.0 serving/day (2.5% and 97.5% quantiles: 0.0 and 5.0 servings/day), whereas for both concentrate (2.5% and 97.5% quantiles: 0.0 and 4.5 servings/day) and soda (2.5% and 97.5% quantiles: 0.0 and 2.5 servings/day) median intake was only 0.1 serving/day. Most of the women who were included were highly educated (60.7%), had never smoked during pregnancy (75.2%), and used folic acid supplementation during pregnancy (88.8%). The mean ± SD birth weight of the children was 3488 ± 561 gram. At six years of age, median BMI of the children was 15.7 (2.5% and 97.5% quantiles: 13.7 and 19.3) with 9.2% of the children being classified as overweight or obese.

4.4.2 Maternal SCB Intake and Child BMI Trajectory

In the linear mixed-model analyses, we observed no relevant associations between energy-adjusted SCB intake during pregnancy and BMI trajectories of children from birth to six years of age in Model 1 (0.01 SD score (SDS); 95% CI: [-0.02, 0.04]) and Model 2 (0.03 SDS; 95% CI [0.00, 0.07]; Table 4.2). After adjustment for child SCB intake, we observed that higher SCB intake during pregnancy was associated with higher BMI of children (Model 3: 0.04 SDS; 95% CI [0.00, 0.07]). This association was not explained by specific intakes of fruit juice, concentrate, or soda during pregnancy.

4.4.3 Maternal SCB Intake and Child Body Composition

In the linear regression analyses, we found in Model 1 that one additional daily serving of SCBs during pregnancy was associated with a 0.05-SDS higher FMI (95% CI [0.02, 0.09]) but not with FFMI of the offspring at age six years (Tables 4.3 and 4.4). This association remained after further adjustment for sociodemographic and lifestyle factors (Model 2: 0.04 SDS, 95% CI [0.01, 0.07]; Model 3: 0.05 SDS, 95% CI [0.01, 0.08]). To study fruit juice, concentrate, and soda intakes as exposures, we replaced total SCB intake for these types of beverages in the analyses. We observed an association between a higher daily serving of fruit

juice, but not of concentrate or soda, during pregnancy and higher FMI (Model 3: 0.04 SDS, 95% CI [0.01, 0.06]; Table 4.3) but not higher FFMI (Model 3: 0.02 SDS, 95% CI [-0.01, 0.05]; Table 4.4). These effect estimates remained similar after adjustment for the potential mediators gestational weight gain, child birth weight, maternal BMI at enrolment, and child insulin concentrations (Table 4.5 in Appendix 4.A).

4.4.4 Additional Analyses

We did not observe interactions between SCB intake and maternal BMI for children's BMI, FMI, or FFMI and consequently our analyses were not stratified by this factor. We did observe an interaction of total SCB and fruit juice intake with child sex on children's FMI but not on BMI or FFMI. Therefore, we stratified analyses for FMI by sex, and we observed higher effect estimates in girls than in boys. In addition, we observed no association between SCB intake and FMI in boys (Table 4.6 in Appendix 4.A). Furthermore, we obtained similar effect estimates as were observed in the whole group after restricting our analyses to women who vomited less than one time per week (n = 2099) (Tables 4.7 and 4.8 in Appendix 4.A) or to women with no comorbidities during pregnancy (n = 2308) (Tables 4.9 and 4.10 in Appendix 4.A). Findings for SCB intake that was unadjusted for total energy intake were similar to those obtained for energy-adjusted SCBs (data not shown).

Table 4.1: Characteristics of the analysis population (n = 3312). Values are means ± SDs for approximately normal variables, medians (2.5%, 97.5% quantile) for continuous non-normal variables, and valid percentages for categorical variables.

Maternal Characteristics during Pregnancy	
age at enrolment in years	31.6 ± 4.3
BMI at enrolment in kg/m^2	23.4 (18.9, 35.1)
nulliparous, %	59.9
gestational age at enrolment in weeks	13.4 (9.8, 22.8)
gestational weight gain in kg	8.1 ± 3.5
educational level, higher, %	60.7
net household income ≥ € 2400/month, %	80.8
alcohol consumption, %	
never	31.0
until pregnancy was known	16.0
occasionally during pregnancy	40.6
frequently during pregnancy	12.4
smoking, %	
never	75.2
until pregnancy was known	9.1
continued during pregnancy	15.6
folic acid supplementation, %	
never	11.2
until pregnancy was known	54.6
started in the first 10 weeks	34.2
total energy intake in kcal/day	2148 ± 506
sugar-containing beverages in servings/day	
total	1.9 (0.1, 7.4)
fruit juice (57.2%)	1.0 (0.0, 5.0)
concentrate (27.6%)	0.1 (0.0, 4.5)
soda (15.1%)	0.1 (0.0, 2.5)
Child Characteristics at Birth	
girls, %	49.6
birth weight in gram	3488 ± 561
gestational age at birth in weeks	13.4 (9.8, 22.8)
Child Characteristics at 6-year Visit (n = 2736)	
age in years	6.0 (5.6, 7.3)
BMI in kg/m^2	15.7 (13.7, 19.3)
overweight or obese, %	9.2
Fat Mass Index in kg/m^2 (n = 2660)	3.6 (2.4, 6.7)
Fat-Free Mass Index in kg/m^2 (n = 2660)	11.9 ± 0.8
participation in sports, % (n = 3312)	50.6
screen time ≥ 2 hours/day, % (n = 3312)	21.3

Table 4.2: Associations of SCB intake during pregnancy with child BMI trajectories until six years of age. Shown are the posterior mean and 95% CIs of the regression coefficient for SCB intake from the Bayesian linear mixed models.

	Model 1 (crude)	Model 2 (confounder)	Model 3 (SCB child)
total, servings per day			
total SCB	0.01 [-0.02, 0.04]	0.03 [-0.00, 0.07]	0.04 [0.00, 0.07]
fruit juice, soda and concentrate, servings per day			
fruit juice	0.01 [-0.02, 0.03]	0.02 [-0.01, 0.04]	0.02 [-0.01, 0.04]
soda	-0.01 [-0.02, 0.01]	0.00 [-0.02, 0.02]	0.00 [-0.01, 0.02]
concentrate	0.00 [-0.02, 0.02]	0.00 [-0.01, 0.02]	0.00 [-0.01, 0.02]

Table 4.3: Associations of SCB intake during pregnancy with child FMI at six years of age. Shown are the posterior mean and 95% CIs of the regression coefficient for SCB intake from the Bayesian linear models.

	Model 1 (crude)	Model 2 (confounder)	Model 3 (SCB child)
total, servings per day			
total SCB	0.05 [0.02, 0.09]	0.04 [0.01, 0.07]	0.05 [0.01, 0.08]
fruit juice, soda and concentrate, servings per day			
juice	0.03 [0.00, 0.06]	0.03 [0.01, 0.06]	0.04 [0.01, 0.06]
soda	0.02 [0.00, 0.03]	-0.01 [-0.02, 0.01]	-0.01 [-0.02, 0.01]
concentrate	-0.02 [-0.04, -0.00]	-0.02 [-0.03, 0.00]	-0.02 [-0.03, 0.00]

Table 4.4: Associations of SCB intake during pregnancy with child FFMI at six years of age. Shown are the posterior mean and 95% CIs of the regression coefficient for SCB intake from the Bayesian linear models.

	Model 1 (crude)	Model 2 (confounder)	Model 3 (SCB child)
total, servings per day			
total SCB	-0.01 [-0.05, 0.03]	0.03 [-0.02, 0.07]	0.03 [-0.01, 0.07]
fruit juice, soda and concentrate, servings per day			
juice	0.01 [-0.02, 0.04]	0.02 [-0.01, 0.05]	0.02 [-0.01, 0.05]
soda	-0.02 [-0.04, -0.00]	-0.00 [-0.02, 0.02]	-0.02 [-0.02, 0.02]
concentrate	0.01 [-0.01, 0.03]	0.01 [-0.01, 0.03]	0.01 [-0.01, 0.03]

4.5 Discussion

The results of this longitudinal observational study suggest that higher intake of SCBs during pregnancy is associated with higher BMI of the child until six years of age. Furthermore, we observed that higher SCB intakes of mothers were associated with higher fat masses, but not higher fat-free masses, of their children, particularly in girls, and that this association was mainly explained by intake of fruit juice. These associations were independent of gestational weight gain, birth weight, child SCB intake, or child insulin concentration.

4.5.1 Interpretation and Comparison with Previous Studies

In our study population, intake of SCBs during pregnancy was associated with trajectories of BMI in children during early childhood. This result is not in line with that of the previously mentioned study that was performed by Phelan et al. (2011) in which they observed that intakes of soft drinks and fruit juice in 285 pregnant women were not associated with birth weight or weight after 6 months of age. However, our findings suggest an association between maternal SCB intake and children's BMI at a later age, which was not explained by a higher birth weight, thereby suggesting that an effect may be stronger later in childhood. Furthermore, BMI has several limitations as a marker of adiposity in children (Freedman et al. 2005). First, BMI does not measure excess fat mass because of the many combinations of FMI and FFMI that result in the same BMI. Second, the amounts of both fat mass and fat-free mass vary within children (Freedman et al. 2005). In our study population, we observed that higher intake of SCBs during pregnancy was associated with higher FMI, but not higher FFMI, of children at six years of age. This finding is in line with results from previous studies that have studied SCB intakes in children and the body compositions of these children (Zheng et al. 2015; Zheng et al. 2014; Ruyter et al. 2012).

Moreover, we observed associations between maternal intake of fruit juice, but not of other beverages, and higher FMI of their six-year-old children. In our study population, intake of fruit juice was 10 times higher (median: 1.0 serving per day) than intakes of concentrate and soda. Consequently, intakes of concentrate and soda in our study may have been too low to detect potential associations with adverse body-composition outcomes of the children. Although fruit juice also provides vitamins and minerals (Caswell 2009), our results suggest that fruit juice could be harmful for a child's body composition.

For FMI, but not for the other outcomes, we observed stronger associations in girls than in boys. We previously also observed stronger associations of early life

nutrition with body fat in girls than in boys at approximately six years of age (Leermakers et al. 2015; Voortman, Braun, et al. 2016). These associations have been suggested to be related to differences in insulin responses between boys and girls (Voortman, van den Hooven, et al. 2016) or to differences in the timing of the adiposity rebound, which usually occurs between the ages of five and seven years (Rolland-Cachera et al. 2006).

4.5.2 Potential Mechanisms

We observed that SCB intake during pregnancy was positively associated with child BMI and FMI during early childhood. One possible mechanism that may explain this relation could be excessive energy intake from the SCBs themselves or from other food sources because caloric intake in a liquid form leads to a lower and shorter feeling of satiation (Pan and Hu 2011; Cassady et al. 2012; Drewnowski and Bellisle 2007; DiMeglio and Mattes 2000). However, in our analyses, effect estimates were similar after adjustment for energy intake, and associations remained between SCB intake and child BMI and FMI, which suggested that the associations were due to other factors than energy intake.

Another possible mechanism could be that the epigenetics of the fetus changes when a mother has frequent intake of SCBs during pregnancy (Mateo-Fernández et al. 2016). These changes could lead to altered gene expression (De Boo and Harding 2006), which may result in children becoming more susceptible to having higher fat mass.

Furthermore, because SCB intake leads to high peak concentration of insulin (Hu and Malik 2010), the role of insulin should be further elucidated. Hyperinsulinemia during the development of the fetus could have long-lasting consequences to the central nervous systems that regulates body weight, possibly resulting in the stimulation of fat mass development (Franke et al. 2005). In addition, SCB intake during pregnancy might affect the insulin sensitivity of the child (Reichetzeder et al. 2016) with the subsequent stimulation of the development of fat mass in the child. However, our findings suggest that a child's insulin concentration is not part of this pathway. Unfortunately, we had no information available on serum insulin during pregnancy.

Our results could also be explained by residual confounding because intake of SCBs is associated with other lifestyle factors (Kvaavik et al. 2005). Although we had information available on several potential lifestyle confounders, including smoking, alcohol intake, and the overall diet quality during pregnancy, we had, for instance, no information on the physical activity of the mother (Schulze et al. 2004). Furthermore, confounding that could have been due to child fac-

tors is difficult to take into account because lifestyle patterns may change during childhood (Lytle et al. 2000). As part of these lifestyle patterns, intake of sweet foods, including SCBs, by the child could have partially determined the child's body composition. Prenatal exposures such as flavours of the maternal diet can transmit to the amniotic fluid, which consequently may lead to greater acceptance by the child of these foods after birth (Mennella et al. 1995). We attempted to take this potential confounding by child diet into account by adjusting for SCB intake of the child, and the association between SCB intake of the mother with her offspring's fat mass remained.

4.5.3 Strengths and Limitations

Strengths of this study are the prospective population-based design, the large sample size, and the available information on numerous confounders of the mothers and children. Additionally, the repeated measurement of child BMI was an important strength in this study. Furthermore, we had extensive measurements on child body composition at six years of age with the use of DXA. This method has been proven to measure fat mass accurately (Svendsen et al. 1993), which allowed us to distinguish the child's body fat mass and fat-free mass. Another strength is that we also studied different types of SCBs rather than only examining overall intake. Finally, we applied the Bayesian approach which allowed us to use all available information of the outcome measurements to impute missing covariate values. The use of this approach provides a better method to deal with bias that is associated with incomplete information on covariates than is achieved with the use of less-sophisticated missing-data methods (Erler, Rizopoulos, Rosmalen, et al. 2016).

This study also has several limitations. One limitation is that dietary intakes were estimated by self-report, which has been shown to be prone to measurement errors (Kipnis et al. 2003b). However, we reduced the magnitude of the measurement errors by adjusting for total energy intake with the use of the residual method. Because all mothers included in our analyses were of Dutch origin and were, on average, highly educated, the generalizability of our findings to other ethnic or socioeconomic groups may be limited. Another limitation may be the single assessment of SCB intake during the first trimester. Repeated measurements of dietary intake would have been better to study whether there may have been an accumulative or trimester-specific effect of maternal SCB intake on child BMI and body composition. Although we adjusted the analyses for numerous sociodemographic and lifestyle factors related to both mother and child, residual confounding was still possible. An example was the absence of information on the energy expenditure or physical activity of the mother (Dewey and McCrory

1994) or on the diet of the father (Li et al. 2016), which might have influenced our results. Finally, although we had information that was available on total SCB intake of the child at six years, we did not collect information on different types of SCB intake, total energy intake, or other components of the diet of the child.

In conclusion, in this prospective cohort, we observe that higher intake of SCBs during pregnancy is associated with higher child BMI until six years of age. Moreover, higher intakes of total SCBs and fruit juice, but not of soda or concentrate, are associated with higher FMI of the child at six years of age. These associations are stronger in girls than in boys. Future studies should further explore whether SCB intake during pregnancy is associated with child body composition, which should preferably be assessed repeatedly to observe whether changes occur in later stages during childhood.

Appendix

4.A Additional Results

Table 4.5: Associations of sugar-containing beverages intake during pregnancy with children's fat mass index (FMI) at six years, after adjustment for potential mediators. Shown are pooled estimates and 95% CIs from linear regression on multiply imputed data.

total, servings per day		
Model 3	+ gestational weight gain	0.04 [0.01, 0.08]
Model 3	+ birth weight z-score	0.04 [0.01, 0.08]
Model 3	+ maternal baseline BMI	0.04 [0.00, 0.07]
Model 3	+ child insulin at six years	0.05 [0.00, 0.09]
Model 3	+ gestational weight gain + birth weight z-scores + maternal BMI at enrolment + child insulin	0.04 [0.00, 0.08]
fruit juice, servings per day		
Model 3	+ gestational weight gain	0.04 [0.01, 0.06]
Model 3	+ birth weight z-score	0.03 [0.01, 0.06]
Model 3	+ maternal baseline BMI	0.03 [0.00, 0.06]
Model 3	+ child insulin at six years	0.04 [0.01, 0.07]
Model 3	+ gestational weight gain + birth weight z-scores + maternal baseline BMI + child insulin	0.03 [0.00, 0.06]

Table 4.6: Associations of sugar-containing beverages intake during pregnancy with children's fat mass index (FMI) at six years in boys and girls separately. Shown are pooled regression coefficients and 95% CIs from linear regression of multiply imputed data.

	all (n = 2660)	girls (n = 1329)	boys (n = 1331)
total, servings per day			
Model 3 (SCB child)	0.05 [0.01, 0.08]	0.09 [0.05, 0.14]	-0.01 [-0.06, 0.04]
fruit juice, servings per day			
Model 3 (SCB child)	0.04 [0.01, 0.06]	0.04 [0.01, 0.08]	0.03 [-0.01, 0.06]

Table 4.7: Associations of sugar-containing beverage intake during pregnancy with children's fat mass index (FMI) at six years in women who vomited less than once a week (n = 2099). Shown are pooled estimates and 95% CIs from linear regression of multiply imputed data.

	Model 1 (crude)	Model 2 (confounder)	Model 3 (SCB child)
total, servings per day			
total SCB	0.05 [0.01, 0.09]	0.04 [-0.00, 0.08]	0.04 [0.00, 0.08]
fruit juice, soda and concentrate, servings per day			
juice	0.03 [0.00, 0.06]	0.03 [0.00, 0.06]	0.04 [0.01, 0.06]
soda	0.01 [-0.01, 0.03]	-0.01 [-0.03, 0.01]	-0.01 [-0.03, 0.01]
concentrate	-0.01 [-0.03, 0.01]	-0.01 [-0.03, 0.01]	-0.01 [-0.03, 0.01]

Table 4.8: Associations of sugar-containing beverage intake during pregnancy with children's fat-free mass index (FFMI) at six years in women who vomited less than once a week (n = 2099). Shown are pooled estimates and 95% CIs from linear regression of multiply imputed data.

	Model 1 (crude)	Model 2 (confounder)	Model 3 (SCB child)
total, servings per day			
total SCB	-0.02 [-0.07, 0.03]	0.01 [-0.04, 0.06]	0.02 [-0.03, 0.06]
fruit juice, soda and concentrate, servings per day			
juice	0.01 [-0.03, 0.04]	0.01 [-0.02, 0.05]	0.01 [-0.02, 0.05]
soda	-0.03 [-0.05, -0.01]	-0.01 [-0.03, 0.02]	-0.01 [-0.03, 0.02]
concentrate	0.01 [-0.01, 0.04]	0.02 [-0.01, 0.04]	0.02 [-0.01, 0.04]

Table 4.9: Associations of sugar-containing beverage intake during pregnancy with children's fat mass index (FMI) at six years in women who had no comorbidities during pregnancy (n = 2308). Shown are the pooled estimates and 95% CIs for SCB intake from linear regression on multiply imputed data.

	Model 1 (crude)	Model 2 (confounder)	Model 3 (SCB child)
total, servings per day			
total SCB	0.05 [0.02, 0.09]	0.03 [0.00, 0.07]	0.04 [0.01, 0.07]
fruit juice, soda and concentrate, servings per day			
fruit juice	0.03 [0.01, 0.06]	0.04 [0.01, 0.06]	0.04 [0.01, 0.07]
soda	0.01 [0.00, 0.03]	-0.01 [-0.03, 0.01]	-0.01 [-0.03, 0.01]
concentrate	-0.02 [-0.03, 0.00]	-0.01 [-0.03, 0.00]	-0.01 [-0.03, 0.00]

Table 4.10: Associations of sugar-containing beverage intake during pregnancy with children's fat-free mass index (FFMI) at six years in women who had no comorbidities during pregnancy (n = 2308). Shown are the pooled estimates and 95% CIs for SCB intake from linear regression on multiply imputed data.

	Model 1 (crude)	Model 2 (confounder)	Model 3 (SCB child)
total, servings per day			
total SCB	-0.01 [-0.06, 0.03]	0.02 [-0.03, 0.06]	0.02 [-0.02, 0.07]
fruit juice, soda and concentrate, servings per day			
fruit juice	0.01 [-0.02, 0.05]	0.02 [-0.01, 0.05]	0.01 [-0.02, 0.04]
soda	-0.03 [-0.05, -0.01]	-0.01 [-0.04, 0.01]	-0.01 [-0.04, 0.01]
concentrate	0.01 [-0.01, 0.03]	0.01 [-0.01, 0.04]	0.00 [-0.01, 0.04]

References

Batch, J. A. and L. A. Baur (2005). "Management and prevention of obesity and its complications in children and adolescents". *The Medical Journal of Australia*, **182**(3):130–5.

Boerschmann, H. et al. (2010). "Prevalence and predictors of overweight and insulin resistance in offspring of mothers with gestational diabetes mellitus". *Diabetes Care*:DOI: 10.2337/dc10-0139.

Cassady, B. A., R. V. Considine, and R. D. Mattes (2012). "Beverage consumption, appetite, and energy intake: what did you expect?" *The American Journal of Clinical Nutrition*, **95**(3):587–593. DOI: 10.3945/ajcn.111.025437.

Caswell, H. (2009). "The role of fruit juice in the diet: an overview". *Nutrition Bulletin*, **34**(3):273–288. DOI: 10.1111/j.1467-3010.2009.01760.x.

Chortatos, A. et al. (2013). "Nausea and vomiting in pregnancy: associations with maternal gestational diet and lifestyle factors in the Norwegian Mother and Child Cohort Study". *BJOG: An International Journal of Obstetrics & Gynaecology*, **120**(13):1642–1653. DOI: 10.1111/1471-0528.12406.

Cole, T. J. et al. (2000). "Establishing a standard definition for child overweight and obesity worldwide: international survey". *BMJ*, **320**(7244):1240. DOI: 10.1136/bmj.320.7244.1240.

Coolman, M. et al. (2010). "Medical record validation of maternally reported history of preeclampsia". *Journal of Clinical Epidemiology*, **63**(8):932–937. DOI: 10.1016/j.jclinepi.2009.10.010.

De Boo, H. A. and J. E. Harding (2006). "The developmental origins of adult disease (Barker) hypothesis". *Australian and New Zealand Journal of Obstetrics and Gynaecology*, **46**(1):4–14. DOI: 10.1111/j.1479-828X.2006.00506.x.

Derogatis, L. R. and N. Melisaratos (1983). "The Brief Symptom Inventory: an introductory report". *Psychological Medicine*, **13**(3):595–605. DOI: 10.1017/S0033291700048017.

Dewey, K. G. and M. A. McCrory (1994). "Effects of dieting and physical activity on pregnancy and lactation". *The American Journal of Clinical Nutrition*, **59**(2):446S–453S. DOI: 10.1093/ajcn/59.2.446S.

DiMeglio, D. P. and R. D. Mattes (2000). "Liquid versus solid carbohydrate: effects on food intake and body weight". *International Journal of Obesity*, **24**(6):794. DOI: 10.1038/sj.ijo.0801229.

Drewnowski, A. and F. Bellisle (2007). "Liquid calories, sugar, and body weight". *The American Journal of Clinical Nutrition*, **85**(3):651–661. DOI: 10.1093/ajcn/85.3.651.

Erler, N. S., D. Rizopoulos, J. van Rosmalen, et al. (2016). "Dealing with missing covariates in epidemiologic studies: a comparison between multiple imputation and a full Bayesian approach". *Statistics in Medicine*, **35**(17):2955–2974. DOI: 10.1002/sim.6944.

Franke, K. et al. (2005). "'Programming'of orexigenic and anorexigenic hypothalamic neurons in offspring of treated and untreated diabetic mother rats". *Brain Research*, **1031**(2):276–283. DOI: 10.1016/j.brainres.2004.11.006.

Freedman, D. S. et al. (2005). "Relation of BMI to fat and fat-free mass among children and adolescents". *International Journal of Obesity*, **29**(1):1–8. DOI: 10.1038/sj.ijo.0802735.

Hu, F. B. and V. S. Malik (2010). "Sugar-sweetened beverages and risk of obesity and type 2 diabetes: Epidemiologic evidence". *Physiology & Behavior*, **100**(1):47–54. DOI: 10.1016/j.physbeh.2010.01.036.

Hutcheon, J. A., S. Lisonkova, and K. Joseph (2011). "Epidemiology of preeclampsia and the other hypertensive disorders of pregnancy". *Best Practice &*

Research Clinical Obstetrics & Gynaecology, **25**(4):391–403. DOI: 10.1016/j.bpobgyn.2011.01.006.

Kalk, P. et al. (2009). "Impact of maternal body mass index on neonatal outcome". *European Journal of Medical Research*, **14**(5):216. DOI: 10.1186/2047-783X-14-5-216.

Kipnis, V. et al. (2003b). "Structure of dietary measurement error: results of the OPEN biomarker study". *American Journal of Epidemiology*, **158**(1):14–21. DOI: 10.1093/aje/kwg091.

Klipstein-Grobusch, K. et al. (1998). "Dietary assessment in the elderly: validation of a semiquantitative food frequency questionnaire". *European Journal of Clinical Nutrition*, **52**(8):588–596. DOI: 10.1038/sj.ejcn.1600611.

Kooijman, M. N. et al. (2016). "The Generation R Study: design and cohort update 2017". *European Journal of Epidemiology*, **31**(12):1243–1264. DOI: 10.1007/s10654-016-0224-9.

Krebs, N. F. and M. S. Jacobson (2003). "Prevention of Pediatric Overweight and Obesity". *Pediatrics*, **112**(2):424–430. DOI: 10.1542/peds.112.2.424.

Kvaavik, E., L. F. Andersen, and K.-I. Klepp (2005). "The stability of soft drinks intake from adolescence to adult age and the association between long-term consumption of soft drinks and lifestyle factors and body weight". *Public Health Nutrition*, **8**(2):149–157. DOI: 10.1079/PHN2004669.

Leermakers, E. T. et al. (2015). "Sugar-containing beverage intake in toddlers and body composition up to age 6 years: The Generation R Study". *European Journal of Clinical Nutrition*, **69**(3):314. DOI: 10.1038/ejcn.2015.2.

Li, J. et al. (2016). "Paternal programming of offspring cardiometabolic diseases in later life". *Journal of Hypertension*, **34**(11):2111. DOI: 10.1097/HJH.0000000000001051.

Lytle, L. A. et al. (2000). "How Do Children's Eating Patterns and Food Choices Change over Time? Results from a Cohort Study". *American Journal of Health Promotion*, **14**(4):222–228. DOI: 10.4278/0890-1171-14.4.222.

Malik, V. S., B. M. Popkin, et al. (2010). "Sugar-Sweetened Beverages, Obesity, Type 2 Diabetes Mellitus, and Cardiovascular Disease Risk". *Circulation*, **121**(11):1356–1364. DOI: 10.1161/CIRCULATIONAHA.109.876185.

Malik, V. S., M. B. Schulze, and F. B. Hu (2006). "Intake of sugar-sweetened beverages and weight gain: a systematic review". *The American Journal of Clinical Nutrition*, **84**(2):274–288. DOI: 10.1093/ajcn/84.2.274.

Mateo-Fernández, M. et al. (2016). "In Vivo and In Vitro Genotoxic and Epigenetic Effects of Two Types of Cola Beverages and Caffeine: A Multiassay Approach". *BioMed research international*, **2016**:DOI: 10.1155/2016/7574843.

Mennella, J. A., A. Johnson, and G. K. Beauchamp (1995). "Garlic Ingestion by Pregnant Women Alters the Odor of Amniotic Fluid". *Chemical Senses*, **20**(2):207–209. DOI: 10.1093/chemse/20.2.207.

Mickey, R. M. and S. Greenland (1989). "The impact of confounder selection criteria on effect estimation". *American Journal of Epidemiology*, **129**(1):125–137. DOI: 10.1093/oxfordjournals.aje.a115101.

Netherlands Nutrition Center (2006). "Nevo: Dutch food composition database 2006". *The Hague, The Netherlands: Netherlands Nutrition Centre*.

Nguyen, A. N. et al. (2017). "Maternal history of eating disorders: Diet quality during pregnancy and infant feeding". *Appetite*, **109**:108–114. DOI: 10.1016/j.appet.2016.11.030.

Niklasson, A. et al. (1991). "An Update of the Swedish Reference Standards for Weight, Length and Head Circumference at Birth for Given Gestational Age (1977-1981)". *Acta Paediatrica*, **80**(8-9):756–762. DOI: 10.1111/j.1651-2227.1991.tb11945.x.

Pan, A. and F. B. Hu (2011). "Effects of carbohydrates on satiety: differences between liquid and solid food". *Current Opinion in Clinical Nutrition & Metabolic Care*, **14**(4):385–390. DOI: 10.1097/MCO.0b013e328346df36.

Phelan, S. et al. (2011). "Maternal Behaviors during Pregnancy Impact Offspring Obesity Risk". *Experimental Diabetes Research*, **2011**:DOI: 10.1155/2011/985139.

Plummer, M. (2003). "JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling". *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*. Ed. by K. Hornik, F. Leisch, and A. Zeileis. ISSN: 1609-395X.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: http://www.R-project.org/.

Raben, A. et al. (2002). "Sucrose compared with artificial sweeteners: different effects on ad libitum food intake and body weight after 10 wk of supplementation in overweight subjects". *The American Journal of Clinical Nutrition*, **76**(4):721–729. DOI: 10.1093/ajcn/76.4.721.

Reichetzeder, C. et al. (2016). "Developmental Origins of Disease – Crisis Precipitates Change". *Cellular Physiology and Biochemistry*, **39**(3):919–938. DOI: 10.1159/000447801.

Rolland-Cachera, M. et al. (2006). "Early adiposity rebound: causes and consequences for obesity in children and adults". *International Journal of Obesity*, **30**(S4):S11. DOI: 10.1038/sj.ijo.0803514.

Ruyter, J. C. de et al. (2012). "A Trial of Sugar-free or Sugar-Sweetened Beverages and Body Weight in Children". *New England Journal of Medicine*, **367**(15):1397–1406. DOI: 10.1056/NEJMoa1203034.

Schulze, M. B. et al. (2004). "Sugar-sweetened beverages, weight gain, and incidence of type 2 diabetes in young and middle-aged women". *JAMA*, **292**(8):927–934. DOI: 10.1001/jama.292.8.927.

Statistics Netherlands (2003). "Standaard Onderwijsindeling 2003." *Den Haag/Heerlen (Netherlands): Statistics Netherlands*.

Svendsen, O. L. et al. (1993). "Accuracy of measurements of body composition by dual-energy x-ray absorptiometry in vivo". *The American Journal of Clinical Nutrition*, **57**(5):605–608. DOI: 10.1093/ajcn/57.5.605.

Tordoff, M. G. and A. M. Alleva (1990). "Effect of drinking soda sweetened with aspartame or high-fructose corn syrup on food intake and body weight". *The American Journal of Clinical Nutrition*, **51**(6):963–969. DOI: 10.1093/ajcn/51.6.963.

Voortman, T., K. Braun, et al. (2016). "Protein intake in early childhood and body composition at the age of 6 years: The Generation R Study". *International Journal of Obesity*, **40**(6):DOI: 10.1038/ijo.2016.29.

Voortman, T., E. H. van den Hooven, et al. (2016). "Protein intake in early childhood and cardiometabolic health at school age: the Generation R Study". *European Journal of Nutrition*, **55**(6):2117–2127. DOI: 10.1007/s00394-015-1026-7.

Willett, W. C., G. R. Howe, and L. H. Kushi (1997). "Adjustment for total energy intake in epidemiologic studies". *The American Journal of Clinical Nutrition*, **65**(4):1220S–1228S. DOI: 10.1093/ajcn/65.4.1220S.

Zheng, M. et al. (2014). "Sugar-sweetened beverages consumption in relation to changes in body fatness over 6 and 12 years among 9-year-old children: the European Youth Heart Study". *European Journal of Clinical Nutrition*, **68**(1):77. DOI: 10.1038/ejcn.2013.243.

Zheng, M. et al. (2015). "Substituting sugar-sweetened beverages with water or milk is inversely associated with body fatness development from childhood to adolescence". *Nutrition*, **31**(1):38–44. DOI: 10.1016/j.nut.2014.04.017.

5 Bayesian Imputation of Time-varying Covariates in Mixed Models

This chapter is based on

Nicole S. Erler, Dimitris Rizopoulos, Vincent W. V. Jaddoe, Oscar H. Franco and Emmanuel M. E. H. Lesaffre. Bayesian imputation of time-varying covariates in linear mixed models. *Statistical Methods in Medical Research*, 2019; **28**(2), 555 – 568. doi:10.1177/0962280217730851

Abstract

Studies involving large observational datasets commonly face the challenge of dealing with multiple missing values. The most popular approach to overcome this challenge, multiple imputation using chained equations, however, has been shown to be sub-optimal in complex settings, specifically in settings with longitudinal outcomes, which cannot be easily and adequately included in the imputation models. Bayesian methods avoid this difficulty by specification of a joint distribution and thus offer an alternative. A popular choice for that joint distribution is the multivariate normal distribution. In more complicated settings, as in our two motivating examples that involve time-varying covariates, additional issues require consideration: the endo- or exogeneity of the covariate and its functional relation with the outcome. In such situations, the implied assumptions of standard methods may be violated, resulting in bias. In this work, we extend and study a more flexible, Bayesian, alternative to the multivariate normal approach, to better handle complex incomplete longitudinal data. We discuss and compare assumptions of the two Bayesian approaches about the endo- or exogeneity of the covariates and the functional form of the association with the outcome, and illustrate and evaluate consequences of violations of those assumptions using simulation studies and two real data examples.

5.1 Introduction

Missing values are a common challenge in the analysis of observational data, especially in longitudinal studies.

This work is motivated by two research questions from the Generation R Study (Kooijman et al. 2016), a large longitudinal cohort study from fetal life onward. Specifically, the questions are: 1) "How is gestational weight associated with maternal blood pressure during pregnancy?", and 2) "How is gestational weight associated with body mass index of the offspring during the first years of life?". Due to the observational nature of the study, there is a considerable amount of incomplete data, with the particular challenge that missing values do not only occur in the outcome but also in baseline and time-varying covariates.

There are several well-established approaches to deal with incomplete data, the most popular being multiple imputation using chained equations (MICE) (Van Buuren 2012), which are readily available in standard statistical software. MICE has been shown to work well in many standard settings but may not be optimal in more complex applications, especially with longitudinal or other multivariate outcomes, which cannot be easily included in the imputation models for incomplete covariates in an appropriate manner (Erler, Rizopoulos, Rosmalen, et al. 2016). Fully Bayesian approaches provide a useful alternative in such complex settings, due to their ability to jointly model multivariate outcomes and incomplete covariates. The most popular omnibus approach in the Bayesian framework postulates a full multivariate normal distribution (Carpenter and Kenward 2013). Although this approach, as well as other approaches, is targeted towards a broad range of applications, in complex settings such as our two motivating research questions, the nature of the data requires careful consideration of the appropriateness of such standard methods and a more dedicated approach may be necessary. Especially with time-varying covariates, imputation and analysis become more demanding and, in order to obtain valid results, require additional considerations about the association between the time-varying covariates and the outcome. Specifically, endogenous covariates, i.e., covariates that are influenced by the outcome, and covariates that have non-standard functional relations with the outcome, can pose challenges that may or may not be adequately handled by standard methods, which usually assume linear associations and implicitly assume exogeneity of the covariates.

In the present paper, we focus on two approaches in the Bayesian framework to deal with covariates that are missing at random. The first approach is described by Carpenter and Kenward (2013). The basic idea is to assume a (latent) normal distribution for each incomplete variable and to connect them in such a way

that the joint distribution is multivariate normal, which allows straightforward sampling to impute missing values. This approach is a common strategy to implement multiple imputation in longitudinal settings, where it can be used as the data generating step. The resulting data are then analysed in a second step with a complete data method, not necessarily Bayesian. While the multivariate normality assumption creates a convenient standardized framework, it thereby also implies linear relations between the variables involved, which may not be the case.

The second approach factorizes the joint distribution of the data into a sequence of conditional distributions, where the first conditional distribution can conveniently be chosen to be the analysis model of interest, allowing simultaneous imputation and analysis within the same procedure. This approach has been described previously for time-constant covariates (Erler, Rizopoulos, Rosmalen, et al. 2016) and we will extend it in the present paper to handle exogenous as well as endogenous time-varying covariates. The specification of separate models for each incomplete covariate requires somewhat more consideration than the specification of a multivariate normal distribution but makes this approach more flexible as well as capable of handling non-linear relationships. We will elucidate the capabilities and limitations of the two approaches with regards to different functional forms for, as well as endo- or exogeneity of, time-varying covariates and demonstrate how the use of an 'off the shelf' approach may be problematic in settings that require a more tailored approach.

The remainder of this chapter is structured as follows. We start with introducing the motivating dataset and describe in more detail the two research questions from the Generation R Study. In Section 5.3 we specify the linear mixed model for time-varying covariates and explore different functional forms as well as the issue of endo- and exogeneity. The two methods of interest are introduced in Section 5.4, where we will also discuss their implied assumptions about endo- or exogeneity of the covariates and their ability to handle different functional forms. We return to the Generation R data in Section 5.5, where we demonstrate how the two methods under investigation can be applied. A more formal evaluation of the methods follows in Section 5.6 where we perform a simulation study. Section 5.7 concludes this paper with a discussion.

5.2 Generation R Data

The Generation R Study is a population-based prospective cohort study from early fetal life onward, conducted in Rotterdam, the Netherlands (Kooijman et al. 2016). An important field of research within the Generation R Study is the exploration of how the mother's condition during pregnancy may affect her own health and that

of her child. Especially weight gain during gestation is of interest as it is closely related to the development of the fetus, as well as to pregnancy comorbidities, such as gestational hypertension, that may adversely affect both mother and child (e.g., Tielemans et al. (2015)). Children's growth and body composition, as for instance measured by BMI, is an important determinant of health throughout childhood and later life. Therefore, current research is concerned with the two questions stated in Section 5.1, i.e., the associations between maternal weight (gain) during pregnancy with maternal blood pressure during pregnancy as well as with child BMI after birth.

To investigate these two research questions, a subset of variables was extracted from the Generation R Study. The dataset contains information on 7643 mothers who had singleton, live births no earlier than 37 weeks of gestation, and their children. Each woman was asked for her pre-pregnancy weight (baseline) and to visit the research centre once in each trimester, during which the weight (`gw`) was measured and the blood pressure taken. Since women were eligible to enter the study at any gestational age, prenatal measurements for the first and second trimester are missing for women who enrolled later in pregnancy. Furthermore, there is some intermittent missingness in the gestational and blood pressure data. There were 3515 women for whom all four weight measurements were recorded, 3094 for whom three weight measurements were observed, 859 women had two measurements, and 175 women had only one measurement of weight. The gestational age at each measurement (`gage`) was recorded and the time point of the baseline measurement was set to be zero for all women. Systolic blood pressure (`bp`) was measured three times in 4755 women, 2403 women had only two measurements of blood pressure and 477 women just one measurement. For 8 women no blood pressures were recorded. Child BMI was measured up to 12 times between the ages of 2 weeks and 5 years, with a median of 7 observations per child; 1848 children had no BMI measurements. The child's age in months (`age`) was recorded at each BMI measurement and age and sex adjusted standard deviation scores were calculated (`bmi`). A graphical summary of the missingness pattern of the gestational weight and systolic blood pressure measurements, and the available child BMI measurements is given in Figures 5.4 – 5.6 in Appendix 5.B.

The trajectories of `gw`, `bp` and `bmi` of a random subset of individuals are visualized in Figure 5.1. Furthermore, we considered a number of potential confounders: maternal age at intake (`agem`, continuous, complete), maternal height (`height`, continuous, 0.38% missing values), parity (`parity`, binary: nulliparous vs multiparous, 1.27% missing values), maternal ethnicity (`ethn`, binary: European vs other, 5.59% missing values), maternal education (`educ`, three ordered categories, 9.29% missing values) and maternal smoking habit during pregnancy (`smoke`, three

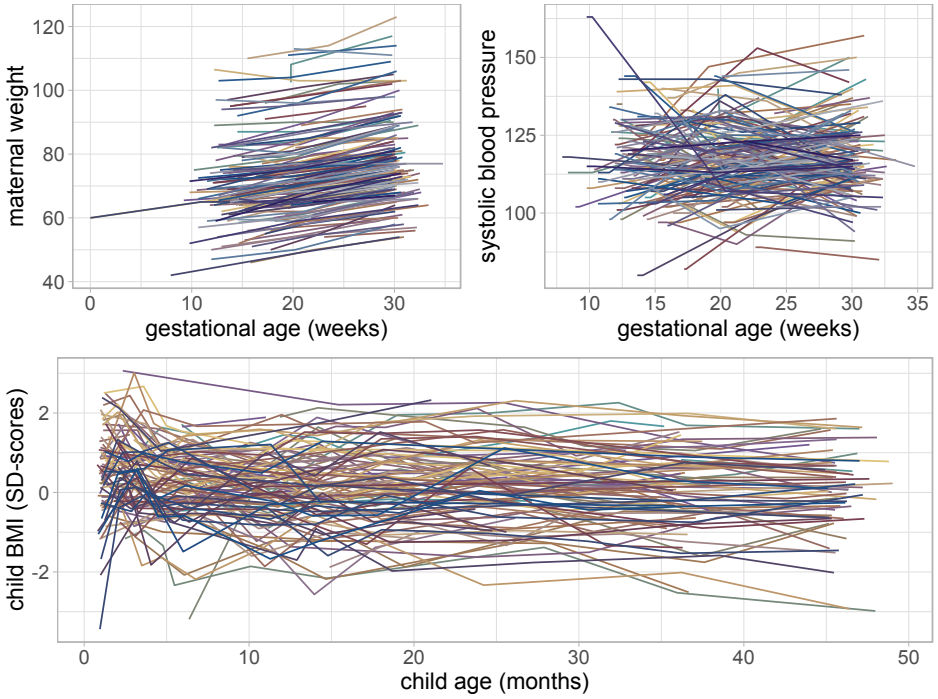


Figure 5.1: Trajectories of maternal weight, maternal systolic blood pressure and child BMI for a random sample of mothers and children from the Generation R data.

ordered categories, 12.17% missing values). Maternal BMI (`bmim`) was calculated as gestational weight (kg), measured at time zero, divided by squared height (in m).

Logistic regression of the complete cases showed that missingness in the baseline covariates (except for parity) was associated with some or all of the other baseline covariates. This indicates that missing values are not completely at random. However, since this study was conducted in the general population subjects are relatively healthy and the practical settings are such that it is reasonable to believe that missing values in the clinical measurements, as for instance `gw` or `bmi`, are at random, given the other variables. It could be argued that missing values in the lifestyle variables, especially in `smoke`, are not missing at random because mothers who are smoking might be more inclined not to report it. If this was the case, the mechanism that led to the missing values had to be included in the imputation procedure since otherwise results would be biased. However, the assumption of

randomly missing data is untestable, and the missing data mechanism is usually unknown, necessitating extensive sensitivity analysis. As this exceeds the purpose of this study, we will focus here on randomly missing data. To make the assumption of randomly missing data more plausible, a number of covariates will be considered in the analysis model, since omission of relevant predictor variables may be another reason of not randomly missing data.

5.3 Modelling Longitudinal Data with Time-varying Covariates

5.3.1 Framework

A standard modelling framework for studying the relation between a longitudinal outcome and predictor variables is mixed effects modelling. As in our motivating case studies, often some of these predictors are time-varying. To facilitate exposition and also for notational simplicity in the following we only consider a single time-varying covariate. In particular, for a continuous longitudinal outcome we postulate the following mixed model

$$y_i(t) = \mathbf{x}_i(t)^\top \boldsymbol{\beta} + f(H_i^s(t), t)^\top \boldsymbol{\gamma} + \mathbf{z}_i(t)^\top \mathbf{b}_i + \varepsilon_i(t),$$

where $y_i(t)$ is the observation of individual i measured at time t, $\boldsymbol{\beta}$ denotes the vector of regression coefficients of the design matrix of the fixed effects $\mathbf{X}_i$, with $\mathbf{x}_i(t)$ being a column vector containing a row of that matrix, $\mathbf{z}_i(t)$, a column vector expressing a row of the design matrix $\mathbf{Z}_i$ of the random effects $\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{D})$, $\boldsymbol{\gamma}$ is a vector of regression coefficients related to the time-varying covariate $\mathbf{s}_i$, and $\varepsilon_i(t) \sim N(0, \sigma_y^2)$ is an error term. Except for time itself, $\mathbf{X}$ does not contain any time-varying covariates. To include $\mathbf{s}$ in the linear predictor of $\mathbf{y}$ assumptions about the relation between the two variables have to be made. These assumptions can be expressed by specifying a function $f(H_i^s(t), t)$ which links the history of the time-varying predictor up to time t, $H_i^s(t) = \{s_i(t_{ij}) : 0 \leq t_{ij} \leq t,\ j = 1, \ldots, n_i^s\}$, to the outcome, where t_{ij} is the time of the j-th measurement of individual i and n_i^s is the number of measurements of $\mathbf{s}$ for that individual.

5.3.2 Functional Forms for Time-varying Covariates

The choice of an appropriate functional form implies that the following two questions need to be addressed. Namely, how are $\mathbf{s}_i$ and $\mathbf{y}_i$ related with regards to their time scales, and which features of $\mathbf{s}_i$ are of interest in the relation with $\mathbf{y}_i$? The first question asks whether $\mathbf{y}_i$ and $\mathbf{s}_i$ have been measured in the same time intervals and whether their time scales have the same origin and unit. To allow for

settings where $\mathbf{y}_i$ and $\mathbf{s}_i$ have been measured on different time scales, for instance maternal weight during pregnancy and child BMI after birth, we use t to denote the time scale of $\mathbf{y}_i$ and $\tilde{t}$ to denote the time scale of $\mathbf{s}_i$. The second question relates to the specific application and is reflected in the choice of $f(\cdot)$. Choices, that represent relevant features of gestational weight in our two motivating research questions are

$$f(H_i^s(t), t) = s_i(t), \tag{5.1}$$
$$f(H_i^s(t), t) = \{\Delta_1(\mathbf{s}_i), \Delta_2(\mathbf{s}_i), \Delta_3(\mathbf{s}_i)\}^\top, \tag{5.2}$$

with

$$\begin{aligned}\Delta_1(\mathbf{s}_i) &= s_i(\tilde{t}_1) - s_i(\tilde{t}_0),\\ \Delta_2(\mathbf{s}_i) &= s_i(\tilde{t}_2) - s_i(\tilde{t}_1),\\ \Delta_3(\mathbf{s}_i) &= s_i(\tilde{t}_3) - s_i(\tilde{t}_2),\end{aligned}$$

where (5.1) represents the commonly chosen linear relation between the value of $\mathbf{s}_i$, e.g., maternal weight, and $\mathbf{y}_i$, e.g., blood pressure, measured at the same time points (i.e., $t = \tilde{t}$). Function (5.2) represents trimester specific weight gain, i.e., the difference of maternal weight over three given time intervals. In a more general notation, (5.1) could be written as $f(H_i^s(t), t) = s_i(g(t))$ and refer to the value of $\mathbf{s}_i$ at a time point that is specified by a function $g(t)$, and (5.2) could be written as $f(H_i^s(t), t) = s_i(g_2(t)) - s_i(g_1(t))$, where the time intervals are specified by the functions $g_1(t)$ and $g_2(t)$ and may not be the same for all t.

In other applications, it is likely that different functional forms will be more appropriate. Such functions may, for instance, represent cumulative effects or use estimates of random effects associated with the individual profiles of the time-varying covariate. In cases where there is not a specific functional form of interest or there is uncertainty about which functional form is most appropriate, multiple functional forms can be included and shrinkage priors used to reduce correlations between parameters or to select the best suited functional form (Andrinopoulou and Rizopoulos 2016).

5.3.3 Endo- and Exogeneity

Another characteristic of the relation between a time-varying covariate and the outcome that needs to be considered is whether the time-varying covariate is exogenous or endogenous. Formally, exogeneity is defined by the following two

conditions (Engle et al. 1983; Diggle et al. 2002)

$$\begin{cases} p(y_i(t), f(H_i^s(t), t) \mid H_i^y(t^-), H_i^s(t^-), \boldsymbol{\theta}) = \\ \qquad\qquad = p(y_i(t) \mid f(H_i^s(t), t), H_i^y(t^-), H_i^s(t^-), \boldsymbol{\theta}_1) \times \\ \qquad\qquad\quad p(s_i(t) \mid H_i^y(t^-), H_i^s(t^-), \boldsymbol{\theta}_2) \\ \\ p(s_i(t) \mid H_i^s(t^-), H_i^y(t^-), \mathbf{x}_i, \boldsymbol{\theta}) = p(s_i(t) \mid H_i^s(t^-), \mathbf{x}_i, \boldsymbol{\theta}) \end{cases}$$

where $\boldsymbol{\theta}$ is a vector of parameters and other unknown quantities, with $\boldsymbol{\theta}^\top = (\boldsymbol{\theta}_1^\top, \boldsymbol{\theta}_2^\top)$ and $\boldsymbol{\theta}_1 \perp\!\!\!\perp \boldsymbol{\theta}_2$, and where $H_i^y(t^-)$ and $H_i^s(t^-)$ denote the history of $\mathbf{y}$ and $\mathbf{s}$, respectively, up to, but excluding measurements at time t. By specifying the functional relation between $\mathbf{y}_i$ and $\mathbf{s}_i$ to be a function of the history of $\mathbf{s}$, we avoid dependence of $y_i(t)$ on future values of $\mathbf{s}_i$, which is an additional requirement for exogeneity, see Diggle et al. (2002). Variables, for which these conditions are not satisfied are called *endogenous*. This may be the case for maternal weight as a predictor variable for blood pressure. Since both variables are measured in the same individual they may be subject to the same unmeasured influences or causal pathways may be reversed, which often entails endogeneity. In the setting where maternal weight is considered as a predictor of child BMI, however, the assumption of exogeneity may be more likely, since the covariate is measured earlier than the outcome and in different subjects.

Most common methods for inference, like generalized linear (mixed) regression models, assume covariates to be exogenous. If that assumption is wrong and the covariate is, in fact, endogenous, estimates may be biased (Diggle et al. 2002; Daniels and Hogan 2008).

5.4 Bayesian Analysis with Incomplete Covariates

As introduced in Section 5.2, the motivating questions from the Generation R Study involve outcomes and covariates that are incomplete. This holds for both the baseline and time-varying covariates. Hence, to appropriately investigate the associations of interest we need to account for missingness. In the Bayesian framework, missing values, whether they are in the outcome or in covariates, can be imputed in a natural and elegant manner. A common assumption, which we make here for the outcome as well as the covariates, is that the missing data mechanism is Missing At Random (MAR), i.e., the probability of a value being unobserved may depend on other observed values but not on values that have not been observed. In addition, the parameters of the analysis model are assumed to be independent of the missingness process. Under these assumptions, the missingness process is ignorable and does not need to be modelled (Little and Rubin 2002). Furthermore,

this assumption entails that explicit imputation of the outcome is not necessary to obtain valid results, and we will, therefore, focus on settings with incomplete covariates. In this section, we adapt and implement two popular Bayesian approaches for analysing data with incomplete covariates, namely, the sequential approach (Ibrahim et al. 2002; Erler, Rizopoulos, Rosmalen, et al. 2016), and the multivariate normal approach (Carpenter and Kenward 2013). In particular, we extend the first approach to settings with time-varying covariates that may be exogenous or endogenous. Both approaches model the joint distribution of the complete data and draw imputations from the posterior full conditional distributions that result from it but differ in the way the joint complete data distribution is specified. These differences influence how the two approaches can handle different functional forms as well as exo- versus endogenous covariates.

We start with some additional notation. As in the motivating data, the time-varying covariate $\mathbf{s}$ is assumed to be incomplete. Missing values in $\mathbf{s}$ occur not only due to missed measurements or drop-out but can also be caused when the functional form $f(H_i^s(t), t)$ depends on values of $\mathbf{s}$ that have not been (scheduled to be) measured. We use $\mathbf{s}_i = (\mathbf{s}_{i,obs}^\top, \mathbf{s}_{i,mis}^\top)^\top$ to distinguish between the observed and missing values of $\mathbf{s}$ for individual i. Analogously, we assume two parts for the baseline covariates $\mathbf{X}$ on the individual level: $\mathbf{x}_{i,obs}$ and $\mathbf{x}_{i,mis}$, which contain the observed and missing values of $\mathbf{x}_i$, respectively. Furthermore, we use the partition $\mathbf{X} = (\mathbf{X}_c, \mathbf{x}_1, \ldots, \mathbf{x}_p)$, where $\mathbf{X}_c$ denotes the subset of covariates that are completely observed for all individuals, and $\mathbf{x}_1, \ldots, \mathbf{x}_p$ are $n \times 1$ vectors of those covariates that contain missing values.

5.4.1 Sequential Approach

The sequential approach to impute missing baseline covariates in models with longitudinal outcomes was previously presented by Erler, Rizopoulos, Rosmalen, et al. (2016) and will be extended here to incomplete time-varying covariates.

In our setting, the posterior distribution of interest (and associated joint distribution) is

$$\begin{aligned} p(\mathbf{s}_{mis}, \mathbf{X}_{mis}, \mathbf{b}, \boldsymbol{\theta} \mid \mathbf{y}, \mathbf{s}_{obs}, \mathbf{X}_{obs}) &\propto p(\mathbf{y}, \mathbf{s}_{obs}, \mathbf{X}_{obs} \mid \mathbf{s}_{mis}, \mathbf{X}_{mis}, \mathbf{b}, \boldsymbol{\theta}) \\ &\quad\; p(\mathbf{s}_{mis}, \mathbf{X}_{mis}, \mathbf{b}, \boldsymbol{\theta}) \\ &= p(\mathbf{y}, \mathbf{s}_{obs}, \mathbf{s}_{mis}, \mathbf{X}_{obs}, \mathbf{X}_{mis}, \mathbf{b}, \boldsymbol{\theta}), \end{aligned}$$

where $\boldsymbol{\theta}$ is a vector of unknown parameters, and can be factorized as

$$p(\mathbf{y} \mid \mathbf{s}, \mathbf{X}, \mathbf{b}, \boldsymbol{\theta})\; p(\mathbf{s} \mid \mathbf{X}, \mathbf{b}, \boldsymbol{\theta})\; p(\mathbf{X} \mid \boldsymbol{\theta})\; p(\mathbf{b} \mid \boldsymbol{\theta})\; p(\boldsymbol{\theta}), \tag{5.3}$$

for which all terms can be specified based on known distributions.

The first term in (5.3), i.e., $p(\mathbf{y} \mid \mathbf{s}, \mathbf{X}, \mathbf{b}, \boldsymbol{\theta})$, is conveniently chosen to be the analysis model of interest,

$$y_i(t) = \mathbf{x}_i(t)^\top \boldsymbol{\beta}_{y|s,x} + f(H_i^s(t), t)^\top \boldsymbol{\gamma} + \mathbf{z}_i^y(t)^\top \mathbf{b}_i^y + \varepsilon_i^y(t), \tag{5.4}$$

with random effects $\mathbf{b}_i^y \sim N(0, \mathbf{D}_y)$ and $\boldsymbol{\varepsilon}_i^y(t) \sim N(0, \sigma_y^2)$, and the second factor, representing the imputation model for the time-varying covariate, can be specified analogously as a linear mixed model

$$s_i(\tilde{t}) = \mathbf{x}_i(\tilde{t})^\top \boldsymbol{\beta}_{s|x} + \mathbf{z}_i^s(\tilde{t})^\top \mathbf{b}_i^s + \varepsilon_i^s(\tilde{t}), \tag{5.5}$$

with $\mathbf{b}_i^s \sim N(0, \mathbf{D}_s)$ and $\boldsymbol{\varepsilon}_i^s(\tilde{t}) \sim N(0, \sigma_s^2)$. All variance matrices $\mathbf{D}$ and parameters σ^2 are assumed to follow vague inverse-Wishart and inverse-gamma distributions, respectively. Inclusion of $f(H_i^s(t), t)$ in the linear predictor for $\mathbf{y}_i$ allows for a large variety of possibly non-linear relations between $\mathbf{y}_i$ and $\mathbf{s}_i$, also when they are measured on different time scales. The joint distribution of the baseline covariates $\mathbf{X}$ is often a multivariate distribution of mixed type variables for which usually no closed form solution is known. It can, however, be specified as a sequence of univariate conditional distributions (Ibrahim et al. 2002; Erler, Rizopoulos, Rosmalen, et al. 2016),

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_p \mid \mathbf{X}_c, \boldsymbol{\theta}_x) = p(\mathbf{x}_1 \mid \mathbf{X}_c, \boldsymbol{\theta}_{x_1}) \prod_{\ell=2}^{p} p(\mathbf{x}_\ell \mid \mathbf{X}_c, \mathbf{x}_1, \ldots, \mathbf{x}_{\ell-1}, \boldsymbol{\theta}_{x_\ell}), \tag{5.6}$$

with $\boldsymbol{\theta}_x^\top = (\boldsymbol{\theta}_{x_1}^\top, \ldots, \boldsymbol{\theta}_{x_p}^\top)$, where $\mathbf{x}_\ell$ denotes the ℓ-th incomplete covariate. The univariate conditional distributions are assumed to be members of the exponential family, extended with distributions for ordinal categorical variables, with linear predictors

$$g_\ell \left\{ E\left(\mathbf{x}_\ell \mid \mathbf{X}_c, \mathbf{x}_1, \ldots, \mathbf{x}_{\ell-1}, \boldsymbol{\theta}_{x_\ell} \right) \right\} \quad = \quad \mathbf{X}_c \boldsymbol{\alpha}_\ell + \sum_{q=1}^{\ell-1} \mathbf{x}_q \xi_{\ell_q},$$

which allows an easy and flexible specification in settings with many covariates of mixed type, since each link function g_ℓ can be chosen separately and appropriately for $\mathbf{x}_\ell$. Factorizing the joint distribution of the data as in (5.3) has the advantage that the parameters of interest, $\boldsymbol{\beta}_{y|s,x}$, are estimated within each iteration of the imputation procedure, conditional on the current value of the imputed covariates. The simultaneity of imputation and analysis leads to a posterior distribution of the parameters, which automatically takes into account the uncertainty due to the missing values and no subsequent analysis and pooling, as in the case of multiple imputation approaches, is necessary. Furthermore, the sequential approach differs from MICE in the specification of the imputation models. MICE requires the

specification of full conditional distributions, i.e., to include all other covariates as well as the outcome in the linear predictor of the imputation models, which is not straightforward when the outcome is longitudinal, and may lead to imputation models that are not compatible with the analysis model (Carpenter and Kenward 2013; Bartlett et al. 2015).

In the specification described above, the sequential approach implies exogeneity of $\mathbf{s}_i$ with regards to the conditions given in Section 5.3.3, which is demonstrated in Appendix 5.A.1. It can be extended to endogenous time-varying covariates by jointly modelling the random effects from models (5.4) and (5.5) as

$$\begin{bmatrix} \mathbf{b}_i^y \\ \mathbf{b}_i^s \end{bmatrix} \sim N\left(\mathbf{0}, \begin{bmatrix} \mathbf{D}_y & \mathbf{D}_{ys} \\ \mathbf{D}_{ys} & \mathbf{D}_s \end{bmatrix}\right), \qquad \mathbf{D}_{ys} \neq \mathbf{0}.$$

When $\mathbf{b}_i^y$ and $\mathbf{b}_i^s$ are correlated, the joint distribution of the random effects $p(\mathbf{b}_i^y, \mathbf{b}_i^s)$ is not equal to the product of the marginal distributions $p(\mathbf{b}_i^y)$ and $p(\mathbf{b}_i^s)$ any more and the exogeneity conditions are no longer satisfied (for details see Appendix 5.A.2). The sequential approach can be further extended to endogenous baseline covariates by relaxing the assumption of independence between the residuals of the covariate and the analysis model, e.g., by assuming a joint distribution of the residuals and the random effects $\mathbf{b}_i^y$.

5.4.2 Multivariate Normal Approach

A popular alternative to handle missing covariates is the multivariate normal approach described in detail by Carpenter and Kenward (2013). The idea behind this approach is to assume (latent) normal distributions for all incomplete variables and the outcome, and to connect them in such a way that the resulting joint distribution is multivariate normal, which eases the sampling of imputed values. Specification of the joint distribution of the data is, hence, not based on a sequence but on a chosen multivariate distribution of known type. In our setting, the posterior distribution of interest can be written and factorized as

$$\begin{aligned} p(\mathbf{s}_{mis}, \mathbf{X}_{mis}, \tilde{\mathbf{b}}, \tilde{\boldsymbol{\theta}} \mid \mathbf{y}, \mathbf{s}_{obs}, \mathbf{X}_{obs}) &\propto p(\mathbf{y}, \mathbf{s}, \mathbf{X}_{mis}, \mathbf{X}_{obs}, \tilde{\mathbf{b}}, \tilde{\boldsymbol{\theta}}) \\ &= p(\mathbf{y}, \mathbf{s}, \mathbf{X}_c, \mathbf{x}_1, \ldots, \mathbf{x}_p, \tilde{\mathbf{b}}, \tilde{\boldsymbol{\theta}}) \\ &= p(\mathbf{y}, \mathbf{s}, \mathbf{x}_1, \ldots, \mathbf{x}_p \mid \mathbf{X}_c, \tilde{\mathbf{b}}, \tilde{\boldsymbol{\theta}}) \\ &\quad p(\tilde{\mathbf{b}} \mid \tilde{\boldsymbol{\theta}})\, p(\tilde{\boldsymbol{\theta}}), \end{aligned} \tag{5.7}$$

where the first factor on the right side of (5.7) is assumed to be a multivariate normal distribution, $\tilde{\mathbf{b}}$ are random effects that are associated with $\mathbf{y}$ and $\mathbf{s}$, and $\tilde{\boldsymbol{\theta}}$ is a vector of parameters. The multivariate normal distribution can be constructed

by specifying linear (mixed) models for the outcome and incomplete covariates, i.e., the time-varying and incomplete baseline covariates,

$$\begin{aligned} y_i(t) &= \mathbf{x}_{i,c}^{y}(t)^\top \tilde{\boldsymbol{\beta}}_y + \tilde{\mathbf{z}}_i^y(t)\tilde{\mathbf{b}}_i^y + \tilde{\varepsilon}_i^y(t) \\ s_i(t) &= \mathbf{x}_{i,c}^{s}(t)^\top \tilde{\boldsymbol{\beta}}_s + \tilde{\mathbf{z}}_i^s(t)\tilde{\mathbf{b}}_i^s + \tilde{\varepsilon}_i^s(t) \\ \hat{x}_{i,\ell} &= \mathbf{x}_{i,c}^{x\ \top} \tilde{\boldsymbol{\beta}}_{x,\ell} + \tilde{\varepsilon}_{i,\ell}^x, \qquad \ell = 1, \ldots, p, \end{aligned}$$

where $\mathbf{x}_{i,c}^y(t)$, $\mathbf{x}_{i,c}^s(t)$ and $\mathbf{x}_{i,c}^x$ are rows of the matrices $\mathbf{X}_c^y$, $\mathbf{X}_c^s$ and $\mathbf{X}_c^x$ which are (possibly different) subsets of $\mathbf{X}_c$, $\hat{x}_{i,\ell}$ denotes the value from a (latent) normal distribution that corresponds to the missing value of the ℓ-th incomplete covariate, for individual i, $\tilde{\boldsymbol{\beta}}_y$, $\tilde{\boldsymbol{\beta}}_s$ and $\tilde{\boldsymbol{\beta}}_x = (\tilde{\boldsymbol{\beta}}_{x_1}^\top, \ldots, \tilde{\boldsymbol{\beta}}_{x_p}^\top)^\top$ are regression coefficients, $\tilde{\mathbf{z}}_i^y(t)$ and $\tilde{\mathbf{z}}_i^s(t)$ are rows of the design matrices $\widetilde{\mathbf{Z}}_i^y$ and $\widetilde{\mathbf{Z}}_i^s$ of the random effects $\tilde{\mathbf{b}}_i^y$ and $\tilde{\mathbf{b}}_i^s$. Note that the models specified here are different from the ones in the sequential approach, since here the predictors only contain the completely observed covariates $\mathbf{X}_c$. The parameters $\tilde{\boldsymbol{\beta}}$ are not the same as the parameters $\boldsymbol{\beta}_{y|s,x}$, used in the sequential approach. To obtain estimates of $\boldsymbol{\beta}_{y|s,x}$ that take into account the uncertainty due to the missing values, multiple imputation may be performed. This involves repeating the imputation a number of times to create multiple imputed datasets, which can then be analysed with appropriate Bayesian or non-Bayesian methods. Pooled estimates from frequentist analyses can be calculated using Rubin's Rules (Rubin 1987). Although imputation with the multivariate normal approach is valid for endogenous covariates, this may not be the case for many standard analysis methods that imply exogeneity of the covariates, which may pose an additional challenge.

To produce the multivariate normal distribution, the models specified above are then connected through their random effects and error terms which are assumed to have a joint multivariate normal distribution

$$\begin{bmatrix} \tilde{\mathbf{b}}_i^y \\ \tilde{\mathbf{b}}_i^s \\ \tilde{\boldsymbol{\varepsilon}}_i^x \end{bmatrix} \sim N\left(\mathbf{0}, \begin{bmatrix} \widetilde{\mathbf{D}}_y & \widetilde{\mathbf{D}}_{y,s} & \operatorname{cov}(\tilde{\mathbf{b}}_i^y, \tilde{\boldsymbol{\varepsilon}}_i^x) \\ \widetilde{\mathbf{D}}_{y,s} & \widetilde{\mathbf{D}}_s & \operatorname{cov}(\tilde{\mathbf{b}}_i^s, \tilde{\boldsymbol{\varepsilon}}_i^x) \\ \operatorname{cov}(\tilde{\mathbf{b}}_i^y, \tilde{\boldsymbol{\varepsilon}}_i^x) & \operatorname{cov}(\tilde{\mathbf{b}}_i^s, \tilde{\boldsymbol{\varepsilon}}_i^x) & \widetilde{\boldsymbol{\Sigma}}^x \end{bmatrix}\right), \tag{5.8}$$

where $\widetilde{\mathbf{D}}_y$ and $\widetilde{\mathbf{D}}_s$ denote the covariance matrices of the random effects $\tilde{\mathbf{b}}_i^y$ and $\tilde{\mathbf{b}}_i^s$, respectively, $\widetilde{\mathbf{D}}_{y,s}$ is a matrix containing parameters that describe the covariance between the two sets of random effects, and $\widetilde{\boldsymbol{\Sigma}}^x$ is the, usually diagonal, covariance matrix of the error terms $\tilde{\boldsymbol{\varepsilon}}_i^x = (\tilde{\varepsilon}_{i,1}^x, \ldots, \tilde{\varepsilon}_{i,p}^x)^\top$.

The error terms of the two longitudinal variables are assumed to be normally distributed as well, and may be modelled jointly as

$$\begin{bmatrix} \tilde{\boldsymbol{\varepsilon}}_i^y \\ \tilde{\boldsymbol{\varepsilon}}_i^s \end{bmatrix} \sim N\left(\mathbf{0}, \begin{bmatrix} \widetilde{\boldsymbol{\Sigma}}^y & \widetilde{\boldsymbol{\Sigma}}^{y,s} \\ \widetilde{\boldsymbol{\Sigma}}^{y,s} & \widetilde{\boldsymbol{\Sigma}}^s \end{bmatrix}\right),$$

where $\widetilde{\boldsymbol{\Sigma}}^{y}$ and $\widetilde{\boldsymbol{\Sigma}}^{s}$ denote the covariance matrices of $\tilde{\boldsymbol{\varepsilon}}_i^{y}$ and $\tilde{\boldsymbol{\varepsilon}}_i^{s}$, respectively, and $\widetilde{\boldsymbol{\Sigma}}^{y,s}$ is a matrix describing the covariance between the error terms of $\mathbf{y}_i$ and the error terms of $\mathbf{s}_i$. Allowing the error terms of the longitudinal variables to be correlated allows for more flexibility. Which covariance structure is appropriate, however, depends on the unknown functional relation between $\mathbf{y}$ and $\mathbf{s}$.

The latent normal model for a binary or ordinal covariate $\mathbf{x}_{mis_\ell}$ with K categories can be written as

$$\begin{aligned}
&\hat{x}_{i,mis_\ell} \leq \kappa_1 && \text{if } x_{i,mis_\ell} = 1,\\
&\hat{x}_{i,mis_\ell} \in (\kappa_{k-1}, \kappa_k] && \text{if } x_{i,mis_\ell} = k,\ k \in (2, \ldots, K-1),\\
&\hat{x}_{i,mis_\ell} > \kappa_{K-1} && \text{if } x_{i,mis_\ell} = k.
\end{aligned}$$

To keep the model identified the variance of $\hat{x}_{i,mis_\ell}$ has to be fixed, e.g., to one, which complicates sampling of $\widetilde{\boldsymbol{\Sigma}}^{x}$. For continuous covariates $\hat{x}_{i,mis_\ell} = x_{i,mis_\ell}$ and no restriction of the variance is necessary.

As in the sequential approach, the use of variable specific random effects design matrices $\widetilde{\mathbf{Z}}_{\mathbf{i}}^{\mathbf{y}}$ and $\widetilde{\mathbf{Z}}_i^{s}$ enables this approach to handle time-varying covariates that are measured on a different time scale than the outcome. The connection of the imputation models by joint random effects and/or error terms, however, implies a linear relation between the variables. When the relation between $\mathbf{y}_i$ and $\mathbf{s}_i$ is non-linear, the true joint distribution is not multivariate normal and does not generally have a closed form (Carpenter and Kenward 2013). The multivariate normal approach may, hence, be less suitable for applications with non-linear relations.

Since $\mathbf{b}_i^{y}$ and $\mathbf{b}_i^{s}$ are modelled jointly and assumed to be correlated the conditions for exogeneity are violated, which can be shown using similar arguments as provided in Appendix 5.A.2 for the sequential approach with correlated random effects. The multivariate normal approach thus implies endogeneity of $\mathbf{s}_i$.

5.5 Analysis of the Generation R Data

We now return to the Generation R data introduced in Section 5.2 and demonstrate how to use the two methods discussed above to investigate the two motivating research questions. As indicated earlier, the first question enables the investigation of the impact of misspecifying the exo-/endogeneity assumption, while the second question requires the use of a non-standard functional form.

5.5.1 Association between Blood Pressure and Gestational Weight

Gestational hypertension is a known risk factor for various health outcomes in mothers as well as their children. One potentially influential factor for this condition is gestational weight, which will be investigated here. Whilst there are several papers exploring the relationship of gestational weight gain and the development of hypertensive conditions during pregnancy, the exact nature and functional form of the relation between these variables, has yet to be explored in detail. Given the information available and the characteristics of the dataset at hand, a reasonable choice of functional form is to assume a linear relation between gestational weight (gw) and systolic blood pressure (bp) at the same time points, i.e., $f(H_i^{\mathtt{gw}}(t), t) = \mathtt{gw}_i(t)$. Furthermore, the relation between these two variables is likely influenced by many unmeasured factors, which makes the standard assumption of exogeneity for gestational weight questionable. To investigate how much the estimates may be influenced by the assumption of exo- or endogeneity in practice, we performed the analysis twice, once under the assumption that gestational weight was endogenous, and once under the common default assumption of exogeneity, and compared the results.

Since both longitudinal variables in this application have non-linear evolutions over time, we modelled their trajectories using natural cubic splines with two degrees of freedom (df) for the effect of gestational age, in the formulas below represented by $\mathtt{ns}_i^{(1)}(t)$, $\mathtt{ns}_i^{(2)}(t)$, and $\widetilde{\mathtt{ns}}_i^{(1)}(\tilde{t})$, $\widetilde{\mathtt{ns}}_i^{(2)}(\tilde{t})$, respectively. Taking into account a number of potential confounding covariates (see Section 5.2), the model of interest in this application can be written as

$$\begin{aligned}\mathtt{bp}_i(t) \;=\;& (\beta_0 + b_{i0}^{\mathtt{bp}}) + \beta_1\mathtt{agem}_i + \beta_2\mathtt{height}_i + \beta_3\mathtt{parity}_i \\ &+\beta_4\mathtt{ethn}_i + \beta_5\mathtt{educ}_i^{(2)} + \beta_6\mathtt{educ}_i^{(3)} + \beta_7\mathtt{smoke}_i^{(2)} \\ &+\beta_8\mathtt{smoke}_i^{(3)} + (\beta_9 + b_{i1}^{\mathtt{bp}})\mathtt{ns}_i^{(1)}(t) + (\beta_{10} + b_{i2}^{\mathtt{bp}})\mathtt{ns}_i^{(2)}(t) \\ &+\gamma\mathtt{gw}_i(t) + \varepsilon_i^{\mathtt{bp}}(t).\end{aligned}$$

In the sequential approach, we used a linear mixed model to impute missing values of gw, specifically,

$$\begin{aligned}\mathtt{gw}_i(\tilde{t}) \;=\;& (\alpha_0 + b_{i0}^{\mathtt{gw}}) + \alpha_1\mathtt{agem}_i + \alpha_2\mathtt{height}_i + \alpha_3\mathtt{parity}_i \\ &+\alpha_4\mathtt{ethn}_i + \alpha_5\mathtt{educ}_i^{(2)} + \alpha_6\mathtt{educ}_i^{(3)} + \alpha_7\mathtt{smoke}_i^{(2)} \\ &+\alpha_8\mathtt{smoke}_i^{(3)} + (\alpha_9 + b_{i1}^{\mathtt{gw}})\widetilde{\mathtt{ns}}_i^{(1)}(\tilde{t}) + (\alpha_{10} + b_{i2}^{\mathtt{gw}})\widetilde{\mathtt{ns}}_i^{(2)}(\tilde{t}) \\ &+\varepsilon_i^{\mathtt{gw}}(\tilde{t}),\end{aligned}$$

and specified the conditional distributions for the missing covariates from Equation (5.6) as linear, logistic and cumulative logistic regression models. The random effects of the models for `gw` and `bp` were modelled jointly as $(b_{i0}^{\texttt{bp}}, b_{i1}^{\texttt{bp}}, b_{i2}^{\texttt{bp}}, b_{i0}^{\texttt{gw}}, b_{i1}^{\texttt{gw}}, b_{i2}^{\texttt{gw}})^\top \sim N(\mathbf{0}, \mathbf{D})$ in the endogenous setting and independently as $(b_{i0}^{\texttt{bp}}, b_{i1}^{\texttt{bp}}, b_{i2}^{\texttt{bp}},)^\top \sim N(\mathbf{0}, \mathbf{D}_{\texttt{bp}})$ and $(b_{i0}^{\texttt{gw}}, b_{i1}^{\texttt{gw}}, b_{i2}^{\texttt{gw}})^\top \sim N(\mathbf{0}, \mathbf{D}_{\texttt{gw}})$ in the exogenous setting. Vague priors were used for all parameters. Following the advice of Garrett and Zeger (2000), we assumed independent normal distributions with mean zero and variance 9/4 for regression coefficients in categorical models (logistic and cumulative logistic) since that choice leads to a prior distribution for the outcome probabilities that is relatively flat between zero and one. All continuous covariates were scaled to have mean zero and standard deviation one, for computational reasons, and the posterior estimates were transformed back to be interpretable on the original scale of the variables. The endogenous as well as the exogenous setting was implemented using R (R Core Team 2016) and JAGS (Plummer 2003). Convergence of the posterior chains was checked using the Gelman-Rubin criterion (Gelman, Meng, et al. 1996) The posterior estimates were considered precise enough if the Monte Carlo error was less than five per cent of the parameter's standard deviation (Lesaffre and Lawson 2012). In the endogenous setting, only 5000 iterations (in each of three posterior chains) were necessary, while in the exogenous setting 20000 iterations were required to satisfy this criterion. Posterior predictive checks were used to evaluate if the assumed model fitted the data appropriately.

In the multivariate normal approach, the imputation models can be specified as

$$
\begin{aligned}
\texttt{bp}_i(t) &= (\tilde{\beta}_0^{\texttt{bp}} + \tilde{b}_{i0}^{\texttt{bp}}) + \tilde{\beta}_1^{\texttt{bp}}\texttt{agem}_i + (\tilde{\beta}_2^{\texttt{bp}} + \tilde{b}_{i1}^{\texttt{bp}})\texttt{ns}_i^{(1)}(t) \\
&\quad +(\tilde{\beta}_3^{\texttt{bp}} + \tilde{b}_{i2}^{\texttt{bp}})\texttt{ns}_i^{(2)}(t) + \tilde{\varepsilon}_i^{\texttt{bp}}(t), \\
\texttt{gw}_i(\tilde{t}) &= (\tilde{\beta}_0^{\texttt{gw}} + \tilde{b}_{i0}^{\texttt{gw}}) + \tilde{\beta}_1^{\texttt{gw}}\texttt{agem}_i + (\tilde{\beta}_2^{\texttt{gw}} + \tilde{b}_{i1}^{\texttt{gw}})\widetilde{\texttt{ns}}_i^{(1)}(\tilde{t}) \\
&\quad +(\tilde{\beta}_3^{\texttt{gw}} + \tilde{b}_{i2}^{\texttt{gw}})\widetilde{\texttt{ns}}_i^{(2)}(\tilde{t}) + \tilde{\varepsilon}_i^{\texttt{gw}}(t), \\
\texttt{height}_i &= \tilde{\beta}_0^{\texttt{height}} + \tilde{\beta}_1^{\texttt{height}}\texttt{agem}_i + \tilde{\varepsilon}_{ij}^{\texttt{height}}, \\
\widehat{\texttt{parity}}_i &= \tilde{\beta}_0^{\texttt{parity}} + \tilde{\beta}_1^{\texttt{parity}}\texttt{agem}_i + \tilde{\varepsilon}_i^{\texttt{parity}}, \\
\widehat{\texttt{ethn}}_i &= \tilde{\beta}_0^{\texttt{ethn}} + \tilde{\beta}_1^{\texttt{ethn}}\texttt{agem}_i + \tilde{\varepsilon}_i^{\texttt{ethn}}, \\
\widehat{\texttt{educ}}_i &= \tilde{\beta}_0^{\texttt{educ}} + \tilde{\beta}_1^{\texttt{educ}}\texttt{agem}_i + \tilde{\varepsilon}_i^{\texttt{educ}}, \\
\widehat{\texttt{smoke}}_i &= \tilde{\beta}_0^{\texttt{smoke}} + \tilde{\beta}_1^{\texttt{smoke}}\texttt{agem}_i + \tilde{\varepsilon}_i^{\texttt{smoke}},
\end{aligned}
$$

and their random effects and error terms modelled jointly as

$$(\tilde{b}_{i0}^{\texttt{bp}}, \tilde{b}_{i1}^{\texttt{bp}}, \tilde{b}_{i2}^{\texttt{bp}}, \tilde{b}_{i0}^{\texttt{gw}}, \tilde{b}_{i1}^{\texttt{gw}}, \tilde{b}_{i2}^{\texttt{gw}}, \tilde{\varepsilon}_i^{\texttt{height}}, \tilde{\varepsilon}_i^{\texttt{parity}}, \tilde{\varepsilon}_i^{\texttt{ethn}}, \tilde{\varepsilon}_i^{\texttt{educ}}, \tilde{\varepsilon}_i^{\texttt{smoke}})^\top \sim N(\mathbf{0}, \widetilde{\mathbf{D}}),$$

where the diagonal elements that correspond to `parity`, `ethn`, `educ` and `smoke` are fixed to 1, and $\left\{\tilde{\varepsilon}_i^{\texttt{bp}}(t), \tilde{\varepsilon}_i^{\texttt{gw}}(t)\right\}^\top \sim N(0, \widetilde{\mathbf{\Sigma}}(t))$.

Using current versions of the software packages JAGS or WinBUGS (D. J. Lunn et al. 2000) it is not possible to sample from such a restricted covariance matrix and we will, therefore, only present results from the sequential approach for the Generation R applications. These results are presented in Figure 5.2. The solid line represents the posterior distribution of the regression coefficients obtained by the sequential approach under the assumption that `gw` was endogenous, while the dashed line depicts the corresponding posterior distributions when `gw` was assumed to be exogenous. The shaded areas in the tails of the distributions mark values outside the 95% credible interval (CI). It can easily be seen that the assumption of exo- or endogeneity had great impact on the posterior distribution. Especially the posterior distribution of the effect of the time-varying covariate, `gw`, differs substantially between the two models. While in the endogenous model the posterior mean of this effect was 0.03 with a 95% CI that includes zero [-0.01, 0.07], this estimate was 0.30 (95% CI [0.29, 0.32]), when `gw` was assumed to be exogenous. Also in other parameters, such as the regression coefficients for `height`, `educ` and the non-linear effect of `gage`, the posterior distributions differed considerably.

A possible explanation for these differences is that in the exogenous model the correlation between `gw` and `bp` is only captured in the parameter γ whereas in the endogenous model it is split between γ and the covariance between the random effects of the model for `bp` and `gw`, i.e., the elements in the upper right quadrant of $\mathbf{D}$. Figure 5.7 in Appendix 5.B.3 shows the posterior density of these elements of the matrix $\mathbf{D}$. Most of the parameters describing the covariance between $\mathbf{b}^{\texttt{gw}}$ and $\mathbf{b}^{\texttt{bp}}$ estimate the respective covariance to be different from zero. The exogenous model implies that these parameters are zero and does not estimate them. Interpreting the results from the endogenous model we may conclude that `gw` and `bp` are correlated, but that there is no evidence that changes in `gw` cause changes in `bp`.

5.5.2 Association between Gestational Weight Gain and Child BMI

Fetal development follows a well-researched course which is influenced by maternal health throughout pregnancy. Specifically, the effect of gestational weight gain may vary between different periods of pregnancy, i.e., different periods of fetal development. Hence, the effect of trimester-specific weight gain is often a predictor of interest. How much weight gain is considered healthy varies with maternal BMI before pregnancy (`bmim`) which, therefore, needs to be considered as predic-

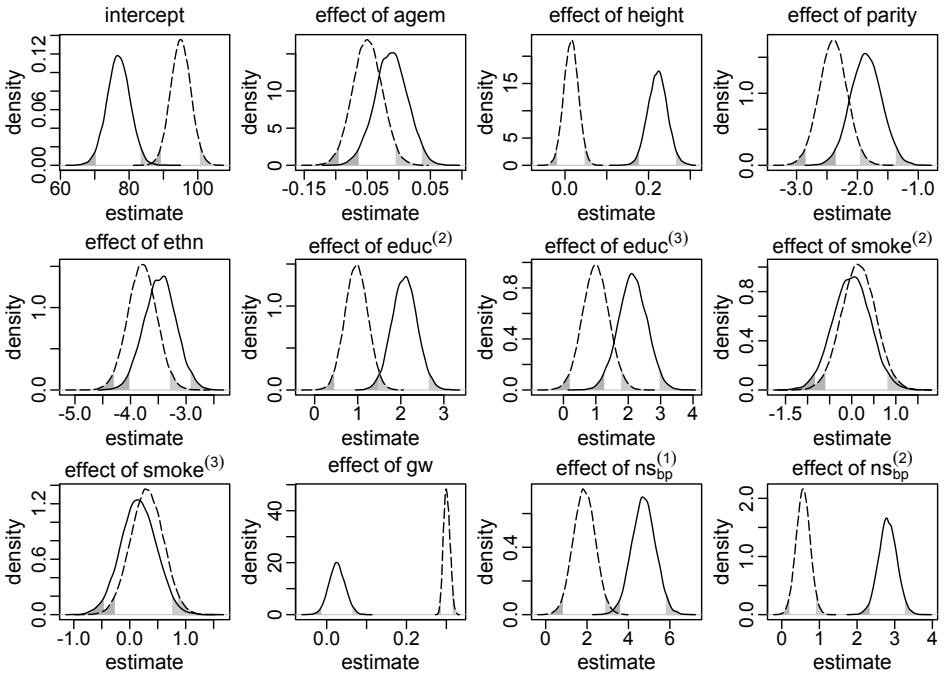


Figure 5.2: Posterior distributions of the main regression coefficients in the first application, derived by the sequential approach. The solid (dashed) line represents the endogenous (exogenous) model. The shaded areas mark values outside the 95% credible interval.

tor variable in this research question. Since `gw` is observed entirely prior to the outcome (`bmi`), it might not be considered to be a time-varying covariate in the narrow sense, i.e., it does not change throughout the time range of the outcome measurements. Nevertheless, it does change over time and an appropriate characterization of this change is essential to obtaining results that allow meaningful conclusions with regards to the research question at hand.

We calculated trimester specific weight gain as the differences between weight before pregnancy, 14 weeks of gestation, 27 weeks of gestation and at (or rather right before) birth (`gestbir`), and scaled these differences to reflect weight gain per week. The functional relation between `gw` and `bmi` can thus be represented as

$$f(H_i^{\mathtt{gw}}(t), t) = \{\Delta_1(\mathtt{gw}_i), \Delta_2(\mathtt{gw}_i), \Delta_3(\mathtt{gw}_i)\}^\top,$$

with

$$\begin{aligned}
\Delta_1(\texttt{gw}_i) &= \frac{\texttt{gw}_i(\texttt{gage}=14)-\texttt{gw}_i(\texttt{gage}=0)}{14},\\
\Delta_2(\texttt{gw}_i) &= \frac{\texttt{gw}_i(\texttt{gage}=27)-\texttt{gw}_i(\texttt{gage}=14)}{27-14},\\
\Delta_3(\texttt{gw}_i) &= \frac{\texttt{gw}_i(\texttt{gage}=\texttt{gestbir}_i)-\texttt{gw}_i(\texttt{gage}=27)}{\texttt{gestbir}_i-27}.
\end{aligned}$$

As in the first application, we assumed a non-linear evolution of `gw` over time, which was modelled using a natural cubic spline for `gage` with 2 df. Also, the trajectories of child BMI, for which age and gender-specific standard deviation scores (SDS) were used, were non-linear and therefore modelled using a natural cubic spline with 3 df for `age`, in the formula below represented by $\texttt{ns}_i^{(1)}(t)$, $\texttt{ns}_i^{(2)}(t)$ and $\texttt{ns}_i^{(3)}(t)$. Since `gw` was measured before birth and `bmi` only after birth, we assumed that `gw` was exogenous in this application.

The analysis model for this research question can be written as

$$\begin{aligned}
\texttt{bmi}_i(t) &= (\beta_0+b_{i0}^{\texttt{bmi}})+\beta_1\texttt{agem}_i+\beta_2\texttt{parity}_i+\beta_3\texttt{ethn}_i\\
&\quad+\beta_4\texttt{educ}_i^{(2)}+\beta_5\texttt{educ}_i^{(3)}+\beta_6\texttt{smoke}_i^{(2)}\\
&\quad+\beta_7\texttt{smoke}_i^{(3)}+\beta_8\texttt{bmim}_i+(\beta_9+b_{i1}^{\texttt{bmi}})\texttt{ns}_i^{(1)}(t)\\
&\quad+(\beta_{10}+b_{i2}^{\texttt{bmi}})\texttt{ns}_i^{(2)}(t)+(\beta_{11}+b_{i3}^{\texttt{bmi}})\texttt{ns}_i^{(3)}(t)\\
&\quad+\gamma_1\Delta_1(\texttt{gw}_i)+\gamma_2\Delta_2(\texttt{gw}_i)+\gamma_3\Delta_3(\texttt{gw}_i)+\varepsilon_i^{\texttt{bmi}}(t),
\end{aligned}$$

The analysis was again performed using the sequential approach, where imputation models for `gw` and the baseline covariates were specified analogous to the first application. To reduce correlation between the elements of $\boldsymbol{\gamma}$, an elastic net shrinkage hyperprior for the variance parameters of $\boldsymbol{\gamma}$ was used (Mallick and Yi 2013).

Results from the analysis of this second research question are presented in Figure 5.3. Only the posterior distributions of the parameters relating to `bmim` and `gw` are shown as these are the parameters of interest here. It can be seen that children of mothers with higher baseline BMI had higher BMIs as well – an increase of one kg/m^2 resulted on average in a 0.03 SDS higher child BMI (95% CI [0.03, 0.04]). Higher gestational weight gain during the first trimester was associated with higher child BMI (0.23 SDS increase per kg weekly weight gain; 95% CI [0.05, 0.40]). Even though the posterior mean of the effect of weekly gestational weight gain during the second trimester was slightly higher (0.26), due to the increased uncertainty of this estimate (95% CI [-0.08, 0.65]) there was no evidence

of an association with the trajectories of child BMI. There was also no evidence that weight gain during the last trimester was a relevant predictor of child BMI (0.12; 95% CI [-0.09, 0.33]).

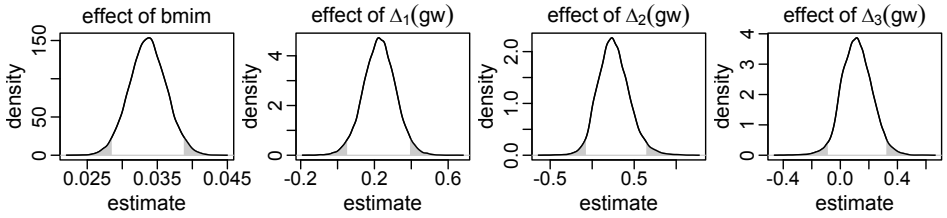


Figure 5.3: Posterior distributions of a selection of regression coefficients from the second application, derived by the sequential approach. The shaded areas mark values outside the 95% credible interval.

5.6 Simulation Study

To evaluate the performance of the two imputation approaches described in Section 5.4 with regards to misspecification of the endo- or exogeneity of a time-varying covariate and the bias introduced by misspecification of the functional form in a more controlled setting, we performed a simulation study in which we compared results from correctly specified models with those that are misspecified, for data generated in a range of different scenarios and different missing mechanisms. Specifically, the key objectives were

1. to confirm that both approaches provide unbiased estimates when the models are correctly specified during imputation and analysis,

2. to investigate how misspecification of the endo- or exogeneity influences the results, and

3. to explore bias due to misspecification of the functional form, specifically

    - the bias introduced during imputation due to the implied linearity assumption of the multivariate normal approach when the true functional form is non-linear, and
    - the bias introduced when the imputation model, as well as the analysis model, are misspecified as linear.

5.6.1 Design

We simulated 200 datasets in each of six scenarios that differed in the endo-/exogeneity of the covariate, the functional form and the model (sequential or multivariate normal) that was used. Common to all scenarios was that 10 repeated measurements of a normally distributed time-varying covariate and a conditionally normal outcome variable, with measurements at the same, unbalanced time points, were created. Under the sequential approach data was generated with a linear or a quadratic relation between covariate and outcome, where the covariate was either exogenous or endogenous. For the multivariate normal model (which always generates data with a linear relation between the outcome and an endogenous covariate) we considered two scenarios with regards to the correlation of the error terms, where in one scenario the error terms of outcome and covariate were independent and in the other correlated.

Missing values were created in the time-varying covariate according to two MAR mechanisms, in which the probability of the time-varying covariate being missing either only depended on the outcome at the same time point or on the outcome at the same time point as well as the covariate at the previous time point.

Details on the exact setup of the simulation study are given in Appendix 5.C.1.

5.6.2 Analysis Models

Each of the datasets was analysed using both approaches with different assumptions regarding the endo- or exogeneity of the covariate and the functional form, before values were deleted, and for both missing mechanisms.

The complete dataset was analysed using function `lmer()` from the R-package **lme4** (R Core Team 2016; Bates et al. 2015) as well as with the sequential approach. Missing data was imputed and analysed with the sequential approach, where the random effects were modelled according to the current assumption of exo- or endogeneity, and the imputation was repeated twice with the multivariate normal approach (once using the model with independent error terms and once assuming correlated error terms). Each time, ten imputed datasets were created by drawing values from the posterior chains of the incomplete covariate and analysed analogously to the analysis of the complete data. When the covariate was assumed exogenous, `lmer()` was used and the coefficients from the ten corresponding analyses pooled using Rubin's Rules (Rubin 1987). When the covariate was assumed to be endogenous, the sequential approach with correlated random effects was used and the ten sets of posterior MCMC chains combined to calculate posterior summary measures.

An overview of the assumptions and models used can be found in Table 5.2 in Appendix 5.C.2. The specific parameter values that were used are given in Tables 5.3 and 5.4 in Appendix 5.C.3.

5.6.3 Results

Firstly, we found that the sequential approach provided unbiased estimates when comparing the results from the analysis of the complete data to the true parameters that were used to generate the data. Secondly, in all scenarios, results were very similar for both MAR mechanisms and we will, hence, not distinguish them during the further description of the results.

Regarding to our first objective, the comparison of the sequential and the multivariate normal approach when exo- or endogeneity and functional form were specified correctly, we found that both approaches were unbiased and their 95% credible/confidence intervals had the desired coverage. However, misspecification of the error terms in the multivariate normal approach as independent had the overall largest impact on the results (estimates were on average half the value of the estimate from the analysis of the complete data and CIs had 0% coverage). Based on this finding, we excluded the multivariate normal approach with independent error terms from further comparisons. Moreover, we saw that misspecification of an endogenous covariate as exogenous resulted in bias while misspecification of an exogenous covariate as endogenous did not. This was the case for both approaches, and linear as well as quadratic (only for the sequential approach) functional form. With respect to our third objective, the simulation study showed that imputation with the multivariate normal approach (with correlated error terms) in a setting where the functional form was correctly assumed to be quadratic during the subsequent analysis had the second largest impact, with a relative bias of approximately 0.8, and resulted in CIs with coverage of close to 0%. The bias that was added due to misspecification of the functional form as linear during imputation as well as analysis, as compared to the results from the analysis of the complete data under the same misspecification, was small and overall comparable between the multivariate normal and the sequential approach. These findings were the same irrespective of the exo- or endogeneity of the covariate. Plots of the results from the simulation study as well as a detailed discussion of these results can be found in Appendix 5.C.4.

In summary, the results of this simulation study demonstrate the impact that imprudent acceptance of default assumptions, like exogeneity, linear relations between variables, or (conditioned on random effects) uncorrelated error terms may have.

5.7 Discussion

Motivated by two research questions from the Generation R study, we investigated two Bayesian approaches to handle missing covariate data in models with longitudinal outcomes and time-varying covariates. Specifically, we compared the multivariate normal approach, a widely known omnibus approach, to the more custom-designed sequential approach, which we extended to handle endogenous time-varying covariates. The focus of this comparison was on the ability to take into account different functional relations between such covariates and the outcome, and the suitability for exogenous as well as endogenous covariates.

The analysis of our real data applications illustrated the necessity for methods that allow for complex functional relations and endogenous covariates. Simulation studies confirmed that in our setting, methods that make the common assumption of exogeneity of a time-varying covariate provide biased estimates. The assumption of endogeneity during imputation and analysis, however, did not introduce any bias, which suggests choosing the endogenous specification, e.g., to model the random effects of the outcome and the time-varying covariate correlated, as a default. The simulation study also demonstrated that imputation with the multivariate normal approach in settings where the implied assumption of linear associations between variables is violated can be biased. Furthermore, great care should be taken when assumptions about the correlation structure of the error terms are made in the multivariate normal approach, as misspecification may result in large bias. Results indicated that the sequential approach is more robust with regards to this type of misspecification, however, future research is required to evaluate this further. Overall, the sequential approach performed well and proved to be a suitable method to impute and analyse longitudinal data with possibly endogenous time-varying covariates.

The ability of the sequential approach to handle various functional forms, and to provide estimates in settings with endogenous time-varying covariates, can be seen as its biggest advantages. Moreover, it can handle non-linear associations of baseline covariates and interaction terms involving incomplete covariates without the need of approximations like the "just another variable" approach or passive imputation via transformation (Bartlett et al. 2015; Seaman, Bartlett, et al. 2012; White et al. 2011). Even in settings where there is no single functional form of interest, but several candidate functions, it may be applied in combination with shrinkage techniques which may help the decision which functional form is most appropriate.

In the present paper, we focused on a single time-varying covariate and ignorable missing data mechanisms, however, extensions of the sequential approach to ac-

commodate more complex settings are possible. Multiple time-varying covariates can be added to the linear predictor of the analysis model. Imputation models for the time-varying covariates could either be specified assuming conditional independence between different time-varying covariates, i.e., excluding them from each other's linear predictor, assuming joint multivariate distributions for their random effects and/or error terms, or by specifying a functional form of one time-varying covariate to be included in the linear predictor of another covariate, analogous to the specification of the analysis model described in Section 5.3. The second option may be a convenient choice if a linear relation between the time-varying covariates is a reasonable assumption. When the missing data mechanism is non-ignorable, this can be taken into account by extending the specification of the joint distribution with terms that either describe the selection mechanism (i.e., the missingness pattern given the data) or specify how the distribution of the data depends on the missingness pattern (Carpenter and Kenward 2013; Van Buuren 2012; Daniels and Hogan 2008).

A reason why the multivariate normal approach may be preferred in practice, is its availability in software packages, as, for instance, the R-package **jomo** (Quartagno and Carpenter 2016) or REALCOM-impute (Carpenter, Goldstein, et al. 2011). Those implementations also provide samplers that can handle restricted covariance matrices. More tailored approaches, like the sequential approach, usually need to be implemented by hand, which, however, can be done in existing Bayesian software packages such as JAGS or WinBUGS in a straightforward way. In Appendix 5.C.5 we give example syntax for both approaches. This syntax can easily be extended to include complete or incomplete baseline covariates (see also the Appendix of Erler, Rizopoulos, Rosmalen, et al. (2016)). Additional example syntax can be provided upon request.

When imputing and analysing complex datasets, researchers need to deliberate if standard methods that are easy to apply meet the requirements of the application at hand, specifically if the assumptions of those methods are met. It is our opinion that too often this is not the case and standard approaches are applied even when they are not adequate. Therefore, we plead for the use of methods that are flexible enough to be adapted to the specific characteristics of a problem. In the context of imputation and analysis of longitudinal data with possibly endogenous time-varying covariates, the sequential approach presented here is such an approach.

Appendix

5.A Details on the Implied Exo- or Endogeneity

In this section, we provide details on the implications that the approaches introduced in Section 5.4 have for the exo- or endogeneity of the time-varying covariate $\mathbf{s}$. For $\mathbf{s}$ to be exogenous, the conditions from Section 5.3.3 need to be fulfilled, i.e.,

$$\begin{cases} p(y_i(t), f(H_i^s(t), t) \mid H_i^y(t^-), H_i^s(t^-), \boldsymbol{\theta}) = \\ \qquad = p(y_i(t) \mid f(H_i^s(t), t), H_i^y(t^-), H_i^s(t^-), \boldsymbol{\theta}_1) \times \\ \qquad\quad p(s_i(t) \mid H_i^y(t^-), H_i^s(t^-), \boldsymbol{\theta}_2) \\ \\ p(s_i(t) \mid H_i^s(t^-), H_i^y(t^-), \mathbf{x}_i, \boldsymbol{\theta}) = p(s_i(t) \mid H_i^s(t^-), \mathbf{x}_i, \boldsymbol{\theta}) \end{cases}$$

with $\boldsymbol{\theta}^\top = (\boldsymbol{\theta}_1^\top, \boldsymbol{\theta}_2^\top)$ and $\boldsymbol{\theta}_1 \perp\!\!\!\perp \boldsymbol{\theta}_2$, where $H_i^y(t^-)$ and $H_i^s(t^-)$ denote the history of $\mathbf{y}$ and $\mathbf{s}$, respectively, up to, but excluding measurements at time t. To abbreviate the notation, we will drop the index i in the following sections.

5.A.1 Exogeneity of the Sequential Approach with Independent Random Effects

As can easily be seen, the first condition is fulfilled in the sequential approach with independent random effects, since the joint distribution of $\mathbf{y}$ and $\mathbf{s}$ is specified as the product of the conditional and the marginal distribution and the parameters of these distributions are usually specified to be a priori independent.

To show that the second condition is fulfilled as well, several steps are necessary. Under the assumption that $s(t)$ is independent from $H^s(t^-)$ and $H^y(t)$ given the random effects we can write

$$\begin{aligned} p(s(t) \mid H^s(t^-), H^y(t), \mathbf{x}, \boldsymbol{\theta}) &= \int p(s(t), \mathbf{b} \mid H^s(t^-), H^y(t), \mathbf{x}, \boldsymbol{\theta})\, d\mathbf{b} \\ &= \int p(s(t) \mid \mathbf{b}, H^s(t^-), H^y(t), \mathbf{x}, \boldsymbol{\theta}) \\ &\qquad p(\mathbf{b} \mid H^s(t^-), H^y(t), \mathbf{x}, \boldsymbol{\theta})\, d\mathbf{b} \\ &= \iint p(s(t) \mid \mathbf{b}^s, \mathbf{x}, \boldsymbol{\theta}) \\ &\qquad p(\mathbf{b}^s, \mathbf{b}^y \mid H^s(t^-), H^y(t), \mathbf{x}, \boldsymbol{\theta})\, d\mathbf{b}^s\, d\mathbf{b}^y. \end{aligned} \tag{5.9}$$

Using Bayes theorem, the conditional distribution of the random effects can be rewritten as

$$\begin{aligned} p(\mathbf{b}^s, \mathbf{b}^y \mid H^s(t^-), H^y(t), \mathbf{x}, \boldsymbol{\theta}) \\ &= \frac{p(\mathbf{b}^s, \mathbf{b}^y, H^s(t^-), H^y(t) \mid \mathbf{x}, \boldsymbol{\theta})}{p(H^s(t^-), H^y(t) \mid \mathbf{x}, \boldsymbol{\theta})} \\ &= \frac{p(H^s(t^-), H^y(t) \mid \mathbf{b}^s, \mathbf{b}^y, \mathbf{x}, \boldsymbol{\theta})\, p(\mathbf{b}^s, \mathbf{b}^y \mid \boldsymbol{\theta})}{p(H^s(t^-), H^y(t) \mid \mathbf{x}, \boldsymbol{\theta})}. \end{aligned} \tag{5.10}$$

With prior independence of $\mathbf{b}^y$ and $\mathbf{b}^s$, i.e., $p(\mathbf{b}^s, \mathbf{b}^y \mid \boldsymbol{\theta}) = p(\mathbf{b}^s \mid \boldsymbol{\theta})\; p(\mathbf{b}^y \mid \boldsymbol{\theta})$, and assuming conditional independence of $H^s(t^-)$ and $H^y(t)$ given the random effects, the denominator of (5.10) can be split in two factors,

$$\begin{aligned} p(H^s(t^-), H^y(t) \mid \mathbf{x}, \boldsymbol{\theta}) &= \iint p(H^s(t^-), H^y(t) \mid \mathbf{b}^y, \mathbf{b}^s, \mathbf{x}, \boldsymbol{\theta})\, p(\mathbf{b}^y, \mathbf{b}^s \mid \boldsymbol{\theta})\, d\mathbf{b}^y d\mathbf{b}^s \\ &= \int p(H^s(t^-) \mid \mathbf{b}^s, \mathbf{x}, \boldsymbol{\theta})\, p(\mathbf{b}^s \mid \boldsymbol{\theta})\, d\mathbf{b}^s \\ &\quad \int p(H^y(t) \mid \mathbf{b}^y, \mathbf{x}, \boldsymbol{\theta})\, p(\mathbf{b}^y \mid \boldsymbol{\theta})\, d\mathbf{b}^y \\ &= p(H^s(t^-) \mid \mathbf{x}, \boldsymbol{\theta})\, p(H^y(t) \mid \mathbf{x}, \boldsymbol{\theta}). \end{aligned} \tag{5.11}$$

Substituting (5.10) and (5.11) into (5.9), and recognizing that

$$p(s(t) \mid \mathbf{b}^s, \mathbf{x}, \boldsymbol{\theta})\; p(H^s(t^-) \mid \mathbf{b}^s, \mathbf{x}, \boldsymbol{\theta}) = p(H^s(t) \mid \mathbf{b}^s, \mathbf{x}, \boldsymbol{\theta}),$$

allows us to factorize the integrand from (5.9) so that each factor only depends on either $\mathbf{b}^y$ or $\mathbf{b}^s$. The two integrals can then be solved separately and (5.9) can be simplified as

$$\begin{aligned} p(s(t) \mid H^s(t^-), H^y(t), \mathbf{x}, \boldsymbol{\theta}) &= \frac{\int p(H^s(t) \mid \mathbf{b}^s, \mathbf{x}, \boldsymbol{\theta})\; p(\mathbf{b}^s \mid \boldsymbol{\theta})\; d\mathbf{b}^s}{p(H^s(t^-) \mid \mathbf{x}, \boldsymbol{\theta})} \\ &\quad \frac{\int p(H^y(t) \mid \mathbf{b}^y, \mathbf{x}, \boldsymbol{\theta})\; p(\mathbf{b}^y \mid \boldsymbol{\theta})\; d\mathbf{b}^y}{p(H^y(t) \mid \mathbf{x}, \boldsymbol{\theta})} \\ &= \frac{p(H^s(t) \mid \mathbf{x}, \boldsymbol{\theta})}{p(H^s(t^-) \mid \mathbf{x}, \boldsymbol{\theta})}\; \frac{p(H^y(t) \mid \mathbf{x}, \boldsymbol{\theta})}{p(H^y(t) \mid \mathbf{x}, \boldsymbol{\theta})} \\ &= p(s(t) \mid H^s(t^-), \mathbf{x}, \boldsymbol{\theta})\; \frac{p(H^s(t^-) \mid \mathbf{x}, \boldsymbol{\theta})}{p(H^s(t^-) \mid \mathbf{x}, \boldsymbol{\theta})} \\ &= p(s(t) \mid H^s(t^-), \mathbf{x}, \boldsymbol{\theta}), \end{aligned}$$

which shows that for the sequential approach with independent random effects also the second condition for exogeneity is fulfilled.

5.A.2 Endogeneity of the Sequential Approach with Correlated Random Effects

Prior independence of the random effects of the models for $\mathbf{y}$ and $\mathbf{s}$ is a crucial assumption for exogeneity. When $p(\mathbf{b}^s, \mathbf{b}^y \mid \boldsymbol{\theta}) \neq p(\mathbf{b}^s \mid \boldsymbol{\theta})\ p(\mathbf{b}^y \mid \boldsymbol{\theta})$, neither the numerator nor the denominator in (5.10) can be factorized into independent factors and (5.9) cannot be simplified as in the case with independent random effects. Either of the two possible factorizations,

$$p(\mathbf{b}^s, \mathbf{b}^y \mid \boldsymbol{\theta}) = p(\mathbf{b}^s \mid \mathbf{b}^y, \boldsymbol{\theta})\ p(\mathbf{b}^y \mid \boldsymbol{\theta}) \quad \text{or} \quad p(\mathbf{b}^s, \mathbf{b}^y \mid \boldsymbol{\theta}) = p(\mathbf{b}^y \mid \mathbf{b}^s, \boldsymbol{\theta})\ p(\mathbf{b}^s \mid \boldsymbol{\theta}),$$

violate the requirements of exogeneity.

Using the former factorization, (5.9) leads to

$$\begin{aligned} p(s(t) \mid H^s(t^-), H^y(t), \mathbf{x}, \boldsymbol{\theta}) &= \dots \\ &= \frac{1}{p(H^s(t^-), H^y(t) \mid \mathbf{x}, \boldsymbol{\theta})} \\ &\quad \int\int p(H^s(t) \mid \mathbf{b}^s, \mathbf{x}, \boldsymbol{\theta})\ p(\mathbf{b}^s \mid \mathbf{b}^y, \boldsymbol{\theta})\ d\mathbf{b}^s \\ &\quad\quad p(H^y(t) \mid \mathbf{b}^y, \mathbf{x}, \boldsymbol{\theta})\ p(\mathbf{b}^y \mid \boldsymbol{\theta})\ d\mathbf{b}^y, \end{aligned}$$

where $\mathbf{s}$ is conditioned on $\mathbf{b}^y$ which, since the random effects depend on all observations of $\mathbf{y}$, implies that $\mathbf{s}$ is not independent of the history of $\mathbf{y}$, thus violating the second condition.

Using the latter factorization leads to

$$\begin{aligned} p(s(t) \mid H^s(t^-), H^y(t), \mathbf{x}, \boldsymbol{\theta}) &= \dots \\ &= \frac{1}{p(H^s(t^-), H^y(t) \mid \mathbf{x}, \boldsymbol{\theta})} \\ &\quad \int\int p(H^y(t) \mid \mathbf{b}^y, \mathbf{x}, \boldsymbol{\theta})\ p(\mathbf{b}^y \mid \mathbf{b}^s, \boldsymbol{\theta})\ d\mathbf{b}^y \\ &\quad\quad p(H^s(t) \mid \mathbf{b}^s, \mathbf{x}, \boldsymbol{\theta})\ p(\mathbf{b}^s \mid \boldsymbol{\theta})\ d\mathbf{b}^s, \end{aligned}$$

where $\mathbf{y}$ depends on all observed values of $\mathbf{s}$ via $\mathbf{b}^s$, i.e., also on future values of $\mathbf{s}$, which also is at conflict with the exogeneity assumption. The sequential model with correlated random effects hence implies that $\mathbf{s}$ is endogenous.

5.B Analysis of the Generation R Data

5.B.1 Missing Data Patterns

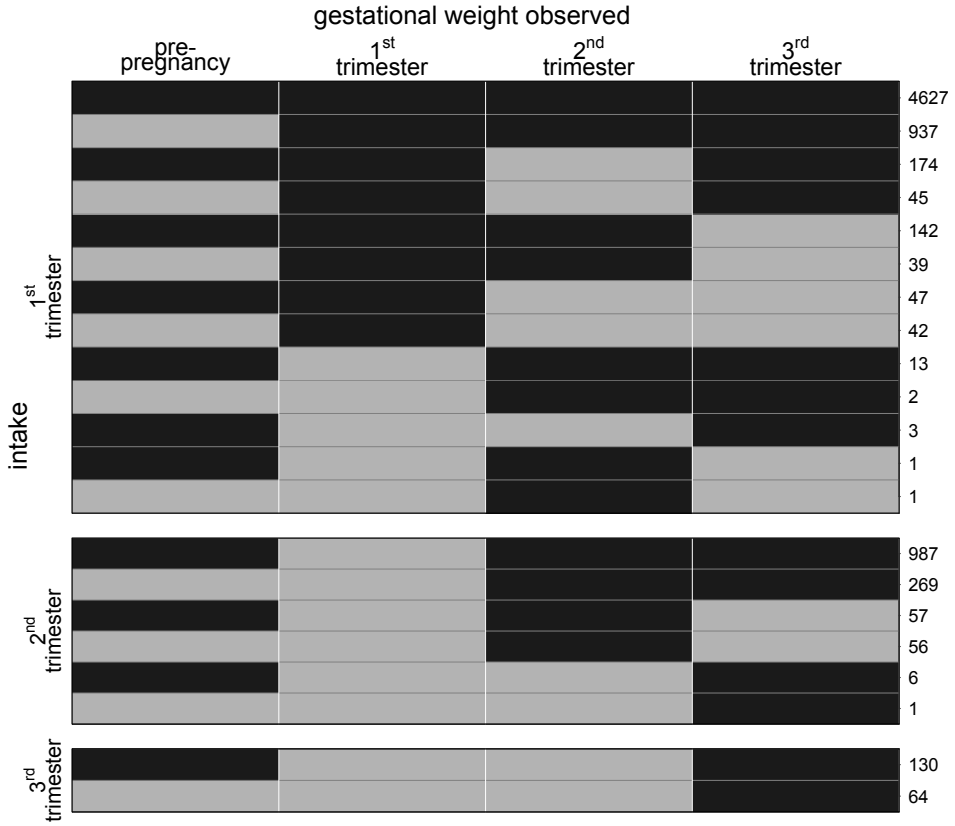


Figure 5.4: Missing data pattern for gestational weight. Dark color depicts observed values, light color missing values. The frequency of each missing data pattern is given on the right.

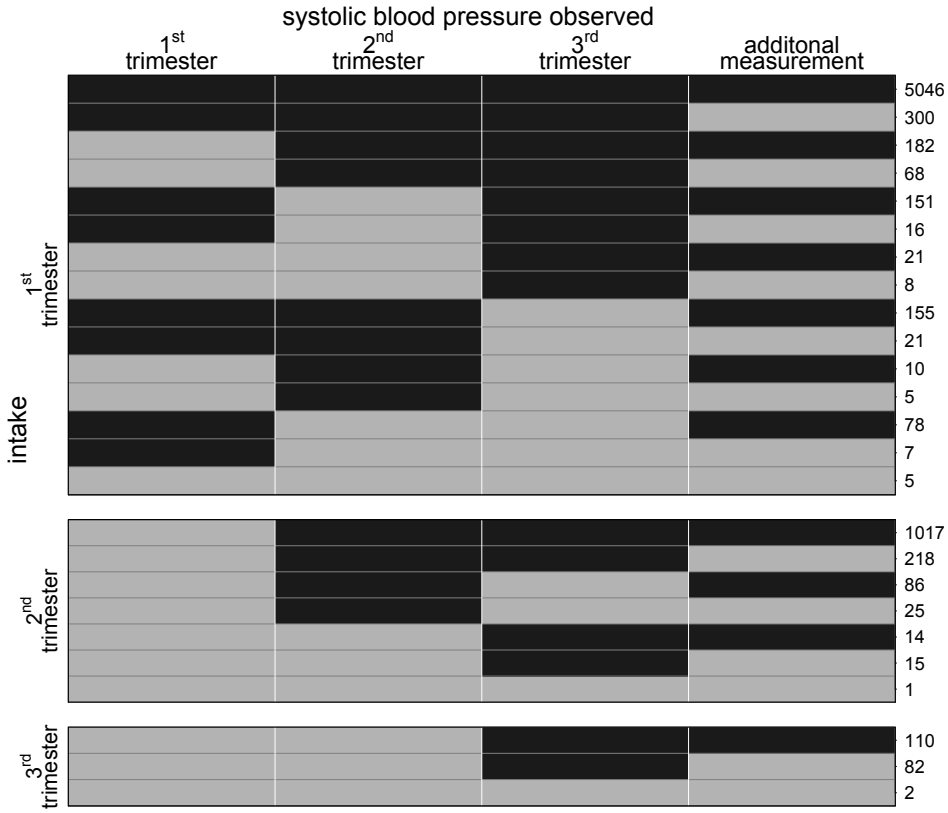


Figure 5.5: Missing data pattern for systolic blood pressure. Dark color depicts observed values, light color missing values. The frequency of each missing data pattern is given on the right.

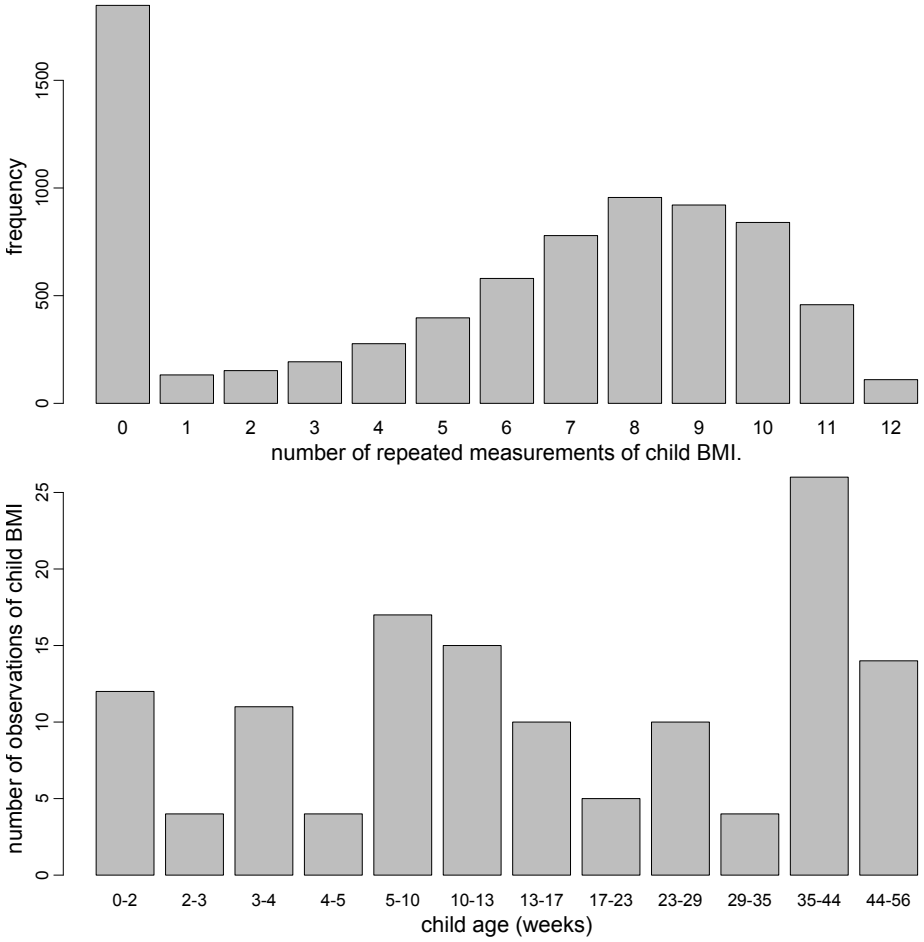


Figure 5.6: Distribution of the number of repeated BMI measurements per child (top) and number of observed values of child BMI per age category (bottom).

5.B.2 Prior Distributions

$$
\begin{aligned}
\beta_k, \gamma &\overset{iid}{\sim} N(0, 1000), \qquad k = 0, \ldots, 10 \\
\sigma^2_{\texttt{bp}}, \sigma^2_{\texttt{gw}} &\overset{iid}{\sim} \text{inv-Gamma}(0.01, 0.01) \\
\alpha_k &\sim N(0, 1000), \qquad k = 0, \ldots, 10 \\
\mathbf{D}_{\texttt{bp}} &\sim \text{inv-Wishart}(R_{\texttt{bp}}, 3) \\
\mathbf{D}_{\texttt{gw}} &\sim \text{inv-Wishart}(R_{\texttt{gw}}, 3) \\
(\text{diag}(R_{\texttt{bp}})^\top, \text{diag}(R_{\texttt{gw}})^\top) &\overset{iid}{\sim} \text{Gamma}(1, 0.001)
\end{aligned}
$$

5.B.3 Additional Graphics

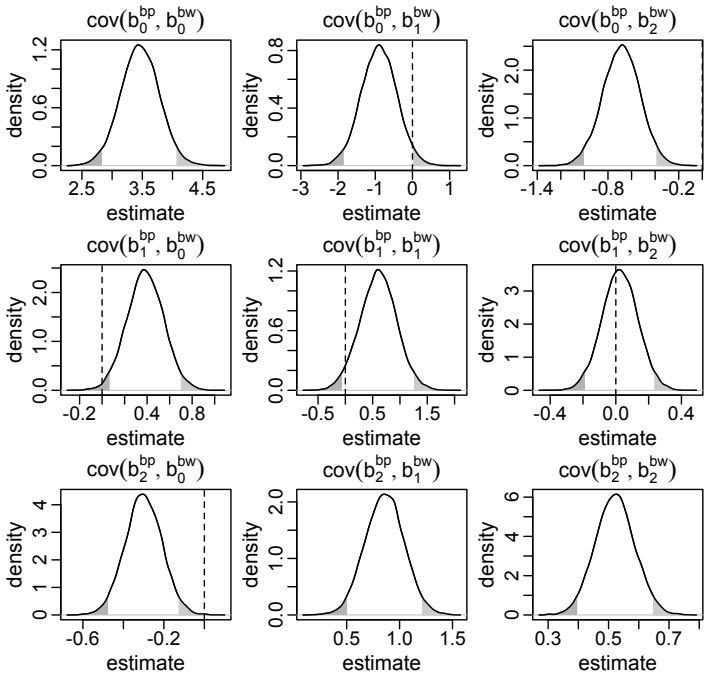


Figure 5.7: Posterior distributions of the covariance between the random effects $b^{\texttt{bp}}$ and $b^{\texttt{gw}}$ from the endogenous setting in the first motivating question from the Generation R data presented in Section 5.5.1. The dashed vertical line marks zero, i.e., the implied covariance in the exogenous setting, the shaded areas mark values outside the 95% credible interval.

5.C Simulation Study

To evaluate the performance of the two imputation approaches described in Section 5.4 with regards to misspecification of the endo- or exogeneity of a time-varying covariate and the bias introduced by misspecification of the functional form in a more controlled setting, we performed a simulation study in which we compared results from correctly specified models with those that are misspecified, for data generated in a range of different scenarios and different missing mechanisms.

The simulation study was set up to

1. confirm that both approaches provide unbiased estimates when the models are correctly specified during imputation and analysis,
2. investigate how misspecification of the endo- or exogeneity influences the results, and
3. to explore bias due to misspecification of the functional form, specifically
   - the bias introduced during imputation due to the implied linearity assumption of the multivariate normal approach when the true functional form is non-linear, and
   - the bias introduced when the imputation model, as well as the analysis model, are misspecified as linear.

5.C.1 Design

We simulated 200 datasets in each of six scenarios that differed in the endo-/exogeneity of the covariate, the functional form and the model (sequential or multivariate normal) that was used. Common in all scenarios was that ten repeated measurements of a normally distributed time-varying covariate and a (conditionally) normal outcome variable, with measurements at the same, unbalanced time points, were created. Table 5.1 gives an overview of the different simulation scenarios. Under the sequential approach data was generated with a linear or a quadratic relation between covariate and outcome, where the covariate was either exogenous or endogenous. The multivariate normal model always generates data with a linear relation between the outcome and an endogenous covariate, however, there we considered two scenarios with regards to the correlation of the error terms, where in one scenario the error terms of outcome and covariate were independent and in the other correlated.

The general model used for simulation from the sequential approach was

$$\begin{aligned} y_{ij} &= (\beta_{y0} + b^y_{i0}) + (\beta_{y1} + b^y_{i1})t_{ij} + \boldsymbol{\gamma} f(s_{ij}) + \varepsilon^y_{ij} \\ s_{ij} &= (\beta_{s0} + b^s_{i0}) + (\beta_{s1} + b^s_{i1})t_{ij} + \varepsilon^s_{ij} \end{aligned} \tag{5.12}$$

Table 5.1: Overview of data generating models used in the simulation study, with their abbreviation written in italic font.

	linear relation	**quadratic relation**
exogenous	• sequential approach *seq. (exo., lin.)*	• sequential approach *seq. (exo., qdr.)*
endogenous	• sequential approach *seq. (endo., lin.)* • multivariate normal approach – independent error terms *mvn. (indep. err.)* – correlated error terms *mvn. (corr. err.)*	• sequential approach *seq. (endo., lin.)*

with

$$\begin{aligned} \varepsilon_{ij}^{y} &\sim N(0, \sigma_y^2) \\ \varepsilon_{ij}^{s} &\sim N(0, \sigma_s^2) \end{aligned}$$

and

$$\begin{bmatrix} (b_{i0}^{y}, b_{i1}^{y})^\top \\ (b_{i0}^{s}, b_{i1}^{s})^\top \end{bmatrix} \sim N\left(\mathbf{0}, \begin{bmatrix} \mathbf{D}_y & \mathbf{D}_{y,s} \\ \mathbf{D}_{y,s} & \mathbf{D}_s \end{bmatrix}\right),$$

where t_{ij} is from a uniform distribution on the interval $[0, 5]$. The specific values of all parameters can be found in Table 5.3 in Appendix 5.C.3. The general model to simulate from a multivariate normal distribution was

$$\begin{aligned} y_{ij} &= (\tilde{\beta}_{y0} + \tilde{b}_{i0}^{y}) + (\tilde{\beta}_{y1} + \tilde{b}_{i1}^{y}) t_{ij} + \tilde{\varepsilon}_{ij}^{y} \\ s_{ij} &= (\tilde{\beta}_{s0} + \tilde{b}_{i0}^{s}) + (\tilde{\beta}_{s1} + \tilde{b}_{i1}^{s}) t_{ij} + \tilde{\varepsilon}_{ij}^{s} \end{aligned} \tag{5.13}$$

with

$$\begin{bmatrix} \tilde{\varepsilon}_{ij}^{y} \\ \tilde{\varepsilon}_{ij}^{s} \end{bmatrix} \sim N\left(\mathbf{0}, \begin{bmatrix} \tilde{\sigma}_y^2 & \tilde{\sigma}_{y,s} \\ \tilde{\sigma}_{y,s} & \tilde{\sigma}_s^2 \end{bmatrix}\right)$$

and

$$\begin{bmatrix} (\tilde{b}_{i0}^{y}, \tilde{b}_{i1}^{y})^\top \\ (\tilde{b}_{i0}^{s}, \tilde{b}_{i1}^{s})^\top \end{bmatrix} \sim N\left(\mathbf{0}, \begin{bmatrix} \widetilde{\mathbf{D}}_y & \widetilde{\mathbf{D}}_{y,s} \\ \widetilde{\mathbf{D}}_{y,s} & \widetilde{\mathbf{D}}_s \end{bmatrix}\right).$$

With these models, the six different data scenarios were specified as

- *seq (exo., lin.)*: simulation from model (5.12), with $\boldsymbol{\gamma} f(s_{ij}) = \gamma_1 s_{ij}$ and $\mathbf{D}_{y,s} = \mathbf{0}$,
- *seq. (exo., qdr.)*: simulation from model (5.12), with $\boldsymbol{\gamma} f(s_{ij}) = \gamma_1 s_{ij} + \gamma_2 s_{ij}^2$ and $\mathbf{D}_{y,s} = \mathbf{0}$,
- *seq (endo., lin.)*: simulation from model (5.12), with $\boldsymbol{\gamma} f(s_{ij}) = \gamma_1 s_{ij}$ and $\mathbf{D}_{y,s} \neq 0$,
- *seq (endo., qdr.)*: simulation from model (5.12), with $\boldsymbol{\gamma} f(s_{ij}) = \gamma_1 s_{ij} + \gamma_2 s_{ij}^2$ and $\mathbf{D}_{y,s} \neq 0$,
- *mvn (indep. err.)*: simulation from model (5.13), with $\widetilde{\mathbf{D}}_{y,s} \neq \mathbf{0}$ and $\tilde{\sigma}^2_{y,s} = 0$,
- *mvn (corr. err.)*: simulation from model (5.13), with $\widetilde{\mathbf{D}}_{y,s} \neq \mathbf{0}$ and $\tilde{\sigma}^2_{y,s} \neq 0$.

Missing values were created in $\mathbf{s}$ according to two MAR mechanisms. In missingness scenario *MAR.1*, the probability of s_{ij} being missing depended on y_{ij} only, while in missingness scenario *MAR.2* this probability depended on y_{ij} as well as s_{ij-1}, specifically

$$\begin{aligned} &\textit{MAR.1:} && \Pr(s_{ij} = \text{NA}) = \text{expit}(y_{ij} + \zeta_1), \\ &\textit{MAR.2:} && \Pr(s_{ij} = \text{NA}) = \text{expit}(y_{ij} + \zeta_1 + \zeta_2 \mathbb{1}(s_{ij-1} < \zeta_3)), \end{aligned}$$

where $\text{expit}(x) = \exp(x)/(1+\exp(x))$ and $\mathbb{1}$ is the indicator function which is one if the statement is true and zero otherwise. The values for the parameters ζ were chosen so that approximately 40% of $\mathbf{s}$ were missing and can be found in Table 5.4 in Appendix 5.C.3.

5.C.2 Analysis Models

Each of the datasets was analysed using both approaches with different assumptions regarding the endo- or exogeneity of the covariate and the functional form, before values were deleted, and for both missing mechanisms. An overview of the analysis methods used under the assumptions of either exo- or endogeneity is given in Table 5.2. The complete data was analysed using function `lmer()` from the R-package **lme4** (R Core Team 2016; Bates et al. 2015) as well as with the sequential approach with independent random effects, $\mathbf{D}_{y,s} = \mathbf{0}$, when the covariate was assumed to be exogenous, and with the sequential approach with correlated random effects, i.e., $\mathbf{D}_{y,s} \neq \mathbf{0}$, when it was assumed to be endogenous. The functional form was specified to be either linear or quadratic, depending on the current assumption. Incomplete data from both missing mechanisms was imputed and analysed with the sequential approach, again with either independent or correlated random effects. The imputation was repeated with the multivariate

normal approach, using the model with independent error terms as well as the model with correlated error terms, and imputed datasets were created by drawing two values that were at least 50 iterations apart from each of the posterior chains of the incomplete covariate. The resulting ten imputed datasets were analysed analogously to the analysis of the complete data. When `lmer()` was used, the coefficients from the ten corresponding analyses were pooled using Rubin's Rules (Rubin 1987) and when the sequential approach was used, the ten sets of posterior Markov chains were combined to calculate posterior summary measures.

Table 5.2: Overview of imputation and analysis methods used under the assumption of exo- or endogeneity, for the complete as well as incomplete data.

assumption	**compl. data**	**MAR.1 & MAR.2**
exogenous	• sequential • `lmer()`	• sequential • multivariate normal (indep. err.) + `lmer()` • multivariate normal (corr. err.) + `lmer()`
endogenous	• sequential	• sequential • multivariate normal (indep. err.) + sequential • multivariate normal (corr. err.) + sequential

All Bayesian analyses used five parallel MCMC chains and were implemented in JAGS (Plummer 2003), using the R-package **rjags** (Plummer 2016). Chains were assumed to have converged when the Gelman-Rubin criterion (Gelman, Meng, et al. 1996) was not more than 1.1 for the parameters of interest, $(\beta_{y0}, \beta_{y1}, \boldsymbol{\gamma}^\top)$, and not more than 1.25 for other parameters. The number of iterations in the posterior sample necessary to obtain sufficient precision of the posterior estimate was determined based on the sampling error of the chains for $(\beta_{y0}, \beta_{y1}, \boldsymbol{\gamma}^\top)$, which was required to be less than 5% of the corresponding standard error.

5.C.3 Parameters Values

Table 5.3: Parameter values that were used in the data generating models.

sequential model	multivariate normal model
coefficients in the model for y	
$\beta_{y0} = -1.00$	$\tilde{\beta}_{y0} = -1.00$
$\beta_{y1} = 0.80$	$\tilde{\beta}_{y1} = 0.80$
$\gamma_1 = 1.20$	
$\gamma_2 = -0.35$	
coefficients in the model for s	
$\beta_{s0} = 2.00$	$\tilde{\beta}_{s0} = 2.00$
$\beta_{s1} = -0.30$	$\tilde{\beta}_{s1} = -0.30$
(co)variances of the error terms	
$\sigma_y^2 = 0.40$	$\tilde{\sigma}_y^2 = 0.40$
$\sigma_s^2 = 0.30$	$\tilde{\sigma}_s^2 = 0.30$
	$\tilde{\sigma}_{y,s} = 0.20$
(co)variances of the random effects	
$\mathbf{D}_y = \begin{bmatrix} 1.413 & 0.166 \\ 0.166 & 0.106 \end{bmatrix}$	$\widetilde{\mathbf{D}}_y = \begin{bmatrix} 1.413 & 0.166 \\ 0.166 & 0.106 \end{bmatrix}$
$\mathbf{D}_s = \begin{bmatrix} 1.500 & -0.202 \\ -0.202 & 0.165 \end{bmatrix}$	$\widetilde{\mathbf{D}}_s = \begin{bmatrix} 1.500 & -0.202 \\ -0.202 & 0.165 \end{bmatrix}$
$\mathbf{D}_{y,s} = \begin{bmatrix} 0.022 & -0.015 \\ -0.015 & -0.083 \end{bmatrix}$	$\widetilde{\mathbf{D}}_{y,s} = \begin{bmatrix} 0.022 & -0.015 \\ -0.015 & -0.083 \end{bmatrix}$

Table 5.4: Values for $\boldsymbol{\zeta}$ that were used to create missing values in each of the six data scenarios.

	ζ_1	ζ_2	ζ_3
seq. (exo., lin.)	-3.20	-0.5	0.0
seq. (endo., lin.)	-3.20	-0.5	0.0
seq. (exo., qdr.)	-1.85	-0.7	-0.3
seq. (endo., qdr.)	-1.85	-0.7	-0.3
mvn. (indep. err.)	-1.60	-0.5	-0.3
mvn. (corr. err.)	-1.60	-0.5	-0.3

5.C.4 Detailed Discussion of the Results

In this section, we present and discuss the results from the simulation study in detail. We adhere to the list of aims of the simulation study stated above and start with the comparison of the sequential and the multivariate normal approach when exo- or endogeneity and functional form are specified correctly. The relevant part of the results is shown in Figure 5.8 and grouped into four figures, according to whether the data generating model implied an exogenous or endogenous covariate and a linear or quadratic functional relation with the outcome. The six data generating models are indicated by different plotting symbols. The different approaches (sequential and multivariate normal, with independent or correlated error terms) are represented by rows, whereas the columns show three evaluation measures, summarized over all 200 simulated datasets. The relative bias was calculated as the median of the ratios of the estimate (REML estimate or posterior mean) from the analysis of the incomplete data and the estimate from the corresponding complete data analysis, the mean squared error (MSE) as the average of the squared differences between the estimates from the missing and complete data analyses, and the CI-coverage as the proportion of CIs from the missing data analysis that covered the estimate from the corresponding complete data analysis. The desired value for each of these measures is indicated by the vertical line. Since results were similar under both missing mechanisms in most settings and to facilitate readability of the figures, we present results under *MAR.1* only. Furthermore, we only present results for the parameter related to the time-varying covariate, $\boldsymbol{\gamma}$.

Figure 5.8a shows that the sequential approach as well as the multivariate normal approach with correlated error terms were both unbiased and had CIs that covered the posterior mean from the complete data analysis in all 200 simulations when data was generated with a linear functional form and exogenous covariate. The multivariate normal approach with independent error terms, however, was

clearly biased and had CIs that did not cover the estimate from the analysis of the complete data for any of the simulated datasets. It performed even worse in the endogenous setting (shown in Figure 5.8b), unless data was generated by this very model. The multivariate normal approach with correlated error terms and the sequential approach performed well in data that was simulated from the sequential model or the multivariate normal model with correlated error terms, but were biased for data that was generated by the multivariate normal model with independent error terms (relative bias approx. 1.05 under *MAR.1* and approx. 1.3 under *MAR.2*, for both approaches), however, coverage of the CIs was above 95%.

Corresponding results in the settings with a quadratic relation are shown in Figures 5.8c and 5.8d for the sequential approach, since a correct specification of a quadratic functional form during the imputation procedure is not possible in the multivariate normal approach. Also in these settings, the sequential approach performed well.

The results from the linear setting demonstrate that misspecification of the correlation structure of the error terms may have great impact on the results and we will, therefore, exclude the multivariate normal model with independent error terms in the subsequent comparisons. Note that, since in our simulation ε_{ij}^s enters the model for y_{ij} through s_{ij}, the sequential approach also implies correlation between the error terms.

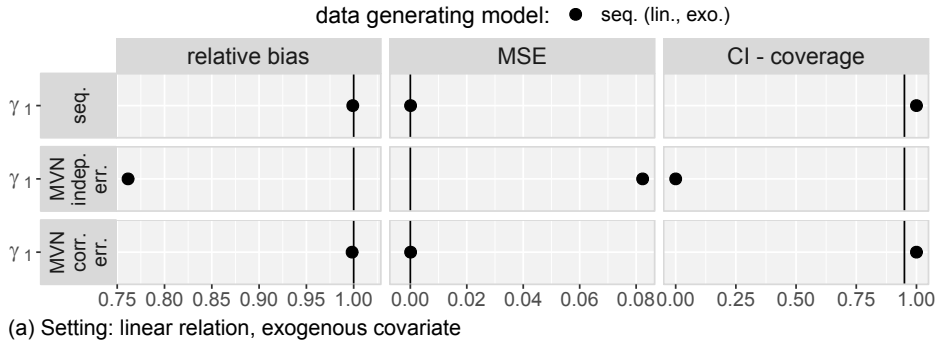
The second aim of this simulation study was to investigate the effect of misspecification of the exo- or endogeneity during imputation and/or analysis. Figure 5.9 summarizes the relevant results, again split into the different settings with regards to the assumption of exo- or endogeneity and the functional form. The setting in which a linear endogenous covariate is misspecified as exogenous is shown in Figure 5.9a. There, imputation with the multivariate normal approach did not add any bias compared to the (in the same way misspecified) analysis of the complete data, irrespective if the data were generated by the multivariate normal model or the (endogenous) sequential model. Imputation and analysis with the sequential approach with independent random effects introduced some additional bias: for data that was simulated from the sequential approach with correlated random effects, the relative bias was 0.98 and the coverage 88%. When data was simulated from the multivariate normal model, however, the relative bias worsened to 0.87 and the coverage to 2.5%. In the reversed case, where a linear, exogenous covariate was misspecified as endogenous, estimates from both the sequential and the multivariate normal approach were unbiased (see Figure 5.9b). Simulations in the quadratic setting led to corresponding results (for the sequential approach), although the estimates of γ_2 were less biased than the estimates of γ_1, as can be

seen in Figures 5.9c and 5.9d.

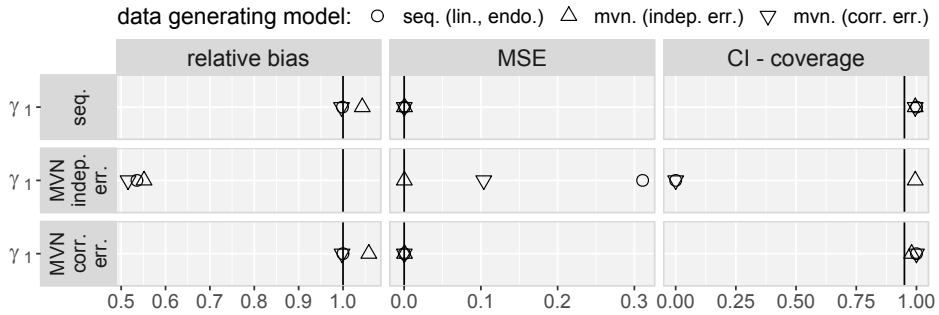
With regards to our third aim, the exploration of bias due to misspecification of the functional form, we investigated two specific issues. First, we explored to what extend misspecification of a quadratic functional form as linear during imputation with the multivariate normal approach, when the functional relation is correctly specified during the subsequent analysis, had an impact on the results. Our findings are summarized in Figure 5.10 and show that this type of misspecification resulted in a relative bias of approximately 0.8 and very bad coverage of the CIs of both parameters, γ_1 and γ_2, which did not cover the true parameters in any of the simulations.

Second, we extended the misspecification to the analysis model, i.e., assumed that the functional form was linear when the true structure was quadratic. In both approaches, this resulted in only little additional bias compared to the misspecified analysis of the complete data, in the exogenous as well as in the endogenous setting, and had no influence on the coverage of the CIs (see Figure 5.10).

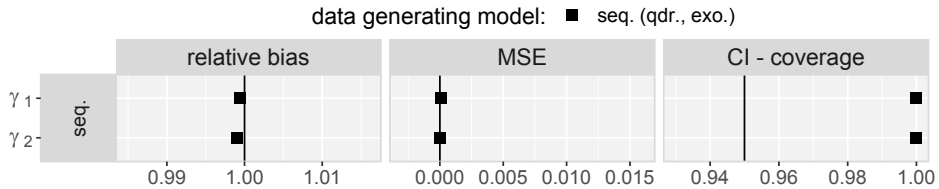
Summarizing the results from our simulation study, we saw that misspecification of the error terms in the multivariate normal approach as independent had the largest impact on the results, leading to estimates that were on average half the value of the estimate from the analysis of the complete data and CIs that had coverage of 0%. For misspecification of an endogenous covariate as exogenous the most severe relative bias that was observed was 0.87 (and a corresponding coverage of the 95% CI of only 2.5%) in the setting where the sequential approach (with independent random effects) was used to impute data that was generated by the multivariate normal model. Imputation with the multivariate normal approach (with correlated error terms) in a setting where the functional form was correctly assumed to be quadratic had the second largest impact, with a relative bias of approximately 0.8, and resulted in CIs that only covered the parameter estimated in the analysis of the complete data in one single simulation for *MAR.2*. The bias that was added due to misspecification of the functional form as linear during imputation and analysis as compared to the results from the analysis of the complete data under the same misspecification was small (between 0.96 and 1.05) and overall comparable between the multivariate normal and the sequential approach.



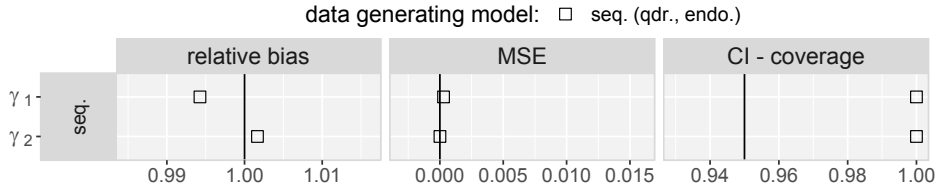
(a) Setting: linear relation, exogenous covariate



(b) Setting: linear relation, endogenous covariate

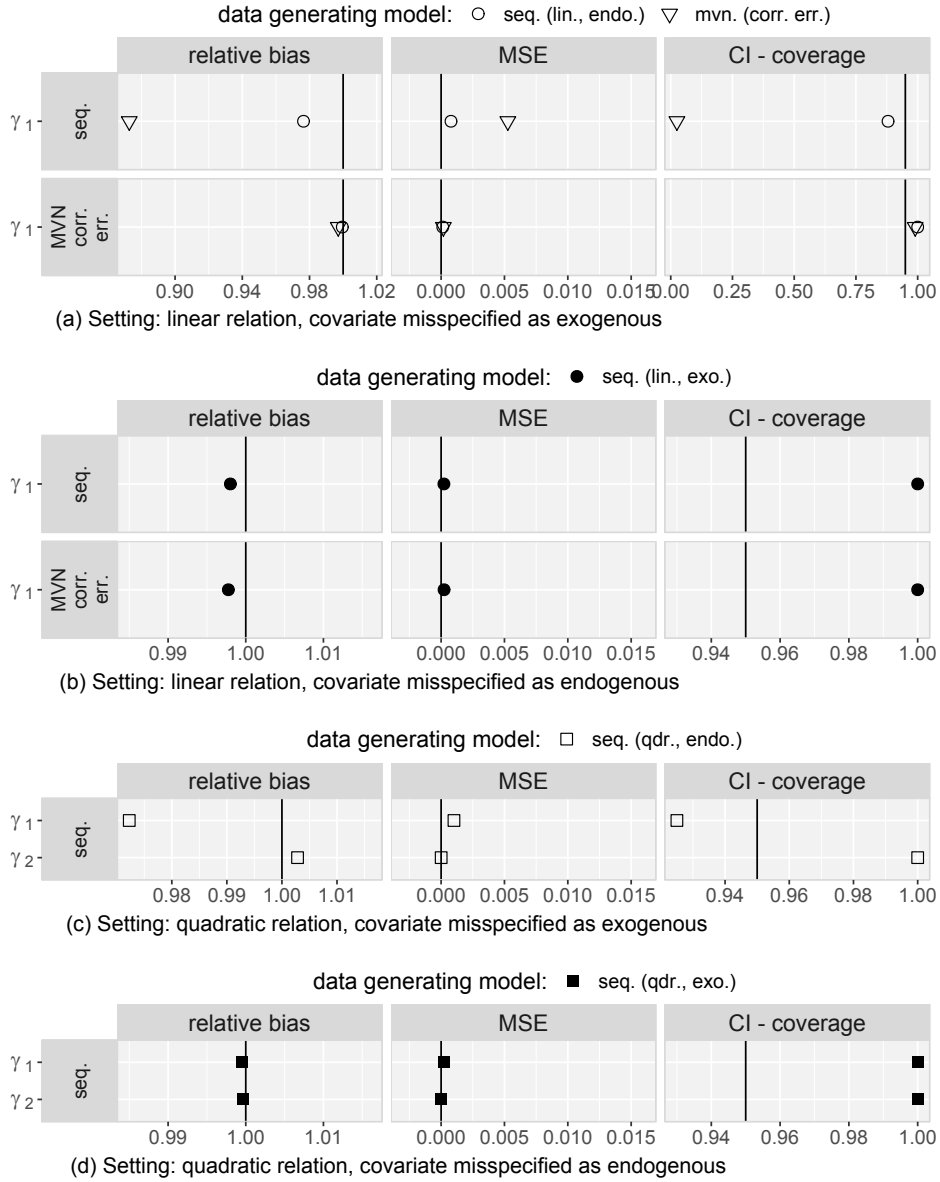


(c) Setting: quadratic relation, exogenous covariate



(d) Setting: quadratic relation, endogenous covariate

Figure 5.8: Relative bias, mean squared error (MSE), and proportion of CIs that covered the estimate from the analysis of the complete data, when imputation and analysis models were correctly specified with regards to exo- or endogeneity and functional form. The vertical lines mark the respective desired values.



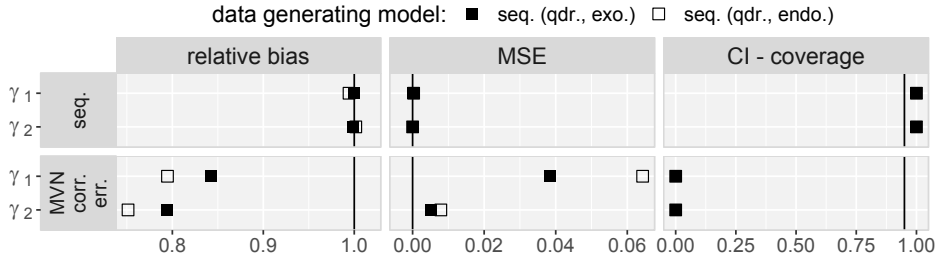
(a) Setting: linear relation, covariate misspecified as exogenous

(b) Setting: linear relation, covariate misspecified as endogenous

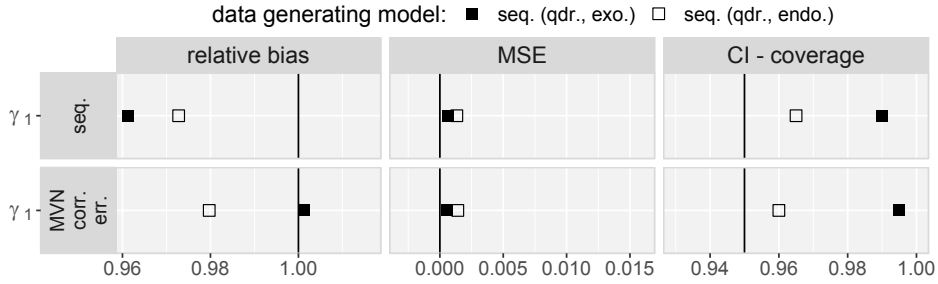
(c) Setting: quadratic relation, covariate misspecified as exogenous

(d) Setting: quadratic relation, covariate misspecified as endogenous

Figure 5.9: Relative bias, mean squared error (MSE), and proportion of CIs that covered the estimate from the analysis of the complete data, when imputation and analysis models were correctly specified with regards to the functional form but misspecified with regards to the exo- or endogeneity of the time-varying covariate. The vertical lines mark the respective desired values.



(a) Misspecification of the functional form as linear during imputation with the multivariate normal approach, while correctly specifying it as quadratic in the analysis model in comparison with the correctly specified sequential model.



(b) Misspecification of the functional form as linear in both, imputation and analysis model.

Figure 5.10: Relative bias, mean squared error (MSE), and proportion of CIs that covered the estimate from the analysis of the complete data, when imputation and analysis models were correctly specified with regards to exo- or endogeneity but misspecified with regards to the functional form. The vertical lines mark the respective desired values.

5.C.5 JAGS Syntax

We provide here an example of the JAGS (Plummer 2003) syntax that was used in the simulation study. The first part of syntax shows the sequential model assuming **s** is exogenous and has a linear relation with **y**. We then show how to adapt that syntax for the endogenous setting. The second part contains example syntax that shows how the multivariate normal approach can be implemented in JAGS. There, we present the syntax used in the setting with correlated error terms.

When implementing mixed models in JAGS, it is convenient to use hierarchical centring. This means that the fixed effects enter the linear predictor through the random effects, i.e., the random effects are not centred around zero but around the fixed effects. The syntax differs from the formulas provided in the previous sections to this regard. Note also that normal distributions have to be specified using the precision instead of the variance when using JAGS.

Data / Notation:

- `TN`: number of observations in the dataset
- `N`: number of individuals
- `priorR`: 4×4 diagonal matrix of NA values

Sequential Approach

```
model {
  for (j in 1:TN) {
    # linear mixed effects models for y and s
    y[j] ~ dnorm(mu.y[j], tau.y)
    s[j] ~ dnorm(mu.s[j], tau.s)

    # linear predictors
    # (hierarchical centring specification for baseline effects)
    mu.y[j] <- inprod(b[subj[j], 1:2], Z[j, ]) + beta[3] * s[j]
    mu.s[j] <- inprod(b[subj[j], 3:4], Z[j, ])
  }

  # priors for the precision of y and s
  tau.y ~ dgamma(0.01, 0.01)
  tau.s ~ dgamma(0.01, 0.01)
```

```
  # specification of the random effects
  for (i in 1:N) {
    # random effects in the model for y
    b[i, 1:2] ~ dmnorm(mu.b[i, 1:2], inv.D.y[ , ])
    mu.b[i, 1] <- beta[1] # random intercept
    mu.b[i, 2] <- beta[2] # random slope

    # random effects in the model for s
    b[i, 3:4] ~ dmnorm(mu.b[i, 3:4], inv.D.s[ , ])
    mu.b[i, 3] <- alpha[1] # random intercept
    mu.b[i, 4] <- alpha[2] # random slope
  }

  # priors for the fixed effects
  for (k in 1:3) {
    beta[k] ~ dnorm(0, 0.001)
  }
  for (k in 1:2) {
    alpha[k] ~ dnorm(0, 0.001)
  }


  # priors for the precision of the random effects
  for (k in 1:4) {
    priorR[k, k] ~ dgamma(0.1, 0.01)
  }

  # precision and covariance matrix of the random effects
  # in the model for y
  inv.D.y[1:2, 1:2] ~ dwish(priorR[1:2, 1:2], 2)
  D.y[1:2, 1:2] <- inverse(inv.D.y[ , ])

  # precision and covariance matrix of the random effects
  # in the model for s
  inv.D.s[1:2, 1:2] ~ dwish(priorR[3:4, 3:4], 2)
  D.s[1:2, 1:2] <- inverse(inv.D.s[ , ])
}
```

In the endogenous setting, the specification of the random effects and related priors changes to:

```
  # specification of the random effects
  for (i in 1:N) {
    b[i, 1:4] ~ dmnorm(mu.b[i, 1:4], inv.D[ , ])

    # random effects in the model for y
    mu.b[i, 1] <- beta[1]
    mu.b[i, 2] <- beta[2]

    # random effects in the model for s
    mu.b[i, 3] <- alpha[1]
    mu.b[i, 4] <- alpha[2]
  }

  # priors for the precision of the random effects
  for(k in 1:4){
    priorR[k, k] ~ dgamma(0.1, 0.01)
  }

  # precision and covariance matrix of the random effects
  inv.D[1:4, 1:4] ~ dwish(priorR[1:4, 1:4], 4)
  D[1:4, 1:4] <- inverse(inv.D[ , ])
}
```

Multivariate Normal Approach

A natural specification for the distribution of $\boldsymbol{y}$ and $\mathbf{s}$ in the setting with correlated random effects would be to specify $\{y(t), s(t)\}^\top$ as multivariate normal. However, JAGS cannot sample from a multivariate normal distribution when only one component is missing. We therefore have to specify separate univariate normal distributions for $y(t)$ and $s(t)$ and additionally a multivariate normal distribution for the error terms $\{\varepsilon_i^y(t), \varepsilon_i^s(t)\}^\top$. The precision in the univariate normal distributions for $y(t)$ and $s(t)$ is then set to a large value to "force" the variation to be modelled in the multivariate normal distribution of the error terms rather than the two univariate normal distributions.

Data / Notation:

- `mu.eps` = c(0, 0)
- `tau` = 10000

```
model {
  for (j in 1:TN) {
    # linear mixed effects models for y and s
    y[j] ~ dnorm(mu.y[j], tau)
    s[j] ~ dnorm(mu.s[j], tau)

    # linear predictors
    # (hierarchical centring specification and "epsilon trick")
    mu.y[j] <- inprod(b[subj[j], 1:2], Z.y[j, ]) + eps[j, 1]
    mu.s[j] <- inprod(b[subj[j], 3:4], Z.s[j, ]) + eps[j, 2]

    eps[j, 1:2] ~ dmnorm(mu.eps[], inv.Sig[ , ])
  }

  # priors for the precision of y and s
  for (k in 1:2) {
    priorR.invSig[k, k] ~ dgamma(0.1, 0.01)
  }
  inv.Sig[1:2, 1:2] ~ dwish(priorR.invSig[1:2, 1:2], 2)
  Sig[1:2, 1:2] <- inverse(inv.Sig[ , ])


  # specification of the random effects
  for (i in 1:N) {
    # random effects in the model for y
    b[i, 1:4] ~ dmnorm(mu.b[i, 1:4], inv.D[ , ])
    mu.b[i, 1] <- beta[1, 1]
    mu.b[i, 2] <- beta[2, 1]

    # random effects in the model for s
    mu.b[i, 3] <- beta[1, 2]
    mu.b[i, 4] <- beta[2, 2]
  }
```

```
  # priors for the precision of the random effects
  for (k in 1:4) {
    priorR.invD[k, k] ~ dgamma(0.1, 0.01)
  }

  # precision and covariance matrix of the random effects
  inv.D[1:4, 1:4] ~ dwish(priorR.invD[1:4, 1:4], 4)
  D[1:4, 1:4] <- inverse(inv.D[ , ])

  # priors for the fixed effects
  for (k in 1:2) {
    beta[k, 1] ~ dnorm(0, 0.001)
    beta[k, 2] ~ dnorm(0, 0.001)
  }
}
```

References

Andrinopoulou, E.-R. and D. Rizopoulos (2016). "Bayesian shrinkage approach for a joint model of longitudinal and survival outcomes assuming different association structures". *Statistics in Medicin*, **35**(26):4813–4823. DOI: 10.1002/sim.7027.

Bartlett, J. W. et al. (2015). "Multiple imputation of covariates by fully conditional specification: accommodating the substantive model". *Statistical Methods in Medical Research*, **24**(4):462–487. DOI: 10.1177/0962280214521348.

Bates, D. et al. (2015). "Fitting Linear Mixed-Effects Models Using lme4". *Journal of Statistical Software*, **67**(1):1–48. DOI: 10.18637/jss.v067.i01.

Carpenter, J. R., H. Goldstein, and M. G. Kenward (2011). "REALCOM-IMPUTE Software for Multilevel Multiple Imputation with Mixed Response Types". *Journal of Statistical Software*, **45**(1):1–14. DOI: 10.18637/jss.v045.i05.

Carpenter, J. R. and M. G. Kenward (2013). *Multiple Imputation and its Application*. John Wiley & Sons, Ltd. DOI: 10.1002/9781119942283.

Daniels, M. J. and J. W. Hogan (2008). *Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis*. Boca Raton, FL: Chapman & Hall/CRC. DOI: 10.1201/9781420011180.

Diggle, P. et al. (2002). *Analysis of Longitudinal Data*. 2nd edition. Oxford, United Kingdom: Oxford University Press. ISBN: 9780199676750.

Engle, R. F., D. F. Hendry, and J.-F. Richard (1983). "Exogeneity". *Econometrica*, **51**(2):277–304. DOI: 10.2307/1911990.

Erler, N. S., D. Rizopoulos, J. van Rosmalen, et al. (2016). "Dealing with missing covariates in epidemiologic studies: a comparison between multiple imputation and a full Bayesian approach". *Statistics in Medicine*, **35**(17):2955–2974. DOI: 10.1002/sim.6944.

Garrett, E. S. and S. L. Zeger (2000). "Latent Class Model Diagnosis". *Biometrics*, **56**(4):1055–1067. DOI: 10.1111/j.0006-341X.2000.01055.x.

Gelman, A., X.-L. Meng, and H. Stern (1996). "Posterior predictive assessment of model fitness via realized discrepancies". *Statistica Sinica*, **6**(4):733–760.

Ibrahim, J. G., M.-H. Chen, and S. R. Lipsitz (2002). "Bayesian methods for generalized linear models with covariates missing at random". *Canadian Journal of Statistics*, **30**(1):55–78. DOI: 10.2307/3315865.

Kooijman, M. N. et al. (2016). "The Generation R Study: design and cohort update 2017". *European Journal of Epidemiology*, **31**(12):1243–1264. DOI: 10.1007/s10654-016-0224-9.

Lesaffre, E. M. and A. B. Lawson (2012). *Bayesian Biostatistics*. John Wiley & Sons. DOI: 10.1002/9781119942412.

Little, R. J. and D. B. Rubin (2002). *Statistical Analysis with Missing Data*. Hoboken, New Jersey: John Wiley & Sons, Inc. DOI: 10.1002/9781119013563.

Lunn, D. J. et al. (2000). "WinBUGS – a Bayesian modelling framework: Concepts, structure, and extensibility". *Statistics and Computing*, **10**(4):325–337. DOI: 10.1023/A:1008929526011.

Mallick, H. and N. Yi (2013). "Bayesian methods for high dimensional linear models". *Journal of Biometrics & Biostatistics*, **4**(S3):S1–005. DOI: 10.4172/2155-6180.S1-005.

Plummer, M. (2003). "JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling". *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*. Ed. by K. Hornik, F. Leisch, and A. Zeileis. ISSN: 1609-395X.

Plummer, M. (2016). *rjags: Bayesian Graphical Models using MCMC*. R package version 4-6. URL: https://CRAN.R-project.org/package=rjags.

Quartagno, M. and J. R. Carpenter (2016). *jomo: A package for Multilevel Joint Modelling Multiple Imputation*. URL: http://CRAN.R-project.org/package=jomo.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: http://www.R-project.org/.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley Series in Probability and Statistics. Wiley. DOI: 10.1002/9780470316696.

Seaman, S. R., J. W. Bartlett, and I. R. White (2012). "Multiple imputation of missing covariates with non-linear effects and interactions: an evaluation

of statistical methods". *BMC Medical Research Methodology*, **12**(1):46. DOI: 10.1186/1471-2288-12-46.

Tielemans, M. J. et al. (2015). "A Priori and a Posteriori Dietary Patterns during Pregnancy and Gestational Weight Gain: The Generation R Study". *Nutrients*, **7**(11):9383–9399. DOI: 10.3390/nu7115476.

Van Buuren, S. (2012). *Flexible Imputation of Missing Data*. Chapman & Hall/CRC Interdisciplinary Statistics. Taylor & Francis.

White, I. R., P. Royston, and A. M. Wood (2011). "Multiple imputation using chained equations: Issues and guidance for practice". *Statistics in Medicine*, **30**(4):377–399. DOI: 10.1002/sim.4067.

6 JointAI: Joint Analysis and Imputation of Incomplete Data in R

This chapter is based on

Nicole S. Erler, Dimitris Rizopoulos and Emmanuel M. E. H. Lesaffre. JointAI: Joint Analysis and Imputation of Incomplete Data in R. (Manuscript in preparation)

Abstract

Missing data occur in many types of studies and typically complicate the analysis. Multiple imputation, either using joint modelling or the more flexible full conditional specification approach, is popular and works well in standard settings. However, in settings involving non-linear associations or interactions, incompatibility of the imputation model with the analysis model is an issue often resulting in bias. Similarly, complex outcomes such as longitudinal or survival outcomes cannot be adequately handled by standard implementations.

In this chapter, we introduce the R package **JointAI**, which utilizes the Bayesian framework to perform simultaneous analysis and imputation in regression models with incomplete covariates. Using a fully Bayesian joint modelling approach it overcomes the issue of uncongeniality while retaining the attractive flexibility of fully conditional specification multiple imputation by specifying the joint distribution of analysis and imputation models as a sequence of univariate models that can be adapted to the type of variable. **JointAI** provides functions for Bayesian inference with generalized linear and generalized linear mixed models as well as survival models, that take arguments analogous to their corresponding and well known complete data versions from base R and other packages. Usage and features of **JointAI** are described and illustrated using various examples and the theoretical background is outlined.

6.1 Introduction

Missing data are a challenge common to the analysis of data from virtually all kinds of studies. Especially when many variables are measured, as in big cohort studies, or when data are obtained retrospectively, e.g., from registries, proportions of missing values up to 50% are not uncommon in some variables.

Multiple imputation, which is often considered the gold standard to handle incomplete data, has its origin in the 1970s and was primarily developed for survey data (Rubin 1987; Rubin 2004). One of its first implementations in R (R Core Team 2018) is the package **norm** (Novo and Schafer 2010), which performs multiple imputation under the joint modelling framework using a multivariate normal distribution (Schafer 1997). Nowadays, multiple imputation using a fully conditional specification (FCS), also called multiple imputation using chained equations (MICE), and its seminal implementation in the R package **mice** (Van Buuren and Groothuis-Oudshoorn 2011; Van Buuren 2012), is more frequently used.

Datasets have gotten more complex compared to the survey data multiple imputation was developed for. Therefore, more sophisticated methods that can adequately handle the features of modern data and comply with the assumptions made in its analysis are required. Modern studies do not only record univariate outcomes, measured in cross-sectional settings, but outcomes that consist of two or more measurements, such as repeatedly measured or survival outcomes. Furthermore, non-linear effects, introduced by functions of covariates, such as transformations, polynomials or splines, or interactions between variables are considered in the analysis and hence need to be taken into account during imputation.

Standard multiple imputation, either using FCS or a joint modelling approach, e.g., under a multivariate normal distribution, assumes linear associations between all variables. Moreover, FCS requires the outcome to be explicitly specified in each of the linear predictors of the full conditional distributions. In settings where the outcome is more complex than just univariate, this is not straightforward and not generally possible without information loss, leading to misspecified imputation models.

Some extensions of standard multiple imputation have been developed and are implemented in R packages and other software, but the larger part of software for imputation is restricted to standard settings such as cross-sectional survey data. R packages that offer extensions frequently focus on particular settings and researchers need to be familiar with a number of different packages, which often require input and specifications in very different forms. Moreover, most R packages dealing with incomplete data implement multiple imputation, i.e., create multiple

imputed datasets, which are then analysed in a second step, followed by pooling of the results.

The R package **JointAI** (Erler 2019), which is presented in this chapter, follows a different, fully Bayesian approach. It provides a unified framework for both simple and more complex models, using a consistent specification most users will be familiar with from commonly used (base) R functions. By modelling the analysis model of interest jointly with the incomplete covariates, analysis and imputation can be performed simultaneously while assuring compatibility between all submodels (Erler, Rizopoulos, Rosmalen, et al. 2016; Erler, Rizopoulos, Jaddoe, et al. 2019). In this joint modelling approach, the added uncertainty due to the missing values is automatically taken into account in the posterior distribution of the parameters of interest, and no pooling of results from repeated analyses is necessary. The joint distribution is specified in a convenient way, using a sequence of conditional distributions that can be specified flexibly according to each type of variable. Since the analysis model of interest defines the first distribution in the sequence, the outcome is included in the joint distribution without the need for it to enter the linear predictor of any of the other models. Moreover, non-linear associations that are part of the analysis model are automatically taken into account for the imputation of missing values. This directly enables our approach to handle complicated models, with complex outcomes and flexible linear predictors.

In the following, we introduce the R package **JointAI**, which performs joint analysis and imputation of regression models with incomplete covariates under the Missing At Random assumption (Rubin 1976), and explain how data with incomplete covariate information can be analysed and imputed with it. The package is available for download at the Comprehensive R Archive Network (CRAN) under https://CRAN.R-project.org/package=JointAI. Section 6.2 briefly describes the theoretical background. An outline of the general structure of **JointAI** is given in Section 6.3, followed by an introduction to example datasets that are used throughout this chapter, in Section 6.4. Details about model specification, settings controlling the Markov Chain Monte Carlo sampling, and summary, plotting and other functions that can be applied after fitting the model are given in Sections 6.5 through 6.7. We conclude the paper with an outlook of planned extensions and discuss the limitations that are introduced by the assumptions made in the sequential imputation approach.

6.2 Theoretical Background

Consider the general setting of a regression model where interest lies in a set of parameters $\boldsymbol{\theta}$ that describe the association between a univariate outcome $\mathbf{y}$ and a

set of covariates $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$. In the Bayesian framework, inference over $\boldsymbol{\theta}$ is obtained by estimation of the posterior distribution of $\boldsymbol{\theta}$, which is proportional to the product of the likelihood of the data $(\mathbf{y}, \mathbf{X})$ and the prior distribution of $\boldsymbol{\theta}$,

$$p(\boldsymbol{\theta} \mid \mathbf{y}, \mathbf{X}) \propto p(\mathbf{y}, \mathbf{X} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta}).$$

When some of the covariates are incomplete, $\mathbf{X}$ consists of two parts, the completely observed variables $\mathbf{X}_{obs}$ and those variables that are incomplete, $\mathbf{X}_{mis}$. If $\mathbf{y}$ had missing values (and this missingness was ignorable), the only necessary change in the formulas below would be to replace $\mathbf{y}$ by $\mathbf{y}_{mis}$. We will, therefore, without loss of generality, consider $\mathbf{y}$ to be completely observed.

The likelihood of the complete data, i.e., observed and unobserved, can be factorized in the following convenient way:

$$p(\mathbf{y}, \mathbf{X}_{obs}, \mathbf{X}_{mis} \mid \boldsymbol{\theta}) = p(\mathbf{y} \mid \mathbf{X}_{obs}, \mathbf{X}_{mis}, \boldsymbol{\theta}_{y|x})\, p(\mathbf{X}_{mis} \mid \mathbf{X}_{obs}, \boldsymbol{\theta}_x),$$

where the first factor constitutes the analysis model of interest, described by a vector of parameters $\boldsymbol{\theta}_{y|x}$, and the second factor is the joint distribution of the incomplete variables, i.e., the imputation part of the model, described by parameters $\boldsymbol{\theta}_x$, and $\boldsymbol{\theta} = (\boldsymbol{\theta}_{y|x}^\top, \boldsymbol{\theta}_x^\top)^\top$.

Explicitly specifying the joint distribution of all data is one of the major advantages of the Bayesian approach, since this facilitates the use of all available information of the outcome in the imputation of the incomplete covariates (Erler, Rizopoulos, Rosmalen, et al. 2016), which becomes especially relevant for more complex outcomes such as repeatedly measured variables (see Section 6.2.2).

In complex models, the posterior distribution can usually not be analytically derived but Markov Chain Monte Carlo (MCMC) methods are used to obtain samples from the posterior distribution. The MCMC sampling in **JointAI** is done using Gibbs sampling, which iteratively samples from the full conditional distributions of the unknown parameters and missing values.

In the following sections, we describe each of the three parts of the model, the analysis model, the imputation part and the prior distributions, in detail.

6.2.1 Analysis Model

The analysis model of interest is described by the probability density function $p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\theta}_{y|x})$. The R package **JointAI** can currently handle analysis models that are either generalized linear regression models (GLM), (generalized) linear mixed models (GLMM), cumulative logit (mixed) models, parametric (Weibull) survival models or Cox proportional hazards models.

For a GLM, the probability density function is chosen from the exponential family and the model has the linear predictor

$$g(E(y_i \mid \mathbf{X}, \boldsymbol{\theta}_{y|x})) = \mathbf{x}_i^\top \boldsymbol{\beta},$$

where $g(\cdot)$ is a link function, y_i the value of the outcome variable for subject i, and $\mathbf{x}_i$ is the row of $\mathbf{X}$ that contains the covariate information for i.

For a GLMM the linear predictor is of the form

$$g(E(y_{ij} \mid \mathbf{X}, \mathbf{b}_i, \boldsymbol{\theta}_{y|x})) = \mathbf{x}_{ij}^\top \boldsymbol{\beta} + z_{ij}^\top \mathbf{b}_i,$$

where y_{ij} is the j-th outcome of subject i, $\mathbf{x}_{ij}$ is the corresponding vector of covariate values, $\mathbf{b}_i$ a vector of random effects pertaining to subject i, and $\mathbf{z}_{ij}$ the row of the design matrix of the random effects, $\mathbf{Z}$, that corresponds to the j-th measurement of subject i. $\mathbf{Z}$ typically contains a subset of the variables in $\mathbf{X}$, and $\mathbf{b}_i$ follows a normal distribution with mean zero and covariance matrix $\mathbf{D}$.

In both cases the parameter vector $\boldsymbol{\theta}_{y|x}$ contains the regression coefficients $\boldsymbol{\beta}$, and potentially additional variance parameters (e.g., for linear (mixed) models), for which prior distributions will be specified in Section 6.2.3.

Cumulative logit mixed models are of the form

$$\begin{aligned}
y_{ij} &\sim \text{Multinom}(\pi_{ij,1}, \ldots, \pi_{ij,K}), \\
\\
\pi_{ij,1} &= P(y_{ij} \leq 1), \\
\pi_{ij,k} &= P(y_{ij} \leq k) - P(y_{ij} \leq k-1), \quad k \in 2, \ldots, K-1, \\
\pi_{ij,K} &= 1 - \sum_{k=1}^{K-1} \pi_{ij,k}, \\
\\
\text{logit}(P(y_{ij} \leq k)) &= \gamma_k + \eta_{ij}, \quad k \in 1, \ldots, K, \\
\eta_{ij} &= x_{ij}^\top \boldsymbol{\beta} + z_{ij}^\top \mathbf{b}_i, \\
\\
\gamma_1, \delta_1, \ldots, \delta_{K-1} &\overset{iid}{\sim} N(\mu_\gamma, \sigma_\gamma^2), \\
\gamma_k &\sim \gamma_{k-1} + \exp(\delta_{k-1}), \quad k = 2, \ldots, K,
\end{aligned}$$

where $\pi_{ij,k} = P(y_{ij} = k)$ and $\text{logit}(x) = \log\left(\frac{x}{1-x}\right)$. A cumulative logit regression model for a univariate outcome y_i can be obtained by dropping the index j and omitting $z_{ij}^\top \mathbf{b}_i$. In cumulative logit (mixed) models, the design matrix $\mathbf{X}$ does not contain an intercept, since outcome category specific intercepts $\gamma_1, \ldots, \gamma_K$ are specified. Here, the parameter vector $\boldsymbol{\theta}_{y|x}$ includes the regression coefficients $\boldsymbol{\beta}$, and the first intercept γ_1 and increments $\delta_1, \ldots, \delta_{K-1}$.

Survival data are typically characterized by the observed event or censoring times, T_i, and the event indicator, D_i, which is equal to one if the event was observed and zero otherwise. **JointAI** provides two types of models to analyse right censored survival data, a parametric model assuming a Weibull distribution for the true (but partially unobserved) survival times T^*, and a semi-parametric Cox proportional hazards model.

The parametric survival model is implemented as

$$\begin{aligned} T_i^* &\sim \text{Weibull}(1, r_i, s), \\ D_i &\sim \mathbb{1}(T_i^* \geq C_i), \\ \log(r_j) &= -\mathbf{x}_i^\top \boldsymbol{\beta}, \\ s &\sim \text{Exp}(0.01), \end{aligned}$$

where $\mathbb{1}$ is the indicator function which is one if $T_i^* \geq C_i$, and zero otherwise.

For the Cox proportional hazards model, following D. Lunn et al. (2012), a counting process representation is implemented, where the baseline hazard is assumed to be piecewise constant and changes only at observed event times. Let $\{N_i(t), t \geq 0\}$ be an event counting process for individual i, where $N_i(t) = 0$ until the individual experiences an event and increases by one at the time of the event. $dN_i(t)$ then denotes the change in $N_i(t)$ in the interval $[t, t + dt)$, where dt is the length of that interval, and can be modelled as a Poisson random variable with time-varying intensity $\lambda_i(t)$. This intensity depends on covariates $\mathbf{x}_i$, the baseline hazard $\lambda_0(t)$, and the risk set indicator $R_i(t)$, which is equal to one if, at time t, subject i is at risk for an event, and zero otherwise.

$$\begin{aligned} dN(t)_i &\sim \text{Poisson}(\lambda_i(t)), \quad t \in 0, \ldots, T \\ \lambda_i(t) &= \exp(\mathbf{x}_i^\top \boldsymbol{\beta})\ \lambda_0(t)\ R_i(t) \\ \lambda_0(t) &\sim \text{Gamma}(c\lambda_0(t)^*, c) \end{aligned}$$

where $\lambda_0(t)^*$ is a prior guess of the failure rate at time t, and c represents the confidence about that prior guess.

6.2.2 Imputation Part

A convenient way to specify the joint distribution of the incomplete covariates $\mathbf{X}_{mis} = (\mathbf{x}_{mis_1}, \ldots, \mathbf{x}_{mis_q})$ is to use a sequence of conditional univariate distributions (Erler, Rizopoulos, Rosmalen, et al. 2016; Ibrahim et al. 2002)

$$\begin{aligned} p(\mathbf{x}_{mis_1}, \ldots, \mathbf{x}_{mis_q} \mid \mathbf{X}_{obs}, \boldsymbol{\theta}_x) &= p(\mathbf{x}_{mis_1} \mid \mathbf{X}_{obs}, \boldsymbol{\theta}_{x_1}) \\ &\quad \prod_{\ell=2}^{q} p(\mathbf{x}_{mis_\ell} \mid \mathbf{X}_{obs}, \mathbf{x}_{mis_1}, \ldots, \mathbf{x}_{mis_{\ell-1}}, \boldsymbol{\theta}_{x_\ell}), \end{aligned} \tag{6.1}$$

with $\boldsymbol{\theta}_x = (\boldsymbol{\theta}_{x_1}^\top, \ldots, \boldsymbol{\theta}_{x_q}^\top)^\top$.

Each of the conditional distributions is a member of the exponential family, extended with distributions for ordinal categorical variables, and chosen according to the type of the respective variable. Its linear predictor is

$$g_\ell \left\{ E\left(x_{i,mis_\ell} \mid \mathbf{x}_{i,obs}, \mathbf{x}_{i,mis_{<\ell}}, \boldsymbol{\theta}_{x_\ell}\right)\right\} = (\mathbf{x}_{i,obs}^\top, x_{i,mis_1}, \ldots, x_{i,mis_{\ell-1}})\boldsymbol{\alpha}_\ell,$$

for $\ell = 1, \ldots, q$, where $\mathbf{x}_{i,mis_{<\ell}} = (x_{i,mis_1}, \ldots, x_{i,mis_{\ell-1}})^\top$.

Factorization of the joint distribution of the covariates in such a sequence yields a straightforward specification of the joint distribution, even when the covariates are of mixed type.

Missing values in the covariates are sampled from their full conditional distributions that can be derived from the full joint distribution of outcome and covariates.

When, for instance, the analysis model is a GLM, the full conditional distribution of an incomplete covariate x_{i,mis_ℓ} can be written as

$$\begin{aligned}
&p(x_{i,mis_\ell} \mid \mathbf{y}_i, \mathbf{x}_{i,obs}, \mathbf{x}_{i,mis_{-\ell}}, \boldsymbol{\theta}) \\
&\quad \propto \; p\left(y_i \mid \mathbf{x}_{i,obs}, \mathbf{x}_{i,mis}, \boldsymbol{\theta}_{y|x}\right)\, p(\mathbf{x}_{i,mis} \mid \mathbf{x}_{i,obs}, \boldsymbol{\theta}_x)\, p(\boldsymbol{\theta}_{y|x})\, p(\boldsymbol{\theta}_x) \\
&\quad \propto \; p\left(y_i \mid \mathbf{x}_{i,obs}, \mathbf{x}_{i,mis}, \boldsymbol{\theta}_{y|x}\right)\, p(x_{i,mis_\ell} \mid \mathbf{x}_{i,obs}, \mathbf{x}_{i,mis_{<\ell}}, \boldsymbol{\theta}_{x_\ell}) \qquad (6.2) \\
&\qquad \left\{ \prod_{k=\ell+1}^{q} p(x_{i,mis_k} \mid \mathbf{x}_{i,obs}, \mathbf{x}_{i,mis_{<k}}, \boldsymbol{\theta}_{x_k}) \right\} p(\boldsymbol{\theta}_{y|x}) p(\boldsymbol{\theta}_{x_\ell}) \prod_{k=\ell+1}^{p} p(\boldsymbol{\theta}_{x_k}),
\end{aligned}$$

where $\boldsymbol{\theta}_{x_\ell}$ is the vector of parameters describing the model for the ℓ-th covariate, and contains the vector of regression coefficients and potentially additional (variance) parameters. The product of distributions enclosed by curly brackets represents the distributions of those covariates that have x_{mis_ℓ} as a predictive variable in the specification of the sequence in (6.1).

Even though (6.2) describes the actual imputation model, i.e., the distribution the imputed values for x_{i,mis_ℓ} are sampled from, we will use the term "imputation model" for the conditional distribution of x_{i,mis_ℓ} from (6.1), since the latter is the distribution that is explicitly specified by the user and, hence, of more relevance when using **JointAI**.

Imputation with Longitudinal Outcomes

Factorizing the joint distribution into the analysis model and imputation part allows a straightforward extension to settings with more complex outcomes, such as repeatedly measured outcomes. In the case where the analysis model is a GLMM,

the conditional distribution of the outcome in (6.2), $p(y_i \mid \mathbf{x}_{i,obs}, \mathbf{x}_{i,mis}, \boldsymbol{\theta}_{y|x})$, has to be replaced by

$$\left\{ \prod_{j=1}^{n_i} p\left(y_{ij} \mid \mathbf{x}_{i,obs}, \mathbf{x}_{i,mis}, \mathbf{b}_i, \boldsymbol{\theta}_{y|x}\right) \right\}. \tag{6.3}$$

Since $\mathbf{y}$ does not appear in any of the other terms in (6.2), and (6.3) can be chosen to be a model that is appropriate for the outcome at hand, the thereby specified full conditional distribution of x_{i,mis_ℓ} allows us to draw valid imputations that use all available information on the outcome.

This is an important difference to standard FCS, where the full conditional distributions used to impute missing values are specified directly, usually as regression models, and require the outcome to be explicitly included in the linear predictor of the imputation model. In settings with complex outcomes, it is not clear how this should be done, and simplifications may lead to biased results (Erler, Rizopoulos, Rosmalen, et al. 2016). The joint model specification utilized in **JointAI** overcomes this difficulty.

When some of the covariates are time-varying, it is convenient to specify models for these variables at the beginning of the sequence of covariate models, so that models for longitudinal variables have other longitudinal and baseline covariates in their linear predictor, but longitudinal covariates do not enter the predictors of baseline covariates.

Note that whenever there are incomplete baseline covariates it is necessary to specify models for all longitudinal variables, even completely observed ones, while models for completely observed baseline covariates can be omitted. This becomes clear when we extend the factorized joint distribution from above with completely and incompletely observed covariates s_{obs} and s_{mis}:

$$\begin{aligned} &p(y_{ij} \mid \mathbf{s}_{ij,obs}, \mathbf{s}_{ij,mis}, \mathbf{x}_{i,obs}, \mathbf{x}_{i,mis}, \boldsymbol{\theta}_{y|x})\, p(\mathbf{s}_{ij,mis} \mid \mathbf{s}_{ij,obs}, \mathbf{x}_{i,obs}, \mathbf{x}_{i,mis}, \boldsymbol{\theta}_{s_{mis}}) \\ &\quad p(\mathbf{s}_{ij,obs} \mid \mathbf{x}_{i,obs}, \mathbf{x}_{i,mis}, \boldsymbol{\theta}_{s_{obs}})\, p(\mathbf{x}_{i,mis} \mid \mathbf{x}_{i,obs}, \boldsymbol{\theta}_{x_{mis}})\, p(\mathbf{x}_{i,obs} \mid \boldsymbol{\theta}_{x_{obs}}) \\ &\qquad\qquad p(\boldsymbol{\theta}_{y|x})\, p(\boldsymbol{\theta}_{s_{mis}})\, p(\boldsymbol{\theta}_{s_{obs}})\, p(\boldsymbol{\theta}_{x_{mis}})\, p(\boldsymbol{\theta}_{x_{obs}}) \end{aligned}$$

Given that the parameter vectors $\boldsymbol{\theta}_{x_{obs}}$, $\boldsymbol{\theta}_{x_{mis}}$, $\boldsymbol{\theta}_{s_{obs}}$ and $\boldsymbol{\theta}_{s_{mis}}$ are a priori independent, and $p(\mathbf{x}_{i,obs} \mid \boldsymbol{\theta}_{x_{obs}})$ is independent of both $\mathbf{x}_{i,mis}$ and $\mathbf{s}_{ij,mis}$, it can be excluded from the model.

Since $p(\mathbf{s}_{ij,obs} \mid \mathbf{x}_{i,obs}, \mathbf{x}_{i,mis}, \boldsymbol{\theta}_{s_{obs}})$, however, has $\mathbf{x}_{i,mis}$ in its linear predictor and will, hence, be part of the full conditional distribution of $\mathbf{x}_{i,mis}$, it cannot be omitted from the model specification.

Non-linear Associations and Interactions

Other settings in which the fully Bayesian approach employed in **JointAI** has an advantage over standard FCS are settings with interaction terms that involve incomplete covariates, or when the association of the outcome with an incomplete covariate is non-linear. In standard FCS such settings lead to incompatible imputation models (White et al. 2011; Bartlett et al. 2015). This becomes clear when considering the following simple example where the analysis model of interest is the linear regression $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$ and x_i is imputed using $x_i = \alpha_0 + \alpha_1 y_i + \tilde{\varepsilon}_i$. While the analysis model assumes a quadratic relationship, the imputation model assumes a linear association between $\mathbf{x}$ and $\mathbf{y}$ and a joint distribution with these imputation and analysis models as its full conditional distributions does not exist. Because, in **JointAI**, the analysis model is a factor in the full conditional distribution that is used to impute x_i, the non-linear association is taken into account. Furthermore, since it is the joint distribution that is specified, and the full conditionals are derived from it, the joint distribution is guaranteed to exist.

6.2.3 Prior Distributions

Prior distributions have to be specified for all (hyper)parameters. A common prior choice for the regression coefficients is the normal distribution with mean zero and large variance. In **JointAI** variance parameters in models for normally distributed variables are specified as inverse-gamma distributions.

The covariance matrix of the random effects in a mixed model, $\mathbf{D}$, is assumed to follow an inverse-Wishart distribution where the degrees of freedom are chosen to be equal to the dimension of the random effects and the scale matrix is diagonal. Since the magnitude of the diagonal elements relates to the variance of the random effects, the choice of suitable values depends on the scale of the variable the random effect is associated with. Therefore, **JointAI** uses independent gamma hyperpriors for each of the diagonal elements. More details about the default hyperparameters and how to change them are given in Section 6.5.8 and Appendix 6.A.

6.3 Package Structure

The package **JointAI** has seven main functions, `lm_imp()`, `glm_imp()`, `clm_imp()`, `lme_imp()`, `glme_imp()`, `clmm_imp()`, `survreg_imp()` and `coxph_imp()`, that perform regression of continuous and categorical, univariate or multi-level data as well as right censored survival data. Model

specification is similar to that of standard regression models in R and is described in detail in Section 6.5.

Based on the specified model formula and other arguments that are provided by the user, **JointAI** does some pre-processing of the data. It checks which variables are incomplete and/or time-varying, and identifies their measurement level in order to specify appropriate (imputation) models. Interactions and functional forms of variables are detected in the model formula and corresponding design matrices for the various parts of the model are created.

MCMC sampling is performed by the program JAGS (Plummer 2003). The JAGS model, data list, containing all necessary parts of the data, and user-specified settings for the MCMC sampling (further described in Section 6.6) are passed to JAGS via the R package **rjags** (Plummer 2018).

The above named main functions, from here on abbreviated as `*_imp()`, all return an object of class `JointAI`. Summary and plotting methods for `JointAI` objects, as well as functions to evaluate convergence and precision of the MCMC samples, to predict from `JointAI` objects and to export imputed values are discussed in Section 6.7.

Currently, the package works under the assumption of a Missing At Random (MAR) missingness process (Rubin 1976; Rubin 1987). When this assumption holds, it is valid to exclude cases with missing values in the outcome in Bayesian inference. Hence, our focus here is on missing covariate values. Nevertheless, **JointAI** can handle missing values in the outcome; they are implicitly imputed using the specified analysis model.

6.4 Example Data

To illustrate the functionality of **JointAI** we use three datasets that are part of this package or the package **survival** (Therneau 2015; Terry M. Therneau and Patricia M. Grambsch 2000). The first dataset, the `NHANES` data, contains data from a cross-sectional cohort study, whereas the second dataset (`simLong`) is a simulated dataset based on a longitudinal cohort study in toddlers. The third dataset (`lung`) contains data on survival of patients with advanced lung cancer.

6.4.1 The NHANES Data

The `NHANES` data is a subset of observations from the 2011 – 2012 wave of the National Health and Nutrition Examination Survey (National Center for Health

Statistics (NCHS) 2011) and contains information on 186 men and women between 20 and 80 years of age. The variables contained in this dataset are

- `SBP`: systolic blood pressure in mmHg; complete
- `gender`: `male` vs `female`; complete
- `age`: in years; complete
- `race`: 5 unordered categories; complete
- `WC`: waist circumference in cm; 1.1% missing
- `alc`: alcohol consumption; `<1` drink per week vs `>= 1` drink per week; 18.3% missing
- `educ`: educational level; `low` vs `high`; complete
- `creat`: creatinine concentration in mg/dL; 4.5% missing
- `albu`: albumin concentration in g/dL; 4.3% missing
- `uricacid`: uric acid concentration in mg/dL; 4.3% missing
- `bili`: bilirubin concentration in mg/dL; 4.3% missing
- `occup`: occupational status; 3 unordered categories; 15.1% missing
- `smoke`: smoking status; 3 ordered categories; 3.8% missing

Figure 6.1 shows the distributions of all variables in the `NHANES` data, together with the proportion of missing values for incomplete variables, and can be obtained with the function `plot_all()`. Arguments `fill` and `border` allow colours to change, the number of rows and columns can be adapted using `nrow` and/or `ncol`, and additional arguments can be passed to `hist()` and `barplot()` via `"..."`.

The pattern of missing values in the `NHANES` data is shown in Figure 6.2. Each row represents a pattern of missing values, where observed (missing) values are depicted with dark (light) colour. The frequency with which each of the patterns is observed is given on the right margin, the number of missing values in each variable is given underneath the plot. Rows and columns are ordered by the number of cases per pattern (decreasing) and the number of missing values (increasing). The first row, for instance, shows that there are 116 complete cases, the second row that there are 29 cases for which only `alc` is missing. Furthermore, it is apparent that `creat`, `albu`, `uricacid` and `bili` are always missing together. Since these variables are all measured in serum this is not surprising.

The plot of the missing data pattern can be obtained with `md_pattern()`. Again, arguments `color` and `border` allow us to change colours, and arguments such as `legend.position`, `print_xaxis` and `print_yaxis` permit further customization.

A matrix representation of the missing data pattern can be obtained by setting `pattern = TRUE`. There, missing and observed values are represented with a `"0"` and `"1"` respectively.

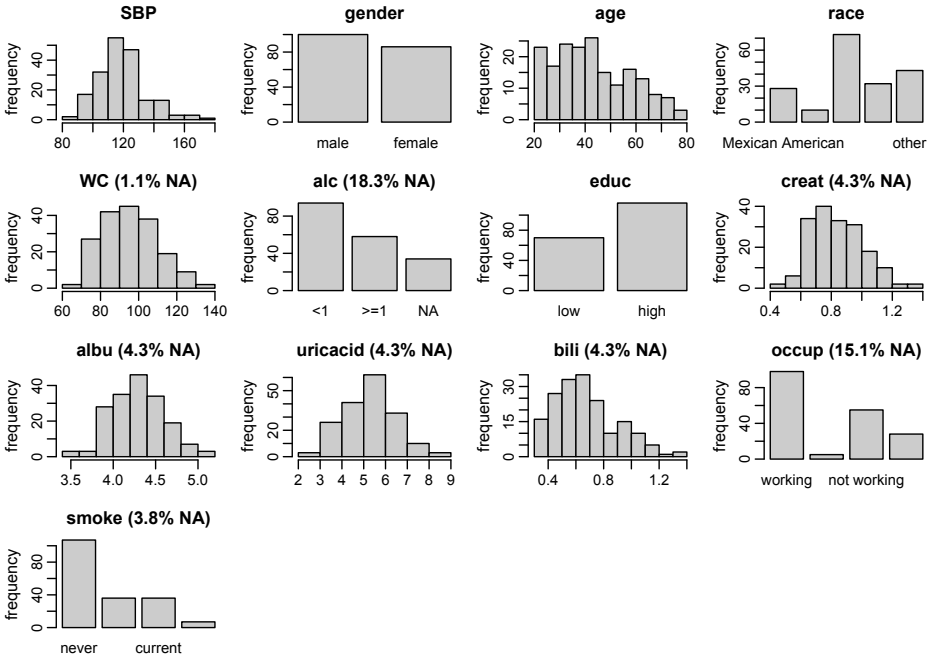


Figure 6.1: Distribution of the variables in the `NHANES` data (with the percentage of missing values given for incomplete variables).

6.4.2 The simLong Data

The `simLong` data is a simulated dataset mimicking a longitudinal cohort study of 200 mother-child pairs. It contains the following baseline (i.e., not time-varying) covariates

- `GESTBIR`: gestational age at birth in weeks; complete
- `ETHN`: ethnicity; `European` vs `other`; 2.8% missing
- `AGE_M`: age of the mother at intake; complete
- `HEIGHT_M`: height of the mother in cm; 2.0% missing
- `PARITY`: number of times the mother has given birth; `0` vs `>=1`; 2.4% missing
- `SMOKE`: smoking status of the mother during pregnancy; 3 ordered categories; 12.2% missing
- `EDUC`: educational level of the mother; 3 ordered categories; 7.8% missing
- `MARITAL`: marital status; 3 unordered categories; 7.0% missing
- `ID`: subject identifier

and seven longitudinal variables:

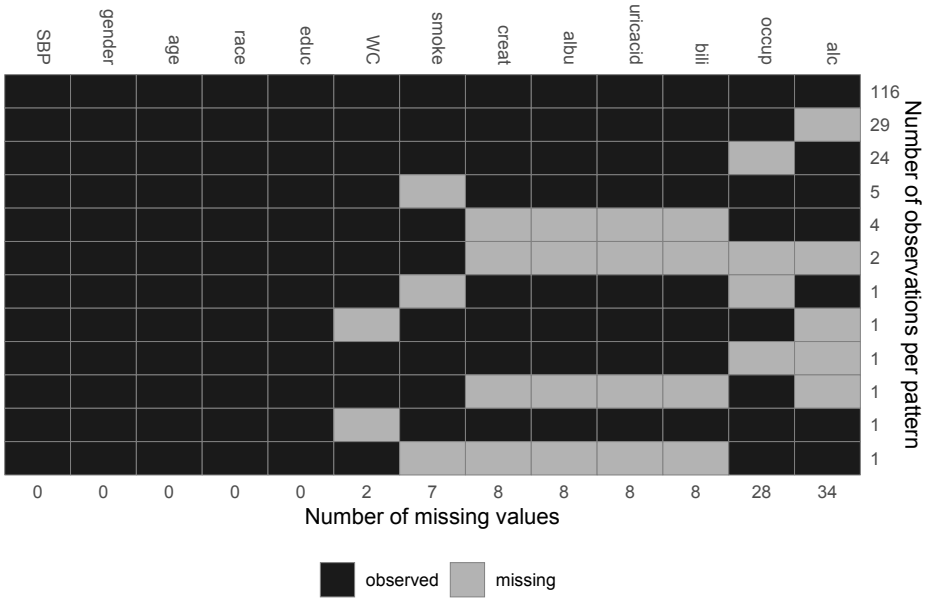


Figure 6.2: Missing data pattern of the `NHANES` data.

- `time`: measurement occasion/visit (by design children should have been measured at 1, 2, 3, 4, 7, 11, 15, 20, 26, 32, 40 and 50 months of age); complete
- `age`: child's age at measurement time in months
- `hgt`: child's height in cm; 20.0% missing
- `wgt`: child's weight in gram; 8.8% missing
- `bmi`: child's BMI (body mass index) in kg/m^2; 21.6% missing
- `hc`: child's head circumference in cm; 23.6% missing
- `sleep`: child's sleeping behaviour; 3 ordered categories; 24.7% missing

Figure 6.3 shows the longitudinal profiles of `hgt`, `wgt`, `bmi` and `hc` over age. All four variables have a non-linear pattern over time. Distributions of all variables in the `simLong` data are displayed in Figure 6.4. Here, arguments `use_level` and `idvar` of the function `plot_all()` are used to display baseline (level-2) covariates on the subject level instead of the observation level:

```
> plot_all(simLong, use_level = TRUE, idvar = "ID", ncol = 4)
```

The missing data pattern of the `simLong` data is shown in Figure 6.5. For readability, it is plotted separately for baseline (left) and longitudinal (right) covariates.

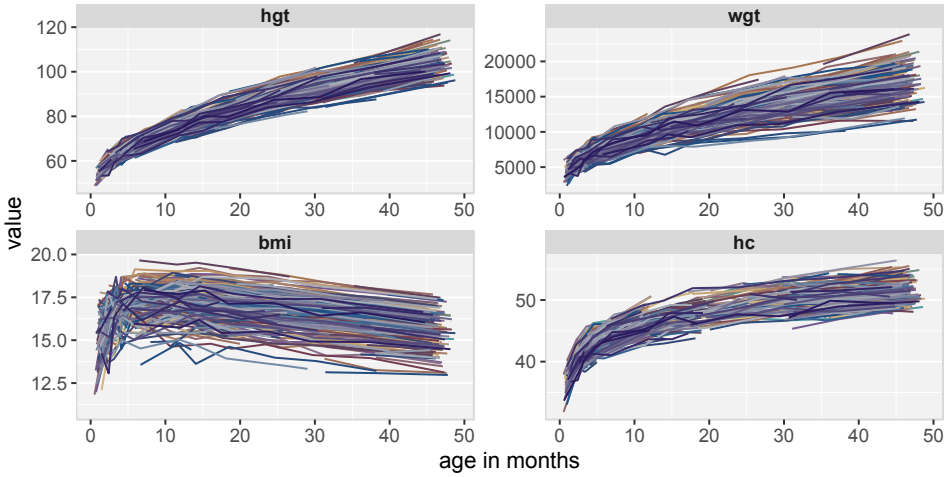


Figure 6.3: Trajectories of height, weight, BMI and head circumference in the `simLong` data.

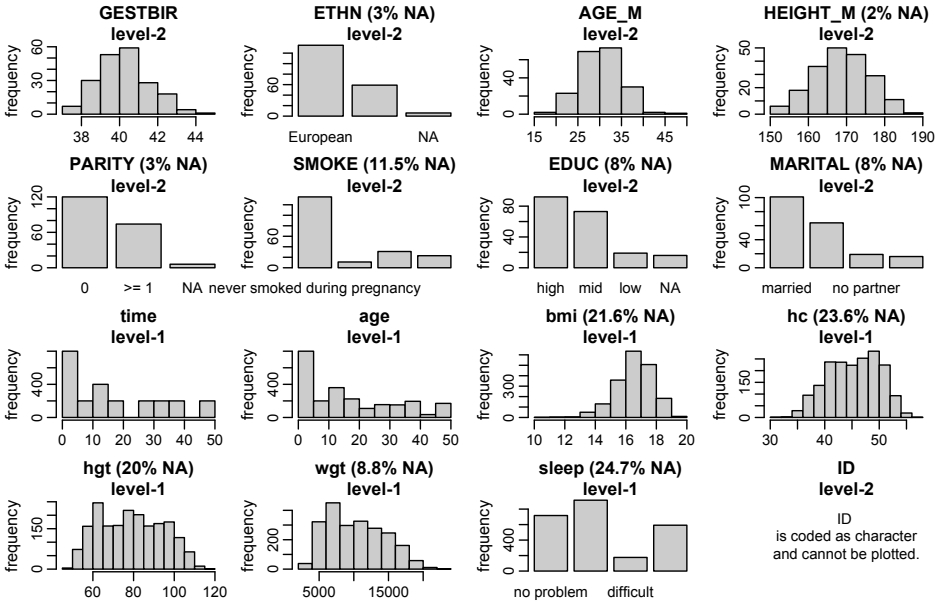


Figure 6.4: Distribution of the variables in the `simLong` data (with percentage of missing values given for incomplete variables).

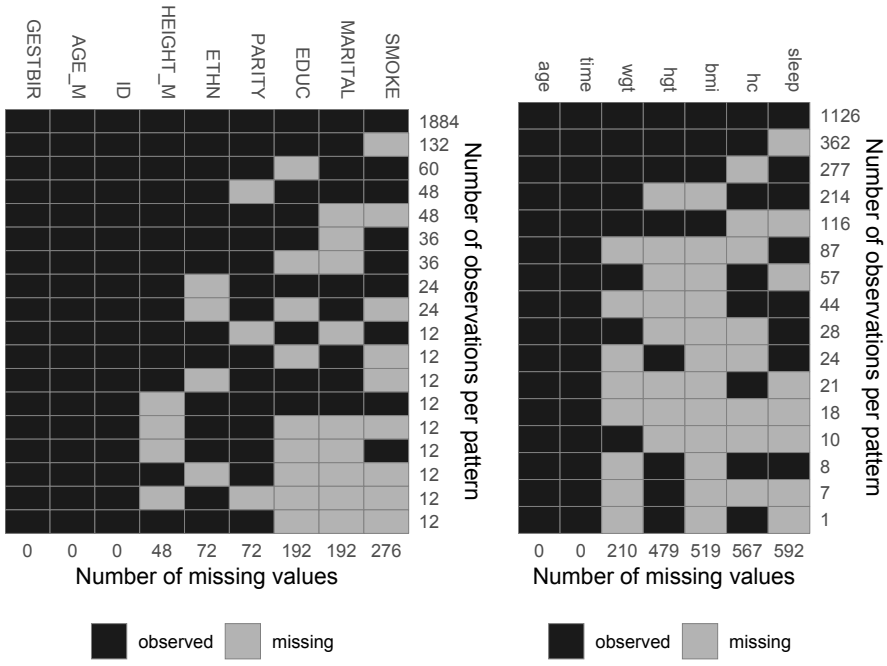


Figure 6.5: Missing data pattern of the `simLong` data.

6.4.3 The lung Data

For demonstration of the use of **JointAI** for the analysis of survival data, we use the dataset `lung` that is included in the R package **survival**. It contains data of 228 patients with advanced lung cancer and includes the following variables:

- `inst`: institution code; complete
- `time`: survival time in days; complete
- `status`: event indicator (`1`: censored, `2`: dead); complete
- `age`: patient's age in years; complete
- `sex`: male (`1`) vs female (`2`); complete
- `ph.ecog`: ECOG performance score (describes how the disease impacts the patient's daily life); scale from `0` (no impact) to `5` (dead); 0.4% missing
- `ph.karno`: Karnofsky performance score as rated by physician (describes the degree of a patient's impairment by the disease); scale from `0` (dead) to `100` (no impairment); 0.4% missing

- `pat.karno`: Karnofsky performance score as rated by patient; 1.3% missing
- `meal.cal`: kilocalories consumed at meals; 20.6% missing
- `wt.loss`: weight loss over the last six months in kg; 6.1% missing

The distribution of the observed values and the missing data pattern of the `lung` data are shown in Figures 6.6 and 6.7.

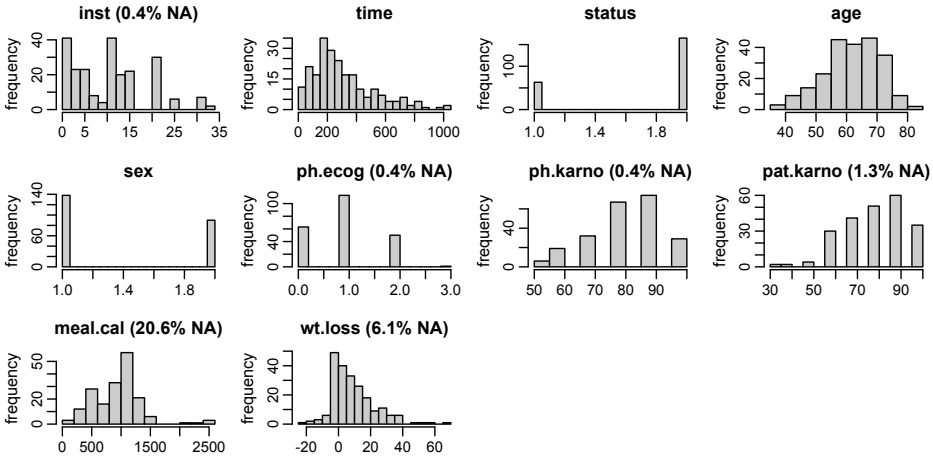


Figure 6.6: Distribution of the variables in the `lung` data (with percentage of missing values given for incomplete variables).

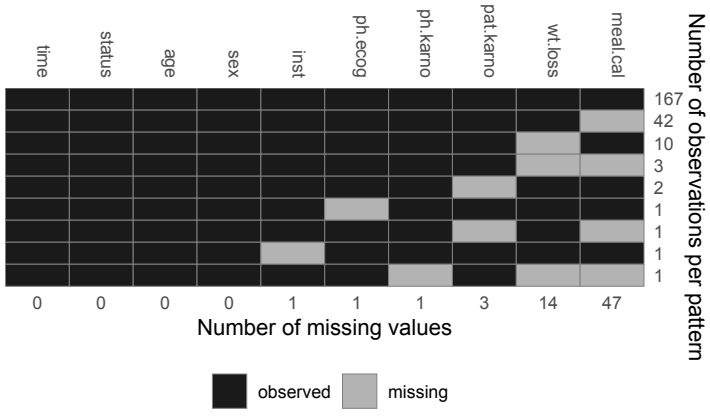


Figure 6.7: Missing data pattern of the `lung` data.

6.5 Model Specification

The main arguments of the functions

```
> lm_imp(formula, data,
+        n.chains = 3, n.adapt = 100, n.iter = 0, thin = 1, ...)
>
> glm_imp(formula, family, data,
+         n.chains = 3, n.adapt = 100, n.iter = 0, thin = 1, ...)
>
> clm_imp(formula, data,
+         n.chains = 3, n.adapt = 100, n.iter = 0, thin = 1, ...)
>
> lme_imp(fixed, data, random,
+         n.chains = 3, n.adapt = 100, n.iter = 0, thin = 1, ...)
>
> glme_imp(fixed, data, random, family,
+          n.chains = 3, n.adapt = 100, n.iter = 0, thin = 1, ...)
>
> clmm_imp(fixed, data, random,
+          n.chains = 3, n.adapt = 100, n.iter = 0, thin = 1, ...)
>
> survreg_imp(formula, data,
+             n.chains = 3, n.adapt = 100, n.iter = 0, thin = 1, ...)
>
> coxph_imp(formula, data,
+           n.chains = 3, n.adapt = 100, n.iter = 0, thin = 1, ...)
```

i.e., `formula`, `data`, `family`, `fixed`, and `random`, are used analogously to the specification in the standard complete data functions `lm()` and `glm()` from package **stats**, `lme()` from package **nlme** (Pinheiro et al. 2018), and `survreg()` and `coxph()` from package **survival**.

For the description of the remaining arguments see Section 6.6.

The arguments `formula` (in `lm_imp()`, `glm_imp()` and `clm_imp()`) and `fixed` (in `lme_imp()`, `glme_imp()` and `clmm_imp()`) take a standard two-sided `formula` object, where an intercept is added automatically. For the use of the argument `random`, see Section 6.5.3.

Survival models expect the left hand side of `formula` to be a survival object (created with the function `Surv()` from package **survival**; see Section 6.5.4).

The argument `family` enables the choice a distribution and link function from a range of options when using `glm_imp()` or `glme_imp()`. The implemented options are given in Table 6.1.

Table 6.1: Possible choices for the model `family` and `link` functions in `glm_imp()` and `glme_imp()`.

distribution	link
gaussian	identity, log, inverse
binomial	logit, probit, log, cloglog
Gamma	inverse, identity, log
poisson	log, identity

6.5.1 Interactions

In **JointAI**, interactions between any type of variables (observed, incomplete, time-varying) can be handled. When an incomplete variable is involved, the interaction term is re-calculated within each iteration of the MCMC sampling, using the imputed values from the current iteration. Interaction terms involving incomplete variables should hence not be pre-calculated as an additional variable since this would lead to inconsistent imputed values of main effect and interaction term.

Interactions between multiple variables can be specified using parentheses; for higher lever interactions the `"^"` operator can be used:

```
> mod1a <- glm_imp(educ ~ gender * (age + smoke + creat),
+                  data = NHANES, family = binomial())
>
> mod1b <- glm_imp(educ ~ gender + (age + smoke + creat)^3,
+                  data = NHANES, family = binomial())
```

6.5.2 Non-linear Functional Forms

In practice, associations between outcome and covariates do not always meet the standard assumption of linearity. Often, assuming a logarithmic, quadratic or other non-linear effect is more appropriate.

For completely observed covariates, **JointAI** can handle any common type of function implemented in R, including splines, e.g., using `ns()` or `bs()` from the package **splines**. Since functions involving variables that have missing values need to be re-calculated in each iteration of the MCMC sampling, currently, only

functions that are available in JAGS can be used for incomplete variables. Those functions include:

- `log()`, `exp()`
- `sqrt()`, polynomials (using `I()`)
- `abs()`
- `sin()`, `cos()`
- algebraic operations involving one or multiple (in)complete variables, as long as the formula can be interpreted by JAGS.

The list of functions implemented in JAGS can be found in the JAGS user manual available at https://sourceforge.net/projects/mcmc-jags/files/Manuals/.

Some examples (that do not necessarily have a meaningful interpretation or good model fit) are:

```
> mod2a <- lm_imp(SBP ~ age + gender + abs(bili - creat), data = NHANES)
```

```
> library("splines")
> mod2b <- lm_imp(SBP ~ ns(age, df = 2) + gender + I(bili^2) + I(bili^3),
+                 data = NHANES)
```

```
> mod2c <- lm_imp(SBP ~ age + gender + I(creat/albu^2), data = NHANES,
+                 trunc = list(albu = c(1e-5, 1e5)))
> # (for explantion of the argument trunc see section below)
```

```
> mod2d <- lm_imp(SBP ~ bili + sin(creat) + cos(albu), data = NHANES)
```

It is also possible to nest a function in another function.

```
> mod2e <- lm_imp(SBP ~ age + gender + sqrt(exp(creat)/2), data = NHANES)
```

Functions with Restricted Support

When a function of an incomplete variable has restricted support, e.g., $\log(x)$ is only defined for $x > 0$, or, as in `mod2c` from above, `I(creat/albu^2)` can not be calculated for `albu = 0`, the distribution used to impute that variable needs to comply with these restrictions. This can either be achieved by truncating the distribution, using the argument `trunc`, or by selecting a distribution that meets the restrictions.

Example: When using a log transformation for the covariate `bili`, we can use the default imputation method `"norm"` (a normal distribution) and truncate it by specifying `trunc = list(bili = c(lower, upper))`, where `lower` and `upper` are the smallest and largest values allowed:

```
> mod3a <- lm_imp(SBP ~ age + gender + log(bili) + exp(creat),
+                 trunc = list(bili = c(1e-5, 1e10)), data = NHANES)
```

Truncation always requires the specification of both limits. Since $-$Inf and Inf are not accepted, a value far enough outside the range of the variable (here: $1e10$) can be selected for the second boundary of the truncation interval.

Alternatively, we may choose a model for the incomplete variable (using the argument `models`; for more details see Section 6.5.5) that only imputes positive values such as a log-normal distribution (`"lognorm"`) or a gamma distribution (`"gamma"`):

```
> mod3b <- lm_imp(SBP ~ age + gender + log(bili) + exp(creat),
+                 models = c(bili = 'lognorm'), data = NHANES)
>
> mod3c <- lm_imp(SBP ~ age + gender + log(bili) + exp(creat),
+                 models = c(bili = 'gamma'), data = NHANES)
```

Functions Not Available in R

It is possible to use functions that have different names in R and JAGS, or that do exist in JAGS, but not in R, by defining a new function in R that has the name of the function in JAGS.

Example: In JAGS the inverse logit transformation is defined in the function `ilogit`. In base R, there is no function `ilogit`, but the inverse logit is available as the distribution function of the logistic distribution `plogis()`. Thus, we can define the function `ilogit()` as

```
> ilogit <- plogis
```

and use it in the model formula

```
> mod4a <- lm_imp(SBP ~ age + gender + ilogit(creat), data = NHANES)
```

A Note on What Happens Inside JointAI

When a function of a complete or incomplete variable is used in the model formula, the main effect of that variable is automatically added as an auxiliary variable (more on auxiliary variables in Section 6.5.6), and only the main effects are used as predictors in the imputation models.

In `mod2b`, for example, the spline of `age` is used as predictor for `SBP`, but in the imputation model for `bili`, `age` enters with a linear effect. This can be checked using the function `list_models()`, which prints a list of the covariate models used in a **JointAI** model. Here, we are only interested in the predictor variables, and, hence, suppress printing of information on prior distributions, regression coefficients and other parameters by setting `priors`, `regcoef` and `otherpars` to `FALSE`:

```
> list_models(mod2b, priors = FALSE, regcoef = FALSE, otherpars = FALSE)
```

```
## Normal imputation model for 'bili'
## * Predictor variables:
##   (Intercept), genderfemale, age
```

When a function of a variable is specified as an auxiliary variable, this function is used in the imputation models. For example, in `mod4b`, waist circumference (`WC`) is not part of the model for `SBP`, and the quadratic term `I(WC^2)` is used in the linear predictor of the imputation model for `bili`:

```
> mod4b <- lm_imp(SBP ~ age + gender + bili, auxvars = "I(WC^2)",
+                 data = NHANES)
>
> list_models(mod4b, priors = FALSE, regcoef = FALSE, otherpars = FALSE)
```

```
## Normal imputation model for 'WC'
## * Predictor variables:
##   (Intercept), age, genderfemale
##
## Normal imputation model for 'bili'
## * Predictor variables:
##   (Intercept), age, genderfemale, I(WC^2)
```

Incomplete variables are always imputed on their original scale, i.e., in `mod2b` the variable `bili` is imputed and the quadratic and cubic versions are calculated

from the imputed values. Likewise, `creat` and `albu` in `mod2c` are imputed separately, and `I(creat/albu^2)` is calculated from the imputed (and observed) values. To ensure consistency between variables, functions involving incomplete variables should be specified as part of the model formula and not be pre-calculated as separate variables.

6.5.3 Multi-level Structure and Longitudinal Covariates

In multi-level models, additional to the fixed effects structure specified by the argument `fixed`, a random effects structure needs to be provided via the argument `random`.

Analogous to the specification of the argument `random` in `lme()`, it takes a one-sided formula starting with a `"~"`, and the grouping variable is separated by `"|"`. A random intercept is added automatically and only needs to be specified in a random intercept only model.

A few examples:

- `random = ~1 | id`: random intercept only, with `id` as grouping variable
- `random = ~time | id`: random intercept and slope for variable `time`
- `random = ~time + I(time^2) | id`: random intercept, slope and quadratic random effect for `time`
- `random = ~time + x | id` random intercept, random slope for `time` and random effect for variable `x`

It is possible to use splines in the random effects structure if there are no missing values in the variables involved, e.g.:

```
> mod5 <- lme_imp(bmi ~ GESTBIR + ETHN + HEIGHT_M + ns(age, df = 2),
+                 random = ~ ns(age, df = 2) | ID, data = simLong)
```

Longitudinal ("time-varying"; level-1) covariates can be used in the fixed or random effects and will be imputed if they contain any missing values. Currently only one level of nesting is possible.

6.5.4 Survival Models

JointAI provides two functions to analyse survival data with incomplete covariates: `survreg_imp()` and `coxph_imp()`. Analogous to the complete data versions of these functions from the package **survival**, the left-hand side of the model formula has to be a survival object specified using the function `Surv()`.

Example: To analyse the `lung` data, we can either use a parametric Weibull model or a Cox proportional hazards model. The **survival** package needs to be loaded for the function `Surv()`.

```
> library(survival)
> mod6a <- survreg_imp(Surv(time, status) ~ age + sex + ph.karno +
+                        meal.cal + wt.loss, data = lung, n.iter = 250)
>
> mod6b <- coxph_imp(Surv(time, status) ~ age + sex + ph.karno + meal.cal +
+                      wt.loss, data = lung, n.iter = 250)
```

Currently only right censored survival data and time-constant covariates can be handled and it is not yet possible to take into account strata, clustering or frailty terms.

6.5.5 Imputation / Covariate Model Types

JointAI automatically selects an (imputation) model for each of the incomplete baseline (level-2) or longitudinal (level-1) covariates, based on the `class` of the variable and the number of levels. The automatically selected types for baseline covariates are:

- `norm`: linear model (default for continuous variables)
- `logit`: binary logistic model (default for factors with two levels)
- `multilogit`: multinomial logit model (default for unordered factors with > 2 levels)
- `cumlogit`: cumulative logit model (default for ordered factors with > 2 levels)

The default methods for longitudinal covariates are:

- `lmm`: linear mixed model (default for continuous longitudinal variables)
- `glmm_logit`: logistic mixed model (default for longitudinal factors with two levels)
- `clmm`: cumulative logit mixed model (default for longitudinal ordered factors with >2 levels)

When a continuous variable has only two different values, it is automatically converted to a factor and imputed using a logistic model, unless an imputation model type is specified by the user. Variables of type `logical` are automatically converted to binary factors.

The imputation models that are chosen by default may not necessarily be appropriate for the data at hand, especially for continuous variables, which often do not comply with the assumptions of (conditional) normality.

Therefore, the following alternative imputation methods are available for continuous baseline covariates:

- `lognorm`: normal model for the log-transformed variable (right-skewed variables > 0)
- `gamma`: gamma model with log-link (right-skewed variables > 0)
- `beta`: beta model with logit-link (continuous variables in $[0, 1]$)

`lognorm` assumes a (conditional) normal distribution for the natural logarithm of the variable, but the variable enters the linear predictor of the analysis model (and possibly other imputation models) on its original scale.

For longitudinal covariates the following alternative model types are available:

- `glmm_gamma`: gamma mixed model with log-link (right-skewed variables > 0)
- `glmm_poisson`: Poisson mixed model with log-link (count variables)

Specification of Imputation/Covariate Model Types

In models `mod3b` and `mod3c` in Section 6.5.2 we have already seen two examples in which the imputation model type was changed using the argument `models`, which takes a named vector. When the vector supplied to `models` only contains specifications for a subset of the incomplete and longitudinal covariates, default models are used for the remaining covariates. As explained in Section 6.2.2, models for completely observed longitudinal covariates only need to be specified when any baseline covariates have missing values.

```
> mod7a <- lm_imp(SBP ~ age + gender + WC + alc + bili + occup + smoke,
+                 models = c(WC = 'gamma', bili = 'lognorm'),
+                 data = NHANES, n.iter = 100)
>
> mod7a$models
```

```
##           WC        smoke         bili        occup          alc
##      "gamma"   "cumlogit"    "lognorm" "multilogit"      "logit"
```

The function `get_models()`, which finds and assigns the default imputation methods, can be called directly. `get_models()` has arguments

- `fixed`: the fixed effects formula

- `random`: the random effects formula (if necessary)
- `data`: the dataset
- `auxvars`: an optional vector of auxiliary variables
- `no_model`: an optional vector of names of covariates for which no model will be specified

```
> mod7b_models <- get_models(bmi ~ GESTBIR + ETHN + HEIGHT_M + SMOKE + hc +
+                              MARITAL + ns(age, df = 2),
+                            random = ~ ns(age, df = 2) | ID,
+                            data = simLong)
> mod7b_models
```

```
## $models
##     HEIGHT_M          ETHN      MARITAL        SMOKE          age
##       "norm"      "logit" "multilogit"   "cumlogit"        "lmm"
##           hc
##        "lmm"
##
## $meth
##     HEIGHT_M          ETHN      MARITAL        SMOKE           hc
##       "norm"      "logit" "multilogit"   "cumlogit"        "lmm"
```

`get_models()` returns a list of two vectors:

- `models`: containing all specified models
- `meth`: containing the models for the incomplete variables only

When there is a "time" variable in the model, such as `age` in our example (which is the age of the child at the time of the measurement), it may not be meaningful to specify a model for that variable. Especially when the "time" variable is pre-specified by the design of the study, it can usually be assumed to be independent of baseline covariates and a model for it has no useful interpretation.

The argument `no_model` (in `get_models()` and in the main functions `*_imp()`) allows for the exclusion of models for such variables (as long as they are completely observed):

```
> mod7c_models <- get_models(bmi ~ GESTBIR + ETHN + HEIGHT_M + SMOKE + hc +
+                              MARITAL + ns(age, df = 2),
+                            random = ~ ns(age, df = 2) | ID,
+                            data = simLong, no_model = "age")
> # mod7c_models
```

By excluding the model for `age` we implicitly assume that incomplete baseline variables are independent of `age`.

Order of the Sequence of Imputation Models

By default, models for level-1 covariates are specified after models for level-2 covariates, so that the level-2 covariates are used as predictor variables in the models for level-1 covariates but not vice versa. Within the two groups, models are ordered by the number of missing values (decreasing), so that the model for the variable with the most missing values has the most variables in its linear predictor.

6.5.6 Auxiliary Variables

Auxiliary variables are variables that are not part of the analysis model but should be considered as predictor variables in the imputation models because they can inform the imputation of unobserved values.

Good auxiliary variables are (Van Buuren 2012)

- associated with an incomplete variable of interest, or are
- associated with the missingness of that variable and
- do not have too many missing values themselves. Importantly, they should be observed for a large proportion of the cases that have a missing value in the variable to be imputed.

In `*_imp()`, auxiliary variables can be specified with the argument `auxvars`, which is a vector containing the names of the auxiliary variables.

Example: We might consider the variables `educ` and `smoke` as predictors for the imputation of `occup`:

```
> mod8a <- lm_imp(SBP ~ gender + age + occup, auxvars = c('educ', 'smoke'),
+                 data = NHANES, n.iter = 100)
```

The variables `educ` and `smoke` are not included in the analysis model (as can be seen when printing the posterior mean of the regression coefficients of the analysis model with `coef()`):

```
> coef(mod8a)
```

```
##           (Intercept)          genderfemale                   age
##           105.7310654            -5.7207836           0.3751877
## occuplooking for work      occupnot working
##             3.3846500            -0.7601980
```

They are, however, used as predictors in the imputation for `occup` and imputed themselves (if they have missing values):

```
> list_models(mod8a, priors = FALSE, regcoef = FALSE, otherpars = FALSE,
+             refcat = FALSE)
```

```
## Cumulative logit imputation model for 'smoke'
## * Predictor variables:
##   genderfemale, age, educhigh
##
## Multinomial logit imputation model for 'occup'
## * Predictor variables:
##   (Intercept), genderfemale, age, educhigh, smokeformer, smokecurrent
```

Functions of Variables as Auxiliary Variables

As shown above in `mod3e`, it is possible to specify functions of auxiliary variables. In that case, the auxiliary variable is not considered as a linear effect but as specified by the function.

Note that omitting auxiliary variables from the analysis model implies that the outcome is independent of these variables, conditional on the other variables in the model. If this is not true, the model is misspecified which may lead to biased results (similar to leaving a confounding variable out of a model).

6.5.7 Reference Values for Categorical Covariates

In **JointAI**, dummy coding is used when categorical variables enter a linear predictor, i.e., a binary variable is created for each category, except the reference category. These binary variables have value one when that category was observed and zero otherwise. Contrary to the behaviour in base R, this coding is also used for ordered factors.

By default, the first category of a categorical variable (ordered or unordered) is used as reference, however, this may not always allow the desired interpretation of the regression coefficients. Moreover, when categories are unbalanced, setting the largest group as reference may result in better mixing of the MCMC chains. Therefore, **JointAI** allows specification of the reference category separately for each variable, via the argument `refcats`. Changes in `refcats` will not impact the imputation of the respective variable, but change categories for which dummy variables are included in the linear predictor of the analysis model or other covariate models.

Setting Reference Categories for All Variables

To specify the choice of reference category globally for all variables in the model, `refcats` can be set as

- `refcats = "first"`
- `refcats = "last"`
- `refcats = "largest"`

For example:

```
> mod9a <- lm_imp(SBP ~ gender + age + race + educ + occup + smoke,
+                 refcats = "largest", data = NHANES)
```

Setting Reference Categories for Individual Variables

Alternatively, `refcats` takes a named vector, in which the reference category for each variable can be specified either by its number or its name, or one of the three global types: `"first"`, `"last"` or `"largest"`. For variables for which no reference category is specified in the list the default is used.

```
> mod9b <- lm_imp(SBP ~ gender + age + race + educ + occup + smoke,
+                 refcats = list(occup = "not working", race = 3,
+                                educ = "largest"), data = NHANES)
```

To facilitate specification of the reference categories, the function `set_refcat()` can be used. It prints the names of the categorical variables that are selected by

- a specified model formula and/or
- a vector of auxiliary variables, or
- a vector naming covariates,

or all categorical variables in the data if only the argument `data` is provided, and asks a number of questions which the user needs to reply to via input of a number.

```
> refs_mod9 <- set_refcat(NHANES, formula = formula(mod9b))

##
## How do you want to specify the reference categories?
##
## 1: Use the first category for each variable.
## 2: Use the last category for each variabe.
## 3: Use the largest category for each variable.
## 4: Specify the reference categories individually.
```

When option 4 is chosen, a choice is given for each categorical variable, for example:

```
> #> The reference category for "race" should be
> #>
> #> 1: Mexican American
> #> 2: Other Hispanic
> #> 3: Non-Hispanic White
> #> 4: Non-Hispanic Black
> #> 5: other
```

After specification of the reference categories, the determined specification for the argument `refcats` is printed:

```
> #> In the JointAI model specify:
> #>  refcats = c(gender = 'female', race = 'Non-Hispanic White',
> #>               educ = 'low', occup = 'not working', smoke = 'never')
> #>
> #> or use the output of this function.
```

`set_refcat()` also returns a named vector that can be passed to the argument `refcats`:

```
> mod9c <- lm_imp(SBP ~ gender + age + race + educ + occup + smoke,
+                 refcats = refs_mod9, data = NHANES)
```

6.5.8 Hyperparameters

In the Bayesian framework, parameters are random variables for which a distribution needs to be specified. These distributions depend on parameters themselves, i.e., on hyperparameters.

The function `default_hyperpars()` returns a list containing the default hyperparameters used in a `JointAI` model (see Appendix 6.A).

To change the hyperparameters in a `JointAI` model, the list obtained from `default_hyperpars()` can be edited and passed to the argument `hyperpars` in `*_imp()`.

`mu_reg_*` and `tau_reg_*` refer to the mean and precision of the prior distribution for regression coefficients. `shape_tau_*` and `rate_tau_*` are the shape and rate parameters of a gamma distribution that is used as prior for precision parameters. `RinvD` is the scale matrix in the Wishart prior for the inverse of the random

effects design matrix `D`, and `KinvD` is the number of degrees of freedom in that distribution. `shape_diag_RinvD` and `rate_diag_RinvD` are the shape and rate parameters of the gamma prior of the diagonal elements of `RinvD`. In random effects models with only one random effect, a gamma prior is used instead of the Wishart distribution for the inverse of `D`.

The hyperparameters `mu_reg_surv` and `tau_reg_surv` are used in `survreg_imp()` as well as `coxph_imp()`. For the Cox proportional hazards model, the hyperparameters `c`, `r` and `eps` refer to the confidence in the prior guess for the hazard function, failure rate per unit time ($\lambda_0(t)^*$ in Section 6.2.1) and time increment (for numerical stability), respectively.

6.5.9 Scaling

When variables are measured on very different scales this can result in slow convergence and bad mixing. Therefore, **JointAI** automatically scales continuous variables to have mean zero and standard deviation one. Results (parameters and imputed values) are transformed back to the original scale when the results are printed or imputed values are exported.

In some settings, however, it is not possible to scale continuous variables. This is the case for incomplete variables that enter a linear predictor in a function and variables that are imputed with models that are defined on a subset of the real line (i.e., with a gamma or a beta distribution). Variables that are imputed with a log-normal distribution are scaled, but not centred.

To prevent scaling, the argument `scale_vars` can be set to `FALSE`. When a vector of variable names is supplied to `scale_vars`, only those variables are scaled.

By default, only the MCMC samples that are scaled back to the scale of the data are stored in a `JointAI` object. When the argument `keep_scaled_mcmc = TRUE`, the scaled sample is also kept.

6.5.10 Ridge Regression

Using the argument `ridge = TRUE` it is possible to impose a ridge penalty on the parameters of the analysis model, shrinking these parameters towards zero. This is done by specification of a Gamma$(0.01, 0.01)$ prior for the precision of the regression coefficients β instead of setting it to a fixed (small) value.

6.6 MCMC Settings

The functions `*_imp()` have a number of arguments that specify settings for the MCMC sampling:

- `n.chains`: number of MCMC chains
- `n.adapt`: number of iterations in the adaptive phase
- `n.iter`: number of iterations in the sampling phase
- `thin`: thinning degree
- `monitor_params`: parameters/nodes to be monitored
- `seed`: optional seed value for reproducibility
- `inits`: initial values
- `parallel`: whether MCMC chains should be sampled in parallel
- `ncores`: how many cores are available for parallel sampling

The first four, as well as the following two parameters, are passed directly to functions from the R package **rjags**:

- `quiet`: should printing of information be suppressed?
- `progress.bar`: type of progress bar (`"text"`, `"gui"` or `"none"`)

In the following sections, the arguments listed above are explained in more detail and examples are given.

6.6.1 Number of Chains, Iterations and Samples

Number of Chains

To evaluate convergence of MCMC chains it is helpful to create multiple chains that have different starting values. More information on how to evaluate convergence and the specification of initial values can be found in Sections 6.7.3 and 6.6.3 respectively.

The argument `n.chains` selects the number of chains (by default, `n.chains = 3`). For calculating the model summary, multiple chains are merged.

Adaptive Phase

JAGS has an adaptive mode, in which samplers are optimized (for example the step size is adjusted). Samples obtained during the adaptive mode do not form a Markov chain and are discarded. The argument `n.adapt` controls the length of this adaptive phase.

The default value for `n.adapt` is 100, which works well in many of the examples considered here. Complex models may require longer adaptive phases. If the adaptive phase is not sufficient for JAGS to optimize the samplers, a warning message will be printed (see example below).

Sampling Iterations

`n.iter` specifies the number of iterations in the sampling phase, i.e., the length of the MCMC chain. How many samples are required to reach convergence and to have sufficient precision (see also Section 6.7.3) depends on the complexity of data and model, and may range from as few as 100 to several million.

Thinning

In settings with high autocorrelation, it may take many iterations before a sample is created that sufficiently represents the whole range of the posterior distribution. Processing of such long chains can be slow and cause memory issues. The parameter `thin` allows the user to specify if and how much the MCMC chains should be thinned out before storing them. By default `thin = 1` is used, which corresponds to keeping all values. A value `thin = 10` would result in keeping every 10th value and discarding all other values.

Example: Default Settings

Using the default settings `n.adapt = 100` and `thin = 1`, and 100 sampling iterations, a simple model would be

```
> mod10a <- lm_imp(SBP ~ alc, data = NHANES, n.iter = 100)
```

The relevant part of the model summary (obtained with `summary()`) shows that the first 100 iterations (adaptive phase) were discarded, the 100 iterations that follow form the posterior sample, thinning was set to "1" and there are three chains.

```
## [...]
## MCMC settings:
## Iterations = 101:200
## Sample size per chain = 100
## Thinning interval = 1
## Number of chains = 3
```

Example: Insufficient Adaptation Phase

```
> mod10b <- lm_imp(SBP ~ alc, data = NHANES, n.adapt = 10, n.iter = 100)
```

```
> ## Warning in rjags::jags.model(file = modelfile, data = data_list, inits
> ## = inits, : Adaptation incomplete
> ## NOTE: Stopping adaptation
```

Specifying `n.adapt = 10` results in a warning message. The relevant part of the model summary from the resulting model is:

```
## [...]
## Iterations = 11:110
## Sample size per chain = 100
## Thinning interval = 1
## Number of chains = 3
```

Example: Thinning

```
> mod10c <- lm_imp(SBP ~ alc, data = NHANES, n.iter = 500, thin = 10)
```

Here, iterations 110 until 600 are used in the output, but due to a thinning interval of ten, the resulting MCMC chains contain only 50 samples instead of 500, that is, the samples from iteration 110, 120, 130, …

```
## [...]
## Iterations = 110:600
## Sample size per chain = 50
## Thinning interval = 10
## Number of chains = 3
```

6.6.2 Parameters to Follow

Since **JointAI** uses JAGS (Plummer 2003) for performing the MCMC sampling, and JAGS only saves the values of MCMC chains for those nodes which the user has specified should be monitored, this is also the case in **JointAI**.

For this purpose, the main functions `*_imp()` have an argument `monitor_params`, which takes a named list (or a named vector) with possible entries given in Table 6.2. This table contains a number of keywords that refer to (groups of) nodes. Each of the keywords works as a switch and can be specified as `TRUE` or `FALSE` (with the exception of `other`).

Parameters of the Analysis Model

The default setting is `monitor_params = c(analysis_main = TRUE)`, i.e., only the main parameters of the analysis model are monitored, and monitoring is switched off for all other parameters. Main parameters are the regression coefficients of the analysis model (`beta`), depending on the model type, the residual

Table 6.2: Keywords and names of (groups of) nodes that can be specified to be monitored using the argument `monitor_params`.

name/keyword	what is monitored
`analysis_main`	`betas` and `sigma_y` (and `D`)
`betas`	regression coefficients of the analysis model
`tau_y`	precision of the residuals from the analysis model
`sigma_y`	std. deviation of the residuals from the analysis model
`analysis_random`	`ranef`, `D`, `invD`, `RinvD`
`ranef`	random effects
`D`	covariance matrix of the random effects
`invD`	inverse of `D`
`RinvD`	scale matrix in Wishart prior for `invD`
`imp_pars`	`alphas`, `tau_imp`, `gamma_imp`, `delta_imp`
`alphas`	regression coefficients in the imputation models
`tau_imp`	precision of the residuals from imputation models
`gamma_imp`	intercepts in ordinal imputation models
`delta_imp`	increments of ordinal intercepts
`imps`	imputed values
`other`	additional parameters

standard deviation (`sigma_y`), and, for mixed models, the random effects variance-covariance matrix `D`.

The function `parameters()` returns the parameters specified to be followed (also for models where no MCMC sampling was performed, i.e., when `n.iter = 0` and `n.adapt = 0`). We use it here to demonstrate the effect that different choices for `monitor_params` have. For example:

```
> mod11a <- lm_imp(SBP ~ gender + WC + alc + creat, data = NHANES,
+                  n.adapt = 0)
>
> parameters(mod11a)
```

```
## [1] "(Intercept)"  "genderfemale" "WC"           "alc>=1"
## [5] "creat"        "sigma_SBP"
```

Parameters of the Covariate Models and Imputed Values

The parameters of the models for the incomplete variables can be selected with `monitor_params = c(imp_pars = TRUE)`. This will set monitors for the regression coefficients (`alpha`) and other parameters, such as precision (`tau_*`) and intercepts and increments (`gamma_*` and `delta_*`) in cumulative logit models.

```
> mod11b <- lm_imp(SBP ~ age + WC + alc + smoke + occup,
+                  data = NHANES, n.adapt = 0,
+                  monitor_params = c(imp_pars = TRUE,
+                                     analysis_main = FALSE))
>
> parameters(mod11b)
```

```
## [1] "alpha"       "tau_WC"      "gamma_smoke" "delta_smoke"
```

To generate (multiple) imputed datasets to be used for further analyses, the imputed values need to be monitored. This can be achieved by setting `monitor_params = c(imps = TRUE)`.

```
> mod11c <- lm_imp(SBP ~ age + WC + alc + smoke + occup,
+                  data = NHANES, n.adapt = 0,
+                  monitor_params = c(imps = TRUE, analysis_main = FALSE))
```

Extraction of multiple imputed datasets from a **JointAI** model is described in Section 6.7.6.

Random Effects

For mixed models, `analysis_main` also includes the random effects variance-covariance matrix `D`. Setting `analysis_random = TRUE` will switch on monitoring for the random effects (`ranef`), random effects variance-covariance matrix (`D`), inverse of the random effects variance-covariance matrix (`invD`) and the diagonal of the scale matrix of the Wishart-prior of `invD` (`RinvD`).

```
> mod11d <- lme_imp(bmi ~ age + EDUC, random = ~age | ID,
+                   data = simLong, n.adapt = 0,
+                   monitor_params = c(analysis_random = TRUE))
>
> parameters(mod11d)
```

```
##  [1] "(Intercept)" "EDUCmid"     "EDUClow"     "age"
##  [5] "sigma_bmi"   "b"           "invD[1,1]"   "invD[1,2]"
##  [9] "invD[2,2]"   "D[1,1]"      "D[1,2]"      "D[2,2]"
## [13] "RinvD[1,1]"  "RinvD[2,2]"
```

It is possible to select only a subset of the random effects parameters by specifying them directly, e.g.,

```
> mod11e <- lme_imp(bmi ~ age + EDUC, random = ~age | ID,
+                   data = simLong, n.adapt = 0,
+                   monitor_params = c(analysis_main = TRUE, RinvD = TRUE))
>
> parameters(mod11e)
```

```
##  [1] "(Intercept)" "EDUCmid"     "EDUClow"     "age"
##  [5] "sigma_bmi"   "D[1,1]"      "D[1,2]"      "D[2,2]"
##  [9] "RinvD[1,1]"  "RinvD[2,2]"
```

or by switching unwanted parts of `analysis_random` off, e.g.,

```
> mod11f <- lme_imp(bmi ~ age + EDUC, random = ~age | ID, data = simLong,
+                   n.adapt = 0, monitor_params = c(analysis_main = TRUE,
+                                                   analysis_random = TRUE,
+                                                   RinvD = FALSE,
+                                                   ranef = FALSE))
>
> parameters(mod11f)
```

```
##  [1] "(Intercept)" "EDUCmid"     "EDUClow"     "age"
##  [5] "sigma_bmi"   "invD[1,1]"   "invD[1,2]"   "invD[2,2]"
##  [9] "D[1,1]"      "D[1,2]"      "D[2,2]"
```

Other Parameters

The element `other` in `monitor_params` allows for the specification of one or multiple additional parameters to be monitored. When `other` is used with more than one element, `monitor_params` has to be a list.

Here, as an example, we monitor the probability of being in the `alc>=1` group for subjects one through three and the expected value of the distribution of `creat` for the first subject.

```
> mod11g <- lm_imp(SBP ~ gender + WC + alc + creat, data = NHANES,
+                  n.adapt = 0,
+                  monitor_params = list(analysis_main = FALSE,
+                                        other = c('p_alc[1:3]',
+                                                  "mu_creat[1]")))
>
> parameters(mod11g)


## [1] "p_alc[1:3]"  "mu_creat[1]"
```

Even though this example may not be particularly meaningful, in cases of convergence issues it can be helpful to be able to monitor any node of the model, not just the ones that are typically of interest.

6.6.3 Initial Values

Initial values are the starting point for the MCMC sampler. Setting good initial values, i.e., values that are likely under the posterior distribution, can speed up convergence. By default, the argument `inits = NULL`, which means that initial values are generated automatically by JAGS. It is also possible to supply initial values directly as a list or as a function.

Initial values can be specified for every unobserved node, that is, parameters and missing values, but it is possible to only specify initial values for a subset of nodes.

When the initial values provided by the user do not have elements named `".RNG.name"` or `".RNG.seed"`, **JointAI** will add those elements, which specify the name and seed value of the random number generator used for each chain.

The argument `seed` allows the specification of a seed value with which the starting values of the random number generator, and, hence, the values of the MCMC sample, can be reproduced.

Initial Values in a List of Lists

A list containing initial values should have the same length as the number of chains, where each element is a named list of initial values. Moreover, initial values should differ between the chains.

For example, to create initial values for the parameter vector `beta` and the precision parameter `tau_SBP` for two chains the following syntax could be used:

```
> init_list <- lapply(1:2, function(i) {
+   list(beta = rnorm(4),
+        tau_SBP = rgamma(1, 1, 1))
+ })
>
> init_list

## [[1]]
## [[1]]$beta
## [1] 0.2995096 0.2123710 0.6478957 0.8952516
##
## [[1]]$tau_SBP
## [1] 1.000624
##
##
## [[2]]
## [[2]]$beta
## [1]  2.2559981  0.9786635 -1.2725176 -0.7251253
##
## [[2]]$tau_SBP
## [1] 0.05501739

> mod12a <- lm_imp(SBP ~ gender + age + WC, data = NHANES, n.chain = 2,
+                  inits = init_list)
```

The user provided lists of initial values are stored in the JointAI object (together with starting values for the random number generator) and can be accessed via `mod11a$mcmc_settings$inits`.

Initial Values as a Function

Initial values can be specified as a function. The function should either take no arguments or a single argument called `chain`, and return a named list that supplies values for one chain.

For example, to create initial values for the parameter vectors `beta` and `alpha`:

```
> inits_fun <- function() {
+   list(beta = rnorm(4),
+        alpha = rnorm(3))
+ }
```

```
> inits_fun()


## $beta
## [1] -1.6045542  0.1872611  1.0167161 -0.4272887
##
## $alpha
## [1] 0.8542140 0.6391477 0.4720952


> mod12b <- lm_imp(SBP ~ gender + age + WC, data = NHANES,
+                  inits = inits_fun)
>
> mod12b$mcmc_settings$inits


## [[1]]
## [[1]]$beta
## [1] -0.07058338  0.41772091 -1.66236440  1.24957652
##
## [[1]]$alpha
## [1] -0.7204577  0.1424769 -1.0114044
##
## [[1]]$.RNG.name
## [1] "base::Wichmann-Hill"
##
## [[1]]$.RNG.seed
## [1] 77704
##
##
## [[2]]
## [[2]]$beta
## [1] -0.50236788 -0.01997157  1.40425944  1.18807193
##
## [[2]]$alpha
## [1] -0.8065902 -0.9709539 -0.7020397
##
## [[2]]$.RNG.name
## [1] "base::Mersenne-Twister"
##
## [[2]]$.RNG.seed
## [1] 29379
##
##
## [[3]]
```

```
## [[3]]$beta
## [1] -0.3516978 -0.9144069 -1.7397631 -0.1083395
##
## [[3]]$alpha
## [1]  0.9067457 -0.4471829  0.1170837
##
## [[3]]$.RNG.name
## [1] "base::Mersenne-Twister"
##
## [[3]]$.RNG.seed
## [1] 83619
```

When a function is supplied, the function is evaluated by **JointAI** and the resulting `list` is stored in the `JointAI` object.

For which Nodes can Initial Values be Specified?

Initial values can be specified for all unobserved stochastic nodes, i.e., parameters or unobserved data for which a distribution is specified in the JAGS model. They have to be supplied in the format of the parameter or unobserved value in the JAGS model. To find out which nodes there are in a model and in which form they have to be specified, the function `coef()` from package **rjags** can be used to obtain a list with the current values of the MCMC chains (by default the first chain) from a JAGS model object. This object is contained in a `JointAI` object under the name `model`. Elements of the initial values should have the same structure as the elements in this list of current values.

Example:

We are using a longitudinal model and the `simLong` data in this example. Here we only show some relevant parts of the output.

```
> mod12c <- lme_imp(bmi ~ time + HEIGHT_M + hc + SMOKE, random = ~ time | ID,
+                   no_model = 'time', data = simLong)
>
> # coef(mod12c$model)
```

`RinvD` is the scale matrix in the Wishart prior for the inverse of the random effects variance-covariance matrix `D`. In the data that is passed to JAGS (which is stored in the element `data_list` in a `JointAI` object), this matrix is specified as a diagonal matrix, with unknown diagonal elements:

```
> mod12c$data_list['RinvD']
```

```
## $RinvD
##      [,1] [,2]
## [1,]   NA    0
## [2,]    0   NA
```

These diagonal elements are estimated in the model and have a gamma prior. The corresponding part of the JAGS model is:

```
## [...]
##   # Priors for the covariance of the random effects
##   for (k in 1:2){
##     RinvD[k, k] ~ dgamma(shape_diag_RinvD, rate_diag_RinvD)
##   }
##   invD[1:2, 1:2] ~ dwish(RinvD[ , ], KinvD)
##   D[1:2, 1:2] <- inverse(invD[ , ])
## [...]
```

The element `RinvD` in the initial values has to be a 2×2 matrix, with positive values on the diagonal and `NA` as off-diagonal elements, since these are fixed in the data:

```
## $RinvD
##          [,1]      [,2]
## [1,] 5.625607        NA
## [2,]       NA 0.6499669
```

Lines 82 through 85 of the design matrix of the fixed effects of baseline covariates, `Xc`, in the data are:

```
> mod12c$data_list$Xc[82:85, ]
```

```
##       (Intercept)  HEIGHT_M SMOKEsmoked until[...] SMOKEcontin[...]
## 172.1           1 0.1148171                     NA               NA
## 173.1           1        NA                     NA               NA
## 174.1           1 0.5045126                     NA               NA
## 175.1           1 1.8822249                     NA               NA
```

The matrix `Xc` in the initial values has the same dimension as `Xc` in the data. It has values where there are missing values in `Xc` in the data, e.g., `Xc[83, 2]`, and is `NA` elsewhere:

```
## $Xc
##        [,1]         [,2] [,3] [,4]
##   [1,]   NA           NA   NA   NA
##
## [...]
##
##  [82,]   NA           NA   NA   NA
##  [83,]   NA  2.06807598   NA   NA
##  [84,]   NA           NA   NA   NA
##
## [...]
```

There are no initial values specified for the third and fourth column, since these columns contain the dummy variables corresponding to the categorical variable `SMOKE` and are calculated from the corresponding column in the matrix `Xcat`, i.e., these are deterministic nodes, not stochastic nodes.

The relevant part of the JAGS model is:

```
## [...]
##     # ordinal model for SMOKE
##     Xcat[i, 1] ~ dcat(p_SMOKE[i, 1:3])
## [...]
##     Xc[i, 3] <- ifelse(Xcat[i, 1] == 2, 1, 0)
##     Xc[i, 4] <- ifelse(Xcat[i, 1] == 3, 1, 0)
## [...]
```

Elements that are completely unobserved, like the parameter vectors `alpha` and `beta`, the random effects `b` or scalar parameters `delta_SMOKE` and `gamma_SMOKE` are entirely specified in the initial values.

6.6.4 Parallel Sampling

To reduce the computational time it is possible to perform sampling of multiple MCMC chains in parallel. This can be specified by setting the argument `parallel = TRUE`. The maximum number of cores to be used can be set with the argument `ncores`. By default this is two less than the number of cores available on the machine, but never more than the number of MCMC chains.

Parallel computation is done using the packages **foreach** (Microsoft and Weston 2017) and **doParallel** (Corporation and Weston 2018). Note that it is currently not possible to display a progress bar when using parallel computation.

6.7 After Fitting

Any of the main functions `*_imp()` will return an object of class `JointAI`. It contains the original data (`data`), information on the type of model (`call`, `analysis_type`, `models`, `fixed`, `random`, `hyperpars`, `scale_pars`) and MCMC sampling (`mcmc_settings`), the JAGS model (`model`) and MCMC sample (`MCMC`; if a sample was generated), the computational time (`time`) of the MCMC sampling, and some additional elements that are used by methods for objects of class `JointAI` but are typically not of interest for the user.

In the remaining part of this section, we describe how the results from a `JointAI` model can be visualized, summarized and evaluated. The functions described here use, by default, the full MCMC sample and show only the parameters of the analysis model. Arguments `start`, `end` and `thin` are available to select a subset of the MCMC samples that is used to calculate the summary. The argument `subset` allows the user to control for which nodes the summary or visualization is returned and follows the same logic as the argument `monitor_params` in `*_imp()`. The use of these arguments is further explained in Section 6.7.4.

6.7.1 Visualizing the Posterior Sample

The posterior sample can be visualized by two commonly used plots: a traceplot, showing samples across iterations, and a plot of the empirical density of the posterior sample.

Traceplot

A traceplot shows the sampled values per chain and node throughout iterations. It allows the visual evaluation of convergence and mixing of the chains and can be obtained with the function `traceplot()`.

```
> mod13a <- lm_imp(SBP ~ gender + WC + alc +  creat,
+                  data = NHANES, n.iter = 500)
>
> traceplot(mod13a)
```

When the sampler has converged the chains show one horizontal band, as in Figure 6.8. Consequently, when traces show a trend, convergence has not been reached and more iterations are necessary (e.g., using `add_samples()`).

Graphical aspects of the traceplot can be controlled by specifying standard graphical arguments via the dots argument `"..."`, which are passed to `matplot()`. This

allows the user to change colour, linetype and -width, limits, etc. Arguments `nrow` and/or `ncol` can be supplied to set specific numbers of rows and columns for the layout of the grid of plots.

With the argument `use_ggplot` it is possible to get a **ggplot2** (Wickham 2016) version of the traceplot. It can be extended using standard **ggplot2** syntax. The output of the following syntax is shown in Figure 6.9.

```
> library(ggplot2)
> traceplot(mod13a, ncol = 3, use_ggplot = TRUE) +
+   theme(legend.position = 'bottom',
+         panel.background = element_rect(fill = grey(0.95)),
+         panel.border = element_rect(fill = NA, color = grey(0.85)),
+         strip.background = element_rect(color = grey(0.85))) +
+   scale_color_manual(values = c("#783D4F", "#60B5BC", "#6F5592"))
```

Densityplot

The posterior distributions can also be visualized using the function `densplot()`, which plots the empirical density per node per chain, or combining chains (when `joined = TRUE`).

The argument `vlines` takes a list of lists, containing specifications passed to the function `abline()` (from package **graphics** which is available with base R), and allows the addition of (vertical) lines to the plot, e.g., marking zero, or marking the posterior mean and 2.5% and 97.5% quantiles (Figure 6.10).

```
> densplot(mod13a, ncol = 3,
+          vlines = list(list(v = summary(mod13a)$stats[, "Mean"], lty = 1,
+                             lwd = 2),
+                        list(v = summary(mod13a)$stats[, "2.5%"], lty = 2),
+                        list(v = summary(mod13a)$stats[, "97.5%"], lty = 2)
+          )
+ )
```

As with `traceplot()` it is possible to use the **ggplot2** version of `densplot()` when setting `use_ggplot = TRUE`. Here, vertical lines can be added as additional layers. Figure 6.11 shows, as an example, the posterior density from `mod13a` to which vertical lines, representing the 95% credible interval and a 95% confidence interval from a complete case analysis, are added. The corresponding syntax is given in Appendix 6.B.

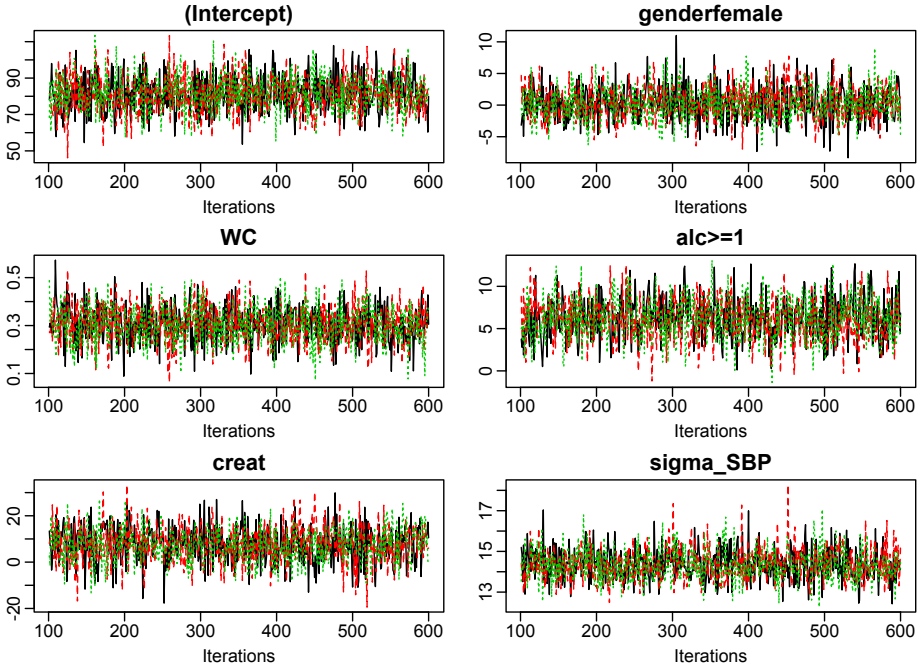


Figure 6.8: Traceplot of `mod13a`.

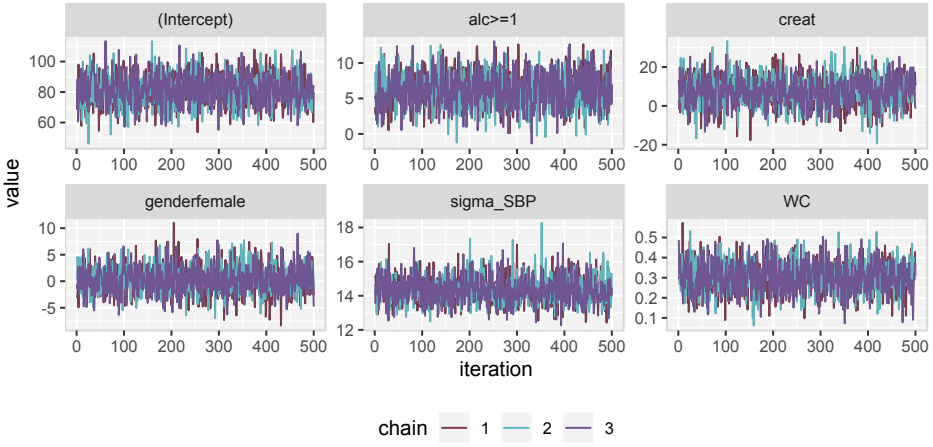


Figure 6.9: **ggplot2** version of the traceplot of `mod13a`.

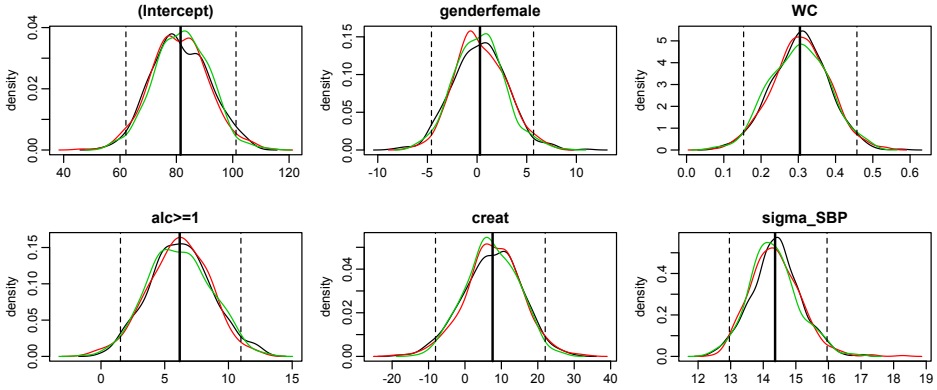


Figure 6.10: Densityplot of `mod13a`.

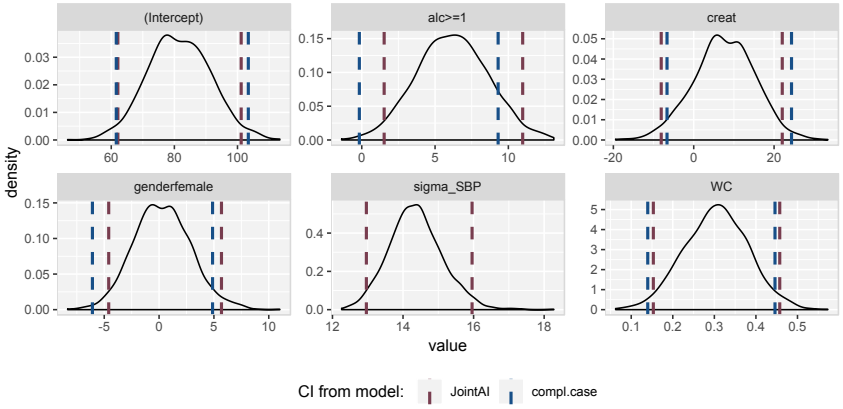


Figure 6.11: Densityplot of model `mod13a`.

6.7.2 Model Summary

A summary of the posterior distribution estimated in a `JointAI` model can be obtained using the function `summary()`.

The posterior summary consists of the mean, standard deviation and quantiles (by default the 2.5% and 97.5% quantiles) of the MCMC samples from all chains combined, as well as the tail probability (see below) and Gelman-Rubin criterion (see Section 6.7.3).

Additionally, some important characteristics of the MCMC samples on which the summary is based, are given. This includes the range and number of iterations

(`Sample size per chain`), thinning interval and number of chains. Furthermore, the number of observations (number of rows in the data) is printed.

```
> summary(mod13a)

##
##  Linear model fitted with JointAI
##
## Call:
## lm_imp(formula = SBP ~ gender + WC + alc + creat, data = NHANES,
##     n.iter = 500)
##
## Posterior summary:
##                Mean     SD   2.5%   97.5% tail-prob. GR-crit
## (Intercept)  81.499 9.9585 62.091 101.167    0.00000       1
## genderfemale  0.293 2.5963 -4.592   5.689    0.94000       1
## WC            0.304 0.0766  0.153   0.457    0.00000       1
## alc>=1        6.175 2.4365  1.528  10.977    0.00933       1
## creat         7.589 7.6414 -8.090  22.037    0.31600       1
##
## Posterior summary of residual std. deviation:
##           Mean    SD 2.5% 97.5% GR-crit
## sigma_SBP 14.4 0.763   13    16       1
##
##
## MCMC settings:
## Iterations = 101:600
## Sample size per chain = 500
## Thinning interval = 1
## Number of chains = 3
##
## Number of observations: 186
```

For mixed models, `summary()` also returns the posterior summary of the random effects covariance matrix `D` and the number of groups:

```
> library(splines)
> mod13b <- lme_imp(bmi ~ GESTBIR + ETHN + HEIGHT_M + ns(age, df = 3),
+                   random = ~ ns(age, df = 3)|ID,
+                   data = subset(simLong, !is.na(bmi)),
+                   n.iter = 500, no_model = 'age')
>
> summary(mod13b, start = 300)
```

```
##  Linear mixed model fitted with JointAI
##
## Call:
## lme_imp(fixed = bmi ~ GESTBIR + ETHN + HEIGHT_M + ns(age, df = 3),
##     data = subset(simLong, !is.na(bmi)), random = ~ns(age, df = 3) |
##         ID, n.iter = 500, no_model = "age")
##
## Posterior summary:
##                       Mean      SD    2.5%   97.5% tail-prob. GR-crit
## (Intercept)      16.11569 2.42592 11.2231 20.9408    0.00000    1.05
## GESTBIR          -0.02724 0.04863 -0.1259  0.0568    0.60908    1.04
## ETHNother        -0.00248 0.14991 -0.2776  0.3303    0.95681    1.02
## HEIGHT_M          0.00478 0.00927 -0.0123  0.0251    0.60687    1.00
## ns(age, df = 3)1 -0.23917 0.08453 -0.4139 -0.0771    0.00443    1.56
## ns(age, df = 3)2  1.74068 0.22251  1.2973  2.1464    0.00000    1.02
## ns(age, df = 3)3 -1.27304 0.06342 -1.4027 -1.1587    0.00000    1.22
##
## Posterior summary of random effects covariance matrix:
##          Mean     SD   2.5%  97.5% tail-prob. GR-crit
## D[1,1]  1.419 0.1726  1.122  1.803               1.05
## D[1,2] -0.801 0.1188 -1.053 -0.597          0    1.17
## D[2,2]  0.764 0.1172  0.564  1.009               1.41
## D[1,3] -2.493 0.3562 -3.243 -1.876          0    1.10
## D[2,3]  2.450 0.2956  1.918  3.089          0    1.07
## D[3,3]  8.107 0.9449  6.451 10.058               1.09
## D[1,4] -0.703 0.1015 -0.911 -0.522          0    1.50
## D[2,4]  0.618 0.0797  0.473  0.790          0    1.26
## D[3,4]  2.008 0.2591  1.561  2.557          0    1.56
## D[4,4]  0.520 0.0839  0.384  0.699               2.52
##
## Posterior summary of residual std. deviation:
##            Mean      SD  2.5% 97.5% GR-crit
## sigma_bmi 0.458 0.00854 0.442 0.475       1
##
##
## MCMC settings:
## Iterations = 300:600
## Sample size per chain = 301
## Thinning interval = 1
## Number of chains = 3
##
## Number of observations: 1881
## Number of groups: 200
```

The summary of parametric Weibull survival models also returns the summary of the posterior sample of the shape parameter of the Weibull distribution:

```
> summary(mod6a)
```

```
##
##  Weibull survival model fitted with JointAI
##
## Call:
## survreg_imp(formula = Surv(time, status) ~ age + sex + ph.karno +
##     meal.cal + wt.loss, data = lung, n.iter = 250)
##
## Posterior summary:
##                   Mean       SD      2.5%    97.5% tail-prob. GR-crit
## (Intercept)  6.157014 0.556472  5.000007 7.161856      0.000    1.04
## age          0.001188 0.005488 -0.009651 0.012562      0.840    1.04
## sex2        -0.158160 0.104307 -0.360187 0.043870      0.144    1.01
## ph.karno    -0.004325 0.003914 -0.011889 0.003265      0.272    1.04
## meal.cal    -0.000119 0.000138 -0.000395 0.000111      0.421    1.04
## wt.loss     -0.001401 0.003836 -0.009482 0.006329      0.699    1.03
##
## Posterior summary of the shape of the Weibull distribution:
##       Mean     SD 2.5% 97.5% GR-crit
## shape 1.45 0.0822  1.3  1.62    1.03
##
##
## MCMC settings:
## Iterations = 101:350
## Sample size per chain = 250
## Thinning interval = 1
## Number of chains = 3
##
## Number of observations: 228
```

Tail Probability

The tail probability, calculated as $2 \times \min\{Pr(\theta > 0), Pr(\theta < 0)\}$, where θ is the parameter of interest, is a measure of how likely the value 0 is under the estimated posterior distribution. Figure 6.12 visualizes three examples of posterior distributions and the corresponding minimum of $Pr(\theta > 0)$ and $Pr(\theta < 0)$ (shaded area).

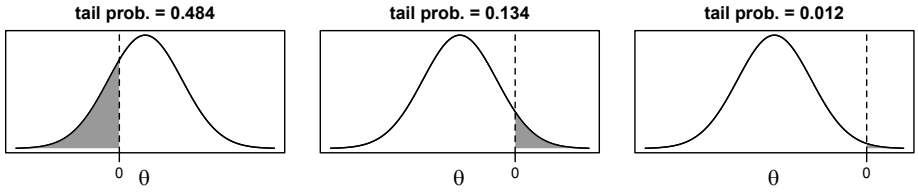


Figure 6.12: Visualization of the tail probability.

6.7.3 Evaluation Criteria

Convergence of the MCMC chains and precision of the posterior sample can also be evaluated in a more formal manner. Implemented in **JointAI** are the Gelman-Rubin criterion for convergence (Gelman and Rubin 1992; Brooks and Gelman 1998) and a comparison of the Monte Carlo Error with the posterior standard deviation.

Gelman-Rubin Criterion for Convergence

The Gelman-Rubin criterion (Gelman and Rubin 1992; Brooks and Gelman 1998), also referred to as "potential scale reduction factor", evaluates convergence by comparing within and between chain variability and, thus, requires at least two MCMC chains to be calculated. It is implemented for `JointAI` objects in the function `GR_crit()`, which is based on the function `gelman.diag()` from the package **coda** (Plummer et al. 2006). The upper limit of the confidence interval should not be much larger than 1.

```
> GR_crit(mod13a)
```

```
## Potential scale reduction factors:
##
##               Point est. Upper C.I.
## (Intercept)           1       1.00
## genderfemale          1       1.00
## WC                    1       1.00
## alc>=1                1       1.01
## creat                 1       1.01
## sigma_SBP             1       1.01
##
## Multivariate psrf
##
## 1.01
```

Besides the arguments `start`, `end`, `thin`, and `subset`, which are explained in Section 6.7.4, `GR_crit()` also takes the arguments `confidence`, `transform` and `autoburnin` of `gelman.diag()`.

Monte Carlo Error

Precision of the MCMC sample can be checked with the function `MC_error()`. It uses the function `mcmcse.mat()` from the package **mcmcse** (Flegal et al. 2017) to calculate the Monte Carlo error (the error that is made since the sample is finite) and compares is to the standard deviation of the posterior sample. A rule of thumb is that the Monte Carlo error should not be more than 5% of the standard deviation (Lesaffre and Lawson 2012). Besides the arguments explained in Section 6.7.4, `MC_error()` takes the arguments of `mcmcse.mat()`.

```
> MC_error(mod13a)
```

```
##                 est   MCSE     SD MCSE/SD
## (Intercept)  81.50 0.2850  9.958   0.029
## genderfemale  0.29 0.0708  2.596   0.027
## WC            0.30 0.0021  0.077   0.028
## alc>=1        6.18 0.0798  2.437   0.033
## creat         7.59 0.2237  7.641   0.029
## sigma_SBP    14.36 0.0233  0.763   0.031
```

`MC_error()` returns an object of class `MCElist`, which is a list containing matrices with the posterior mean, estimated Monte Carlo error, posterior standard deviation and the ratio of the Monte Carlo error and posterior standard deviation, for the scaled (if this MCMC sample was included in the `JointAI` object) and unscaled (transformed back to the scale of the data) posterior samples. The associated print method prints only the latter.

To facilitate quick evaluation of the Monte Carlo error to posterior standard deviation ratio, plotting of an object of class `MCElist` using `plot()` shows this ratio for each (selected) node and automatically adds a vertical line at the desired cut-off (by default 5%; see Figure 6.13).

```
> par(mar = c(3, 5, 0.5, 0.5), mgp = c(2, 0.6, 0), mfrow = c(1, 2))
> plot(MC_error(mod13a))  # left panel
> plot(MC_error(mod13a, end = 250))  # right panel
```

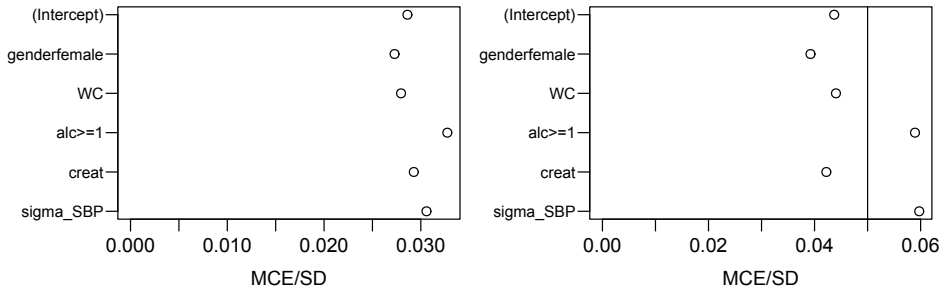


Figure 6.13: Plot of the `MCElist` object from `mod13a`. Left: including all iterations, right: using only the first 250 iterations of the MCMC sample.

6.7.4 Subset of Output

By default, the functions `traceplot()`, `densplot()`, `summary()`, `predict()`, `GR_crit()` and `MC_Error()` use all iterations of the MCMC sample and consider only the parameters of the analysis model (if they were monitored). In this section we describe how the set of iterations and parameters/nodes to display can be changed using the arguments `subset`, `start`, `end` and `thin`.

Subset of Parameters

When the main parameters of the analysis model have been monitored in a `JointAI` object, i.e., when `monitor_params` was set to `TRUE`, only these parameters are returned in the model summary, plots and criteria shown above. If the main parameters of the analysis model were not monitored and the argument `subset` is not specified, all parameters that were monitored are displayed.

To display output for nodes other than the main parameters of the analysis model or for a subset of nodes, the argument `subset` needs to be specified.

Example:

To display only the parameters of the imputation models, we set `subset = c(analysis_main = FALSE, imp_pars = TRUE)` (after re-estimating the model with the monitoring for these parameters switched on):

```
> mod13c <- update(mod13a, monitor_params = c(imp_pars = TRUE))
> summary(mod13c, subset = c(analysis_main = FALSE, imp_pars = TRUE))
```

```
##  Linear model fitted with JointAI
##
## Call:
## lm_imp(formula = SBP ~ gender + WC + alc + creat, data = NHANES,
##     n.iter = 500, monitor_params = c(imp_pars = TRUE))
##
## Posterior summary:
##              Mean     SD    2.5%   97.5% tail-prob. GR-crit
## alpha[1]   0.1659 0.0987 -0.0214  0.3617     0.0973    1.00
## alpha[2]  -0.3533 0.1463 -0.6423 -0.0759     0.0187    1.00
## alpha[3]   0.4934 0.0871  0.3176  0.6566     0.0000    1.00
## alpha[4]  -1.0418 0.1312 -1.3043 -0.7844     0.0000    1.00
## alpha[5]   0.0768 0.0654 -0.0521  0.2027     0.2427    1.00
## alpha[6]  -0.1200 0.2590 -0.6277  0.3597     0.7027    1.04
## alpha[7]  -0.9330 0.4688 -1.8326 -0.0294     0.0467    1.02
## alpha[8]   0.0922 0.1730 -0.2439  0.4305     0.5760    1.01
## alpha[9]  -0.2704 0.2243 -0.7497  0.1542     0.2080    1.01
## tau_WC     1.0290 0.1094  0.8236  1.2587     0.0000    1.00
## tau_creat  1.3970 0.1510  1.1177  1.7119     0.0000    1.00
##
##
## MCMC settings:
## Iterations = 101:600
## Sample size per chain = 500
## Thinning interval = 1
## Number of chains = 3
##
## Number of observations: 186
```

Example:

To select only some of the parameters, they can be specified directly by name via the `other` element of `subset`.

```
> densplot(mod13a, subset = list(other = c('creat', 'alc>=1')))
```

Example:

This also works when a subset of the imputed values should be displayed:

```
> # re-fit the model and monitor the imputed values
> mod13d <- lm_imp(SBP ~ gender + age + albu + occup + alc, n.iter = 200,
+                  data = NHANES, monitor_params = c(imps = TRUE))
```

```
>
> # select all imputed values for 'albu' (4th column of Xc)
> sub3 <- grep('Xc\\[[[:digit:]]+,4\\]', parameters(mod13d), value = TRUE)
> sub3

## [1] "Xc[10,4]"  "Xc[14,4]"  "Xc[18,4]"  "Xc[25,4]"  "Xc[80,4]"
## [6] "Xc[118,4]" "Xc[172,4]" "Xc[180,4]"

> # pass "sub3" to "subset" via "other", for example in a traceplot:
> # traceplot(mod13d, subset = list(other = sub3), ncol = 2)
```

Example:

When the number of imputed values is large or in order to check convergence of random effects, it may not be feasible to plot and inspect all traceplots. In that case, a random subset of, for instance, the random effects, can be selected (output not shown):

```
> # re-fit the model monitoring the random effects
> mod13e <- update(mod13a, monitor_params = c(ranef = TRUE))
>
> # extract random intercepts and random slopes
> ri <- grep('^b\\[[[:digit:]]+,1\\]$', colnames(mod13e$MCMC[[1]]),
+            value = TRUE)
> rs <- grep('^b\\[[[:digit:]]+,2\\]$', colnames(mod13e$MCMC[[1]]),
+            value = TRUE)
>
> # to plot the chains of 12 randomly selected random intercepts & slopes:
> traceplot(mod13e, subset = list(other = sample(ri, size = 12)), ncol = 4)
> traceplot(mod13e, subset = list(other = sample(rs, size = 12)), ncol = 4)
```

Subset of MCMC Samples

With the arguments `start`, `end` and `thin` it is possible to select which iterations from the MCMC sample are included in the summary. `start` and `end` specify the first and last iterations to be used, `thin` the thinning interval. Specification of `start` thus allows the user to discard a "burn-in", i.e., the iterations before the MCMC chain had converged.

6.7.5 Predicted Values

Often, the aim of an analysis is not only to estimate the association between outcome and covariates but to predict future outcomes or outcomes for new subjects.

The function `predict()` allows us to obtain predicted values and corresponding credible intervals from `JointAI` objects. Note that for mixed models, currently only marginal prediction is implemented, not prediction conditional on the random effects. A dataset containing data for which the prediction should be performed for is specified via the argument `newdata`. The argument `quantiles` allows specification of the quantiles of the posterior sample that are used to obtain the prediction interval (by default the 2.5% and 97.5% quantile). Arguments `start`, `end` and `thin` control the subset of MCMC samples used.

```
> predict(mod13a, newdata = NHANES[27, ])
```

```
## $dat
##          SBP gender age             race   WC alc educ creat albu
## 392 126.6667   male  32 Mexican American 94.1  <1  low  0.83  4.2
##     uricacid bili occup  smoke      fit     2.5%    97.5%
## 392      8.7    1  <NA> former 116.4369 112.3891 120.2579
##
## $fit
## [1] 116.4369
##
## $quantiles
##           [,1]
## 2.5%  112.3891
## 97.5% 120.2579
```

`predict()` returns a list with elements `dat` (a dataset consisting of `newdata` with the predicted values and quantiles appended), `fit` (the predicted values) and `quantiles` (the quantiles that form the prediction interval).

Prediction to Visualize Non-linear Effects

Another reason to obtain predicted values is the visualization of non-linear effects (see Figure 6.14). To facilitate the generation of a dataset for such a prediction, the function `predDF()` can be used. It generates a `data.frame` that contains a sequence of values through the range of observed values for a covariate specified by the argument `var`, and the median or reference value for all other continuous and categorical variables.

```
> # create dataset for prediction
> newDF <- predDF(mod13b, var = "age")
>
```

```
> # obtain predicted values
> pred <- predict(mod13b, newdata = newDF, start = 300)
>
> # plot predicted values and prediction interval
> matplot(pred$dat$age, pred$dat[, c('fit', '2.5%', '97.5%')],
+         lty = c(1,2,2), type = 'l', col = 1,
+         xlab = 'age in months', ylab = 'predicted value')
```

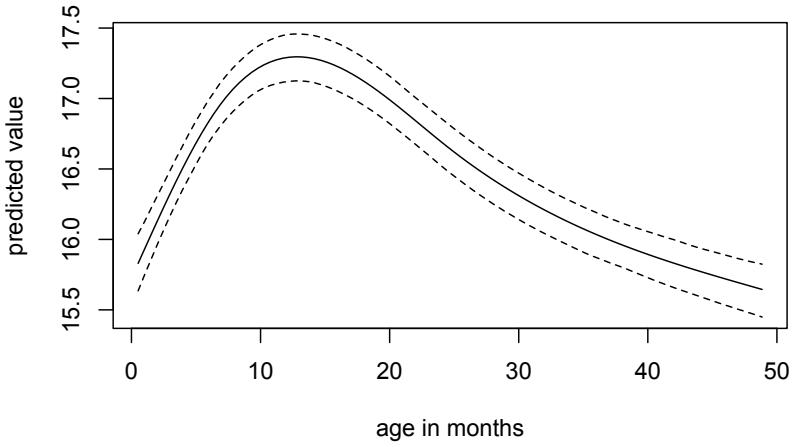


Figure 6.14: Predicted values of BMI and corresponding 95% credible intervals from `mod13b`.

6.7.6 Export of Imputed Values

Imputed datasets can be extracted from a `JointAI` object (in which a monitor for the imputed values has been set, i.e., `monitor_params = c(imps = TRUE)`) with the function `get_MIdat()`.

A completed dataset is created by taking the imputed values from a randomly chosen iteration of the MCMC sample, transforming them back to the original scale (if scaling was performed before the MCMC sampling) and filling them into the original incomplete data.

The argument `m` specifies the number of imputed datasets to be created, `include` controls whether the original data are included in the long format `data.frame`

(default is `include = TRUE`), `start` specifies the first iteration that may be used, and `minspace` is the minimum number of iterations between iterations eligible for selection. To make the selection of iterations reproducible, a seed value can be specified via the argument `seed`.

When `export_to_SPSS = TRUE` the imputed data is exported to SPSS (IBM SPSS Statistics, IBM Corp.), i.e., a `.txt` file containing the data and a `.sps` file containing SPSS syntax to convert the data into an SPSS data file (with ending `.sav`) are written. Arguments `filename` and `resdir` allow specification of the name of the `.txt` and `.sps` file and the directory they are written to.

`get_MIdat()` returns a long-format `data.frame` containing the imputed datasets (and by default the original data) stacked on top of each other. The imputation number is given in the variable `Imputation_`, column `.id` contains a newly created id variable for each observation in cross-sectional data (multi-level data should already contain an id variable) and the column `.rownr` identifies rows of the original data (which is relevant in multi-level data).

```
> impDF <- get_MIdat(mod13d, m = 5, seed = 2019)
```

The function `plot_imp_distr()` allows visual comparison of the distributions of the observed and imputed values (Figure 6.15).

```
> plot_imp_distr(impDF, nrow = 1)
```

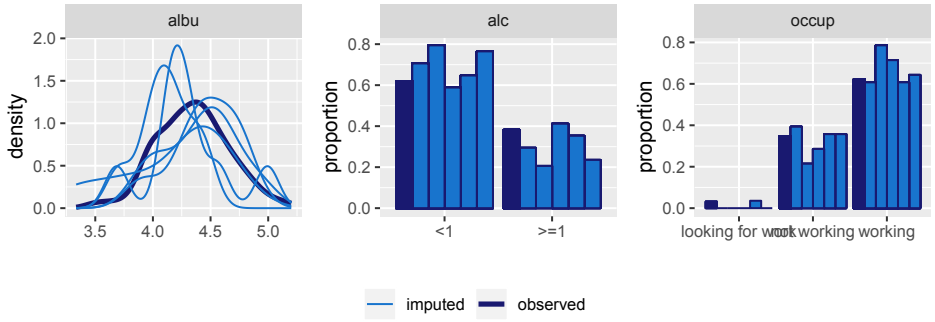


Figure 6.15: Distribution of observed and imputed values from `mod13d`.

6.8 Assumptions and Extensions

Like any statistical model, the approach to imputation followed in **JointAI** relies on assumptions that need to be satisfied in order to obtain valid results.

A commonly made assumption that is also required for **JointAI** is that the missing data mechanism is ignorable, i.e., that data is Missing At Random (MAR) or Missing Completely At Random (MCAR) (Rubin 1976) and that parameters in the model of the missingness mechanism are independent of the parameters in the data model (Schafer 1997). It is the task of the researcher to critically evaluate whether this assumption is satisfied for a given dataset and model.

Furthermore, all models involved in the imputation and analysis need to be correctly specified. In current implementations of imputation procedures in software, imputation models are typically automatically specified, using standard assumptions like linear associations and default model types. In **JointAI**, the arguments `models` and `auxvar` permit tailoring of the automatically chosen models to some extent, by allowing the user to chose non-normal imputation models for continuous variables and to include variables or functional forms of variables in the linear predictor of the imputation models that are not used in the analysis model.

When using auxiliary variables in **JointAI**, it should be noted that due to the ordering of the conditional distributions in the sequence of models it is implied that the auxiliary variable is independent of the outcome, since neither the model for the auxiliary variable (if the auxiliary variable has missing values) has the outcome in its linear predictor nor vice versa.

Moreover, in order to make software usable, default values have to be chosen for various parameters. These default values are chosen to work well in certain settings, but can not be guaranteed to be appropriate in general and it is the task of the user to make the appropriate changes. In **JointAI** this concerns, for example, the choice of hyperparameters and automatically chosen types of imputation models.

To expand the range of settings in which **JointAI** provides a valid and user-friendly way to simultaneously analyse and impute data, several extensions are planned, for example:

- Implement the use of (penalized) splines for incompletely observed covariates, thereby improving model fit.
- Increase the flexibility of imputation models by optional inclusion of interaction terms and non-parametric Bayesian models that allow imputation under non-standard distributions.

- Evaluation of model fit of the analysis and imputation models to help the user prevent bias due to misspecification.
- Implementation of subject-specific prediction from mixed models.
- Extend the analysis models to handle endogenous covariates by modelling random effects (and error terms) jointly.

Appendix

6.A Default Hyperparameters

```
> default_hyperpars()
```

```
## $norm
##    mu_reg_norm   tau_reg_norm shape_tau_norm  rate_tau_norm
##          0e+00          1e-04          1e-02          1e-02
##
## $gamma
##    mu_reg_gamma   tau_reg_gamma shape_tau_gamma  rate_tau_gamma
##           0e+00           1e-04           1e-02           1e-02
##
## $beta
##    mu_reg_beta   tau_reg_beta shape_tau_beta  rate_tau_beta
##          0e+00          1e-04          1e-02          1e-02
##
## $logit
##  mu_reg_logit tau_reg_logit
##         0e+00         1e-04
##
## $poisson
##  mu_reg_poisson tau_reg_poisson
##           0e+00           1e-04
##
## $probit
##  mu_reg_probit tau_reg_probit
##          0e+00          1e-04
##
## $multinomial
##  mu_reg_multinomial tau_reg_multinomial
##               0e+00               1e-04
##
## $ordinal
```

```
##    mu_reg_ordinal   tau_reg_ordinal  mu_delta_ordinal
##             0e+00             1e-04             0e+00
## tau_delta_ordinal
##             1e-04
##
## $Z
## function (nranef)
## {
##     if (nranef > 1) {
##         RinvD <- diag(as.numeric(rep(NA, nranef)))
##         KinvD <- nranef
##     }
##     else {
##         RinvD <- KinvD <- NULL
##     }
##     list(RinvD = RinvD, KinvD = KinvD, shape_diag_RinvD = 0.1,
##         rate_diag_RinvD = 0.01)
## }
## <bytecode: 0x00000000180f0a38>
## <environment: 0x000000001fb09a18>
##
## $surv
##  mu_reg_surv tau_reg_surv
##        0.000        0.001
##
## $coxph
##     c     r   eps
## 1e-03 1e-01 1e-10
```

6.B Density Plot using ggplot2

```
> # fit the complete-case version of the model
> mod13a_cc <- lm(formula(mod13a), data = NHANES)
>
>
> # make a dataset containing the quantiles of the posterior sample and
> # confidence intervals from the complete case analysis:
> quantDF <- rbind(data.frame(variable = rownames(summary(mod13a)$stat),
+                             type = '2.5%',
+                             model = 'JointAI',
+                             value = summary(mod13a)$stat[, c('2.5%')]
+                             ),
+                  data.frame(variable = rownames(summary(mod13a)$stat),
```

```
+                             type = '97.5%',
+                             model = 'JointAI',
+                             value = summary(mod13a)$stat[, c('97.5%')]
+                   ),
+                   data.frame(variable = names(coef(mod13a_cc)),
+                             type = '2.5%',
+                             model = 'cc',
+                             value = confint(mod13a_cc)[, '2.5 %']
+                   ),
+                   data.frame(variable = names(coef(mod13a_cc)),
+                             type = '97.5%',
+                             model = 'cc',
+                             value = confint(mod13a_cc)[, '97.5 %']
+                   )
+ )
>
>
> # ggplot version:
> p13a <- densplot(mod13a, ncol = 3, use_ggplot = TRUE, joined = TRUE) +
+   theme(legend.position = 'bottom',
+         panel.background = element_rect(fill = grey(0.95)),
+         panel.border = element_rect(fill = NA, color = grey(0.85)),
+         strip.background = element_rect(color = grey(0.85)))
>
>
> # add vertical lines for the:
> # - confidence intervals from the complete case analysis
> # - quantiles of the posterior distribution
> p13a +
+   geom_vline(data = quantDF, aes(xintercept = value, color = model),
+              lty = 2, lwd = 1) +
+   scale_color_manual(name = 'CI from model: ',
+                      limits = c('JointAI', 'cc'),
+                      values = c("#783D4F", "#174F88"),
+                      labels = c('JointAI', 'compl.case'))
```

References

Bartlett, J. W. et al. (2015). "Multiple imputation of covariates by fully conditional specification: accommodating the substantive model". *Statistical Methods in Medical Research*, **24**(4):462–487. DOI: 10.1177/0962280214521348.

Brooks, S. P. and A. Gelman (1998). "General Methods for Monitoring Convergence of Iterative Simulations". *Journal of Computational and Graphical Statistics*, **7**(4):434–455. DOI: 10.1080/10618600.1998.10474787.

Corporation, M. and S. Weston (2018). *doParallel: Foreach Parallel Adaptor for the 'parallel' Package*. R package version 1.0.14. URL: https://CRAN.R-project.org/package=doParallel.

Erler, N. S. (2019). *JointAI: Joint Analysis and Imputation of Incomplete Data*. R package version 0.5.1. URL: https://github.com/nerler/JointAI.

Erler, N. S., D. Rizopoulos, V. W. Jaddoe, et al. (2019). "Bayesian imputation of time-varying covariates in linear mixed models". *Statistical Methods in Medical Research*, **28**(2):555–588. DOI: 10.1177/0962280217730851.

Erler, N. S., D. Rizopoulos, J. van Rosmalen, et al. (2016). "Dealing with missing covariates in epidemiologic studies: a comparison between multiple imputation and a full Bayesian approach". *Statistics in Medicine*, **35**(17):2955–2974. DOI: 10.1002/sim.6944.

Flegal, J. M. et al. (2017). *mcmcse: Monte Carlo Standard Errors for MCMC*. R package version 1.3-2. Riverside, CA, Denver, CO, Coventry, UK, and Minneapolis, MN.

Gelman, A. and D. B. Rubin (1992). "Inference from Iterative Simulation Using Multiple Sequences". *Statistical Science*, **7**(4):457–472. DOI: 10.1214/ss/1177011136.

Ibrahim, J. G., M.-H. Chen, and S. R. Lipsitz (2002). "Bayesian methods for generalized linear models with covariates missing at random". *Canadian Journal of Statistics*, **30**(1):55–78. DOI: 10.2307/3315865.

Lesaffre, E. M. and A. B. Lawson (2012). *Bayesian Biostatistics*. John Wiley & Sons. DOI: 10.1002/9781119942412.

Lunn, D. et al. (2012). *The BUGS book: A practical introduction to Bayesian analysis*. Chapman and Hall/CRC. ISBN: 9781584888499.

Microsoft and S. Weston (2017). *foreach: Provides Foreach Looping Construct for R*. R package version 1.4.4. URL: https://CRAN.R-project.org/package=foreach.

National Center for Health Statistics (NCHS) (2011). *National Health and Nutrition Examination Survey Data*. URL: https://www.cdc.gov/nchs/nhanes/.

Novo, A. A. and J. L. Schafer (2010). *norm: Analysis of Multivariate Normal Datasets with Missing Values*. Version R package version 1.0-9.5. URL: http://CRAN.R-project.org/package=norm.

Pinheiro, J. et al. (2018). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-137. URL: https://CRAN.R-project.org/package=nlme.

Plummer, M. (2003). "JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling". *Proceedings of the 3rd International Workshop on*

Distributed Statistical Computing (DSC 2003). Ed. by K. Hornik, F. Leisch, and A. Zeileis. ISSN: 1609-395X.

Plummer, M. (2018). *rjags: Bayesian Graphical Models using MCMC*. R package version 4-8. URL: https://CRAN.R-project.org/package=rjags.

Plummer, M. et al. (2006). "CODA: Convergence Diagnosis and Output Analysis for MCMC". *R News*, **6**(1):7–11. URL: https://journal.r-project.org/archive/.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Rubin, D. B. (1976). "Inference and Missing Data". *Biometrika*, **63**(3):581–592. DOI: 10.2307/2335739.

— (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley Series in Probability and Statistics. Wiley. DOI: 10.1002/9780470316696.

— (2004). "The Design of a General and Flexible System for Handling Nonresponse in Sample Surveys". *The American Statistician*, **58**(4):298–302. DOI: 10.1198/000313004X6355.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis. ISBN: 9780412040610.

Terry M. Therneau and Patricia M. Grambsch (2000). *Modeling Survival Data: Extending the Cox Model*. New York: Springer. ISBN: 0-387-98784-3.

Therneau, T. M. (2015). *A Package for Survival Analysis in S*. version 2.38. URL: https://CRAN.R-project.org/package=survival.

Van Buuren, S. (2012). *Flexible Imputation of Missing Data*. Chapman & Hall/CRC Interdisciplinary Statistics. Taylor & Francis.

Van Buuren, S. and K. Groothuis-Oudshoorn (2011). "mice: Multivariate Imputation by Chained Equations in R". *Journal of Statistical Software*, **45**(3):1–67. DOI: 10.18637/jss.v045.i03.

White, I. R., P. Royston, and A. M. Wood (2011). "Multiple imputation using chained equations: Issues and guidance for practice". *Statistics in Medicine*, **30**(4):377–399. DOI: 10.1002/sim.4067.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN: 978-3-319-24277-4. URL: http://ggplot2.org.

7 General Discussion

In this thesis, we have presented a fully Bayesian approach for simultaneous analysis and imputation of incomplete data. Here, we summarize some of its advantages, highlight important assumptions made by the approach and its current implementation in the R package **JointAI**, and discuss possible extensions and directions for future work.

7.1 Summary of Advantages

Modelling the joint distribution as a sequence of univariate distributions yields several advantages. Contrary to MICE, the joint distribution always exists and, since the predictive distributions used to draw imputations are derived from this joint distribution, compatibility among imputation models and congeniality with the analysis model is assured. Factorization of the joint distribution using univariable conditional distributions facilitates its specification in settings with variables of mixed type, where the joint multivariate distribution does not have a known form. This avoids the need to make any general approximation, as is the case with joint model MI, thereby improving model fit. Moreover, since the analysis model is part of the factorization, non-linear associations between outcome and covariates, or between covariates, can be handled adequately. Choosing the analysis model as the first factor in the sequence of conditional distributions, i.e., the model for the outcome conditional on all covariates, prevents the need to include the outcome in any of the linear predictors of the other distributions, and makes analysis and imputation in settings with complex data structures straightforward.

In Chapters 3 and 4 we demonstrated that analysis with the sequential fully Bayesian approach also allows obtaining imputed values, which may be used for secondary analyses of the same data, and can thus be used in a multiple imputation framework. Data were imputed during the main analysis, and used in subsequent analyses of outcomes derived from the longitudinal outcome, or sensitivity analysis in subsets of the data. Naturally, congeniality of the imputation models with these secondary analysis models is then no longer assured, but bias is likely to be small, provided that the secondary analysis models are in a sense contained in the analysis model assumed during imputation. In Chapter 3 this is the case, since the outcome of the secondary models, gestational weight gain during different periods of pregnancy, is calculated from the longitudinal trajectories of weight, which were taken into account during imputation, and the same set of baseline covariates is used.

In settings where multiple related outcomes are of interest and their models share incomplete covariates, it may be desirable to model these outcomes jointly and to perform imputation in this joint model. This was the case in Chapter 4, where

the association of the consumption of sugar containing beverages of the mother during pregnancy with child body composition, consisting of repeated measures of BMI, and measurements of the fat mass index (FMI) and fat-free mass index (FFMI) at age six, were modelled jointly. Correlation between different outcomes can be taken into account by specification of a joint distribution for error terms of univariate outcomes and random effects in models for longitudinal outcomes, following the same principle used in (5.8) in the multivariate normal approach to multiple imputation applied in Chapter 5.

Another advantage of the Bayesian approach is its ability to take into account endogeneity of, usually longitudinal, covariates by jointly modelling their random effects, and potentially also the error terms, with the corresponding parts of the analysis model. Standard methods that are commonly applied after multiple imputation only specify models for the outcome, which implicitly assumes that all covariates are exogenous, and may lead to bias.

The implementation of the sequential Bayesian approach in the R package **JointAI** provides a convenient tool for researchers from various backgrounds, who are familiar with commonly used base R functions and do not have specific experience in Bayesian methodology or software for Bayesian inference, such as JAGS or WinBUGS.

7.2 Assumptions

Like all statistical methods, the sequential fully Bayesian approach is based on certain assumptions that must be met in order to obtain correct results. In the following, we highlight two assumptions that are crucial to the validity of the analysis: correctness of the model specification and the assumption of ignorable missingness.

7.2.1 Model Specification

An implicit assumption of the sequential Bayesian approach to handle missing data is that the model is correctly specified. This assumption is not particular for our approach but made in all parametric models. Even though in practice it may usually not be possible to specify a correct model, and the specified model can only approximate the "truth", it is important that this approximation is precise enough to avoid relevant bias.

In the model presented in the previous chapters, it is the joint distribution $p(\mathbf{y}, \mathbf{X}_{obs}, \mathbf{X}_{mis} \mid \boldsymbol{\theta})$ that needs to be specified correctly or at least needs to fit

the data well. Since the joint distribution is specified as a sequence of conditional distributions, this requirement translates to the conditional distributions.

When there are more than just a few incomplete variables, choosing conditional distributions that fit the data sufficiently well can become a tedious task. Variables in the linear predictor of the conditional models may have non-linear effects and interact with other variables, and which variables are included in a particular linear predictor depends on the order of the sequence of conditional distributions. Moreover, except for the conditional distribution for the first incomplete covariate in the sequence, models have incomplete variables in their linear predictor, complicating evaluation of model fit. For continuous variables, additionally, different choices for the error distribution need to be considered. It is unrealistic to expect such time-consuming model building to happen in practice. Many researchers will rely on the default choices set in the software implementation and, at most, consider alternative distributions for continuous, non-normal variables.

A useful implementation in a software package should, therefore, make it easy to use models that are flexible enough to fit a wide range of data sufficiently well to provide valid results. When each of the conditional distributions in the sequential fully Bayesian approach is specified correctly, the order chosen for the sequence is irrelevant for the validity of the results. In its current version (0.5.1) the R package **JointAI** provides multiple parametric options for imputation of continuous variables (using a normal, log-normal, gamma or beta distribution), but makes the default assumption that all associations between covariates are linear and do not interact with each other. Even though this assumption is appropriate in many settings, relaxing them may be necessary for some analyses.

7.2.2 Ignorable Missingness

The second crucial assumption necessary to obtain valid results when using the approach presented in the preceding chapters is that the missing data process is ignorable.

Imputation under the MAR assumption is attractive since MAR implies that non-responders are the same as responders in the sense that the conditional distribution of a variable $\mathbf{x}$ is the same for cases where $\mathbf{x}$ is missing and for cases where $\mathbf{x}$ is observed. This facilitates a straightforward specification of the posterior predictive distribution to impute missing values.

Throughout this thesis we have made the ignorability assumption, however, for some variables, like maternal smoking or alcohol use during pregnancy, the assumption of MAR, i.e., that the probability of the smoking or drinking status

being missing is independent of the true status, is questionable. It is widely known that the use of alcohol or cigarettes during pregnancy is not advised, and mothers disregarding this advice may feel guilty or do not dare to respond to the question due to fear of judgement.

When missing data are indeed MNAR conditional on the available information, ignoring the missingness process when analysing or imputing the data may lead to severe bias and faulty conclusions. This issue is not limited to missing values in covariates. Primarily in clinical trials, missingness of outcome values is often associated with what is investigated in the study. For instance, critically ill patients may feel too sick to fill in questionnaires regarding their quality of life, or patients who believe that the treatment or intervention they have been assigned to does not have the desired effect may leave the study. Especially in conjunction with the usually small number of covariates that is (available to be) included in the analysis, it is unlikely that the probability of a value being missing can be fully explained by the recorded information. In such settings, where MNAR is the more plausible assumption, omitting missing outcome values would lead to biased results.

7.3 Directions for Future Work

To reduce bias due to violation of the assumptions implied by the approach described in this thesis, future work should focus on extensions that help researchers to model their data appropriately. In this section, we briefly discuss some approaches that address the aforementioned assumptions of model fit and ignorability of the missingness mechanism.

7.3.1 Implementation of Flexible Models

A possible extension to improve the fit of the conditional distributions for incomplete covariates is to allow for a more general specification of the mean structure, relaxing the assumption of linear associations. One way to do this in an automated way that can be implemented in JAGS would be to provide the option to include pair-wise interactions between all covariates in the linear predictors of the models for incomplete covariates. Since this may lead to a large number of coefficients and, hence, likely leads to overfitting and convergence issues, it may be necessary to shrink parameters that contribute little to the model fit. In the Bayesian framework, shrinkage can be applied through specification of the prior distribution. A popular choice is the use of Bayesian ridge regression, where instead of using a vague normal prior for the regression coefficients $\boldsymbol{\alpha}$, a hyperprior

$\sigma_\alpha^2 \sim Inv - \chi^2(\nu, u^2)$, where u^2 has a gamma prior (Mallick and Yi 2013), is specified for the variance of this prior distribution. Alternatively, Bayesian elastic net priors, like the one used in Chapter 5, may be chosen.

To relax the assumption of a parametric error distribution for continuous variables, non-parametric Bayesian methods can be applied. In this framework, uncertainty about the probability density function G of an incompletely observed covariate $\mathbf{x}$ can be reflected by not specifying G directly, but assuming a prior probability model for G. A convenient and popular choice for such a prior probability model is the Dirichlet process prior, which allows the specification of an infinite mixture of simple parametric distributions, often normal distributions, where the values of the parameters of these parametric distributions are determined by the Dirichlet process (Escobar and West 1995; Müller et al. 2015). Due to the clustering property of the Dirichlet process, in practice, only a finite number of clusters are used, but since the number of clusters needed is determined by the data and does not need to be pre-specified, this approach facilitates automated and very flexible density estimation.

7.3.2 Evaluation of Model Fit

Even when flexible models are used, model fit should be evaluated to ensure the assumption of a correctly specified model is not violated. In Chapters 2 and 5, the fit of the conditional distributions was evaluated using posterior predictive checks (Gelman, Meng, et al. 1996).

A common approach to posterior predictive checks is to calculate a χ^2-type statistic

$$\sum_{i=1}^{N} \frac{\{x_i - \mathrm{E}(x_i \mid \boldsymbol{\theta})\}^2}{\mathrm{Var}(x_i \mid \boldsymbol{\theta})}.$$

The statistic is calculated twice, once for the observed values and once for the corresponding values sampled from the estimated posterior predictive distribution. If the specified model fits the observed data well, comparison of the two statistics should show that on average (over all iterations) one is not larger than the other.

To advocate and facilitate the use of posterior predictive checks for model fit when using the sequential Bayesian approach, functionality to automatically perform such model evaluation will be added to **JointAI** in the future.

7.3.3 Non-ignorable Missingness

Most popular software for handling incomplete data is limited to settings with MAR mechanisms, posing an often insurmountable hurdle for applied researchers

to perform imputation or analysis under the MNAR assumption.

It is not possible to provide a completely automated procedure since the handling of non-randomly missing data requires external information on the missingness mechanism. Under MNAR the unobserved data are assumed to have a different distribution (conditional on covariates) than the observed data, but since this distribution cannot be obtained from the observed data, assumptions about this unknown distribution need to be made by the researcher. Nevertheless, the use of MNAR analysis can be supported by providing software that allows the researcher to introduce his or her assumption into the analysis.

By factorizing the likelihood $p(\mathbf{X}, \mathbf{R} \mid \boldsymbol{\psi}, \boldsymbol{\theta}) = p(\mathbf{X} \mid \mathbf{R}, \boldsymbol{\theta})\ p(\mathbf{R} \mid \boldsymbol{\psi})$, i.e., using a *pattern mixture model* specification instead of the *selection model* factorization used in Chapter 1, MNAR can be modelled as a deviation from MAR:

$$\begin{aligned} p(\mathbf{X}, \mathbf{R} \mid \boldsymbol{\psi}, \boldsymbol{\theta}) &= p(\mathbf{X}_{obs}, \mathbf{X}_{mis} \mid \mathbf{R}, \boldsymbol{\theta})\ p(\mathbf{R} \mid \boldsymbol{\psi}) \\ &= p(\mathbf{X}_{mis} \mid \mathbf{X}_{obs}, \mathbf{R}, \boldsymbol{\theta}_{mis})\ p(\mathbf{X}_{obs} \mid \mathbf{R}, \boldsymbol{\theta}_{obs})\ p(\mathbf{R} \mid \boldsymbol{\psi}), \end{aligned}$$

where $(\boldsymbol{\theta}_{mis}, \boldsymbol{\theta}_{obs}) = g(\boldsymbol{\theta})$, $\boldsymbol{\theta}_{mis} = f(\boldsymbol{\theta}_{obs}, \boldsymbol{\Delta})$ and $\boldsymbol{\Delta}$ represents the deviation from MAR. Since no information about $\boldsymbol{\Delta}$ is available from the data, $\boldsymbol{\Delta}$ or a prior distribution $p(\boldsymbol{\Delta})$ must be specified, reflecting the analyst's hypothesis about the missing data mechanism.

The R package **JointAI** could be extended to enable users to specify for which incomplete variables they assume MNAR, and allow them to provide a value or distribution for $\boldsymbol{\Delta}$.

Offering researchers the opportunity to perform analysis and imputation under the assumption of MNAR with only slightly more effort than analysis under the assumption of MAR would require, will substantially improve the quality of research; researchers will be more likely to make appropriate assumptions and perform the sensitivity analyses required for those (untestable) assumptions.

7.4 Conclusion

Missing data occur in a wide range of studies but in practice their treatment is often not given adequate time and consideration, as they are regarded a frustrating complication that needs to be resolved before the actual research question can be answered. Applied researchers, hence, often prefer off-the-shelf solutions that require minimal knowledge about specific statistical approaches and little time to apply. This is in conflict with the careful consideration that is necessary to make appropriate assumptions about the missing data mechanism and the shape of the models used to impute missing values.

To support the correct use of statistical methodology, software is necessary that provides an accessible interface to valid statistical approaches. Such software needs to be flexible enough to handle data with different characteristics, non-restrictive enough to make it more likely that assumptions are met, and should provide options allowing straightforward evaluation of potential violations. Since the analysis of incomplete data involves untestable assumptions, additional functionality allowing sensitivity analyses about these assumptions is needed.

The implementation of the fully Bayesian approach to analysis of incomplete data in the R package **JointAI** is a first step to provide such software. Extensions, as discussed above and in Chapter 6 are necessary to fully reach this goal.

References

Escobar, M. D. and M. West (1995). "Bayesian density estimation and inference using mixtures". *Journal of the American Statistical Association*, **90**(430):577–588. DOI: 10.2307/2291069.

Gelman, A., X.-L. Meng, and H. Stern (1996). "Posterior predictive assessment of model fitness via realized discrepancies". *Statistica Sinica*, **6**(4):733–760.

Mallick, H. and N. Yi (2013). "Bayesian methods for high dimensional linear models". *Journal of Biometrics & Biostatistics*, **4**(S3):S1–005. DOI: 10.4172/2155-6180.S1-005.

Müller, P. et al. (2015). *Bayesian nonparametric data analysis*. Springer. DOI: 10.1007/978-3-319-18968-0.

8 Summary
Samenvatting
Zusammenfassung

Summary

Missing values are a pervasive problem in almost all kinds of studies. In large cohort studies, the type of study most often conducted in the field of epidemiology, missing observations in covariates pose the major challenge. Since measurements are taken in an uncontrolled environment, typically many covariates need to be considered as potential confounders to filter out unwanted influences that environmental factors may have on the estimates of interest. Due to the large number of variables measured and the fact that measurement often relies on participants recalling and reporting detailed information, large proportions of missing data are common in these types of studies. In light of the above, the research that forms this thesis focuses on the analysis of incomplete cohort study data where missingness is in the covariates.

Chapter 1 provides a brief introduction into the history of the most popular approach to handle incomplete data, multiple imputation (MI), and gives an overview of common approaches to perform multiple imputation or to directly perform inference with incomplete data. Moreover, concepts relevant to the analysis under the Bayesian paradigm are outlined.

The focus of Chapter 2 is the analysis of longitudinal data with incomplete baseline covariates. We describe a fully Bayesian approach to analyse and impute data in this setting and discuss a number of naive and more sophisticated approaches to impute such data in wide format using multiple imputation with chained equations (MICE). Results from the analysis of the motivating dataset from the Generation R Study as well as two simulation studies demonstrate that with MICE omission of the outcome from the imputation models, or even the use of simple summaries, can lead to severe bias. Only when more sophisticated summaries of the outcome, which captured important features associated with the missingness, were used, bias was negligible. Since it is generally not known which features of the outcome are relevant to the missingness, the fully Bayesian approach, in which the outcome is included automatically and implicitly without the need to summarize it, and which provided unbiased results throughout all analyses, is the preferred method for imputation of incomplete baseline covariates in longitudinal data.

The fully Bayesian approach is applied to data from the Generation R Study in Chapter 3, in which the association between gestational weight (gain) and dietary patterns is investigated. In the primary analysis, a stratified Bayesian linear mixed model is fitted to repeated measures of gestational weight, and missing covariate values are imputed. Extracting these imputed values and creating multiple completed datasets allows performing secondary analyses of gestational weight gain

during different periods of pregnancy as well as sensitivity analyses using the idea of multiple imputation.

Chapter 4 provides another application of the proposed approach to data from the Generation R Study. The association between child body composition and maternal sugar containing beverage consumption during pregnancy is investigated. Three different measures of child body composition are of interest in the primary analysis: BMI, measured repeatedly until six years of age, and fat mass index (FMI) and fat free mass index (FFMI), measured at the age of six. The three outcomes are modelled jointly in the Bayesian framework using a linear mixed model for BMI and linear models for FMI and FFMI, and missing values in covariates are imputed simultaneously. Again, imputed values are extracted to perform additional analyses, in this case analyses on subgroups within the data.

The Bayesian approach is further extended to settings with time-varying covariates in Chapter 5. Additional challenges that arise with time-varying covariates, such as the functional form of the association between outcome and covariate, and potential endogeneity, are investigated. The previously described Bayesian approach, extended to settings with time-varying covariates, is compared to joint model multiple imputation using a multivariate normal distribution, with regards to its ability to handle the additional challenges. Simulation studies show that misspecification of the functional form or misspecification of an endogenous covariate as exogenous can lead to severe bias. Even though joint model MI assumes endogeneity, the subsequent analysis of the imputed data usually assumes exogeneity for all covariates, and associations between outcome and time-varying covariates are assumed to be linear during imputation. Since the (extended) fully Bayesian approach allows inclusion of endogenous covariates, flexible non-linear associations and performs simultaneous analysis and imputation, it is the superior approach in this setting.

Chapter 6 describes the implementation of the fully Bayesian approach in the R package **JointAI** and illustrates the use of the package by means of various examples.

This thesis is concluded in Chapter 7 with a short summary of the advantages of the fully Bayesian approach, discusses implications by the assumptions made by the approach and explores extensions and directions for future work.

Samenvatting

Ontbrekende waardes zijn een veelvoorkomend probleem in bijna alle soorten studies. In grote cohortstudies, het meest gebruikte type studie in de epidemiologie, vormen ontbrekende waardes in de covariabelen het grootste probleem. Omdat metingen in een niet gecontroleerde omgeving worden gedaan moet er doorgaans met veel covariabelen rekening gehouden worden om ongewenste invloed van omgevingsfactoren op de schattingen van belangrijke parameters te voorkomen. Door het grote aantal gemeten variabelen en door het feit dat metingen vaak vereisen dat deelnemers zich gedetailleerde informatie herinneren en deze informatie ook rapporteren, komen grote hoeveelheden ontbrekende waardes vaak voor in dit type studies. Naar aanleiding van het bovenstaande, focust het onderzoek in deze dissertatie zich op de analyse van incomplete data uit cohortstudies met ontbrekende waardes in de covariabelen.

Hoofdstuk 1 geeft een korte inleiding in de geschiedenis van de meest populaire manier om onvolledige gegevens te benaderen; multiple imputatie (MI), en geeft een overzicht van gebruikelijke methodes om MI uit te voeren of incomplete data direct te analyseren. Bovendien worden concepten geschetst die relevant zijn in het kader van het Bayesiaanse paradigma.

De focus van hoofdstuk 2 is de analyse van longitudinale gegevens met onvolledige "baseline" covariaten. We beschrijven een volledig Bayesiaanse benadering om data in deze situatie te analyseren en imputeren en bespreken zowel een aantal naïeve, als ook meer verfijnde benaderingen om dergelijke gegevens in het zogenaamde "wide format" te imputeren met "Multiple Imputation with Chained Equations" (MICE). Resultaten van de analyse van de motiverende dataset uit de Generation R studie en twee simulatiestudies tonen aan dat met MICE het niet includeren van de uitkomst in de imputatiemodellen, of zelfs het gebruik van eenvoudige samenvattingen van de uitkomst, kan leiden tot ernstige systematische fouten. Alleen wanneer meer uitgebreide samenvattingen van de uitkomst werden gebruikt, welke belangrijke karakteristieken van de "missingness" bevatten, was de systematische fout verwaarloosbaar. Omdat in het algemeen niet bekend is welke kenmerken van de uitkomst relevant zijn voor de "missingness", heeft de volledig Bayesiaanse benadering, waarin de uitkomst automatisch en impliciet in de imputatie geïncludeerd is, zonder de noodzaak om het samen te vatten, en die tijdens alle analyses resultaten zonder systematische fouten opleverde, de voorkeur voor imputatie van onvolledige baseline gegevens in longitudinale studies.

De Bayesiaanse benadering wordt toegepast op gegevens van de Generation R studie in hoofdstuk 3, waarin de associatie tussen (de toename van) gewicht tijdens zwangerschap en voedingspatronen wordt onderzocht. In de primaire anal-

yse wordt een gestratificeerd Bayesiaans lineair gemengd model toegepast op herhaalde metingen van zwangerschapsgewicht en worden ontbrekende waardes in de covariabelen geïmputeerd. Het extraheren van deze geïmputeerde waardes en het creëren van meerdere volledige datasets maakt het mogelijk om in het kader van MI secundaire analyses van gewichtstoename tijdens verschillende periodes van de zwangerschap, evenals gevoeligheidsanalyses, uit te voeren.

Hoofdstuk 4 laat een verdere toepassing zien van de voorgestelde methode op data uit de Generation R studie. De associatie tussen lichaamssamenstelling van kinderen en de consumptie van suikerhoudende dranken door hun moeders tijdens de zwangerschap wordt onderzocht. Er zijn drie belangrijke maten van lichaamssamenstelling in de primaire analyse: BMI, herhaald gemeten tot zes jaar, vetmassa index (FMI) en vetvrije massa index (FFMI), beiden gemeten op zesjarige leeftijd. De drie uitkomsten worden gezamenlijk gemodelleerd in het Bayesiaanse kader met behulp van een lineair gemengd model voor BMI en lineaire modellen voor FMI en FFMI, waarbij ontbrekende waardes in covariaten binnen dezelfde procedure worden geïmputeerd. Wederom worden deze geïmputeerde waarden geëxtraheerd om extra analyses, in dit geval in subgroepen, uit te voeren.

De Bayesiaanse benadering wordt in hoofdstuk 5 verder uitgebreid naar situaties met covariaten die in de tijd variëren. Bijkomende uitdagingen die horen bij tijds variërende covariaten, zoals de functionele vorm van de associatie tussen uitkomst en covariaat en mogelijke endogeniteit, worden onderzocht. De eerder beschreven Bayesiaanse benadering, uitgebreid naar situaties met in de tijd varieerde covariaten, wordt vergeleken met "joint model" MI met behulp van een multivariate normale verdeling, op het vermogen om de additionele uitdagingen aan te gaan. Simulatiestudies tonen aan dat een verkeerde specificatie van de functionele vorm of een specificatie van een endogene covariabel als exogeen, tot ernstige systematische fouten kan leiden. Hoewel "joint model" MI endogeniteit veronderstelt, neemt de daaropvolgende analyse van de geïmputeerde gegevens meestal exogeniteit aan voor alle covariaten, en er wordt tijdens de imputatie van uitgegaan dat associaties tussen uitkomst en tijdsvariërende covariaten lineair zijn. Omdat de (uitgebreide) volledig Bayesiaanse benadering zowel rekening kan houden met endogeniteit van covariabelen en flexibele niet-lineaire associaties, als analyse en imputatie simultaan uitvoert, is dit de superieure methode in deze situatie.

Hoofdstuk 6 beschrijft de implementatie van de volledig Bayesiaanse benadering in het R pakket **JointAI** en illustreert het gebruik van het pakket aan de hand van verschillende voorbeelden.

Dit proefschrift wordt afgesloten in hoofdstuk 7 met een korte samenvatting van de voordelen van de volledig Bayesiaanse methode, de implicaties van de aannames

die gemaakt worden in deze benadering en mogelijke verbeteringen en onderwerpen voor toekomstig werk.

Zusammenfassung

Fehlende Werte sind ein allgegenwärtiges Problem in vielerlei Studien. In großen Kohortenstudien, dem Studientyp, der in der Epidemiologie am häufigsten durchgeführt wird, stellen fehlende Werten in Kovariablen die größte Herausforderung dar. Da Daten nicht in einem kontrollierten Umfeld erhoben werden, müssen üblicherweise viele Kovariablen als potentielle Störfaktoren berücksichtigt werden, um so zumindest einen Teil der unerwünschten Einflüsse herauszufiltern, die das Ergebnis verzerren können. Aufgrund der großen Anzahl von Werten, die deshalb erhoben werden, und da es oft notwendig ist, dass Probanden sich an detaillierte Informationen erinnern und diese auch wiedergeben, ist ein hoher Prozentsatz fehlender Werte keine Seltenheit. Die wissenschaftlichen Arbeiten, die dieser Dissertation zugrunde liegen, konzentrieren sich daher auf die Analyse unvollständiger Daten aus Kohortenstudien, wobei die fehlenden Werte in den Kovariablen auftreten.

Kapitel 1 gibt eine kurze Einführung in die Geschichte der derzeit bekanntesten Methode zum Umgang mit fehlenden Werten: „Multiple Imputation“ (MI), und eine Übersicht häufig verwendeter Methoden, mit denen entweder MI oder direkt Inferenz aus unvollständigen Daten durchgeführt werden können. Außerdem werden Konzepte skizziert, die im Zusammenhang mit bayesianischer Statistik relevant sind.

Der Fokus in Kapitel 2 liegt auf der Analyse longitudinaler Daten mit unvollständigen zeitunabhängigen Kovariablen. Wir beschreiben eine vollständig bayesianische Methode um derartige Daten zeitgleich zu analysieren und zu imputieren, und erörtern einige einfache sowie kompliziertere Ansätze um diese Daten, in horizontalem Format, mit der gebräuchlicheren Methode „Multiple Imputation using Chained Equations“ (MICE), d.h. durch eine Reihe von univariaten Modellen zu imputieren. Ergebnisse aus der Analyse des motivierenden Datensatzes aus der „Generation R“ Studie sowie aus zwei Simulationsstudien zeigen, dass bei MICE das Weglassen der Zielvariablen aus den Imputationsmodellen, oder auch die Verwendung einfacher Zusammenfassungen der Zielvariablen zu schwerwiegenden systematischen Fehlern führen kann. Nur wenn komplexere Zusammenfassungen der Zielvariablen verwendet wurden, die die im Zusammenhang mit den fehlenden Werten wichtigen Merkmale erfassen, war der systematische Fehler vernachlässigbar. Bei der vollständig bayesianischen Methode wird die Zielvariable automatisch und implizit berücksichtigt ohne dass eine Zusammenfassung

notwendig ist. Da im Allgemeinen nicht bekannt ist welche Merkmale relevant sind und die bayesianische Methode in allen Analysen korrekte Ergebnisse lieferte, ist sie die überlegene Methode, um fehlende Werte in zeitkonstanten Kovariablen in longitudinalen Datensätzen zu imputieren bzw. solche Daten zu analysieren.

Die vollständig bayesianische Methode wird in Kapitel 3 auf eine Fragestellung aus der „Generation R“ Studie angewendet, in der die Beziehung zwischen Ernährungsmustern und Gewicht bzw. Gewichtszunahme während der Schwangerschaft untersucht wird. Für die primäre Fragestellung wird ein stratifiziertes, bayesianisches, lineares gemischtes Model für die wiederholten Messungen des Gewichts der Mutter aufgestellt, womit gleichzeitig fehlende Werte in Kovariablen imputiert werden. Dem Konzept von MI folgend werden mit diesen imputierten Werten vervollständigte Versionen des ursprünglichen Datensatzes generiert, und somit die Analyse der sekundären Frage nach der Beziehung zwischen Ernährung und Gewichtszunahme in verschiedenen Phasen der Schwangerschaft sowie Sensitivitätsanalysen ermöglicht.

Kapitel 4 stellt eine weitere Anwendung der vorgeschlagenen bayesianischen Methode auf Daten aus der „Generation R“ Studie dar. Der Zusammenhang zwischen der Menge von mit Zucker gesüßten Getränken, die Frauen während der Schwangerschaft zu sich nehmen, und dem Körperbau ihrer Kinder nach der Geburt wird untersucht. Drei verschiedene Kennzahlen des Körperbaus sind in der primären Analyse von Interesse: BMI, mehrfach gemessen von der Geburt ab bis zum Alter von sechs Jahren, Fett-Masse-Index (FMI) und Fettfreie-Masse-Index (FFMI), gemessen im Alter von sechs Jahren. Zugleich mit der gemeinsamen Modellierung der drei Zielvariablen in einem bayesianischen Model werden fehlende Werte in Kovariablen imputiert. Dabei werden ein lineares gemischtes Model für BMI sowie lineare Regressionsmodelle für FMI und FFMI verwendet. Zur Durchführung weitere Analysen, in diesem Fall in Untergruppen, werden auch hier die imputierten Werte extrahiert.

In Kapitel 5 wird die bayesianische Methode auf Situationen mit zeitabhängigen Kovariablen erweitert. Dabei werden Herausforderungen untersucht, wie sie bei zeitabhängigen Kovariablen entstehen. In diesem Falle die funktionelle Form der Beziehung zwischen Zielvariable und Kovariable, sowie die mögliche Endogenität der Kovariable. Anhand von Simulationsstudien wird die zuvor beschriebenen, für zeitabhängigen Kovariablen erweiterten, bayesianische Methode mit einer alternativen Methode verglichen, in der fehlende Werte mit Hilfe einer gemeinsamen multivariaten Normalverteilung imputiert werden („joint model“ MI). Es ergibt sich, dass Fehlspezifikation der funktionellen Form oder Fehlspezifikation einer endogenen Kovariablen als exogen zu schwerwiegenden systematischen Fehlern führen

kann. Obwohl "joint model" MI Endogenität annimmt, ist dies bei den meisten Methoden, die im Anschluss auf die imputierten Daten angewendet werden nicht der Fall. Außerdem wird die Beziehung zwischen Kovariable und Zielvariabel bei "joint model" MI als linear angenommen. Da bei der erweiterten vollständig bayesianischen Methode endogenen Kovariablen sowie flexible Formen für die funktionelle Beziehung zwischen Zielvariabel und Kovariablen berücksichtigt werden können, ist sie die überlegene Methode für derartige Daten.

Kapitel 6 beschreibt die Implementierung der vorgestellten vollständig bayesianischen Methode im R Paket **JointAI** und illustriert die Verwendung dieses Paketes anhand zahlreicher Beispiele.

Das abschließende Kapitel 7 fasst die Vorteile der von uns verwendeten bayesianischen Methode zusammen, und erläutert wesentliche Konsequenzen der Annahmen dieser Methode. Zusätzlich werden Ideen aufgezeigt wie die Methode und ihre Implementierung erweitert und verbessert werden kann.

Appendix: PhD Portfolio, Curriculum Vitae & Acknowledgements

PhD Portfolio

Name PhD student:	Nicole S. Erler
Department:	Department of Biostatistics
	Erasmus Medical Center Rotterdam
PhD period:	2014 – 2018
Promotors:	Prof.dr. Dimitris Rizopoulos
	Prof.dr. Emmanuel M.E.H. Lesaffre

Presentations

At International Conferences	year	ECTS
35th ISCB conference, Vienna, Austria	2014	1.0
5th IBC Channel Meeting, Nijmegen, the Netherlands	2015	1.0
8th IBS EMR conference, Kapadokya, Turkey	2015	1.0
36th ISCB conference, Utrecht, the Netherlands	2015	1.0
28th IBS conference, Victoria, Canada	2016	1.0
37th ISCB conference, Birmingham, United Kingdom	2016	1.0
Joint Modeling and Beyond, Hasselt, Belgium	2016	1.0
9th IBS EMR conference, Thessaloniki, Greece	2017	1.0
38th ISCB conference, Vigo, Spain	2017	1.0
29th IBS conference, Barcelona, Spain	2018	1.0

Other	year	ECTS
Heidelberger Kolloquium, Heidelberg, Germany	2015	1.0
Erasmus Statistics Day, Rotterdam, the Netherlands	2016	1.0

Teaching

Teaching	year	ECTS
Missing Values in Clinical Research (EP16), NIHES	2018	15

Assisting	year	ECTS
SPSS practical, MSc Medicine, Erasmus MC	2014 – 2018	1.3
Biostatistical Methods II (EP03), NIHES	2014 – 2018	1.4
Repeated Measurements (CE08), NIHES	2016	1.0

Courses

	year	ECTS
GAMLSS in Action (Erasmus MC, Rotterdam)	2014	0.6
INLA (IBC Channel Meeting, Nijmegen)	2015	0.1
Causal Questions and Principled Answers (ISCB, Vigo, Spain)	2017	0.3
Microbiomics I (MolMed, Erasmus MC)	2017	0.6
Bayesian non-parametric methods (Leuven, Belgium)	2017	1.0
Scientific Integrity (Erasmus MC)	2017	0.3
Network Meta Analysis with R (IBC, Barcelona, Spain)	2018	0.3
Bayesian Parametric and Nonparametric Methods for Missing Data and Causal Inference (Leuven, Belgium)	2018	1.0
DSMB training: To stop or not to stop, is that the question? (Erasmus MC)	2018	0.1

Seminars and Workshops

	year	ECTS
Erasmus MC bi-weekly Biostatistics / CQM seminar	2014 – 2018	1.0
Erasmus MC bi-weekly ErasmusAGE seminar	2014 – 2017	1.0
BMS-Aned Meeting, Rotterdam, the Netherlands	2014	0.2
BMS-Aned Meeting, Leiden, the Netherlands	2017	0.2
BMS-Aned Meeting, Amsterdam, the Netherlands	2017	0.2
BMS-Aned Meeting, Rotterdam, the Netherlands	2018	0.2

Consulting

	year	ECTS
Statistical consultant in the Erasmus MC CPO system	2014 – 2018	16.0

Curriculum Vitae

Education

since 2014	PhD in Biostatistics Erasmus University Rotterdam, Rotterdam, the Netherlands
2013 – 2014	DSc Epidemiology Netherlands Institute of Health Sciences, Erasmus Medical Center, Rotterdam, the Netherlands
2006 – 2012	Diplom Statistics Ludwig-Maximilians-Universität München, Munich, Germany

Work Experience

since 03/2013	Scientific Researcher Department of Biostatistics, Erasmus Medical Center, Rotterdam, the Netherlands
12/2012 – 02/2013	Scientific Researcher Rotterdam Ophthalmic Institute, Rotterdam, the Netherlands
11/2010 – 06/2011	Student Assistant Central Examination Office for Natural Sciences, Ludwig-Maximilians-Universität München, Munich, Germany
07/2008 – 11/2010	Student Assistant to Prof. Dr. Augustin Department of Statistics, Ludwig-Maximilians-Universität München, Munich, Germany

Awards

2016	Student Award of the International Society for Clinical Biostatistics
2015	Student Award of the International Biometric Society (Eastern Mediterranean Region)

Collaborations and Consulting

since 2017	Collaboration with the Department of Gastroenterology and Hepatology, Erasmus Medical Center statistical support for the planning and analysis of clinical trials, preparation of scientific manuscripts, analysis of data with advanced statistical methods
since 2014	Statistical consultant in the Consultation Center for Patient Oriented research, Erasmus Medical Center providing statistical support to researchers within the Erasmus Medical Center
2013 – 2017	Collaboration with ErasmusAGE (research group within the Department of Epidemiology, Erasmus Medical Center) providing statistical support and collaboration on various research projects with focus on nutrition, lifestyle and healthy ageing

Publications

- Alferink, L. J., Kiefte-de Jong, J. C., Erler, N. S., de Knegt, R. J., Hoorn, E. J., Ikram, M. A., Janssen, H. L., Metselaar, H. J., Franco, O. H., and Darwish Murad, S. Diet-dependent acid load – the missing link between an animal protein-rich diet and non-alcoholic fatty liver disease? In press, 2019.

- Alferink, L. J., Radjabzadeh, D., Erler, N. S., Vojinovic, D., Gomez, C. M., Uitterlinden, A. G., de Knegt, R. J., Amin, N., Ikram, M. A., Janssen, H. L., Metselaar, H. J., van Duijn, C. M., Kraaij, R., and Murad, S. D. Microbiomics, metabolomics, predicted metagenomics and hepatic steatosis in a population-based study of 1355 adults. Manuscript submitted for publication, 2019.

- Alferink, L. J., Trajanoska, K., Erler, N. S., Schoufour, J. D., de Knegt, R. J., Ikram, M. A., Janssen, H. L., Franco, O. H., Metselaar, H. J., Rivadeneira, F., and Murad, S. D. Non-alcoholic fatty liver disease in The Rotterdam Study: about muscle mass, sarcopenia, fat mass and fat distribution. *Journal of Bone and Mineral Research*, 2019.

- Boor, P. P., Sideras, K., Biermann, K., Levink, I. J., Mancham, S., Erler, N. S., van Eijck, C. H., Bruno, M. J., Sprengers, D., Zang, X., and Kwekkeboom, J. HHLA2 is expressed in pancreatic and ampullary cancers and increased expression is associated with better post-surgical prognosis. Manuscript in preparation, 2019.

- Coopmans, E. C., Schneiders, J. J., El-Sayed, N., Erler, N. S., Muhammad, A., Hofland, L. J., Petrossians, P., van der Lely, A.-J., and Neggers, S. J. T2-signal intensity, SST receptor expression and first-generation somatostatin analogues efficacy predict hormone and tumor responses to pasireotide in acromegaly. Manuscript in preparation, 2019.
- De Wijs, L. E., Bosma, A., Erler, N. S., Hollestein, L. M., Nijsten, T., Spuls, P. I., and Hijnen, D.-J. Dupilumab treatment for moderate-to-severe atopic dermatitis: daily practice data. Manuscript submitted for publication, 2019.
- Erler, N. S., Rizopoulos, D., Jaddoe, V. W., Franco, O. H., and Lesaffre, E. M. Bayesian imputation of time-varying covariates in linear mixed models. *Statistical Methods in Medical Research*, 28(2):555 – 568, 2019.
- Erler, N. S., Rizopoulos, D., and Lesaffre, E. M. JointAI: Joint Analysis and Imputation of Incomplete Data in R. Manuscript in preparation, 2019.
- Hu, C., Duijts, L., Erler, N. S., Elbert, N. J., Piketty, C., Bourdès, V., Blanchet-Réthore, S., de Jongste, J. C., Jaddoe, V. W., Pasmans, S. G., Felix, J. F., and Nijsten, T. Most associations of early life environmental exposures and genetic risk factors poorly differentiate between eczema phenotypes. The Generation R Study. *British Journal of Dermatology*, 2019.
- Koolhaas, C. M., Kocevska, D., te Lindert, B. H. W., Erler, N. S., Franco, O. H., Luik, A. I., and Tiemeier, H. Objectively measured sleep and body mass index: a prospective bidirectional study in middle-aged and older adults. *Sleep Medicine*, 57:43 – 50, 2019.
- Oey, R. C., Aarts, P., Erler, N. S., Metselaar, H. J., Lakenman, P. L., van der Ree, S. R. B., van Kemenade, M. C., van Buuren, H. R., and de Man, R. A. Limited efficacy of recently proposed easl criteria for identifying patients with malnutrition and sarcopenia in a population screened for liver transplantation. Manuscript submitted for publication, 2019.
- Oey, R. C., Buck, L. E., Erler, N. S., van Buuren, H. R., and de Man, R. A. The efficacy and safety of rifaximin-α: a 2-year observational study of overt hepatic encephalopathy. Manuscript submitted for publication, 2019.
- Oey, R. C., van Tilburg, L., Erler, N. S., Metselaar, H. J., Spaander, M. C., van Buuren, H. R., and de Man, R. A. The yield and safety of screening colonoscopy in patients evaluated for liver transplantation. *Hepathology*, 2019.
- Overbeek, K. A., Kamps, A., van Riet, P. A., Marco, M. D., Zerboni, G., van Hooft, J. E., Carrara, S., Ricci, C., Gonda, T. A., Schoon, E., Polkowski, M., Beyer, G., Honkoop, P., van der Waaij, L. A., Casadei, R., Capurso, G., Erler, N. S., Bruno, M. J., Bleiker, E. M., and Cahen, D. L. On behalf of the

Pancreatic CYst Follow-up: an International Collaboration (PACYFIC) study work group. Pancreatic cyst surveillance imposes low psychological burden. Manuscript submitted for publication, 2019.

- Rauwers, A. W., Voor in 't holt, A. F., Buijs, J. G., de Groot, W., Erler, N. S., Bruno, M. J., and Vos, M. C. A nationwide risk analysis of duodenoscope and linear echoendoscope contamination. Manuscript submitted for publication, 2019.
- Van der Veer, M. A., Nangrahary, N., Hesselink, D. A., Erler, N. S., Metselaar, H. J., van Gelder, T., and Darwish Murad, S. High Intra-Patient Variability in Tacrolimus Exposure Is Not Associated with Immune-Mediated Graft Injury after Liver Transplantation. *Transplantation*, 2019.
- Van Riet, P. A., Larghi, A., Attili, F., Rindi, G., Nguyen, N. Q., et al. A multicenter randomized trial comparing a 25-gauge EUS fine-needle aspiration device with a 20-gauge EUS fine-needle biopsy device. *Gastrointestinal Endoscopy*, 89(2):329 – 339, 2019.
- Alferink, L. J., Kiefte-de Jong, J. C., Erler, N. S., Veldt, B. J., Schoufour, J. D., de Knegt, R. J., Ikram, A. M., Metselaar, H. J., Janssen, H. L., Franco, O. H., and Darwish Murad, S. Association of dietary macronutrient composition and non-alcoholic fatty liver disease in an ageing population: The Rotterdam Study. *Gut*, 2018.
- Beelen, E. M., van der Woude, C. J., Pierik, M., Hoentjen, F., de Boer, K., Oldenburg, B., van der Meulen, A., Ponsioen, C., Dijkstra, G., Gijsbers, A., Erler, N. S., Schouten, W., de Vries. On behalf of the Dutch Initiative on Crohn, A. C., and (ICC), C. Decreasing trends in ileocecal resections and postoperative surgical recurrence in Crohn's disease: a nationwide cohort study in the Netherlands. Manuscript submitted for publication, 2018.
- Buschow, S. I., Biesta, P. J., Groothuismink, Z. M., Erler, N. S., Vanwolleghem, T., Ho, E., Najera, I., Ait-Goughoulte, M., de Knegt, R. J., Boonstra, A., and Woltman, A. M. TLR7 polymorphism, sex and chronic HBV infection influence plasmacytoid DC maturation by TLR7 ligands. *Antiviral Research*, 157:27 – 37, 2018.
- De Jonge, E. A., Rivadeneira, F., Erler, N. S., Hofman, A., Uitterlinden, A. G., Franco, O. H., and Kiefte-de Jong, J. C. Dietary patterns in an elderly population and their relation with bone mineral density: The Rotterdam Study. *European Journal of Nutrition*, 57(1):61–73, 2018.
- Didden, P., Reijm, A. N., Erler, N. S., Wolters, L., Tang, T., ter Borg, P. C., Leeuwenburgh, I., Bruno, M. J., and Spaander, M. C. Fully vs. partially covered

selfexpandable metal stent for palliation of malignant esophageal strictures: a randomized trial (the COPAC study). *Endoscopy*, 50(10):961–971, 2018.

- Kamst, N. W., Mathijssen, I. M., Erler, N. S., and van Veelen, M.-L. C. Patient Reported Outcome Measurements in patients after scaphocephaly correction (PROM). Manuscript submitted for publication, 2018.
- Kanis, S. L., Modderman, S. C., Escher, J., Erler, N. S., Beukers, R., et al. Long-term health outcomes of 1000 children born to mothers with Inflammatory Bowel Disease in the anti-TNF-α era. Manuscript submitted for publication, 2018.
- Oey, R. C., de Man, R. A., Erler, N. S., Verbon, A., and van Buuren, H. R. Microbiology and antibiotic susceptibility patterns in spontaneous bacterial peritonitis: A study of two Dutch cohorts at a 10-year interval. *United European Gastroenterology Journal*, 6(4), 2018.
- Oey, R. C., de Wit, K., Moelker, A., Atalik, T., Van Delden, O. M., Maleux, G., Erler, N. S., Takkenberg, B., De Man, R. A., Nevens, F., and van Buuren, H. R. Variable efficacy of transjugular intrahepatic portosystemic shunt (TIPSS) in the management of ectopic variceal bleeding: a multicenter retrospective study. *Alimentary Pharmacology & Therapeutics*, 48(9):975 – 983, 2018.
- Oey, R. C., van Buuren, H. R., de Jong, D. M., Erler, N. S., and de Man, R. A. Bacterascites: A study of clinical features, microbiological findings, and clinical significance. *Liver International*, 38(12):2199 – 2209, 2018.
- Ottenhof, M. J., Fani, L., Erler, N. S., Castricum, J., Obdam, I. F., van der Vaart, T., Kushner, S. A., de Wit, M.-C. Y., Elgersma, Y., and Tulen, J. Within-subject consistency of paired associative stimulation as assessed by linear mixed models. *bioRxiv*, 2018.
- Santosaningsih, D., Erikawati, D., Hakim, I. A., Santoso, S., Hidayat, M., Suwendha, A. H., Puspitasari, V., Irhamni, I., Kuntaman, K., van Arkel, A., Terlouw, L., Oudenes, N., Willemse-Erix, D., Snijders, S., Erler, N. S., Verbrugh, H. A., and Severin, J. Reducing transmission of methicillin-resistant Staphylococcus aureus in a surgical ward of a resource-limited hospital in Indonesia: an intervention study. Manuscript submitted for publication, 2018.
- Tielemans, M. J., Erler, N. S., Franco, O. H., Jaddoe, V. W., Steegers, E. A., and Kiefte-de Jong, J. C. Dietary acid load and blood pressure development in pregnancy: the Generation R Study. *Clinical Nutrition*, 37(2):597–603, 2018.
- Vergouwe, F. W., IJsselstijn, H., Biermann, K., Erler, N. S., Wijnen, R. M., Bruno, M. J., and Spaander, M. C. High Prevalence of Barrett's Esophagus and Esophageal Squamous Cell Carcinoma after Repair of Esophageal Atresia. *Clinical Gastroenterology and Hepatology*, 16(4), 2018.

- Brouwer, J., Dolhain, R. J., Hazes, J. M., Erler, N. S., Visser, J. A., and Laven, J. S. Decline of ovarian function in patients with rheumatoid arthritis: serum anti-Müllerian hormone levels in a longitudinal cohort. Manuscript submitted for publication, 2017.
- Cepeda, M., Schoufour, J. D., Erler, N. S., Marques-Vidal, P., and Franco, O. H. Influence of lifestyle markers and meteorological factors on the seasonality of cardiovascular risk factors: The Rotterdam Study. Manuscript submitted for publication, 2017.
- Garcia, A., Erler, N. S., Jadooe, V. W., Tiemeier, H., van den Hooven, E. H., Franco, O. H., Rivadeneira, F., and Voortman, T. 25-hydroxyvitamin D concentrations during fetal life and bone health in children aged 6 years: a population-based prospective cohort study. *The Lancet Diabetes & Endocrinology*, 5(5):367–376, 2017.
- Jaspers, L., Erler, N. S., Fauser, B. C., Laven, J. S., Franco, O. H., and Kavousi, M. Towards a life-course approach in women's healthy ageing: fertile lifespan characteristics are associated with a healthy ageing score in postmenopausal women of the Rotterdam Study. Manuscript submitted for publication, 2017.
- Jaspers, L., Kavousi, M., Erler, N. S., Hofman, A., Laven, J. S., and Franco, O. H. Fertile lifespan characteristics and all-cause and cause-specific mortality among postmenopausal women: The Rotterdam Study. *Fertility and Sterility*, 107(2):448–456.e1, 2017.
- Jaspers, L., Schoufour, J. D., Erler, N. S., Darweesh, S. K., Portegies, M. L., Sedaghat, S., Lahousse, L., Brusselle, G. G., Stricker, B. H., Tiemeier, H., Ikram, A. M., Laven, J. S., Franco, O. H., and Kavousi, M. Development of a Healthy Aging Score in the Population-Based Rotterdam Study: Evaluating Age and Sex Differences. *Journal of the American Medical Directors Association*, 18(3):276.e1–276.e7, 2017.
- Jen, V., Erler, N. S., Tielemans, M. J., Braun, K. V., Jaddoe, V. W., Franco, O. H., and Voortman, T. Mothers' intake of sugar-containing beverages during pregnancy and body composition of their children during childhood: The Generation R Study. *The American Journal of Clinical Nutrition*, 105(4):834–841, 2017.
- Muka, T., Blekkenhorst, L. C., Lewis, J. R., Prince, R. L., Erler, N. S., Hofman, A., Franco, O. H., Rivadeneira, F., and Kiefte-de Jong, J. C. Dietary fat composition, total body fat and regional body fat distribution in two Caucasian populations of middle-aged and older adult women. *Clinical Nutrition*, 36(5):1411 – 1419, 2017.

- Schoufour, J. D., Erler, N. S., Jaspers, L., Kiefte-de Jong, J. C., Voortman, T., Ziere, G., Lindemans, J., Klaver, C. C., Tiemeier, H., Stricker, B., Ikram, A. M., Laven, J. S., Brusselle, G. G., Rivadeneira, F., and Franco, O. H. Design of a frailty index among community living middle-aged and older people: The Rotterdam Study. *Maturitas*, 97:14–20, 2017.
- Braun, K. V., Erler, N. S., Kiefte-de Jong, J. C., Jaddoe, V. W., van den Hooven, E. H., Franco, O. H., and Voortman, T. Dietary intake of protein in early childhood is associated with growth trajectories between 1 and 9 years of age. *The Journal of Nutrition*, 146(11):2361–2367, 2016.
- Egal, M., Erler, N. S., De Geus, H. R., Van Bommel, J., and Groeneveld, J. A. Targeting oliguria reversal in goal-directed hemodynamic management does not reduce renal dysfunction in perioperative and critically ill patients: A systematic review and meta-analysis. *Anesthesia & Analgesia*, 122(1):173–185, 2016.
- Erler, N. S., Rizopoulos, D., van Rosmalen, J., Jaddoe, V. W., Franco, O. H., and Lesaffre, E. M. Dealing with missing covariates in epidemiologic studies: A comparison between multiple imputation and a full Bayesian approach. *Statistics in Medicine*, 35(17):2955–2974, 2016.
- Ophuis, C. M. O., van Akkooi, A. C., Rutkowski, P., Voit, C. A., Stepniak, J., Erler, N. S., Eggermont, A. M., Wouters, M. W., Grünhagen, D. J., and Verhoef, C. Effects of time interval between primary melanoma excision and sentinel node biopsy on positivity rate and survival. *European Journal of Cancer*, 67:164–173, 2016.
- Leermakers, E. T., Felix, J. F., Erler, N. S., Ćerimagić, A., Wijtzes, A. I., Hofman, A., Raat, H., Moll, H. A., Rivadeneira, F., Jaddoe, V. W., Franco, O. H., and Kiefte-de Jong, J. C. Sugar-containing beverage intake in toddlers and body composition up to age 6 years: The Generation R Study. *European Journal of Clinical Nutrition*, 69(3):314–321, 2015.
- Tielemans, M. J., Erler, N. S., Leermakers, E. T., van den Broek, M., Jaddoe, V. W., Steegers, E. A., Kiefte-de Jong, J. C., and Franco, O. H. A Priori and a Posteriori Dietary Patterns during Pregnancy and Gestational Weight Gain: The Generation R Study. *Nutrients*, 7(11):9383–9399, 2015.
- Erler, N. S., Bryan, S. R., Eilers, P. H., Lesaffre, E. M., Lemij, H. G., and Vermeer, K. A. Optimizing structure–function relationship by maximizing correspondence between glaucomatous visual fields and mathematical retinal nerve fiber models. *Investigative Ophthalmology & Visual Science*, 55(4):2350–2357, 2014.

Acknowledgements