# Applying Psychometric Validation Methodology to Longitudinal Clinical Trial Data

Proefschrift voorgelegd tot het behalen van de graad van
Doctor in de Wetenschappen, richting Wiskunde
te verdedigen door

Tony Vangeneugden

Promotor: Prof. dr. Geert Molenberghs

# Preface

I joined Janssen Pharmaceutica as a project statistician in 1990 and since 2004, I am working at Tibotec, another J&J company specialized in developing anti-virological drugs. During these interesting 18 years, I have mostly been involved in late stage clinical development, in what is called the "Global Clinical Development" department. This type of departments are sometimes also disrespectfully referred to as the "production department". Selected molecules that have stand the tests in pre-clinical research and who have shown promise in the early clinical evaluation should have a high probability of reaching the marking when pharmaceutical companies engage in resource costly large scale phase 2B and phase 3 trials. While statisticians in the earlier phases of development have to deal with smaller numbers and more creative designs and analyses, the focus during the late stages is more on reaching the market as soon as possible using more standardized analyses in larger numbers of subjects. As a result, the analysis is focussed on proving the phase 3 confirmatory hypothesis, did we reach the p-value to prove the primary objective and was the drug safe, resulting in a positive risk benefit? The danger however is that we miss the chance to learn more, and to improve efficiency in future clinical studies. There is indeed an incredible wealth of information present in longitudinal clinical trial data. For instance, instead of focusing on mean difference between treatments and its significance, what can we learn about *variability* and *correlation* within and between the different measurements? During the years I was confronted with questions from clinicians and regulators related to this topic of correlation and variability. I remember for instance a discussion with the Dutch Authorities related to a trial in demented subjects with psychotic and aggressive behavior. The phase 3 trial showed that the experimental drug was superior to control in terms of an aggregate score of a scale with multiple items. The experimental arm had a larger decrease in the total score as compared to

the control arm. But the scale was not so well known and established as compared to for instance the Hamilton Depression Scale, and then the pertinent question from the clinician was: How does this difference in a total score translate into clinical more tangible difference, in other words is the difference clinically relevant. Another observation, especially in the CNS area, was that similar studies can result in different conclusions: why is it that some studies reach highly significant difference and other trials fail to show effect with the same drug and similar design?

These questions are of course related to the area of psychometric scale validation, a field which I was confronted with while working in a drug development project in schizophrenia. Indeed, the area of "psychiatric health sciences" has developed tools to evaluate measurements properties because of the inherent subjectivity of the measures employed in this field. These psychometric validation tools can be applied to all types of measurements and can provide insight into the performance of measurements in clinical trials.

Tony Vangeneugden

# Contents

# List of Abbreviations

| | |
|---|---|
| ACMV | Available-Case Missing Values |
| AIC | Akaike's Information Criterion |
| AR1 | First-Order Autoregressive Process |
| BPRS | Brief Psychiatric Rating Scale |
| CGI | Clinician's Global Impression |
| CT | Classical Theory |
| FDA | Food and Drug Administration |
| GEE | Generalized Estimating Equations |
| GT | Generalizability |
| GLM | General Linear Model |
| GLMM | General Linear Mixed Model |
| ICC | Intraclass Correlation Coefficient |
| LMM | Linear Mixed Model |
| MAR | Missing At Random |
| MCAR | Missing Completely At Random |
| NFMV | Non-Future Missing Values |
| ML | Maximum Likelihood |
| MLE | Maximum Likelihood Estimator |
| MNAR | Missing Not At Random |
| MNFD | Missing Non-Future Dependence |
| PANSS | Positive And Negative Syndrome Scale |
| PE | Proportion Explained |
| PMM | Pattern-Mixture Model |
| RB | Relative Bias |
| RE | Relative Effect |
| SD | Standard Deviation |
| SE | Standard Error |
| SEM | Selection Model |

# List of Tables

# List of Figures

# Main References for each Chapter

| Chapter | Reference |
|---|---|
| 3 | Verbeke and Molenberghs (2000) |
| | Molenberghs and Verbeke (2005) |
| | Molenberghs and Kenward (2007) |
| 4 | Dunn (1989) |
| | Streiner and Norman (1995) |
| | Cronbach, Rajaratnam, and Gleser (1963) |
| 5 | Buyse, Molenberghs, Burzykowski, Renard, and Geys (2000) |
| | Alonso, Geys, Molenberghs, and Vangeneugden (2002) |
| 6 | Vangeneugden, Laenen, Geys, Renard, and Molenberghs (2004) |
| 7 | Vangeneugden, Laenen, Geys, Renard, and Molenberghs (2005) |
| 8 | Vangeneugden, Molenberghs, Laenen, Geys, Beunckens, and Sotto (2008) |
| 9 | Vangeneugden, Molenberghs, Laenen, Alonso, and Geys (2008) |
| 10 | Vangeneugden, Molenberghs, Verbeke, and Demétrio (2008) |
| 11 | Vangeneugden, Molenberghs, Laenen, Geys, Beunckens, and Sotto (2008) |
| | Alonso, Geys, Molenberghs, and Vangeneugden (2002) |
| 12 | Michiels, Molenberghs, Bijnens, Vangeneugden, and Thijs (2002) |

# Additional References by the Author


Alonso, A., Geys, H. , Kenward, M., Molenberghs, G., and Vangeneugden, T. (2003) Validation of surrogate markers in multiple randomized clinical trials with repeated measurements. *Biometrical Journal*, **45**(8), 1–15.

Alonso, A., Geys, H., and Vangeneugden, T. (2005) Repeated measures and surrogate endpoint validation. Published in *The evaluation of surrogate endpoints*, Springer-Verlag, p. 231-251.

Alonso, A., Geys, H., Molenberghs, G., Kenward, M.G., and Vangeneugden, T. (2004) Validation of surrogate markers in multiple randomized clinical trials with repeated measurements: Canonical correlation approach. *Biometrics*, **60**(4), 845–853.

Buyse, M., Vangeneugden, T., Bijnens, L., Renard, D., Burzykowski, T., Geys, H., and Molenberghs, G. (2003) Validation of biomarkers as surrogates for clinical endpoints. Published in: Bloom, J. C. & Dean, R. A. (Ed.) *Biomarkers in Clinical Drug Development*, New York : Marcel Dekker.

Carman, J., Peuskens, J., and Vangeneugden, T. (1995) Risperidone in the treatment of negative symptoms of schizophrenia: a meta-analysis. *International Clinical Psychopharmacology*, **10**(4), 207–214.

Debruyne, F., Murray, R., Fradet, J., Johansson, J., Tyrrell, C., Boccardo, F. , Denis, L., Marberger, J., Brune, D., Rassweiler, J., Vangeneugden, T., Bruynseels, J., Janssens, M., and De Porre, P. (1998) Liarozole-A novel treatment approach

for advanced prostate cancer: results of a large randomized trial versus cyproterone acetate. *Urology*, **52**(1), 72–81.

Katlama, C., Esposito, R., Gatell, J., Goffard, J.C., Grinsztejn, B., Pozniak, A., Rockstroh, J., Stoehr, A., Vetter, N., Yeni, P., Parys, W., and Vangeneugden, T., on behalf of the POWER 1 study group (2007) Efficacy and safety of TMC114/ritonavir in treatment-experienced HIV patients: 24-week results of POWER 1. *AIDS*, **21**(4), 395–402.

Laenen, A., Vangeneugden, T., Geys, H., and Molenberghs, G. (2006). Generalized reliability estimation using repeated measurements. *British Journal Of Mathematical and Statistical Psychology.* **59**, 113–131.

Madruga, J.V., Berger, D., McMurchie, M., Suter, F., Banhegyi, D., Ruxrungtham, K., Norris,D., Lefebvre, E., de Bthune, M.P., Tomaka, F., De Pauw, M., Vangeneugden, T., and Spinosa-Guzman, S., on behalf of the TITAN study group (2007) Efficacy and safety of darunavir-ritonavir compared with that of lopinavir-ritonavir at 48 weeks in treatment- experienced, HIV-infected patients in TITAN: a randomised, controlled phase III trial. *The Lancet*, **370**, 49–58.

Molina, J.M., Cohen, C., Katlama, C., Grinsztejn, B., Timerman, A., Rogerio de Jesus, P., Vangeneugden, T., Miralles, D., De Meyer, S., Parys, W., and Lefebvre, E., on Behalf of the TMC114-C208 and -C215 Study Groups (2007) Safety and Efficacy of Darunavir (TMC114) With Low-Dose Ritonavir in Treatment-Experienced Patients: 24-Week Results of POWER 3. *Journal of Acquired Immune Deficiency Syndromes*, **46**(1), 24-31.

Renard, D., Geys, H., Molenberghs, G., Burzykowski, T., Buyse, M., Vangeneugden, T., and Bijnens, L. (2003). Validation of a longitudinally measured surrogate marker for a time-to-event endpoint. *Journal of Applied Statistics*, **30**, 235–247.

Sekar, V., Kestens, D., Spinosa-Guzman, G., De Pauw, M., De Paepe, E., Vangeneugden, T., Lefebvre, E., and Hoetelmans, R. (2007) The effect of different meal types on the pharmacokinetics of darunavir (TMC114)/ritonavir in HIV-negative healthy volunteers. *Journal of Clinical Pharmacology*, **47**, 479–484.

Sekar, V., Lefebvre, E., De Marez, T., Spinosa-Guzman, G., De Pauw, M., De Paepe, E., Vangeneugden, T., and Hoetelmans, R. (2007) Pharmacokinetics of

darunavir (TMC114) and atazanavir during coadministration in HIV-negative, healthy volunteers. *Drugs R D*, **8**(4), 241–248.

Sekar, V., Lefebvre, E., Marien, K., De Pauw, M., Vangeneugden, T., and Hoetelmans, R. (2007) Pharmacokinetic interaction between darunavir and saquinavir in HIV-negative volunteers *Therapeutic Drug Monitoring*, **29**(6), 795–801.

Sekar, V., Spinosa-Guzman, S., De Paepe, E., De Pauw, M., Vangeneugden, T., Lefebvre, E., and Hoetelmans, R. (2008) Darunavir/ritonavir pharmacokinetics following coadministration with clarithromycin in healthy volunteers. *Journal of Clinical Pharmacology*, **48**(1), 60–65.

# Chapter 1

# Introduction

Two important properties in psychometric validation are *reliability* and *validity*. Reliability consists in determining the extent to which the measurement is free from random error. This can be performed through analyzing *internal consistency* and *reproducibility* of the questionnaire. Internal consistency is the extent to which individual items are consistent with each other and reflect a single underlying construct. Intra-observer or test-retest reliability is the degree to which a measure yields stable scores at different points in time for patients who are assumed not to have changed clinical status on the domains being assessed. The calculation of intraclass correlation coefficients is one of the most commonly used methods. For interviewer-administered questionnaires, the inter-observer reliability is the degree to which a measurement yields stable scores when administered by different interviewers, rating the same patients. The calculation of interclass correlation coefficients is also one of the most commonly used methods. The validity of a questionnaire is defined as the degree to which the questionnaire measures what it purports to measure. This can be performed through the analysis of *content*, *construct* and *criterion validity*. Content validity can be defined as the extent to which the instrument assesses all the relevant or important content or domains. Also, the term *face validity* is used to indicate whether the instrument appears to be assessing the desired qualities at face. This form of validity consists of a judgment by experts in the field. Construct validity refers to a wide range of approaches which are used when what we are trying to measure is a "hypothetical construct" (e.g., anxiety, irritable bowel syndrome, ...) rather than

something that can be readily observed. The most commonly used methods to explore construct validity are: extreme groups (apply instrument for example to cases and non-cases), convergent and discriminant validity testing (correlate with other measures of this construct and not correlate with dissimilar or unrelated constructs) and multi-trait-multi method matrix. Criterion validity can be divided into two types: *concurrent validity* and *predictive validity*. With concurrent validity we correlate the measurement with a criterion measure (gold standard), both of which are given at the same time. In predictive validity, the criterion will not be available until some time in the future. The most commonly used method to assess the validity is by calculation of the Pearson correlation coefficient.

Whenever a mental measurement scale is developed or translated or used in a new population, its psychometric properties have to be assessed. This assessment is usually done on a relatively small and separate sample of stable subjects. In this thesis we want to show that these psychometric validation techniques can also be applied to data resulting from longitudinal clinical data. This work aims to provide a flexible framework to evaluate the actual performance of the scales in terms of reliability and validity in the specific clinical trial, or meta analysis of multiple trials. The goal is certainly not to replace up-front psychometric evaluation, but rather offer methodology to evaluate reliability and validity in the specific trial population. Indeed, these properties are relative, and linked to the population to which the scale is measured. The population in the trial might not be the same as the population in which the scale was evaluated. By looking at these properties, we can indeed learn how certain scales are correlated with more clinical endpoint as posed in the question above via the techniques offered in criterion validity. By study test-retest and inter-rater reliability we can learn more about the performance of the scale in the studied population. This is very important for the pharmaceutical industry because reliability is related to reproducibility and measurement error. The higher the reliability, the better patients and treatment groups can be discriminated, and the lower the sample size needs to be. Reliability can be extended to *generalizability*. In essence, the aim there is to investigate which factors do impact reliability. Are there subgroups with poor reliability?

In 2006, the FDA issued a draft guidance on the use of PRO measurements in medical product development to support labeling claims. This guidance will undoubtedly lead to the emergence of more scales and measurement tools to measure the patients' health status. If it is the intention to make labeling claims, then obviously the validity

of the scale must be clearly established and additional confirmation of reliability and validity could support potential label claims.

It is important to emphasize that the concepts of psychometric validation can also be applied to any measurements. Indeed, not only psychiatric scales can suffer form measurement error and poor reliability, but also other clinical measurements

## Structure of the Thesis

First, the we will introduce the longitudinal clinical data which will be used throughout this thesis in **Chapter 2**. Next, in **Chapter 3** we will briefly review the statistical modeling framework that will be employed in later chapters including the General Linear Model, the Generalized Linear Mixed Model, combined models to deal with count data and we finish with terminology, taxonomy and frameworks for missing data analyses. **Chapter 4** provides a summary of concepts of psychometric validation of measurements and scales including *reliability* and its extension to *generalizability* and also a brief discussion of the evaluation on *validity*. The subsequent **Chapter 5** ends the introductory part of this thesis by summarizing concepts in surrogate marker evaluation such as *individual-* and *trial-level surrogacy*.

In **Chapter 6** we derive a general formula for test-retest intraclass correlation coefficient of reliability for interval scales longitudinal clinical data. This general formula is worked out for four different models with different levels of complexity using pooled data of 5 clinical trials in schizophrenia. Subsequently, in **Chapter 7** we extent reliability testing to the evaluation of generalizability using the same continuous longitudinal clinical data as in Chapter 6. The purpose here is to develop a framework to evaluate if factors influence reliability and reproducibility. **Chapter 8** is an extension to Chapter 6, where we use the General Linear Mixed Model framework to obtain an approximative formula to derive the intraclass correlation coefficient. The derivations allow for any type of data is also flexible in terms of model complexity, e.g. allowing for serial correlation or not. An example is worked out for the binary response parameter CGI response using 4 of the clinical trials described in Chapter 2. The following **Chapter 9** again extends reliability testing to generalizability theory using the same general framework. **Chapter 10** addresses the special case of count data. Similar as for Gaussian data, a closed form can be derived for the intraclass correlation coefficient of reliability. Data from a clinical study in epilepsy will be used to derive the reliability using the closed form and the approximate formula derived

in Chapter 7. **Chapter 11** explores methods to investigate correlation between joint longitudinal sequences of different measures. Similar as in criterion validity, we evaluate the correlation between the PANSS total score, the BPRS total score and CGI response using schizophrenia data introduced in the Chapter 2. Finally, **Chapter 12** presents a case study in incomplete data, exploring missing data analysis techniques for a Quality of Life Questionnaire in a clinical study in breast cancer. While previous chapters focus on correlations between observations of the same measurement or from different measurements within a patient, this chapter focuses on the evaluation of treatment effects, accounting for incomplete data. We end this thesis with **Chapter 13** with discussion and topics for further research.

# Chapter 2

# Motivating Case Studies

The following three sections briefly introduce the main data used in this thesis and at the same time demonstrate the versatility of the type of data which can be used. Indeed, the first case study in schizophrenia includes a pooling of 5 studies and focuses on data resulting from psychiatric measurement scales, one interval scaled and one binary response parameter derived from an ordinal scale. And the second case study in epilepsy originates from a single randomized study and focuses on the number of seizures which can be considered as a pure clinical endpoint. In Section 2.3 we introduce the Vorozole data which was analyzed in Chapter 12, a case study in the analysis of incomplete data.

## 2.1 The Schizophrenia Data

Individual patient data from five double-blind randomized clinical trials, comparing the effects of risperidone to conventional anti psychotic agents for the treatment of chronic schizophrenia. Schizophrenia has long been recognized as a heterogeneous disorder with patients suffering from both "negative" and "positive" symptoms. Negative symptoms are characterized by deficits in social functions such as poverty of speech, apathy and emotional withdrawal. Positive symptoms entail more florid symptoms such as delusions and hallucinations, which are superimposed on the mental status. Several measures can be considered to assess a patient's global condition. The *Positive and Negative Syndrome Scale* (PANSS) consists of 30 items that provide an

operationalized, drug-sensitive instrument, which is highly useful for both typological and dimensional assessment of schizophrenia (Kay, Fiszbein, and Opler, 1987). Each of the 30 items can be scored ordinal from 1 (symptom absent) to 7 (symptom extreme). Classical reliability of the PANSS has been studied previously (Kay, Opler, and Lindenmayer, 1988; Bell *et al.*, 1992; Peralta and Cuesta, 1994). The PANSS is actually an extension of the Brief Psychiatric Rating Scale (BPRS), an 18-item scale (Overall and Gorham, 1962). Since the 18 items from the BPRS are also included in the PANSS, and therefore, the BPRS can be derived when the PANSS is available.

The *Clinical Global Impression* (CGI) of overall change versus baseline is a 7-grade scale used by the treating physician to characterize how well a subject has clinically improved versus baseline. The levels are: "very much improved", "much improved", "minimally improved", "no change", "minimally worse", "much worse", "very much worse". The binary CGI response is often defined as a CGI score of "very much improved" or "much improved". Figure 2.1 summarizes both CGI response and the mean total PANSS score over time. Note that patients enter the study with an acute worsening of symptoms, as witnessed by the higher (i.e., worse) mean total PANSS score at baseline. This figure also graphically depicts the correlation between the decreasing, improving trend in psychotic symptom score as measured by the PANSS and the more clinical evaluation of improvement as measured by the CGI.

Since the label in most countries recommends doses ranging from 4-6 mg/day, we include in our analysis only patients who received either these doses of risperidone or an active control (haloperidol, perphenazine, or zuclopenthixol). Depending on the trial, treatment was administered for a duration of 4-8 weeks. For example, in the international trials by Peuskens *et al.* (1995), Marder and Meibach (1994), and Hoyberg *et al.* (1993) patients received treatment for 8 weeks; in the study by Blin, Azorin, and Bouhours (1996) patients received treatment for 4 weeks, while in the study by Huttunen *et al.* (1995) patients were treated over a period of 6 weeks. The sample sizes were 453, 176, 74, 49 and 71, respectively. Measurements were taken at Week 1, 2, 4, 6, and 8.

## 2.2 The Epilepsy Data

In this section, we introduce data obtained from a single randomized, double-blind, parallel group multi-center study for the comparison of placebo with a new anti-epileptic drug (AED), in combination with one or two other AED's. The study

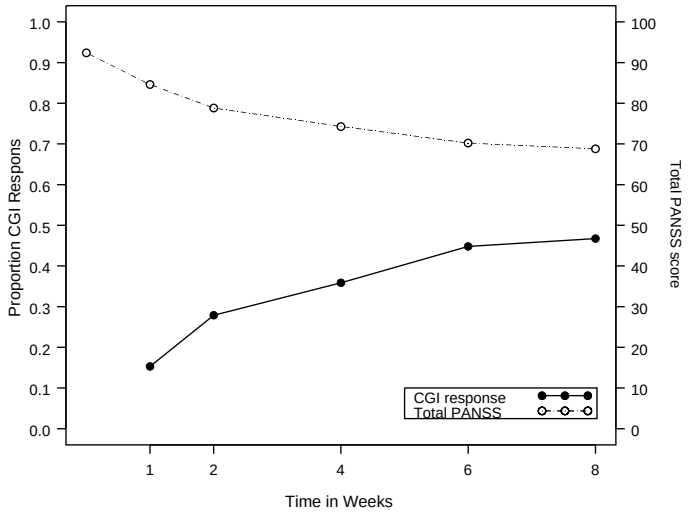


Figure 2.1: *Schizophrenia Data. Proportion CGI response and mean total PANSS score, over all visits.*

is described in full detail in Faught *et al.* (1996). The randomization of epilepsy patients took place after a 12-week baseline period that served as a stabilization period for the use of AED's, and during which the number of seizures were counted. After that period, 45 patients were assigned to the placebo group, 44 to the active (new) treatment group. Patients were then measured weekly. Patients were followed (double-blind) during 16 weeks, after which they were entered into a long-term open-extension study. Some patients were followed for up to 27 weeks. The outcome of interest is the number of epileptic seizures experienced during the last week, i.e., since the last time the outcome was measured. The key research question is whether or not the additional new treatment reduces the number of epileptic seizures. Figure 2.2 shows a frequency plot, over all visits, over both treatment groups. We observe a very skewed distribution, with largest observed value equal to 73 seizures in one week time. Average and median evolutions are shown in Figure 2.3. The unstable behavior can be explained by the presence of extreme values, but is also the result of the fact that very little observations are available at some of the time-points, especially past week

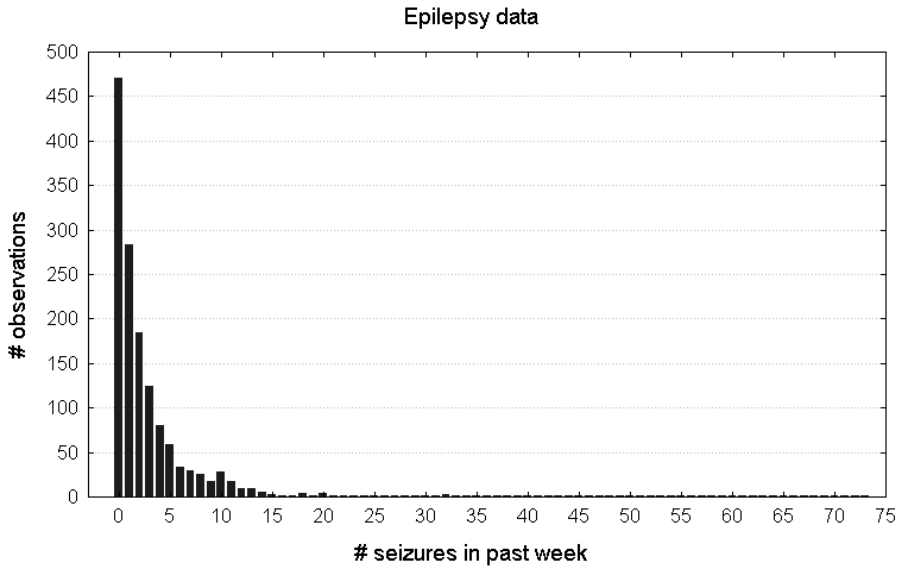


Figure 2.2: *Epilepsy Data. Frequency plot, over all visits, over both treatment groups.*

20. This is also reflected in Table 2.1, which shows the number of measurements at a selection of time-points. Note the serious drop in number of measurements past the end of the actual double-blind period, i.e., past week 16.

## 2.3 The Vorozole Data

This study was an open-label, multicenter, parallel group design conducted at 67 North American centers. Patients were randomized to either vorozole (2.5 mg taken once daily) or megestrol acetate (40 mg four times daily). The patient population consisted of post-menopausal patients with histologically confirmed estrogen-receptor positive metastatic breast carcinoma. All 452 randomized patients were followed until disease progression or death. The main objective was to compare the treatment group with respect to response rate while secondary objectives included a comparison relative to duration of response, time to progression, survival, safety, pain relief, performance status and quality of life. This paper focuses on overall quality of life, measured by the total Functional Living Index: Cancer (Schipper *et al.*, 1984). Precisely, a higher FLIC score is the more desirable outcome. Full details of the Vorozole

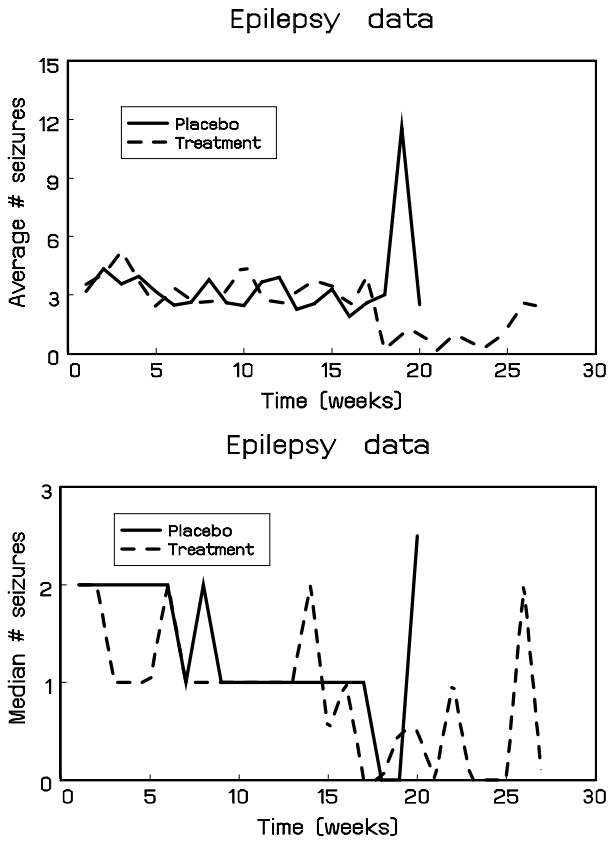


Figure 2.3: *Epilepsy Data. Average and median evolutions over time.*

study are reported in Goss *et al.* (1999).

Patients underwent screening and for those deemed eligible a detailed examination at baseline (occasion 0) took place. Further measurement occasions were month 1, then from month 2 at bi-monthly intervals until month 44.

The median age was 66 years for vorozole, and 67 for megestrol acetate, and the means were respectively 65.1 (SD 9.8) and 65.6 (SD 10.0) years. The mean duration of breast cancer when entering the study was 6.8 (SD 5.4) years for vorozole, and 6.9 (SD 5.5) years for megestrol acetate. The average total FLIC score at baseline was 116.3 (SD 21.2) for vorozole, and 117.1 (SD 19.0) for megestrol acetate. These total FLIC scores were calculated based on 199 and 213 patients, respectively.

Goss *et al.* (1999) analyzed the data and found no significant differences: the

Table 2.1: *Epilepsy Data. Number of measurements available at a selection of time-points, for both treatment groups separately.*

| | # Observations | | |
|---|---|---|---|
| Week | Placebo | Treatment | Total |
| 1 | 45 | 44 | 89 |
| 5 | 42 | 42 | 84 |
| 10 | 41 | 40 | 81 |
| 15 | 40 | 38 | 78 |
| 16 | 40 | 37 | 77 |
| 17 | 18 | 17 | 35 |
| 20 | 2 | 8 | 10 |
| 27 | 0 | 3 | 3 |

response rate was 9.7% for vorozole, versus 6.8% for megestrol acetate (p=0.24); clinical benefit from treatment was demonstrated in 23.5% of vorozole-treated patients versus 27.2% of megestrol acetate-treated patients (p=0.42). They also performed an endpoint analysis of change in the total FLIC score using a two-way ANOVA model with effects for treatment, disease status, as well as their interaction. Again, no significant difference was found.

Dropout rates are displayed in Figure 2.4 Precisely, dropout is presented w.r.t. total FLIC score. The main reasons for dropout are disease progression (152 patients in the vorozole arm; 134 in the megestrol acetate arm), adverse events (5 and 13) and death during treatment (5 in each arm). More detailed information on discontinuation and dropout can be found in Goss *et al.* (1999).

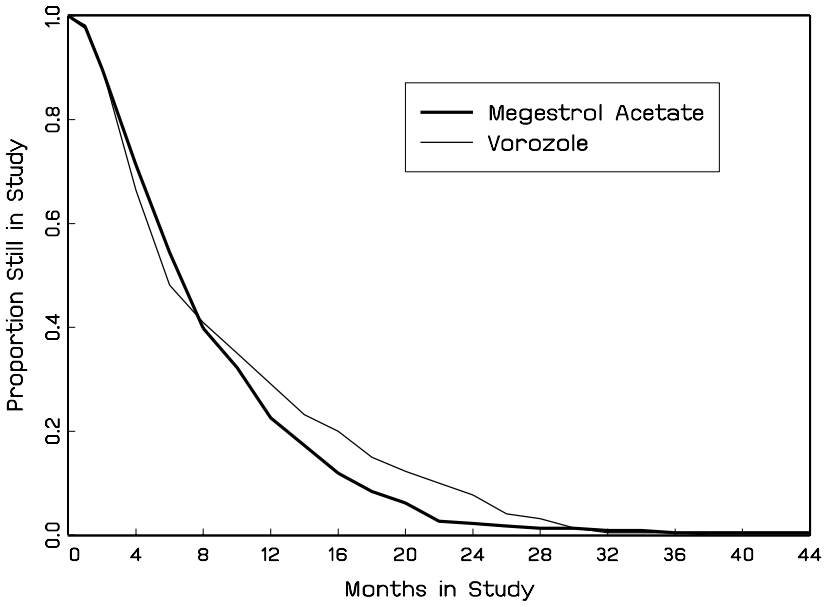


Figure 2.4: *Vorozole study. Representation of dropout.*

# Chapter 3

# Concepts in Repeated Measures

First we will briefly review the well known exponential family for univariate data and the general linear model (GLM) based on it in Section 3.1. Subsequently we will provide a short review of the Linear Mixed Model (LMM) for continuous longitudinal data in Section 3.2 and extend the GLM to the General Linear Mixed Model (GLMM) for longitudinal data in Section 3.3. While the LMM will serve as a modeling tool to derive estimates for the reliability and generalizability coefficients specifically for normal distributed data (Chapter 6–7), the GLMM will be used to derive a general modelling framework for any type of data including binary data (Chapter 8–9). In Section 3.4 we introduce models for count data. These will serve as a modelling tool for count data addressed in Chapter 10. Finally Section 3.5 is devoted to missing values. This section provides an introduction to Chapter 12, a case study in incomplete data.

## 3.1 Generalized Linear Models

A random variable $Y$ follows an exponential family distribution if the density is of the form

$$f(y) \equiv f(y|\theta,\phi) = \exp\left\{\phi^{-1}[y\theta - \psi(\theta)] + c(y,\phi)\right\}, \tag{3.1}$$

for a specific set of unknown parameters $\theta$ and $\phi$, and for known functions $\psi(\cdot)$ and $c(\cdot,\cdot)$. Often, $\theta$ and $\phi$ are termed 'natural (canonical) parameter' and 'scale parameter,' respectively. The mean and variance follow from $\psi(\cdot)$ as $\mathrm{E}(Y) = \mu = \psi'(\theta)$ and $\mathrm{Var}(Y) = \sigma^2 = \phi\psi''(\theta)$, leading to the so-called mean-variance relationship $\sigma^2 = \phi\psi''[\psi^{'-1}(\mu)] = \phi v(\mu)$, with $v(\cdot)$ the variance function. Conventionally, inference is conducted through either quasi-likelihood, restricting model specification to the first two moments, or through full likelihood, with (3.1) as its basis (McCullagh and Nelder, 1989, Molenberghs and Verbeke, 2005). In practice, a sample of $N$ independent outcomes $Y_1, \ldots, Y_N$ is collected, together with $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ $p$-dimensional vectors of covariate values. It is assumed that all $Y_i$ have densities $f(y_i|\theta_i, \phi)$ which belong to the exponential family, and with $\theta_i$ depending on the covariates. Specification of the generalized linear model is completed by modeling the means $\mu_i$ as functions of the covariate values: $\mu_i = h(\theta_i) = h(\boldsymbol{x}_i'\boldsymbol{\beta})$, for a known inverse link function $h(\cdot)$, and with $\boldsymbol{\beta}$ a vector of $p$ fixed unknown regression coefficients. The so-called natural link function is given by $h(\cdot) = \psi'(\cdot)$.

When Y is normally distributed with mean $\mu$ and variance $\sigma^2$, the density can be written as

$$f(y) = \exp\left\{\frac{1}{\sigma^2}\left[y\mu - \frac{\mu^2}{2}\right] + \left[\frac{\ln(2\pi\sigma^2)}{2} - \frac{y^2}{2\sigma^2}\right]\right\}, \tag{3.2}$$

and hence the normal distribution belongs to the exponential family, with natural parameter $\theta$ equal to $\mu$, scale parameter $\phi$ equal to $\sigma^2$ and variance function $v(\mu) = 1$. Hence the normal distribution is very particular in the sense that there is no mean-variance relation. The natural link function equals the identity function, leading to the classical linear regression model $\boldsymbol{Y}_i \sim N(\mu_i, \sigma^2)$ with $\mu_i = \boldsymbol{x_i}'\boldsymbol{\beta}$.

If $Y$ is Bernoulli distributed with success probability $P(Y = 1) = \pi$, the density can be written as

$$f(y) = \exp\left\{y \ln\left(\frac{\pi}{1-\pi}\right) + \ln(1-\pi)\right\}, \tag{3.3}$$

which implies that the Bernoulli distribution belongs to the exponential family, with natural parameter $\theta$ equal to the logit, i.e., $\ln[\pi/(1-\pi)]$ of $\pi$, scale parameter $\phi = 1$, with mean $\mu = \pi$ and variance function $v(\pi) = \pi(1-\pi)$. The natural link function is the logit link, leading to the classical logistic regression model $\boldsymbol{Y}_i \sim Bernoulli(\pi_i)$ with $\ln[\pi_i/(1-\pi_i)] = \boldsymbol{x_i}'\boldsymbol{\beta}$ or equivalently

$$\pi_i = \frac{\exp(\boldsymbol{x_i}'\boldsymbol{\beta})}{[1 + \exp(\boldsymbol{x_i}'\boldsymbol{\beta})]}.$$

In case of a Poisson distribution with mean $\lambda$: $Y \sim \text{Poi}(\lambda)$. The density can be written as

$$f(y) = \frac{e^{-\lambda}\lambda^y}{y!} = \exp\{y \ln \lambda - \lambda - \ln y!\}, \tag{3.4}$$

with natural parameter $\theta = \ln \lambda$, mean $\mu = \lambda$, scale parameter $\phi = 1$, and variance function $v(\mu) = \mu = \lambda$. The logarithm is the natural link function, leading to the classical Poisson regression model $Y_i \sim \text{Poisson}(\lambda_i)$, with $\ln \lambda_i = \boldsymbol{x}_i'\boldsymbol{\beta}$.

## 3.2 Linear Mixed Models

Among the models for longitudinal data, methods for continuous data form the best developed and most advanced body of research, while the same is true for software implementation. This is natural, since the special status and the elegant properties of the multivariate normal distribution simplify model building and ease software development. It is in this area that the linear mixed model is situated (Laird and Ware, 1982, Verbeke and Molenberghs, 2000). Gaussian data can be modeled entirely in terms of their means, variances and covariances. The parameters of the mean model are referred to as *fixed-effects* parameters, and the parameters of the variance-covariance model as *covariance parameters*. The fixed-effects parameters capture the influence of explanatory variables on the mean structure, exactly as in the standard linear model. However, the occurrence of random effects and a structured covariance matrix distinguishes the linear mixed model from the standard linear model. The need for covariance modeling arises quite frequently in applications such as when repeated measurements are taken on the same experimental unit, with spatially correlated data, or when experimental units can be grouped into clusters and data from a cluster are correlated. One can distinguish between three components of variability. Part of the covariance structure arises from so-called *random effects*, i.e., additional covariate effects with random parameters. These are effects which arise from the characteristics of individual subjects. The variances of the random-effects parameters are commonly referred to as *variance components* (Searle, Casella, and McCullogh, 1992). Another component of the variability is the serial correlation which captures that measurements taken close together in time are typically more strongly correlated than those taken further apart in time. On a sufficiently small time-scale, this kind of structure is almost inevitable. The last component is the measurement error: when the measurement process involves fuzzy determinations, the results may

show substantial variation even when two measurements are taken at the same time from the same subject.

A linear mixed-effects model (LMM) with serial correlation can be written as

$$\boldsymbol{Y}_i = X_i\boldsymbol{\beta} + Z_i\boldsymbol{b}_i + \boldsymbol{W}_i + \boldsymbol{\varepsilon}_i, \tag{3.5}$$

where $\boldsymbol{Y}_i$ is the $n_i$ dimensional response vector for subject $i$, $1 \leq i \leq N$, $N$ is the number of subjects, $X_i$ and $Z_i$ are $(n_i \times p)$ and $(n_i \times q)$ known design matrices, $\boldsymbol{\beta}$ is the $p$ dimensional vector containing the fixed effects, $\boldsymbol{b}_i \sim N(\mathbf{0}, D)$ is the $q$ dimensional vector containing the random effects, $\boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \sigma^2 I_{n_i})$ is a $n_i$ dimensional vector of measurement error components, and $\boldsymbol{b_1}, \ldots, \boldsymbol{b_N}, \boldsymbol{\varepsilon_1}, \ldots, \boldsymbol{\varepsilon_N}$ are assumed to be independent. Serial correlation is captured by the realization of a Gaussian stochastic process, $\boldsymbol{W}_i$, which is assumed to follow a $N(\mathbf{0}, \tau^2 H_i)$ law. The serial covariance matrix $H_i$ only depends on $i$ through the number $n_i$ of observations and through the time points $t_{ij}$ at which measurements are taken. The structure of the matrix $H_i$ is determined through the autocorrelation function $\rho(t_{ij} - t_{ik})$. A first simplifying assumption is that it depends only on the time interval between two measurements $Y_{ij}$ and $Y_{ik}$, i.e., $\rho(t_{ij} - t_{ik}) = \rho(|t_{ij} - t_{ik}|)$, where $u = |t_{ij} - t_{ik}|$ denotes time lag. This function decreases such that $\rho(0) = 1$ and $\rho(+\infty) = 0$. Finally, $D$ is a general $(q \times q)$ covariance matrix with $(i, j)$ element $d_{ij} = d_{ji}$. Inference is based on the marginal distribution of the response $\boldsymbol{Y_i}$ which, after integrating over random effects, can be expressed as

$$\boldsymbol{Y}_i \ \sim \ N(X_i\boldsymbol{\beta}, Z_i D Z_i' + \Sigma_i). \tag{3.6}$$

Here, $\Sigma_i = \sigma^2 I_{n_i} + \tau^2 H_i$ is a $(n_i \times n_i)$ covariance matrix grouping the measurement error and serial components.

Fitting of mixed models described in (3.5) is based upon maximum likelihood methods (maximum likelihood or restricted maximum likelihood). These methods can be sensitive to peculiar observations that can have an unusually large influence on the results of the analysis. Many diagnostic tools have been developed for linear regression models but the generalization of these methods is far from obvious. First, several kinds of residuals could be defined: the *marginal residuals* $\boldsymbol{Y}_i - X_i\hat{\boldsymbol{\beta}}$, reflecting how a specific profile deviates from the overall population mean, the *subject-specific residuals* $\boldsymbol{Y}_i - X_i\hat{\boldsymbol{\beta}} - Z_i\hat{\boldsymbol{b_i}}$, measuring how much the observed values deviate from the subject's own predicted regression line, and the *estimated random effects* $\hat{\boldsymbol{b_i}}$ reflecting how much

specific profiles deviate from the population average profile. Further, the linear mixed model involves two kinds of covariates. The matrix $X_i$ represents the design matrix for the fixed effects, and $Z_i$ is the design matrix for the random effects. Therefore, it is not clear how leverages should be defined, partially because the matrices $X_i$ and $Z_i$ are not necessarily of the same dimension. A classification of influential subjects can be based on Cook's distance, which measures how much parameter estimates change when a specific individual has been removed from the dataset. In classical regression, closed-form expression exist, allowing easy calculation and also ascribing influence to the specific characteristics of the subjects (leverage, outlying). Unfortunately, this is no longer the case in linear mixed models. For exact Cook's distances, the iterative estimation procedure has to be used $N + 1$ times, which can be extremely time-consuming. The local influence approach was first introduced by Cook (1986). The general idea is to give every individual its own weight in the calculation of the parameter estimates and to investigate how these estimates depend on the weights, locally around the equal-weight case, which is the ordinary maximum likelihood case. We restrict the discussion to models which assume conditional independence, hence no serial correlation and $\Sigma_i = \sigma^2 I_{n_i}$. Denote $\hat{\theta}$ as the maximum likelihood estimate for $\theta$, obtained after maximizing $\ell(\theta)$ and $\hat{\theta}_\omega$ the estimate for $\theta$ after maximizing $\ell(\theta|\omega)$, any perturbed version of $\ell(\theta)$. The weight vector $\omega$ is $N$ dimensional and the original log-likelihood corresponds to $\omega = \omega_0 = (1, 1, ..., 1)'$. Cook (1986) proposed to measure the distance between $\hat{\theta}$ and $\hat{\theta}_\omega$ by the so-called likelihood displacement, defined by

$$LD(\omega) = 2[\ell(\hat{\theta}) - \ell(\hat{\theta}_\omega)] \tag{3.7}$$

$LD(\omega)$ will be large if $\ell(\theta)$ is strongly curved at $\hat{\theta}$, which means that $\theta$ is estimated with high precision and $LD(\omega)$ will be small if $\theta$ is estimated with high variability. From this perspective, a graph of $LD(\omega)$ versus $\omega$ contains essential information. Ideally, we would like a complete influence graph to assess influence for a particular model and a particular data set. However, this is very difficult in high dimensional situations. One method to extract the most relevant information from an influence graph is *local influence*, which uses normal curvatures, see Verbeke and Molenberghs (2000) for more detail. Denote $C_h$ as the normal curvature at the surface of $(\omega, LD(\omega))$ at $\omega_0$, in the direction $h$. Large values of $C_h$ indicate sensitivity to the induced perturbations in the direction $h$. There are several choices for h. One evident choice correspond to the perturbation of the $i$the weight only. This is obtained by taking $h$ equal to the vector $h_i$ which contains zeros everywhere except on the $i$the position. One can prove

that $C_i$ can be approximated by

$$C_i = -2[\hat{\theta} - \hat{\theta}^1_{(i)}]' \ddot{L}_{(i)} \ddot{L}^{-1} \ddot{L}_{(i)} [\hat{\theta} - \hat{\theta}^1_{(i)}],$$

where $\ddot{L}$ and $\ddot{L}_{(i)}$ are respectively the matrix of second-order derivatives of full log-likelihood and of the log-likelihood calculated after deleting the $i$ case and where $\hat{\theta}^1_{(i)}$ is the one-step approximation of $\hat{\theta}_{(i)}$ from a single Newton-Raphson step in the maximization procedure of $\ell_{(i)}(\theta)$, using $\hat{\theta}$ as starting value. One can also show that for sufficiently large $N$, $C_i$ can be interpreted as an approximation to the global case-deletion diagnostics. Lesaffre and Verbeke (1998) have shown that $C_i$ can be decomposed into five interpretable components: the "length" of the standardized covariate in the mean structure, the overall measure for how well the observed data for the $i$the subject are predicted by the mean structure $X_i\boldsymbol{\beta}$, two similar components for the covariance structure, and finally the size of the variability of the $i$the subject.

## 3.3 Generalized Linear Mixed Model

The generalized linear mixed model (GLMM, Breslow and Clayton, 1993) is the most frequently used random effects model for discrete outcomes and is a straightforward extension of the general linear model introduced in Section 3.1. As before, $Y_{ij}$ is the $j$th outcome measured for subjects $i, i = 1, \ldots, N, j = 1, \ldots, n_i$ and $\boldsymbol{Y}_i$ is the $n_i$-dimensional vector of all measurements available for cluster $i$. This model assumes that, conditionally on $q$-dimensional random effects $\boldsymbol{b}_i$, assumed to be drawn independently from the $N(\boldsymbol{0}, \boldsymbol{D})$, the outcomes $Y_{ij}$ are independent with densities of the form

$$f_i(y_{ij}|\boldsymbol{b}_i, \boldsymbol{\beta}, \phi) = \exp\left[\frac{y_{ij}\theta_{ij} - \psi(\theta_{ij})}{\phi} + c(y_{ij}, \phi)\right], \tag{3.8}$$

where the mean $\mu_{ij}$ is modeled through a linear predictor containing fixed regression parameters $\boldsymbol{\beta}$ as well as subject-specific parameters $\boldsymbol{b}_i$, i.e., $g(\mu_{ij}) = g(E(Y_{ij}|\boldsymbol{b}_i)) = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{b}_i$ for a known link function $g(.)$, with $\boldsymbol{x}_{ij}$ and $\boldsymbol{z}_{ij}$ $p$-dimensional and $q$-dimensional vectors of known covariate values, with $\boldsymbol{\beta}$ a $p$-dimensional vector of unknown fixed regression coefficients, and with $\phi$ a scale parameter. With a natural link function this becomes $\theta_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{b}_i$. The random effects $\boldsymbol{b}_i$ are assumed to be sampled from a (multivariate) normal distribution with mean $\boldsymbol{0}$ and covariance matrix $\boldsymbol{D}$.

In this GLMM setting, we can write the general model as follows: $\boldsymbol{Y}_i = \boldsymbol{\mu}_i + \boldsymbol{\varepsilon}_i$, where $\boldsymbol{\mu}_i$, the conditional mean, given the random effects, can be written as $\boldsymbol{\mu}_i =$

$\boldsymbol{\mu}_i(\boldsymbol{\eta}_i) = h(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i)$, $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are known design matrices, $\boldsymbol{\beta}$ are fixed-effect parameters, $\boldsymbol{b}_i$ are random effects, and $h$ is a known link function. Finally, $\boldsymbol{\varepsilon}_i$ is the residual error component.

## 3.4 Combined Model

It is clear from (3.4) that the standard Poisson model forces equality between mean and variance. However, comparing the sample average with the sample variance might already reveal that this assumption is not in line with a particular set of data. Therefore, a number of extensions have been proposed (Breslow 1984, Lawless 1987). A straightforward step is to allow the overdispersion parameter $\phi$ to differ from one, so that the mean-variance relationship produces $\text{Var}(Y) = \phi\mu$. This is in line with the moment-based approach mentioned in Section 3.1, although one can still think of such moments as arising from a random sum of Poisson variables, a point made by Hinde and Demétrio (1998ab).

An elegant way forward is through a two-stage, or random-effects, approach. A commonly encountered instance is by assuming that $Y_i|\lambda_i \sim \text{Poi}(\lambda_i)$ and then further that $\lambda_i$ is a random variable with $\text{E}(Y_i) = \mu_i$ and $\text{Var}(Y_i) = \sigma_i^2$. Using iterated expectations, it follows that $\text{E}(Y_i) = \mu_i$ and $\text{Var}(Y_i) = \mu_i + \sigma_i^2$. Note that we have not assumed a particular distributional form for the random effects $\lambda_i$. A common choice is the gamma distribution, leading to the so-called negative-binomial model.

This model is straightforward to extend to repeated measurements. We then assume a hierarchical data structure, where now $Y_{ij}$ denotes the $j$th outcome measured for cluster (subject) $i$, $(i = 1, \ldots, N; j = 1, \ldots, n_i)$ and $\boldsymbol{Y_i}$ is the $n_i$-dimensional vector of all measurements available for cluster $i$. Then, the scalar $\lambda_i$ becomes a vector $\boldsymbol{\lambda}_i = (\lambda_{i1}, \ldots, \lambda_{in_i})'$, with $\text{E}(\boldsymbol{\lambda}_i) = \boldsymbol{\mu}_i$ and $\text{Var}(\boldsymbol{\lambda}_i) = \boldsymbol{\Sigma}_i$. Similar logic as in the univariate case produces $\text{E}(\boldsymbol{Y_i}) = \boldsymbol{\mu}_i$ and $\text{Var}(\boldsymbol{Y_i}) = M_i + \Sigma_i$, where $M_i$ is a diagonal matrix with the vector $\boldsymbol{\mu}_i$ along the diagonal, the diagonal structure of $M_i$ reflecting the conditional independence assumption, i.e., all dependence between measurements on the same unit stem from the random effects. A versatile class of models results. For example, assuming the components of $\boldsymbol{\lambda}_i$ are independent, a pure overdispersion model follows, without correlation between the repeated measures. On the other hand, assuming $\lambda_{ij} = \lambda_i$, i.e., all components are equal, then $\text{Var}(\boldsymbol{Y_i}) = M_i + \sigma_i^2 J_{n_i}$, where $J_{n_i}$ is an $n_i \times n_i$ dimensional matrix of ones, as a Poisson version of compound symmetry. Of course, one can also combine general correlation structures between

the components of $\boldsymbol{\lambda}_i$.

Alternatively, this repeated version of the overdispersion model can be combined with normal random effects in the linear predictor, as proposed by Molenberghs, Verbeke, and Demétrio 2007 (MVD). We first review the GLMM (Section 3.4.1) and then move on to the combined model (Section 3.4.2).

### 3.4.1 Normal Random Effects: the Poisson-normal Model

In general, the GLMM assumes that, conditionally on $q$-dimensional random effects $\boldsymbol{b_i}$, commonly assumed to be drawn independently from the $N(\mathbf{0}, D)$, the outcomes $Y_{ij}$ to be independent with exponential-family densities of the form

$$f_i(y_{ij}|\boldsymbol{b_i}, \boldsymbol{\beta}, \phi) = \exp\left\{\phi^{-1}[y_{ij}\theta_{ij} - \psi(\theta_{ij})] + c(y_{ij}, \phi)\right\}, \tag{3.9}$$

with $h^{-1}(\mu_{ij}) = h^{-1}[\mathrm{E}(Y_{ij}|\boldsymbol{b_i})] = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{b_i}$, and with $\boldsymbol{x}_{ij}$ and $\boldsymbol{z}_{ij}$ $p$-dimensional and $q$-dimensional vectors of known covariate values, with $\boldsymbol{\beta}$ a $p$-dimensional vector of unknown fixed regression coefficients, and with $\phi$ a scale parameter.

For the specific case of Poisson data, the assumptions are

$$Y_{ij} \sim \mathrm{Poi}(\lambda_{ij}), \tag{3.10}$$

$$\ln(\lambda_{ij}) = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{b_i}, \tag{3.11}$$

$$\boldsymbol{b_i} \sim N(\mathbf{0}, D). \tag{3.12}$$

When normality for the random effects is assumed, the mean vector and variance-covariance matrix of $\boldsymbol{Y_i}$ can be derived relatively easily, as shown by MVD; see also Section 10.1. The mean vector $\boldsymbol{\mu}_i = \mathrm{E}(\boldsymbol{Y_i})$ has components:

$$\mu_{ij} = \exp\left(\boldsymbol{x}'_{ij}\boldsymbol{\beta} + \frac{1}{2}\boldsymbol{z}_{ij}D\boldsymbol{z}'_{ij}\right), \tag{3.13}$$

while the variance-covariance matrix equals

$$\mathrm{Var}(\boldsymbol{Y_i}) = M_i + M_i\left(e^{Z_i D Z'_i} - J_{n_i}\right)M_i, \tag{3.14}$$

with $M_i$ as in the previous section. In the special case of univariate data with a single normal random intercept $b_i \sim N(0, d)$ and $z_i = 1$, expressions (3.13) and (3.14) simplify to:

$$\mu_i = \exp\left(x'_i\boldsymbol{\beta} + \frac{1}{2}d\right), \qquad \mathrm{Var}(Y_i) = \mu_i + \mu_i^2(e^d - 1),$$

the latter being of the well-known quadratic form (Hinde and Demétrio 1998ab).

### 3.4.2 Combining Overdispersion With Normal Random Effects

Combining ideas from the overdispersion models in Section 3.4 and the Poisson-normal model of Section 3.4.1, MVD specified, in line with Booth *et al.* (2003), a model for repeated Poisson data with overdispersion, by extending (3.10)–(3.12) to

$$\begin{aligned} Y_{ij} &\sim \text{Poi}(\lambda_{ij}), && (3.15)\\ \lambda_{ij} &= \theta_{ij} \exp\left(\boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{b_i}\right), && (3.16)\\ \boldsymbol{b_i} &\sim N(\mathbf{0}, D), && (3.17)\\ \text{E}(\boldsymbol{\theta}_i) = \text{E}[(\theta_{i1}, \ldots, \theta_{in_i})'] &= \Phi_i, && (3.18)\\ \text{Var}(\boldsymbol{\theta}_i) &= \Sigma_i. && (3.19) \end{aligned}$$

The mean vector $\boldsymbol{\mu}_i = \text{E}(\boldsymbol{Y_i})$ now has components:

$$\mu_{ij} = \phi_{ij} \exp\left(\boldsymbol{x}'_{ij}\boldsymbol{\beta} + \frac{1}{2}\boldsymbol{z}_{ij}D\boldsymbol{z}'_{ij}\right), \tag{3.20}$$

and the variance-covariance matrix is given by

$$\text{Var}(\boldsymbol{Y_i}) = M_i + M_i\left(P_i - J_{n_i}\right)M_i, \tag{3.21}$$

where $M_i$ is still defined as before and the $(j,k)^{\text{th}}$ element of $P_i$ equals

$$p_{i,jk} = \exp\left(\frac{1}{2}\boldsymbol{z}_{ij}D\boldsymbol{z}'_{ik}\right) \cdot \frac{\sigma_{i,jk} + \phi_{ij}\phi_{ik}}{\phi_{ij}\phi_{ik}} \cdot \exp\left(\frac{1}{2}\boldsymbol{z}_{ik}D\boldsymbol{z}'_{ij}\right). \tag{3.22}$$

An expression for the joint marginal probabilities is presented in Section 10.3.

## 3.5 Missing Data

In this section we briefly review concepts of missing data handling. The text follows the development layed out in Chapter 3 form Molenberghs and Kenward (2007).

### 3.5.1 Modeling Incompleteness

It is very common for sets of quantitative data to be incomplete, in the sense that not all planned observations are actually made This is especially true when studies are conducted on human subjects. Often a distinction is made between missingness *patterns*. *Dropout* or *attrition* refers to the specific situation, arising in longitudinal

studies, where subjects are observed uninterruptedly from the beginning of the study until a given point in time, perhaps prior to the scheduled end of the study, when they drop out and do not return to the study. The general mechanism, where subjects can be observed and missing on any partition of the set of planned measurement occasions, is often called *non-monotone missingness*. In clinical trials, dropout is not only a common occurrence, there are also specific procedures for reporting and subsequently dealing with it. Patients who drop out of a clinical trial are usually listed on a separate withdrawal sheet of the case record form with the reasons for withdrawal entered by the authorised investigator. Reasons frequently encountered are adverse events, illness not related to study medication, an uncooperative patient, protocol violation, and ineffective study medication. Further specifications may include so-called *loss to follow-up*. Based on this medically inspired typology, Gould (1980) proposed specific methods to handle this type of incompleteness. Even though the primary focus of such trials is often on a specific time of measurement, usually the last, the outcome of interest is recorded in a longitudinal fashion, and dropout is a common occurrence. While dropout, in contrast to non-monotone missingness, may simplify model formulation and manipulation, the causes behind it can be more problematic. For example, dropout may derive from lack of efficacy, or from potentially serious and possible treatment-related side effects. In contrast, an intermittently missing endpoint value may be due more plausibly to the patient skipping a visit for practical or administrative reasons, to measurement equipment failure, and so on. In addition, one often sees that incomplete sequences in clinical trials are, for the vast majority, of a dropout type, with a relatively minor fraction of incompletely observed patients producing non-monotone sequences.

To incorporate incompleteness into the modeling process, we need to reflect on the nature of the missing value mechanism and its implications for statistical inference. Rubin (1976) and Little and Rubin (2002, Chapter 6) distinguish between different missing values processes. A process is termed *missing completely at random* (MCAR) if missingness is independent of both unobserved and observed outcomes, and *missing at random* (MAR) if, conditional on the observed data, missingness is independent of the unobserved outcomes; otherwise, the process is termed *missing not at random* (MNAR). A more formal definition of these concepts is given in Section 3.5.4.

Given MAR, a valid analysis can be obtained through a likelihood-based analysis that ignores the missing value mechanism, provided the parameters describing the measurement process are functionally independent of the parameters describing the

missingness process, the so-called parameter distinctness condition. This situation is termed ignorable by Rubin (1976) and Little and Rubin (2002) and leads to considerable simplification in the analysis (Diggle, 1989, Verbeke and Molenberghs, 2000). See also Section 3.5.5.

In practice, the reasons for missingness are likely to be manifold and it is therefore difficult to justify solely on *a priori* grounds the assumption of missingness at random. Arguably, under MNAR, a wholly satisfactory analysis of the data is not feasible, and it should be noted that the data alone cannot distinguish between MAR and MNAR mechanisms.

In the light of this one approach could be to estimate from the available data the parameters of a model representing a MNAR mechanism. It is typically difficult to justify the particular choice of missingness model, and it does not necessarily follow that the data contain information on the parameters of the particular model chosen. These points have been studied in Jansen et al. (2006b) and are discussed in Chapter 20 of Molenberghs and Kenward (2007). Where such information exists (and as we emphasize below this is normally derived from untestable modeling assumptions), the fitted model can be seen as providing some insight into the fit of the missing at random model to the observed data. Only through external assumptions can we use this subsequently to make inferences about the missing value process. Consequently the approach is potentially useful for assessing the sensitivity of the conclusions to assumptions about the missing value process, but not for making definitive statements about it. Several authors have used MNAR models that explicitly model the dropout process, and attempted from these to draw conclusions about the missing value mechanism. These included include Diggle and Kenward (1994) in the context of continuous longitudinal data and by Molenberghs, Kenward, and Lesaffre (1997) for ordinal outcomes. Overviews of and extensive discussion on this topic is found in Little (1995), Diggle *et al.* (2002), Verbeke and Molenberghs (2000) and Molenberghs and Verbeke (2005). Further early approaches for continuous data were proposed by Laird, Lange, and Stram (1987), Wu and Bailey (1988, 1989), Wu and Carroll (1988), and Greenlees, Reece, and Zieschang (1982). Proposals for categorical data were made by Baker and Laird (1988), Stasny (1986), Baker, Rosenberger, and DerSimonian (1992), Conaway (1992, 1993), Park and Brown (1994).

A feature common to all complex (MNAR) modeling approaches is that they rely on untestable assumptions about the relationship between the measurement and missing value processes. An obvious consequence of this is that one should therefore

avoid missing data as much as possible and, when the problem arises, ensure that all practicable efforts are made to collect information on the reasons for this. As an example, consider a clinical trial where outcome and missingness are both strongly related to a specific covariate $X$ and where, conditionally on $X$, the response $Y$ and the missing data process $R$ are independent. In the selection framework (Section 3.5.4), we then have that $f(Y, R|X) = f(Y|X)f(R|X)$, implying MCAR, whereas omission of $X$ from the model may imply MAR or even MNAR, which has important consequences for selecting valid statistical methods.

Different MNAR models may fit the observed data equally well, but have quite different implications for the unobserved measurements, and hence for the conclusions to be drawn from the respective analyses. Without additional information we can only distinguish between such models using their fit to the observed data, and so goodness-of-fit tools typically do not provide a relevant means of choosing between such models. It follows that there is an important role for sensitivity analysis in assessing inferences from incomplete data.

In Section 3.5.2 we introduce summarize terminology and in Section 3.5.3 we sketch the broad frameworks for incomplete data modeling. Missing data patterns is formalized in Section 3.5.4. Ignorability is the subject of Section 3.5.5. Section 3.5.6 sketches a general pattern-mixture model framework.

### 3.5.2 Terminology

The following terminology is based on the standard framework of Rubin (1976) and Little and Rubin (2002). It allows us to place formal conditions on the missing value mechanism which determine how the mechanism may influence subsequent inferences. Assume that for each independent unit $i = 1, \ldots, N$ in the study, it is planned to collect a set of measurements $Y_{ij}$ $(j = 1, \ldots, n_i)$. In a longitudinal study, $i$ indicates subject and $j$ the measurement occasion. For multivariate studies, $j$ refers to the particular outcome variable. In a hierarchical data setting, with more than two levels, $j$ can be taken to refer generically to all sub-levels, in which case it would become a vector-valued indicator. The index $i$ is always reserved for units (blocks) of independent replication.

We group the outcomes into a vector $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})'$. In addition, for each

occasion $j$ we define

$$R_{ij} \quad = \quad \begin{cases} 1 & \text{if } Y_{ij} \text{ is observed} \\ 0 & \text{otherwise.} \end{cases}$$

These *missing data indicators* $R_{ij}$ are organized into a vector $\boldsymbol{R}_i$ of parallel structure to $\boldsymbol{Y}_i$.

We then partition $\boldsymbol{Y}_i$ into two subvectors such that $\boldsymbol{Y}_i^o$ is the vector containing those $Y_{ij}$ for which $R_{ij} = 1$, and $\boldsymbol{Y}_i^m$ contains the remaining components. These subvectors are referred to as the *observed* and *missing* components, respectively. The following terminology is adopted. *Complete data* refers to the vector $\boldsymbol{Y}_i$ of planned measurements. This is the outcome vector that would have been recorded if no data had been missing. The vector $\boldsymbol{R}_i$ and the process generating $\boldsymbol{R}_i$ is referred to as the *missing data process*. The *full data* $(\boldsymbol{Y}_i, \boldsymbol{R}_i)$ consist of the complete data, together with the missing data indicators. Note that, unless all components of $\boldsymbol{R}_i$ equal one, the full data components are never jointly observed but rather one observes the measurements $\boldsymbol{Y}_i^o$ together with the dropout indicators $\boldsymbol{R}_i$, which we refer to as the *observed data.*

When missingness is restricted to dropout or attrition, we can replace the vector $\boldsymbol{R}_i$ by a scalar variable $D_i$, the *dropout indicator*. In this case, each vector $\boldsymbol{R}_i$ is of the form $(1, \ldots, 1, 0, \ldots, 0)$ and we can define the scalar dropout indicator

$$D_i \quad = \quad 1 + \sum_{j=1}^{n_i} R_{ij}. \tag{3.23}$$

For an incomplete sequence, $D_i$ denotes the occasion at which dropout occurs. For a complete sequence, $D_i = n_i + 1$. In both cases, $D_i$ is equal to one plus the length of the measurement sequence, whether complete or incomplete. Sometimes, it is convenient to define an alternative dropout indicator, $T_i = D_i - 1$, that indicates the number of measurements actually taken, rather then the first occasion at which the planned measurement has not been taken.

Dropout, or attrition, is an example of a *monotone* pattern of missingness. Missingness is termed monotone when there exists a permutation of the measurement occasions such that a measurement earlier in the permuted sequence is observed for at least those subjects that are observed at later measurements. Note that, for this definition to be meaningful, we need to have a balanced design in the sense of a common set of designed measurement occasions. Other patterns are called *non-monotone*.

### 3.5.3 Missing Data Frameworks

We now consider in turn the so-called *selection, pattern-mixture,* and *shared-parameter* modeling frameworks.

When data are incomplete due to the operation of a random (missing value) mechanism the appropriate starting point for a model is the full data density

$$f(\boldsymbol{y}_i, \boldsymbol{r}_i | X_i, W_i, \boldsymbol{\theta}, \boldsymbol{\psi}), \tag{3.24}$$

where $X_i$ and $W_i$ denote design matrices for the measurement and missingness mechanism, respectively. The corresponding parameter vectors are $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$, respectively.

The *selection model* factorization is based on

$$f(\boldsymbol{y}_i, \boldsymbol{r}_i | X_i, W_i, \boldsymbol{\theta}, \boldsymbol{\psi}) = f(\boldsymbol{y}_i | X_i, \boldsymbol{\theta}) f(\boldsymbol{r}_i | \boldsymbol{y}_i, W_i, \boldsymbol{\psi}), \tag{3.25}$$

where the first factor is the marginal density of the measurement process and the second one is the density of the missingness process, conditional on the outcomes. The name is chosen because $f(\boldsymbol{r}_i | \boldsymbol{y}_i, W_i, \boldsymbol{\psi})$ can be seen as describing a unit's self-selection mechanism to either continue or leave the study. The term originates from the econometric literature (Heckman 1976) and it can be thought of that a subject's missing values are "selected" through the probability model, given their measurements, whether observed or not.

An alternative family is based on so-called *pattern-mixture models* (Little 1993, 1994, 1995). These are based on the factorization

$$f(\boldsymbol{y}_i, \boldsymbol{r}_i | X_i, W_i, \boldsymbol{\theta}, \boldsymbol{\psi}) = f(\boldsymbol{y}_i | \boldsymbol{r}_i, X_i, \boldsymbol{\theta}) f(\boldsymbol{r}_i | W_i, \boldsymbol{\psi}). \tag{3.26}$$

The pattern-mixture model allows for a different response model for each pattern of missing values, the observed data being a mixture of these weighted by the probability of each missing value or dropout pattern.

The third family is referred to as *shared-parameter models*:

$$\begin{aligned} & f(\boldsymbol{y}_i, \boldsymbol{r}_i | X_i, W_i, \boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{b}_i) \\ = \;& f(\boldsymbol{y}_i | \boldsymbol{r}_i, X_i, \boldsymbol{\theta}, \boldsymbol{b}_i) f(\boldsymbol{r}_i | W_i, \boldsymbol{\psi}, \boldsymbol{b}_i), \end{aligned} \tag{3.27}$$

where we explicitly include a vector of unit-specific latent (or random) effects $\boldsymbol{b}_i$ of which one or more components are shared between both factors in the joint distribution. Early references to such models are Wu and Carroll (1988) and Wu and Bailey

(1988, 1989). A sensible assumption is that $\boldsymbol{Y}_i$ and $\boldsymbol{R}_i$ are conditionally independent, given the random effects $\boldsymbol{b}_i$. The random effects $\boldsymbol{b}_i$ can be used to define an appropriate hierarchical model. The same vector can then be used to describe the missing data process. The shared parameter $\boldsymbol{b}_i$ can be thought of as referring to a latent trait driving both the measurement and missingness processes.

The natural parameters of selection models, pattern-mixture models, and shared-parameter models have different interpretations, and transforming one statistical model from one of the frameworks to another is generally not straightforward, and these three models can indeed lead to different results and conclusions.

### 3.5.4 Missing Data Mechanisms

Rubin's taxonomy of missing value processes (Rubin 1976, Little and Rubin, 2002), referred to in Section 3.5.1, is fundamental to the modeling of incomplete data. It is perhaps most naturally expressed within the selection modeling framework for which it is based on the second factor of (3.25):

$$f(\boldsymbol{r}_i|\boldsymbol{y}_i, W_i, \boldsymbol{\psi}) = f(\boldsymbol{r}_i|\boldsymbol{y}_i^o, \boldsymbol{y}_i^m, W_i, \boldsymbol{\psi}). \tag{3.28}$$

**Missing Completely at Random (MCAR).** Under an MCAR mechanism, the probability of an observation being missing is independent of the responses:

$$f(\boldsymbol{r}_i|\boldsymbol{y}_i, W_i, \boldsymbol{\psi}) = f(\boldsymbol{r}_i|W_i, \boldsymbol{\psi}) \tag{3.29}$$

and therefore (3.25) simplifies to

$$f(\boldsymbol{y}_i, \boldsymbol{r}_i|X_i, W_i, \boldsymbol{\theta}, \boldsymbol{\psi}) = f(\boldsymbol{y}_i|X_i, \boldsymbol{\theta}) f(\boldsymbol{r}_i|W_i, \boldsymbol{\psi}), \tag{3.30}$$

implying that both components are independent. The implication is that the joint distribution of $\boldsymbol{y}_i^o$ and $\boldsymbol{r}_i$ becomes

$$f(\boldsymbol{y}_i^o, \boldsymbol{r}_i|X_i, W_i, \boldsymbol{\theta}, \boldsymbol{\psi}) = f(\boldsymbol{y}_i^o|X_i, \boldsymbol{\theta}) f(\boldsymbol{r}_i|W_i, \boldsymbol{\psi}). \tag{3.31}$$

Under MCAR, the observed data can be analyzed as though the pattern of missing values was predetermined. In whatever way the data are analyzed, whether using a frequentist, likelihood, or Bayesian procedure, the process(es) generating the missing values can be ignored. For example, in this situation simple averages of the observed data at different occasions provide unbiased estimates of the corresponding population

averages. The observed data can be regarded as a random sample of the complete data.

Note that this definition and the ones to follow are made conditionally on the covariates. When the covariates, assembled into $X_i$ and $W_i$, are removed, the nature of a mechanism may change. In defining these mechanisms some authors distinguish between those made conditionally or not on the covariates.

**Missing at Random (MAR).** Under an MAR mechanism, the probability of an observation being missing is *conditionally* independent of the unobserved outcome(s), given the values of the observed outcome(s):

$$f(\boldsymbol{r}_i|\boldsymbol{y}_i, W_i, \boldsymbol{\psi}) = f(\boldsymbol{r}_i|\boldsymbol{y}_i^o, W_i, \boldsymbol{\psi}). \tag{3.32}$$

and again the joint distribution of the observed data can be partitioned:

$$f(\boldsymbol{y}_i, \boldsymbol{r}_i|X_i, W_i, \boldsymbol{\theta}, \boldsymbol{\psi}) = f(\boldsymbol{y}_i|X_i, \boldsymbol{\theta}) f(\boldsymbol{r}_i|\boldsymbol{y}_i^o, W_i, \boldsymbol{\psi}), \tag{3.33}$$

and hence at the observed data level:

$$f(\boldsymbol{y}_i^o, \boldsymbol{r}_i|X_i, W_i, \boldsymbol{\theta}, \boldsymbol{\psi}) = f(\boldsymbol{y}_i^o|X_i, \boldsymbol{\theta}) f(\boldsymbol{r}_i|\boldsymbol{y}_i^o, W_i, \boldsymbol{\psi}). \tag{3.34}$$

Given the simplicity of (3.34), handling of MAR processes is typically easier than handling MNAR.

Although the MAR assumption is particularly convenient in that it leads to considerable simplification in the issues surrounding the analysis of incomplete longitudinal data, an investigator is rarely able to justify its adoption, and so in many situations the final class of missing value mechanisms cannot be ruled out.

**Missing Not at Random (MNAR).** In this case, neither MCAR nor MAR hold. Under MNAR, the probability of a measurement being missing depends on unobserved outcome(s). No simplification of the joint distribution is possible and the joint distribution of the observed measurements and the missingness process has to be written as:

$$f(\boldsymbol{y}_i^o, \boldsymbol{r}_i|X_i, W_i, \boldsymbol{\theta}, \boldsymbol{\psi}) = \int f(\boldsymbol{y}_i|X_i, \boldsymbol{\theta}) f(\boldsymbol{r}_i|\boldsymbol{y}_i, W_i, \boldsymbol{\psi}) d\boldsymbol{y}_i^m. \tag{3.35}$$

Inferences can only be made by making further assumptions, about which the observed data alone carry no information. Ideally, if such models are to be used, the choice of such assumptions should be guided by external information, but the degree to which this is possible varies greatly across application areas and applications. Such models

can be formulated within each of the three main families: selection, pattern-mixture, and shared-parameter models. The differences between the families are especially important in the MNAR case, and lead to quite different, but complementary, views of the missing value problem. Little (1995), Hogan and Laird (1997), and Kenward and Molenberghs (1999) provide detailed reviews. See also Verbeke and Molenberghs (2000) and Molenberghs and Verbeke (2005).

It has been shown, for dropout in longitudinal studies, how Rubin's classification can be applied in the pattern-mixture framework as well (Molenberghs *et al.*, 1998, Kenward, Molenberghs, and Thijs, 2003). We will discuss these points in Section 3.5.6.

The MCAR–MAR–MNAR terminology is independent of the inferential framework chosen. This is different for the concept of *ignorability*, which depends crucially on this framework (Rubin, 1976). We will turn to this issue in the next section.

### 3.5.5 Ignorability

In this section we focus on likelihood-based estimation. The full data likelihood contribution for unit $i$ takes the form

$$L^*(\boldsymbol{\theta}, \boldsymbol{\psi} | X_i, W_i, \boldsymbol{y}_i, \boldsymbol{r}_i) \quad \propto \quad f(\boldsymbol{y}_i, \boldsymbol{r}_i | X_i, W_i, \boldsymbol{\theta}, \boldsymbol{\psi}).$$

Because inference has to be based on what is observed, the full data likelihood $L^*$ needs to be replaced by the observed data likelihood $L$:

$$L(\boldsymbol{\theta}, \boldsymbol{\psi} | X_i, W_i, \boldsymbol{y}_i^o, \boldsymbol{r}_i) \quad \propto \quad f(\boldsymbol{y}_i^o, \boldsymbol{r}_i | X_i, W_i, \boldsymbol{\theta}, \boldsymbol{\psi}) \tag{3.36}$$

with

$$\begin{aligned} f(\boldsymbol{y}_i^o, \boldsymbol{r}_i | \boldsymbol{\theta}, \boldsymbol{\psi}) &= \int f(\boldsymbol{y}_i, \boldsymbol{r}_i | X_i, W_i, \boldsymbol{\theta}, \boldsymbol{\psi}) d\boldsymbol{y}_i^m \\ &= \int f(\boldsymbol{y}_i^o, \boldsymbol{y}_i^m | X_i, \boldsymbol{\theta}) f(\boldsymbol{r}_i | \boldsymbol{y}_i^o, \boldsymbol{y}_i^m, W_i, \boldsymbol{\psi}) d\boldsymbol{y}_i^m. \end{aligned} \tag{3.37}$$

Under an MAR process, we obtain

$$\begin{aligned} f(\boldsymbol{y}_i^o, \boldsymbol{r}_i | \boldsymbol{\theta}, \boldsymbol{\psi}) &= \int f(\boldsymbol{y}_i^o, \boldsymbol{y}_i^m | X_i, W_i, \boldsymbol{\theta}) f(\boldsymbol{r}_i | \boldsymbol{y}_i^o, W_i, \boldsymbol{\psi}) d\boldsymbol{y}_i^m \\ &= f(\boldsymbol{y}_i^o | X_i, W_i, \boldsymbol{\theta}) f(\boldsymbol{r}_i | \boldsymbol{y}_i^o, W_i, \boldsymbol{\psi}). \end{aligned} \tag{3.38}$$

Thus, the likelihood factors into two components of the same functional form as the general factorization (3.25) of the complete data. If, further, $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ are disjoint

in the sense that the parameter space of the full vector $(\boldsymbol{\theta}', \boldsymbol{\psi}')'$ is the product of the parameter spaces of $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$, then inference can be based solely on the marginal observed data density. This technical requirement is referred to as the separability condition. However, still some caution should be used when constructing precision estimators. This point is discussed in detail in Chapter 12 of Molenberghs and Kenward (2007).

In conclusion, when the separability condition is satisfied, *within the likelihood framework*, ignorability is equivalent to the union of MAR and MCAR. A formal derivation is given in Rubin (1976), where it is also shown that the same requirements hold for Bayesian inference, but that for frequentist inference to be ignorable, MCAR is the corresponding sufficient condition. Of course, it is possible that at least part of the scientific interest is directed towards the missing data process. Then still, ignorability is useful since the measurement model and missingness model questions can be addressed through separate models, rather than jointly.

Classical examples of the more stringent condition with frequentist methods are ordinary least squares and the generalized estimating equations (GEE) approach of Liang and Zeger (1986). The latter produce unbiased estimators in general only under MCAR. Robins, Rotnitzky, and Zhao (1995) and Rotnitzky and Robins (1995) have established that some progress can be made under MAR and that, even under MNAR processes, these methods can be applied (Rotnitzky and Robins, 1997, Robins, Rotnitzky, and Scharfstein, 1998).

### 3.5.6 Pattern-mixture Models

Pattern-mixture models (PMM) were introduced in Section 3.5.3 as one of the three major frameworks within which missing data models can be developed.

Little (1993, 1994, 1995) originally proposed the use of pattern-mixture models as a viable alternative to selection models.

An important issue is that pattern-mixture models are by construction under-identified, *i.e.,* overspecified. Little (1993, 1994) solves this problem through the use of identifying restrictions: inestimable parameters of the incomplete patterns are set equal to (functions of) the parameters describing the distribution of the completers. Identifying restrictions are not the only way to overcome under-identification alternative approaches are discussed in Chapter 17 of Molenberghs and Kenward (2007). Although some authors perceive this under-identification as a drawback, it can be

viewed as an asset because it forces one to reflect on the assumptions being made, and the assumptions are necessarily transparent. This can serve as a useful starting point for sensitivity analysis.

Little (1993, 1994) advocated the use of identifying restrictions and presented a number of examples. One of those, ACMV (available case missing values), is the natural counterpart of MAR in the PMM framework, as was established by Molenberghs *et al.* (1998). Specific counterparts to MNAR selection models were studied by Kenward, Molenberghs, and Thijs (2003). These will be discussed in what follows.

In line with Molenberghs *et al.* (1998), we restrict attention to monotone patterns, dropping the unit index $i$ from the notation, for simplicity. In general, let us assume that there are $t = 1, \ldots, n = T$ dropout patterns where the dropout indicator, introduced in (3.23), is $d = t + 1$. The indices $j$ for measurements occasions and $t$ for dropout patterns assume the same values, but using both simplifies notation.

For pattern $t$, the complete data density, is given by

$$f_t(y_1, \ldots, y_T) = f_t(y_1, \ldots, y_t) f_t(y_{t+1}, \ldots, y_T | y_1, \ldots, y_t). \tag{3.39}$$

The first factor is clearly identified from the observed data, while the second factor is not. It is assumed that the first factor is known or, more realistically, modeled using the observed data. Then, identifying restrictions are applied in order to identify the second component.

Although, in principle, completely arbitrary restrictions can be used by means of any valid density function over the appropriate support, strategies that imply links back to the observed data are likely to have more practical relevance. One can base identification on all patterns for which a given component, $y_s$ say, is identified. A general expression for this is

$$f_t(y_s | y_1, \ldots y_{s-1}) = \sum_{j=s}^{T} \omega_{sj} f_j(y_s | y_1, \ldots y_{s-1}), \quad s = t+1, \ldots, T. \tag{3.40}$$

We will use $\boldsymbol{\omega}_s$ as shorthand for the set of $\omega_{sj}$'s used, the components of which are typically positive. Every $\boldsymbol{\omega}_s$ that sums to one provides a valid identification scheme.

Let us incorporate (3.40) into (3.39):

$$\begin{aligned} &f_t(y_1, \ldots, y_T) \\ &\quad = f_t(y_1, \ldots, y_t) \prod_{s=0}^{T-t-1} \left[ \sum_{j=T-s}^{T} \omega_{T-s,j} f_j(y_{T-s} | y_1, \ldots, y_{T-s-1}) \right]. \end{aligned} \tag{3.41}$$

We will consider three special but important cases, associated with such choices of $\boldsymbol{\omega}_s$ in (3.40). Little (1993) proposes CCMV (complete case missing values) which uses the following identification:

$$f_t(y_s|y_1,\ldots y_{s-1}) = f_T(y_s|y_1,\ldots y_{s-1}), \quad s = t+1,\ldots,T, \tag{3.42}$$

corresponding to $\omega_{sT} = 1$ and all others zero. In other words, information which is unavailable is always borrowed from the completers. Alternatively, the nearest identified pattern can be used:

$$f_t(y_s|y_1,\ldots y_{s-1}) = f_s(y_s|y_1,\ldots y_{s-1}), \qquad s = t+1,\ldots,T, \tag{3.43}$$

corresponding to $\omega_{ss} = 1$ and all others zero. We will refer to these restrictions as *neighboring case missing values* or NCMV.

The third special case of (3.40) is ACMV. ACMV is reserved for the counterpart of MAR in the PMM context. The corresponding $\boldsymbol{\omega}_s$ vectors can be shown (Molenberghs *et al.* 1998) to have components:

$$\omega_{sj} = \frac{\alpha_j f_j(y_1,...,y_{s-1})}{\sum'_{\ell=s} \alpha_\ell f_\ell(y_1,\ldots,y_{s-1})}, \tag{3.44}$$

$(j = s,\ldots,T)$ where $\alpha_j$ is the fraction of observations in pattern $j$ (Molenberghs *et al.* 1998).

This MAR–ACMV link connects the selection and pattern-mixture families. It is of further interest to consider specific sub-families of the MNAR family. In the context of selection models for longitudinal data, one typically restricts attention to a class of mechanisms where dropout may depend on the current, possibly unobserved, measurement, but not on future measurements. The entire class of such models will be termed missing non-future dependent (MNFD). Although they are natural and easy to consider in a selection model situation, there exist important examples of mechanisms that do not satisfy MNFD, such as shared-parameter models (Wu and Bailey, 1989, Little, 1995).

Kenward, Molenberghs, and Thijs (2003) have shown there is a counterpart to MNFD in the pattern-mixture context. The conditional probability of pattern $t$ in the MNFD selection models obviously satisfies

$$f(r = t|y_1,\ldots,y_T) = f(r = t|y_1,\ldots,y_{t+1}). \tag{3.45}$$

Within the PMM framework, we define non-future dependent missing value restrictions (NFMV) as follows:

$$\begin{aligned} & \forall t \geq 2, \forall j < t-1 \quad : \\ f(y_t|y_1,\ldots,y_{t-1}, r=j) \quad &= \quad f(y_t|y_1,\ldots,y_{t-1}, r \geq t-1). \end{aligned} \tag{3.46}$$

NFMV is not a single set of restrictions, but rather leaves one conditional distribution per incomplete pattern unidentified:

$$f(y_{t+1}|y_1,\ldots,y_t, r=t). \tag{3.47}$$

In other words, the distribution of the "current" unobserved measurement, given the previous ones, is unconstrained. Note that (3.46) excludes such mechanisms as CCMV and NCMV. Kenward, Molenberghs, and Thijs (2003) have shown that, for longitudinal data with dropouts, MNFD and NFMV are equivalent.

For pattern $t$, the complete data density is given by

$$\begin{aligned} f_t(y_1,\ldots,y_T) \quad &= \quad f_t(y_1,\ldots,y_t) f_t(y_{t+1}|y_1,\ldots,y_t) \\ & \quad \times f_t(y_{t+2},\ldots,y_T|y_1,\ldots,y_{t+1}). \end{aligned} \tag{3.48}$$

It is assumed that the first factor is known or, more realistically, modeled using the observed data. Then, identifying restrictions are applied to identify the second and third components. First, from the data, estimate $f_t(y_1,\ldots,y_t)$. Second, the user has full freedom to choose

$$f_t(y_{t+1}|y_1,\ldots,y_t). \tag{3.49}$$

Substantive considerations could be used to identify this density. Alternatively, a family of densities might be considered by way of sensitivity analysis. Third, using (3.46), the densities $f_t(y_j|y_1,\ldots,y_{j-1})$, $(j \geq t+2)$ are identified. This identification involves not only the patterns for which $y_j$ is observed, but also the pattern for which $y_j$ is the current and hence the first unobserved measurement. An overview of the connection between selection and pattern-mixture models is given in Figure 3.1.

Two obvious mechanisms, within the MNFD family but outside MAR, are NFD1 (NFD standing for 'non-future dependent'), *i.e.,* choose (3.49) according to CCMV, and NFD2, *i.e.,* choose (3.49) according to NCMV. NFD1 and NFD2 are strictly different from CCMV and NCMV.

$$
\begin{array}{lcccccccc}
\text{SEM} & : & \text{MCAR} & \subset & \text{MAR} & \subset & \text{MNFD} & \subset & \text{general MNAR} \\
 & & \updownarrow & & \updownarrow & & \updownarrow & & \updownarrow \\
\text{PMM} & : & \text{MCAR} & \subset & \text{ACMV} & \subset & \text{NFMV} & \subset & \text{general MNAR} \\
 & & & & & \curvearrowleft & \neq & \curvearrowright & \\
 & & & & & & \text{interior} & &
\end{array}
$$

Figure 3.1: *Relationship between nested families within the selection model (SEM) and pattern-mixture model (PMM) families. MCAR: missing completely at random; MAR: missing at random; MNAR: missing not at random; MNFD: missing non-future dependence; ACMV: available-case missing values; NFMV: non-future missing values; interior: restrictions based on a combination of the information available for other patterns. The '$\subset$' symbol here indicates 'is a special case of.' The '$\updownarrow$' symbol indicates correspondence between a class of SEM models and a class of PMM models.*

# Chapter 4

# Concepts in Psychometric Methodology

In this Chapter we will introduce concepts in Psychometric Validation methodology. First we will give a brief overview on the Classical Theory of Reliability. The ICC of reliability, defined in Section 4.1, will be developed in Chapter 6, 8 and 10 using the case studies decribed in Chapter 2. Subsequently, we will expand on Generalizability Theory in Section 4.2 which is a natural extension of reliability. Similarly, the concepts of Generalizability Theory will be developed and applied in Chapter 7 and 9. We will end this Chapter with the concept of Validity testing in Section 4.3. This will serve as a basis for Chapter 11 where we develop methods to evaluate criterion validity on parallel measurements from the case studies.

## 4.1 Reliability

The terms observer *reliability* and *agreement* are often used interchangeably, but in theory they are different concepts. Reliability coefficients express the ability to differentiate among subjects. They are ratios of variances: in general, the variance attributed to the difference among subjects divided by the total variance (Shrout and Fleiss, 1979). Reliability concerns the consistency of repeated measures, whether the same value is achieved if a measurement is performed twice. The repetitions might

be repeated measures by the same rater, also referred to *test-retest reliability*, or alternatively a subject might be measured by multiple raters, also referred to *interrater reliability*. Also consistency between questions or measures within a subscale is a form of reliability also called *internal consistency*

The parameters for assessment of observer reliability and agreement differ according to the scale of measurement. For nominal and ordinal categorical measurements, respectively the $\kappa$-coefficient and the weighted $\kappa$-coefficient ($\kappa_W$) are measures of agreement (Dunn, 1989, 2000 and Shoukri, 2004). In case of continuous data, the intraclass correlation coefficient (ICC) is used to measure observer reliability, although the ICC also can be used for ordinal categorical data.

As stated by Fleiss: 'The most elegant design of a clinical study will not overcome the damage by unreliable or imprecise measurement' (Fleiss, 1986). In clinical trials, one typically wants to differentiate among treatments. If reliability is low, the ability to differentiate between the different subjects in the different treatment arm decreases. Fleiss describes a number of consequences of *unreliability*. He brings up attenuation of correlation in studies designed to estimate correlation between variables with poor reliability, biased sample selection in clinical studies where patients are selected with a minimum level of a certain measurement with low reliability, and last but not least, an increased sample size for trials with a primary parameter with low reliability. For the latter, one can easily show that for a paired t-test, the required sample size becomes $n = \frac{n^*}{R}$ where $R$ denotes the reliability coefficient and $n^*$ is the required sample size for the true score, i.e., the required sample size when responses are measured without error. It is very clear that a high reliability is important to the clinical trialist. Investigators in the mental disorders traditionally have been more concerned with the reliability of their measures than have their colleagues in other medical specialties.

In the classical test theory, the outcome of an interval scaled test is modeled as

$$Y = \tau + \varepsilon, \tag{4.1}$$

where $Y$ represents an observation or measurement, $\tau$ is the true score and $\varepsilon$ the corresponding measurement error. It is assumed that the measurement errors are mutually uncorrelated as well as with the true scores. If this assumption is correct, we obtain

$$\mathrm{Var}(Y) = \mathrm{Var}(\tau) + \mathrm{Var}(\varepsilon).$$

The reliability of a measuring instrument is defined as the ratio of the true score

variance to the observed score variance, i.e.,

$$R = \frac{\text{Var}(\tau)}{\text{Var}(Y)} = \frac{\text{Var}(\tau)}{\text{Var}(\tau) + \text{Var}(\varepsilon)}. \tag{4.2}$$

One can easily show that the reliability coefficient is in fact an intraclass correlation coefficient. Suppose we have two measurements of the same patient, either from two raters or from the same rater, taken at two instances not too far apart, $Y_1 = \tau + \varepsilon_1$ and $Y_2 = \tau + \varepsilon_2$, with $\text{Var}(Y_1) = \text{Var}(Y_2) = \text{Var}(Y)$ and $\text{Var}(\varepsilon_1) = \text{Var}(\varepsilon_2) = \text{Var}(\varepsilon)$ , i.e., parallel measurements. Further, the covariance of the two measurements equals

$$\text{Cov}(Y_1, Y_2) = \text{Cov}(\tau + \varepsilon_1, \tau + \varepsilon_2) = \text{Var}(\tau),$$

and the correlation between the two measurements can be written as

$$\text{Corr}(Y_1, Y_2) = \frac{\text{Cov}(Y_1, Y_2)}{\sqrt{\text{Var}(\boldsymbol{Y}_1)}\sqrt{\text{Var}(\boldsymbol{Y}_2)}} = \frac{\text{Var}(\tau)}{\text{Var}(\tau) + \text{Var}(\varepsilon)} = R. \tag{4.3}$$

This shows that reliability is in fact an intraclass correlation coefficient with patient taken as the class. Stronger, as Bartko stated, a reliability coefficient defined as a ratio of variances which are estimated by a linear model can only be correct when it can be interpreted as a "correlation coefficient" (Bartko, 1966).

Note that the assumption of steady state behavior of the measurements, i.e., the assumption that measurements are parallel (same mean and same variance), is crucial. If for instance the patients are rated by the same investigator on two occasions which are too far apart, the patient's condition can have changed, translating into a low intraclass correlation coefficient, even in the case of highly reliable measures.

In the classical approach, reliability is estimated by the intraclass correlation coefficient (Fleiss, 1986, Bartko, 1966, Dunn, 1989). For a simple replication study, this can be derived from a one way analysis of variance with patient as factor (Table 4.1).

The estimate for the intraclass correlation coefficient of reliability in (Bartko, 1966) then is:

$$\widehat{R}_c = \frac{\widehat{\sigma}_p^2}{\widehat{\sigma}_p^2 + \widehat{\sigma}_e^2} = \frac{\text{BMS} - \text{WMS}}{\text{BMS} + (k-1)\text{WMS}}.$$

## 4.2 Generalizability

The classical theory behind the estimation of reliability can be extended to *Generalizability Theory* by estimating the magnitude of multiple sources of measurement error

Table 4.1: *Classical Theory. ANOVA table to derive reliability coefficient from a simple replication study.*

| Source of variation | Df | Mean sum of sq. | Exp. sum of sq. |
|---|---|---|---|
| Between patient | $n-1$ | BMS | $\sigma_e^2 + k\sigma_p^2$ |
| Within patients (error) | $n(k-1)$ | WMS | $\sigma_e^2$ |
| Total | $nk-1$ | | |

where $k$ is the number of measurements per patient, $n$ is the number of patients.

Table 4.2: *Classical Theory. Analysis of variance table for Person by Occasion design.*

| Source of variation | Df (*) | Mean Sum of Sq. | Exp. Sum of Sq. |
|---|---|---|---|
| Person | $n_P - 1$ | $MS_P(BMS)$ | $\sigma_E^2 + n_O\sigma_P^2$ |
| Occasion | $n_O - 1$ | $MS_O(WMS)$ | $\sigma_E^2 + n_P\sigma_O^2$ |
| Person x Occasion (err.) | $(n_P-1)\times(n_O-1)$ | $MS_E$ | $\sigma_E^2$ |

(*)$n_O$ is the number of measurements per patient, $n_P$ the number of persons

and providing reliability and generalizability coefficients tailored to the proposed use of the measurement and isolating major sources of error so that a cost efficient measurement design can be built (Shavelson, Webb, and Rowley, 1989). By investigating other sources of error such as for instance country or sub category of diagnosis, the clinical trialist could learn a lot about performance of scales or other measurements in certain subgroups and what the impact of such factors is on reliability.

In a $Person \times Occasion$ design, where occasion could be two time points (test-retest) or two raters (interrater), there are 3 sources of variation (Table 4.2), Person, Occasion and Person $\times$ Occasion confounded with error. Model (4.1) could be written as

$$Y_{PO} = \mu + \mu_P + \mu_O + \varepsilon. \tag{4.4}$$

For the Person $\times$ Occasion design, the reliability coefficient is estimated with two

of the three sources of variation, Person and Residual of the ANOVA model described in Table 4.2:

$$\widehat{R} = \frac{\widehat{\sigma}_P^2}{\widehat{\sigma}_P^2 + \widehat{\sigma}_E^2} = \frac{\text{BMS} - \text{WMS}}{\text{BMS} + (n_O - 1)\text{WMS}}. \tag{4.5}$$

The occasion effect should be zero because Classical Theory assumes strictly parallel measurements. Only Person and Residual variation give rise to differences among individuals; the occasion effect is constant for all individuals in the $P \times O$ design (Shavelson, Webb, and Rowley, 1989). If we consider patient to be random in model (4.4), one can also easily show that the reliability coefficient $R$ is the correlation coefficient between measurements of the same patient, on different occasions, conditioning for occasion, i.e., keeping occasion fixed:

$$R = \text{Corr}(Y_{PO}, Y_{PO'} \mid O, O'). \tag{4.6}$$

In classical test theory (4.1), an observation is assumed to be a combination of an individual's *true* score and random measurement error. The assumption that all variance in scores can be divided into true and error variance is rather simplistic. Furthermore, there are many approaches to estimate reliability, each of which generates a different coefficient: inter-rater reliability, test-retest reliability, internal consistency. This will lead to different estimates of the true scores for each study, with no logical way to combine them. In addition to the true score of an individual, multiple potential sources of error can exist. The goal is to obtain the most precise estimate of the score that person should have if there were no sources of error contaminating our results; each of the multiple forms of reliability we have mentioned above identifies and quantifies only one source error variance at the time. What we really need is some way of combining all the sources of variability in a single study, using all the data to estimate the variance between subjects and the various components of error variance. This can provide a lot of information on observer reliability and can determine the relative importance of each component. This broad approach was originally devised by Cronbach (1963) and his associates and is known as *Generalizability Theory (GT)*. The essence of the theory is the recognition that in any measurement situation, there are multiple sources of error variance. The goal is to attempt to identify, measure, and thereby possibly find strategies to reduce the influence of these sources on the measurement in question. Imagine that we could identify the most likely sources of error in a measurement of some characteristic of a person. We then have defined our "universe" of possible observations. If we then proceed to average each person's score

over all of these possible conditions, this would be an unbiased estimate of that person's score over the universe as we have defined it. Note that there is no pretense that this is the "true" score, since we may well have guessed wrong about the universe. If we can reasonably identify possible sources of error, we can incorporate them into a *generalizability study* or G-study. Consider the following design (Table 4.3) of a Generalizability study investigating both inter-rater and intra-rater variability, assuming that rater and day of observation are the most important sources of error.

Table 4.3: *Generalizability Theory. Test-retest and interrater design.*

| | Observer 1 | | Observer 2 | |
|---|---|---|---|---|
| Patient | Day 1 | Day 7 | Day 1 | Day 7 |
| 1 | $Y_{111}$ | $Y_{112}$ | $Y_{121}$ | $Y_{122}$ |
| 2 | $Y_{211}$ | $Y_{212}$ | $Y_{221}$ | $Y_{222}$ |
| 3 | $Y_{311}$ | $Y_{312}$ | $Y_{321}$ | $Y_{322}$ |
| . | . | . | . | . |
| . | . | . | . | . |
| 10 | $Y_{1011}$ | $Y_{1012}$ | $Y_{1021}$ | $Y_{1022}$ |

Instead of the simple CT decomposition (4.1), GT decomposes the observed score as follows

$$\begin{aligned} Y_{prd} \quad &= \mu && \text{Grand mean} \\ &+\mu_P && \text{Person effect} \\ &+\mu_R && \text{Rater effect} \\ &+\mu_D && \text{Day effect} \\ &+\mu_{PR} && \text{Person x Rater effect} \\ &+\mu_{PD} && \text{Person x Day effect} \\ &+\mu_{RD} && \text{Rater x Day effect} \\ &+\varepsilon && \text{Residual Error.} \end{aligned} \tag{4.7}$$

The associated sources of variability are denoted by $\sigma_P^2$ for the person effect, $\sigma_R^2$ for

the rater effect, etc. This approach will enable us to estimate the magnitude of the variance in observed scores due to universe-score variance and to multiple sources of error, rater and day in the example above. If the sources that we have identified are trivial, and we have missed some important source of error, then there will be a large amount of variance due to random error or residual. In terms of GT, *Person* is a *facet of differentiation* and *rater* and *day* are called *facets of generalization*. The levels of the facets of generalization are called *conditions*. ANOVA (Table 4.4) is mostly used to study and estimate the various variance components, often ignoring intra subject correlation. From these estimated variance components, a *generalizability coefficient*, analogous to a reliability coefficient, can be calculated by dividing the estimated person variance component by an estimated observed score variance. GT distinguishes between decisions based on the relative standing of individuals and decisions based on the absolute value of a score (Shavelson, Webb, and Rowley, 1989).

Table 4.4: *Generalizability Theory. ANOVA model for Test-retest and inter-rater design.*

| Source of variation | Df. | SS | MSS | Estimated variance component |
|---|---|---|---|---|
| Patient P | 9 | $SS_P$ | $MS_P$ | $\sigma^2_P = (MS_P - MS_{PO} - MS_{PD} + MS_E)/4$ |
| Observer O | 1 | $SS_O$ | $MS_O$ | $\sigma^2_O = (MS_O - MS_{PO} - MS_{OD} + MS_E)/20$ |
| Day D | 1 | $SS_D$ | $MS_D$ | $\sigma^2_D = (MS_D - MS_{PD} - MS_{OD} + MS_E)/20$ |
| P x O | 9 | $SS_{PO}$ | $MS_{PO}$ | $\sigma^2_{PO} = (MS_{PO} - MS_E)/2$ |
| P x D | 9 | $SS_{PD}$ | $MS_{PD}$ | $\sigma^2_{PD} = (MS_{PD} - MS_E)/2$ |
| O x D | 1 | $SS_{OD}$ | $MS_{OD}$ | $\sigma^2_{OD} = (MS_{OD} - MS_E)/10$ |
| P x O x D(Error) | 9 | $SS_E$ | $MS_E$ | $\sigma^2_E = MS_E$ |

Error in *relative decisions* arises from all nonzero variance components associated with rank ordering of individuals, other than the component for the object of measurement (persons). Specifically, variance components associated with the interaction of person with each facet or combinations of facets define error. For the example above we have $\sigma^2_{PR}$, $\sigma^2_{PD}$, $\sigma^2_{PRD} = \sigma^2_E$. So if one wishes to generalize from a rating by one rater on one day to a rating by a different rater at another point in time, the following generalizability coefficient can be constructed as the ratio of the universe-score

variance to the expected rater-score variance, i.e., an $ICC$:

$$\begin{aligned} E_{\rho^2 Rel} &= \text{Corr}(Y_{PRD}, Y_{PR'D'} \mid R, R', D, D') \\ &= \frac{\sigma_P^2}{\sigma_P^2 + \sigma_{Rel.Error}^2} \\ &= \frac{\sigma_P^2}{\sigma_P^2 + \sigma_{PR}^2 + \sigma_{PD}^2 + \sigma_{PRD}^2}. \end{aligned} \tag{4.8}$$

Indeed, it is easy to show that equation (4.8) can be derived as a conditional correlation coefficient for model (4.7) where we condition on rater and day, but where rater and day can be different. Alternatively, we can derive a test-retest or an interrater reliability coefficient, by only generalizing over day of observation and fixing rater and generalizing respectively over rater and fixing day of observation:

$$\begin{aligned} R_{test-retest,Rel} &= \text{Corr}(Y_{PRD}, Y_{PRD'} \mid R, D, D') \\ &= \frac{\sigma_P^2 + \sigma_{PR}^2}{\sigma_P^2 + \sigma_{PR}^2 + \sigma_{PD}^2 + \sigma_{PRD}^2} \end{aligned} \tag{4.9}$$

$$\begin{aligned} R_{inter-rater,Rel} &= corr(Y_{PRD}, Y_{PR'D} \mid R, R', D) \\ &= \frac{\sigma_P^2 + \sigma_{PD}^2}{\sigma_P^2 + \sigma_{PR}^2 + \sigma_{PD}^2 + \sigma_{PRD}^2}. \end{aligned} \tag{4.10}$$

Decisions based on the level of observed score, without regards to the performance of others, are called *absolute decisions*. All variance components associated with this score, except the component for the object of measurement are defined as error. Then (4.8) becomes

$$\begin{aligned} E_{\rho^2 Abs} &= \text{Corr}(Y_{PRD}, Y_{PR'D'}) \\ &= \frac{\sigma_P^2}{\sigma_P^2 + \sigma_{Abs.Error}^2} \\ &= \frac{\sigma_P^2}{\sigma_P^2 + \sigma_R^2 + \sigma_D^2 + \sigma_{PR}^2 + \sigma_{PD}^2 + \sigma_{RD}^2 + \sigma_{PRD}^2} \end{aligned} \tag{4.11}$$

It is easy to show that equation (4.11) is indeed an ICC, this time conditioned neither on rater nor on day. Similar to the above, we can derive an *absolute* test-retest or

interrater reliability coefficient:

$$\begin{aligned} R_{test-retest,Abs} &= \text{Corr}(Y_{PRD}, Y_{PRD'}) \\ &= \frac{\sigma_P^2 + \sigma_R^2 + \sigma_{PR}^2}{\sigma_P^2 + \sigma_R^2 + \sigma_D^2 + \sigma_{PR}^2 + \sigma_{PD}^2 + \sigma_{RD}^2 + \sigma_{PRD}^2} \end{aligned} \tag{4.12}$$

$$\begin{aligned} R_{inter-rater,Abs} &= \text{Corr}(Y_{PRD}, Y_{PR'D}) \\ &= \frac{\sigma_P^2 + \sigma_D^2 + \sigma_{PD}^2}{\sigma_P^2 + \sigma_R^2 + \sigma_D^2 + \sigma_{PR}^2 + \sigma_{PD}^2 + \sigma_{RD}^2 + \sigma_{PRD}^2}. \end{aligned} \tag{4.13}$$

The purpose of this example is to provide insight into the nature of generalizability theory. First, significant sources of observational error are determined, and then these are incorporated into an experiment and components of variance are derived. Different coefficients can then be calculated depending on which facets will remain fixed and which ones will vary.

The example we used was based on a simple *crossed* design, in which there were two factors, each factor occurring at all levels of the other factors. This method can be used with more complex designs as well as including more factors, and even in *nested* designs, where the factor structure is more complex. As discussed in the book of Streiner and Norman (1995), the general approach remains the same, i.e., to begin by isolating the various sources of variance in the scores, and then generating a family of coefficients that depend on the particular factors that are allowed to vary and remain fixed.

A study (e.g., Table 4.3) that is designed to estimate variance components underlying a measurement process is called a *G-study*. Having generated the variance estimates, we can then determine the effect of changing the number of observations for instance, or what will happen to the generalizability coefficient if we add a third rater, or decrease the number of days of observations. Since these explore the impact of certain decisions, they are called *Decisions* or *D* studies. These "studies" are done using only paper and pencil (or a computer). In planning a D study, the decision maker defines the universe of generalization and specifies the proposed interpretation of the measurement. The goal is to identify important sources of variability in a particular measurement situation from the outset, and then one attempts to quantify these sources or error.

These developments made are already quite general. In the Chapter 7, we will show

how this can be embedded in the flexible linear mixed model framework introduced in Section 3.2, which, in turn, will allow for further extensions, such as, for example, incorporating serial correlation.

## 4.3 Validity

The *validity* of a questionnaire is defined as the degree to which the questionnaire measures what it purports to measure. This can be performed through the analysis of *content*, *construct*, and *criterion validity* (Carmines and Zeller, 1979). Content validity can be defined as the extent to which the instrument assesses all the relevant or important content or domains. Also the term *face validity* is used to indicate whether the instrument appears to be assessing the desired qualities at face. This form of validity consists of a judgment by experts in the field. Construct validity refers to a wide range of approaches which are used when what we are trying to measure is a "hypothetical construct" (e.g., anxiety, irritable bowel syndrome,...) rather than something that can be readily observed. The most commonly used methods to explore construct validity are: extreme groups (apply instrument for example to cases and non-cases), convergent and discriminant validity testing (correlate with other measures of this construct and not correlate with dissimilar or unrelated constructs) and the multitrait-multimethod matrix. Criterion validity can be divided into two types: *concurrent validity* and *predictive validity*. With concurrent validity we correlate the measurement with a criterion measure (gold standard), both of which are given at the same time. In predictive validity, the criterion will not be available until some time in the future at which time the true endpoint is actually observed. This also clearly links validity testing to surrogate marker validation as shown in Alonso *et al.* (2002). Of course, while measures of correlation are an important aspect of surrogacy evaluation, there is more to it than this (Baker and Kramer 2003, Burzykowski, Molenberghs, and Buyse 2005). The most commonly used method to assess the validity is by calculation of the Pearson correlation coefficient.

# Chapter 5

# Concepts in Surrogate Marker Evaluation

In this last introductory Chapter we present recently developed criteria initially meant to investigate the validity of using one endpoint as a "surrogate" for another (surrogate endpoints can be referred to as endpoints that are used instead of other endpoints in the evaluation of experimental treatments or interventions). This will provide a basis for Chapter 11, where we will apply the concepts in surrogate marker evaluation to study Criterion Validity.

Many attempts have been made in the literature to establish the validation of surrogate endpoints (Prentice, 1989; Freedman, Graubard, and Schatzkin, 1992; Buyse and Molenberghs, 1998). However, Molenberghs *et al.* (2002) point to the difficulties accompanying all these approaches and note that a sensible validation strategy can only be expressed in full in a multi-trial setting.

Therefore, Buyse *et al.* (2000) adopted an alternative approach based on a meta-analysis of several trials which led to a definition of validity in terms of the quality of both trial level and individual level association between the surrogate and the true endpoint. These authors concentrated on continuous responses. We will summarize their methodology below since we believe that it may be useful for the validation of psychiatric symptom scales.

In cases where a gold standard scale can be assigned, we can almost directly

apply their methodology for the validation of surrogate markers with the standard scale playing the role of true endpoint. In many psychiatric situations however, a more "symmetric" situation is encountered where different scales are in conjunction without knowing their relationships. In that case we need to "symmetrize" the validation technique.

Let us therefore present their hierarchical approach for two normally distributed scales $S_1$ and $S_2$ and a binary indicator variable for treatment ($Z$=0 or 1). At the first stage they consider

$$S_{1ij}|Z_{ij} = \mu_{S_{1i}} + \beta_i Z_{ij} + \varepsilon_{S_{1ij}}, \tag{5.1}$$

$$S_{2ij}|Z_{ij} = \mu_{S_{2i}} + \alpha_i Z_{ij} + \varepsilon_{S_{2ij}}, \tag{5.2}$$

where $\alpha_i$ and $\beta_i$ are trial-specific effects of treatment $Z$ on the endpoints in a trial, $\mu_{S_{1i}}$ and $\mu_{S_{2i}}$ are trial-specific intercepts, and $\varepsilon_{S_{1i}}$ and $\varepsilon_{S_{2i}}$ are correlated error terms, assumed to be mean-zero normally distributed with covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_{S_1S_1} & \sigma_{S_1S_2} \\ \sigma_{S_1S_2} & \sigma_{S_2S_2} \end{pmatrix}.$$

Due to the replication at the trial level, they can impose a further model on the trial-specific parameters. At the second stage, they then assume

$$\begin{pmatrix} \mu_{S_{1i}} \\ \mu_{S_{2i}} \\ \beta_i \\ \alpha_i \end{pmatrix} = \begin{pmatrix} \mu_{S_1} \\ \mu_{S_2} \\ \beta \\ \alpha \end{pmatrix} + \begin{pmatrix} m_{S_{1i}} \\ m_{S_{2i}} \\ b_i \\ a_i \end{pmatrix} \tag{5.3}$$

where the second term on the right hand side of (5.3) is assumed to follow a zero-mean normal distribution with dispersion matrix

$$D = \begin{pmatrix} d_{S_1S_1} & d_{S_1S_2} & d_{S_1b} & d_{S_1a} \\ d_{S_2S_1} & d_{S_2S_2} & d_{S_2b} & d_{S_2a} \\ d_{bS_1} & d_{bS_2} & d_{bb} & d_{ba} \\ d_{aS_1} & d_{aS_2} & d_{ab} & d_{aa} \end{pmatrix}.$$

Hence, a linear mixed model results. When the effects in (5.3) are assumed to be fixed, then a so-called fixed-effects model follows. The setting described above naturally lends itself for the validation of two scales at both the trial level as well as the individual level.

## 5.1 Trial-level Surrogacy

To investigate the trial-level concurrent and/or predictive validity of two psychiatric scales, it is of interest to investigate how a change in treatment effect on one measurement scale can be translated into the other psychiatric measurement instrument. Therefore, it is essential to explore the quality of the prediction of the treatment effect on $S_1$ in trial $i$ by (a) information obtained in the validation process based on trials $i = 1, \ldots, N$, and (b) the estimate of the effect of $Z$ on $S_2$ in a new trial $i = 0$. Whenever there is no clear standard but simply relationship are studied, as is often the case with psychometric instruments, the reverse prediction (on $S_2$ based on the effect on $S_1$) is also important.
To this end, observe that $(\beta + b_0 | m_{S10}, a_0)$ follows a normal distribution with mean and variance

$$\begin{aligned} & E(\beta + b_0 | m_{S_2 0}, a_0) \\ &= \beta + \begin{pmatrix} d_{S_2 b} \\ d_{ab} \end{pmatrix}^T \begin{pmatrix} d_{S_2 S_2} & d_{S_2 a} \\ d_{S_2 a} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} \mu_{S_2 0} - \mu_{S_2} \\ \alpha_0 - \alpha \end{pmatrix}, \end{aligned} \tag{5.4}$$

$$\begin{aligned} & \mathrm{Var}(\beta + b_0 | m_{S_2 0}, a_0) \\ &= d_{bb} - \begin{pmatrix} d_{S_2 b} \\ d_{ab} \end{pmatrix}^T \begin{pmatrix} d_{S_2 S_2} & d_{S_2 a} \\ d_{S_2 a} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} d_{S_2 b} \\ d_{ab} \end{pmatrix}. \end{aligned} \tag{5.5}$$

Similarly, $(\alpha + a_0 | m_{S0}, \alpha_0)$ follows a normal distribution with mean and variance:

$$\begin{aligned} & E(\alpha + a_0 | m_{S_1 0}, b_0) \\ &= \alpha + \begin{pmatrix} d_{S_1 a} \\ d_{ab} \end{pmatrix}^T \begin{pmatrix} d_{S_1 S_1} & d_{S_1 b} \\ d_{S_1 b} & d_{bb} \end{pmatrix}^{-1} \begin{pmatrix} \mu_{S_1 0} - \mu_{S_1} \\ \beta_0 - \beta \end{pmatrix}, \end{aligned} \tag{5.6}$$

$$\begin{aligned} & \mathrm{Var}(\alpha + a_0 | m_{S_1 0}, b_0) \\ &= d_{aa} - \begin{pmatrix} d_{S_1 a} \\ d_{ab} \end{pmatrix}^T \begin{pmatrix} d_{S_1 S_1} & d_{S_1 b} \\ d_{S_1 b} & d_{bb} \end{pmatrix}^{-1} \begin{pmatrix} d_{S_1 a} \\ d_{ab} \end{pmatrix}. \end{aligned} \tag{5.7}$$

To assess the validity of $S_2$ with respect to $S_1$ we propose to follow the suggestion

of Buyse *et al.* (2000) and look at the coefficient of determination:

$$R^2_{trial(f)S_2S_1} = R^2_{b_i|m_{S_2i},a_i} = \frac{1}{d_{bb}} \begin{pmatrix} d_{S_2b} \\ d_{ab} \end{pmatrix}^T \begin{pmatrix} d_{S_2S_2} & d_{S_2a} \\ d_{S_2a} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} d_{S_2b} \\ d_{ab} \end{pmatrix}. \quad (5.8)$$

Again, when none of the two scales can be assumed standard, we may also have to look at the second coefficient of determination:

$$\begin{aligned} R^2_{trial(f)S_1S_2} &= R^2_{a_i|m_{S_1i},b_i} \\ &= \frac{1}{d_{aa}} \begin{pmatrix} d_{S_1a} \\ d_{ab} \end{pmatrix}^T \begin{pmatrix} d_{S_1S_1} & d_{S_1b} \\ d_{S_1b} & d_{bb} \end{pmatrix}^{-1} \begin{pmatrix} d_{S_1a} \\ d_{ab} \end{pmatrix}. \end{aligned} \quad (5.9)$$

These coefficients are unitless and range in the unit interval, two desirable features for interpretation. Whenever these quantities are sufficiently close to 1, we can say that scales are strongly correlated at trial level.

An attractive special case of (5.8) applies when the prediction of the treatment effect can be done independently of the trial-specific random intercept $m_{S0}$. In that case formulas (5.4)–(5.7) respectively reduce to:

$$E(\beta + b_0|a_0) = \beta + \frac{d_{ab}}{d_{aa}}(\alpha_0 - \alpha), \quad (5.10)$$

$$\mathrm{Var}(\beta + b_0|a_0) = d_{bb} - \frac{d^2_{ab}}{d_{aa}}, \quad (5.11)$$

$$E(\alpha + a_0|b_0) = \alpha + \frac{d_{ab}}{d_{bb}}(\beta_0 - \beta), \quad (5.12)$$

$$\mathrm{Var}(\alpha + a_0|b_0) = d_{aa} - \frac{d^2_{ab}}{d_{bb}}, \quad (5.13)$$

leading to a simplified coefficient of determination

$$R^2_{trial(r)} = R^2_{b_i|a_i} = \frac{d^2_{ab}}{d_{aa}d_{bb}}, \quad (5.14)$$

which is now symmetric on both scales. Clearly this is a very attractive property when validating two psychometric scales for which in many cases no gold standard can be assigned. A simple quantity suffices to assess the validity.

## 5.2 Individual-level Surrogacy

To validate two scales at the individual level, we follow the suggestion by Buyse *et al.* (2000) and consider the squared correlation between the two instruments after adjustment for both the trial effects as well as the treatment effect:

$$R^2_{\text{indiv}} = R^2_{\varepsilon_{S_{1i}}|\varepsilon_{S_{2i}}} = \frac{\sigma^2_{S_1S_2}}{\sigma_{S_1S_1}\sigma_{S_2S_2}}.$$

## 5.3 Validation Criteria in Case of Mixed Continuous-ordinal Endpoints

The paper of Alonso *et al.* (2002) extended the methodology described above and used in validation of surrogate markers to validate psychiatric scales. More specifically, they assumed that $S_1$ and $S_2$ are random variables that represent two scales for which we want to assess the criterion validity.

Traditional approaches investigate the concurrent validity by correlating one measurement scale ($S_2$) with the other, assumed to be a gold standard ($S_1$).

In the previous section we described a method that is suitable for normally distributed endpoints. However, psychiatric scales are often ordinal in nature. Although it may seem reasonable for the data described in Section 2 to assume a normal distribution for the PANSS scale, this might be unrealistic for the 7-response CGI scale. Therefore, we propose an extension of the methodology by Buyse *et al.* (2000) to the situation where one of the endpoints is continuous (say $S_1$) and the other ordinal ($S_2 \in \{1, 2, \ldots, k\}$).

To this end, we assume there exists, for each ordinal variable, an underlying latent continuous variable $\tilde{S}_2$ such that

$$S_2 = r \quad \Leftrightarrow \quad \theta_{r-1} < \widetilde{S}_2 \leq \theta_r,$$

where $r = \{1, 2 \ldots, k\}$, $\theta_0 = -\infty$ and $\theta_k = +\infty$.

Next, we propose the following joint model at the first stage:

$$\begin{aligned} S_{1ij}|Z_{ij} &= \mu_{S_{1i}} + \beta_i Z_{ij} + \varepsilon_{S_{1ij}}, \\ \widetilde{S}_{2ij}|Z_{ij} &= \mu_{\widetilde{S}_{2i}} + \alpha_i Z_{ij} + \varepsilon_{\widetilde{S}_{2ij}}, \end{aligned}$$

where $\widetilde{S}_{2ij}$ is a latent unobservable variable, $\mu_{S_{1i}}$ and $\mu_{\widetilde{S}_{2i}}$ are trial specific intercepts and $\beta_i$, $\alpha_i$ are trial specific treatment effects. Further $\varepsilon_{S_{1ij}}$, $\varepsilon_{\widetilde{S}_{2ij}}$ are correlated error

terms, assumed to satisfy

$$\begin{pmatrix} \varepsilon_{S_{1ij}} \\ \varepsilon_{\widetilde{S}_{2ij}} \end{pmatrix} \sim N\left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma \right),$$

where

$$\Sigma = \begin{pmatrix} \sigma^2 & \dfrac{\rho\sigma}{\sqrt{1-\rho^2}} \\ \dfrac{\rho\sigma}{\sqrt{1-\rho^2}} & \dfrac{1}{1-\rho^2} \end{pmatrix}.$$

The variance of $\varepsilon_{\widetilde{S}_{2ij}}$ is chosen for reasons that will be made clear. From this model it is easily seen that the density of $S_{1ij}$ is univariate normal with mean $(\mu_{S_{1i}} + \beta_i Z_{ij})$ and variance $\sigma^2$, implying that the parameters $\mu_{S_{1i}}$, $\beta_i$ and $\sigma^2$ can be determined using linear regression software with response $S_{1i}$ and a single covariate $Z_i$.

In addition, the conditional density of $\widetilde{S}_2$ given $Z_i$ and $S_{1i}$ is

$$\widetilde{S}_{2i} | S_{1i}, Z_i \sim N\left(\mu_i; 1\right), \tag{5.15}$$

where

$$\mu_i = \left( \mu_{\widetilde{S}_{2i}} - \frac{\rho}{\sigma\sqrt{1-\rho^2}} \mu_{S_{1i}} \right) + \left( \alpha_i - \frac{\rho}{\sigma\sqrt{1-\rho^2}} \beta_i \right) Z_i + \frac{\rho}{\sigma\sqrt{1-\rho^2}} S_{i1}$$

From (5.15), it follows

$$P\left(S_{2i} \leq r\right) = \Phi\left(\lambda^i_{0r} + \lambda^i_z Z_i + \lambda^i_{S_1} S_{1i}\right).$$

Using standard software we can fit this proportional odds probit model and obtain estimates for

$$\begin{aligned} \lambda^i_{0r} &= \theta_{ir} - \mu_{\widetilde{S}_{2i}} + \frac{\rho}{\sigma\sqrt{1-\rho^2}} \mu_{S_{1i}}, \\ \lambda^i_z &= \frac{\rho}{\sigma\sqrt{1-\rho^2}} \beta_i - \alpha_i, \\ -\lambda^i_{S_1} &= \frac{\rho}{\sigma\sqrt{1-\rho^2}}, \end{aligned}$$

Note that also the logit and the complementary log-log link could be used here. Given these parameters, together with the parameters from the linear regression on $S_{1i}$, we

obtain

$$\begin{aligned} \alpha_i &= \frac{\rho}{\sigma\sqrt{1-\rho^2}}\beta_i - \lambda_z^i, \\ \rho^2 &= \frac{\bar{\lambda}_{S_1}^2\sigma^2}{\sqrt{1+\bar{\lambda}_{S_1}^2\sigma^2}}. \end{aligned}$$

Without loss of generality, we can assume that the cutpoints lie at equidistant intervals symmetric around zero so that $\sum_{r=1}^{k-1}\theta_{ir} = 0$ implying

$$\mu_{\tilde{S}_{2i}} = -\bar{\lambda}_0^i + \frac{\rho}{\sigma\sqrt{1-\rho^2}}\mu_{S_{1i}},$$

with $\bar{\lambda}_0^i = \frac{1}{k-1}\sum_{r=1}^{k-1}\lambda_{0r}^i$. For all of these parameters, bootstrap confidence intervals can be calculated.

While the method described above could equally well be applied to investigate the predictive validity (where one of the two criteria will not be available until some time in the future), this falls beyond the scope of the data analyses presented here.

# Chapter 6

# Reliability Estimation in Case of Interval Scaled Data

In this Chapter we will derive a general and closed formula for test-retest intraclass correlation coefficient of reliability for interval scales longitudinal clinical data resulting from clinical trials. The results of this chapter are published in the paper of Vangeneugden *et al.* 2004).

## 6.1 Estimation of Reliability in the Linear Mixed Models Framework

The general formula to calculate the intra-class correlation coefficient for model (3.5) can be derived via (4.3). Denote $Y_{it}$ the observed measurement of subject $i$ on time point $t$; $s$ will also be used to denote (a second) time point. Then

$$\begin{aligned} \text{Var}(Y_{is}) &= \boldsymbol{z}_s D \boldsymbol{z}_s' + \tau^2 + \sigma^2, \\ \text{Var}(Y_{it}) &= \boldsymbol{z}_t D \boldsymbol{z}_t' + \tau^2 + \sigma^2, \\ \text{Cov}(Y_{is}, Y_{it}) &= \boldsymbol{z}_s D \boldsymbol{z}_t' + \tau^2 (H_i)_{st}, \end{aligned} \tag{6.1}$$

where $z_s$ and $z_t$ denote the design matrices for the random effects at time $s$ and $t$ respectively, $D$ the covariance matrix for the random effects, $\tau^2(H_i)$ the covariance

matrix for the serial effects for subject $i$, and $\sigma^2$ the residual error variance. Therefore, reliability in this general setting with multiple time points is time or lag dependent. Denote the test-retest reliability between time point $s$ and $t$ by $R(s,t)$. From (6.1) we have

$$R(s,t) = \text{Corr}(Y_{is}, Y_{it}) = \frac{\boldsymbol{z}_s D \boldsymbol{z}_t' + \tau^2 (H_i)_{st}}{\sqrt{\boldsymbol{z}_s D \boldsymbol{z}_s' + \tau^2 + \sigma^2}\sqrt{\boldsymbol{z}_t D \boldsymbol{z}_t' + \tau^2 + \sigma^2}}. \tag{6.2}$$

In Section 6.2, in different settings, we will apply the formula (6.2) above and derive the reliability of psychiatric symptom scales from such models, thereby generalizing the classical developments as outlined previously.

## 6.2 Data Analyses

Let us now apply the previously developed methodology on the pooled Schizophrenia data described in Section 2.1. We will assess the reliability for the PANSS, using the SAS procedure MIXED. As mentioned earlier, the PANSS scale is a continuous response aggregating 30 items scored from 1 to 7. For this response we considered in turn four different models and calculated the corresponding reliability measures.

### Model 1

First, we assume a linear mixed model with a random intercept and with time, treatment and their interaction as fixed effects. Time is modeled as a factor with seven levels such that we obtain a saturated cell means model for time and treatment. In that case (3.5) becomes $Y_i = X_i\boldsymbol{\beta} + Z_i b_i + \boldsymbol{\varepsilon}_i$, with $Z_i$ a $n_i$ dimensional vector of ones, $\boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \sigma^2 I)$ and $b_i \sim N(\mathbf{0}, d)$. This can be rewritten as:

$$Y_{ijk} = \mu_{jk} + b_i + \varepsilon_{ijk},$$

where $Y_{ijk}$ is the measure at time point $j$ for subject $i$ under treatment $k$; $\mu_{jk}$ groups the fixed-effects structure, $b_i$ is still the random intercept and $\varepsilon_{ijk}$ is the measurement error. The fitted variance components are $\widehat{d} = 311.00$ and $\widehat{\sigma}^2 = 125.14$. From (6.2) we can easily derive the formula for the reliability for this simple model. Since $\tau$ equals 0 (no serial correlation) and since $\boldsymbol{z}_s$ is 1 and $D = d$, the variability of the random intercept model, we have:

$$R = \frac{d}{d + \sigma^2}. \tag{6.3}$$

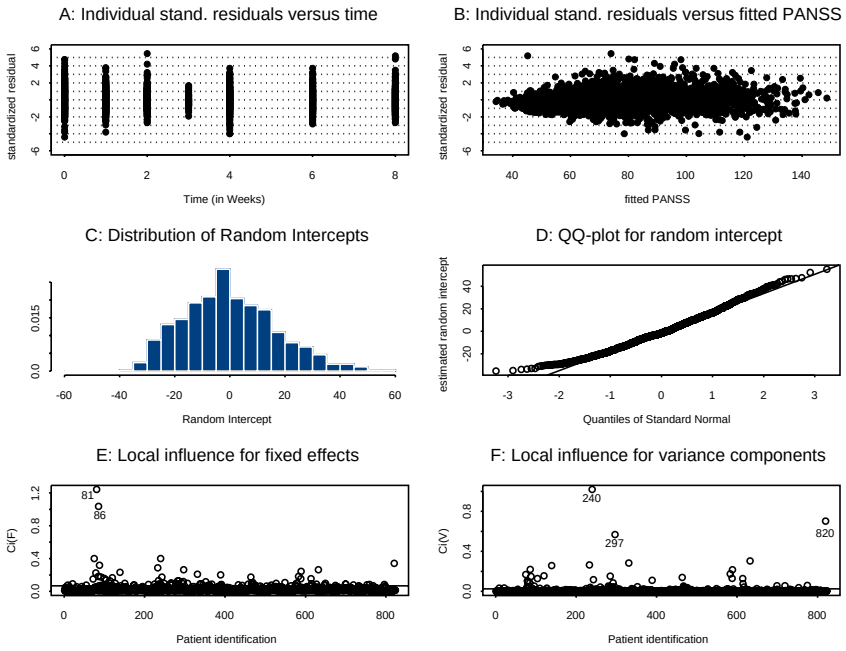


Figure 6.1: *Schizophrenia PANSS Data. Diagnostic plots for model 1.*

The reliability expresses the ratio of the variance explained by the model to the total observed variance. The link of (6.3) with the intuitive definition of reliability as we have expressed in (4.2) is obvious. For data containing two measurements per subject, this value equals the test-retest reliability of the instrument. For any series of repeated measurements, this value gives an overall measure of the intraclass correlation between all the measurements within subjects. For the PANSS data this global reliability measure yields a value of $\widehat{R} = 0.713$ (SE 0.012). The standard error is calculated using the delta method. If we first apply Fisher's variance-stabilizing transformation on $R$, $Z = 0.5[ln(1+R) - ln(1-R)]$ and the delta method, the 95% confidence interval is [0.688;0.736].

Fig. (6.1) displays the standardized subject-specific residuals (6.1A and B) for this model to assess the model fit and also investigates the distribution of the random intercept (6.1C and D) and identifies influential observations (6.1E and F).

The local influence method described above revealed five influential observations

(6.1 E and F), two on the estimation of the fixed effects (81,86) and three on the estimation of the variance components (240, 297 and 820). If we omit observations 81 and 86, this has little or no influence on estimation of variance components and the reliability coefficient remains $R = 0.71$. If we omit 240, 297 and 820, the reliability increases to $R = 0.72$, which shows that the most influential measurements have little or no impact on the estimation of the overall reliability coefficient.

Note that the assumption of parallel measurements is not met. The mean PANSS decreases from 92.4 at baseline to 68.8 at endpoint. Even though classical reliability studies usually require the assumption of parallel measurements, our approach, due to the flexibility of the linear mixed model, obviates the need for this, since the mean and variability structures can be clearly separated. In particular, the linear mixed model will account for time and treatment effects by including them into the fixed effects component of the model. Although steady state is not taken care of by design as it would be in classical test-retest designs in psychometrics, steady state is provided through modeling at the analysis stage. A conceptually useful way to think about this is through the two-stage approach as the mixed effects model has been introduced, historically, by Laird and Ware (1982). If we derive the individual residuals for the model above and subsequently apply a random intercept model on these residuals without a fixed effect component($\mu_{jk} = 1$), the same estimates for $\widehat{d}$ and $\widehat{\sigma}^2$ would be obtained. Furthermore, there are three additional advantages: the mixed model approach can be applied when (1) there are more than two measurement occasions, (2) not all subjects have the same number of measurements (due to missingness or irregularly spaced measurement times) and (3) more complicated variance-covariance structures within subjects exist. To study these advantages further, we will consider more elaborate models in subsequent sections.

## Model 2

The use of random effects in the assessment of reliability dates back to Bartko (1966) and has been described by Dunn (1989). Model 1 builds upon this work. In addition, we will introduce serial correlation and then generalize the calculation of reliability to this situation. Explicitly, the second model combines a random intercept with serial correlation. Typical choices for such serial correlation structures are based on exponentially or Gaussian decaying processes. These are standardly available in the SAS procedure MIXED (Littell and Milliken, 1996). In order to choose the covariance

structure that best fits the data, an empirical variogram was created which is shown in Fig. 6.2. For a formal introduction to the variogram in the context of longitudinal data, we refer to Diggle, Liang, and Zeger (1994) or Verbeke and Molenberghs (2000). The value of the variogram at time lag zero is an indication for the relative importance of the measurement error, the discrepancy between the variogram at the largest time lag, and the process variance (represented as a level straight line at the top of the plot) is an indication for the importance of the random intercept. The shape of the variogram describes the serial correlation process. The strength of the process is indicated by the amount of increase between zero and maximum time lags, while the shape of the curve is indicative for the shape of the process of serial decay.

The variogram is essentially flat. This implies that the largest component of variability is attributable to a random intercept, i.e., the within-unit correlation comes from a subject-specific intercept rather than from a serial correlation. However, there is a hint that a perhaps small serial component may be present; we opt for a Gaussian serial process. Then $\Sigma_i$, the covariance matrix grouping the measurement error and serial components in (3.5), is defined by the matrix with elements

$$\begin{aligned} \Sigma_{ss} &= \sigma_{ss} = \tau^2 + \sigma^2, \\ \Sigma_{st} &= \sigma_{st} = \tau^2 \exp(-u_{st}^2/\rho^2), \qquad s \neq t, \end{aligned}$$

where $\sigma^2$ denotes the measurement error variance and the remaining part is the serial variance component with $u_{st}$ the time lag between measurements $Y_{isk}$ and $Y_{itk}$ for subject $i$ and treatment $k$. The estimated covariance parameters of this model, applied to the PANSS data, are $\widehat{d} = 103.21$, $\widehat{\tau}^2 = 274.97$, $\widehat{\rho} = 6.38$, and $\widehat{\sigma}^2 = 65.21$.

The reliability can again be derived from (6.2) and is a function of time lag $u_{st}$ between two observations measured at time point $s$ and $t$

$$R(u_{st}) = \frac{d + \tau^2 \exp\left(\frac{-u_{st}^2}{\rho^2}\right)}{d + \tau^2 + \sigma^2}. \tag{6.4}$$

After correction for the fixed-time and treatment effects, the covariance parameter estimates show a considerable remaining serial component in the PANSS data. As can be seen from formula (6.4), a strong serial effect will lead to a fast decreasing reliability for increasing time lags. Fig. 6.3 shows that reliability is 0.80 or higher for measurements no further apart than 2 weeks but declines rapidly thereafter. This is consistent with the general consensus regarding the appropriate interval: generally speaking a retest interval of 2 days to 2 weeks is appropriate (Streiner and Norman

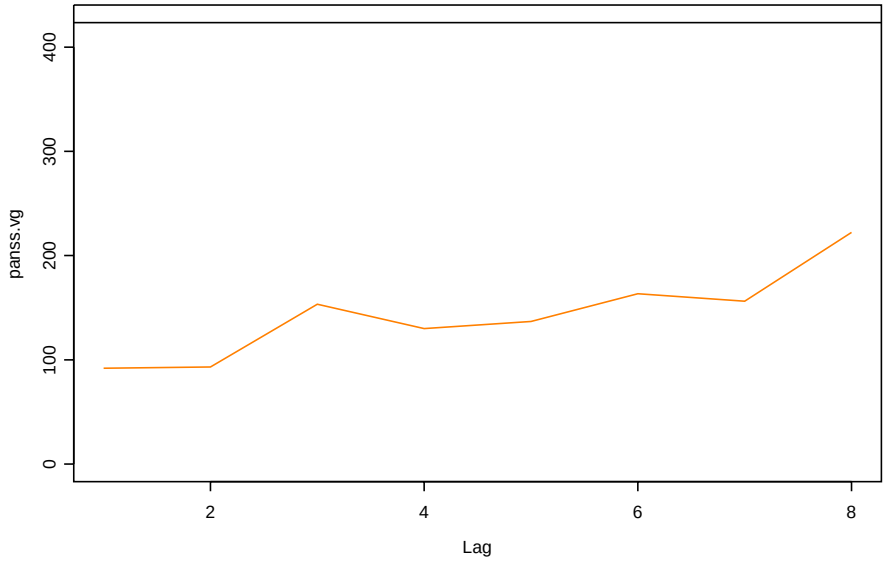


Figure 6.2: *Schizophrenia PANSS Data. Empirical variogram of the total PANSS data.*

1995): if the interval is too short, the patients may remember their previous responses, if the interval is too long, things may have changed. A big advantage of model 2 is that this type of model allows to study the effect of lag time on the reliability.

The individual, subject-specific residuals of this model as well as the distribution of the random effects are displayed in Fig. 6.4. Although the standardized residuals are not as large as for model 1, Fig. 6.4 B shows that the standardized residuals tend to increase with higher fitted PANSS values.

Influential observations determined by means of likelihood displacement instead of local influence due to the presence of serial correlation. Fig. 6.4 E determines five influential observations: 79, 80, 240, 297 and 775. Removing these influential observations has little or no impact on estimation of reliability; the reliability does not differ more than 0.014 with or without the five influential observations.

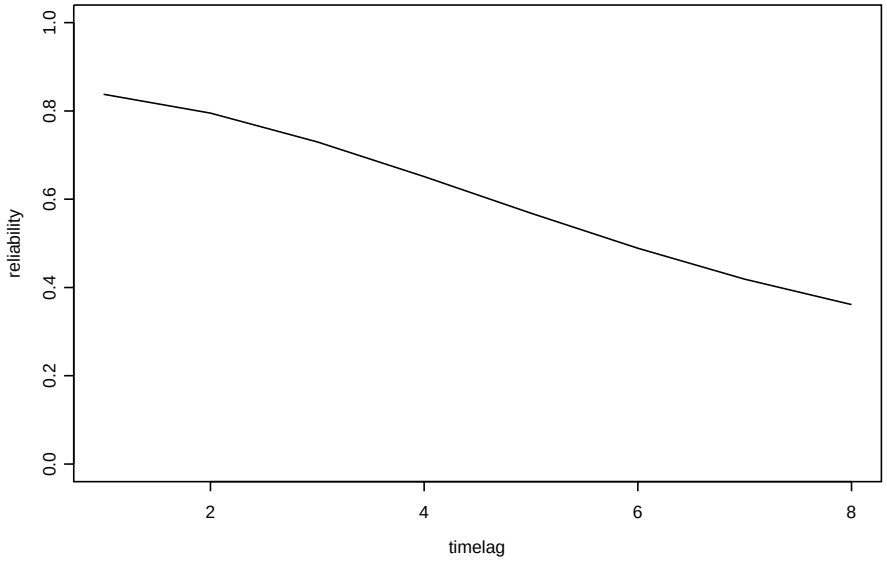


Figure 6.3: *Schizophrenia PANSS Data. Reliability of the total PANSS as a function of the time-lag u between any two measurements.*

## Model 3

After adding serial correlation in model 2 to the random-intercept model 1, we now add random slope in time as well. The random-effects variance then equals

$$D = \begin{pmatrix} d_{11} & d_{12} \\ d_{12} & d_{22} \end{pmatrix}.$$

The estimated covariance parameters for the PANSS data are $\widehat{d}_{11} = 47.24$, $\widehat{d}_{12} = 13.65$, $\widehat{d}_{22} = -0.10$, $\widehat{\tau}^2 = 247.39$, $\widehat{\rho} = 5.82$, and $\widehat{\sigma}^2 = 63.96$. The residuals shown in Fig. 6.5 display a clear trend, variance of the residuals increase for increasing PANSS values and decrease in time, indicating an non optimal fit.

The model can now be written as follows:

$$Y_{ijk} = \mu_{jk} + \begin{pmatrix} b_{i0} & b_{i1} \end{pmatrix} \begin{pmatrix} 1 \\ j \end{pmatrix} + w_{ij} + \varepsilon_{ijk}. \tag{6.5}$$

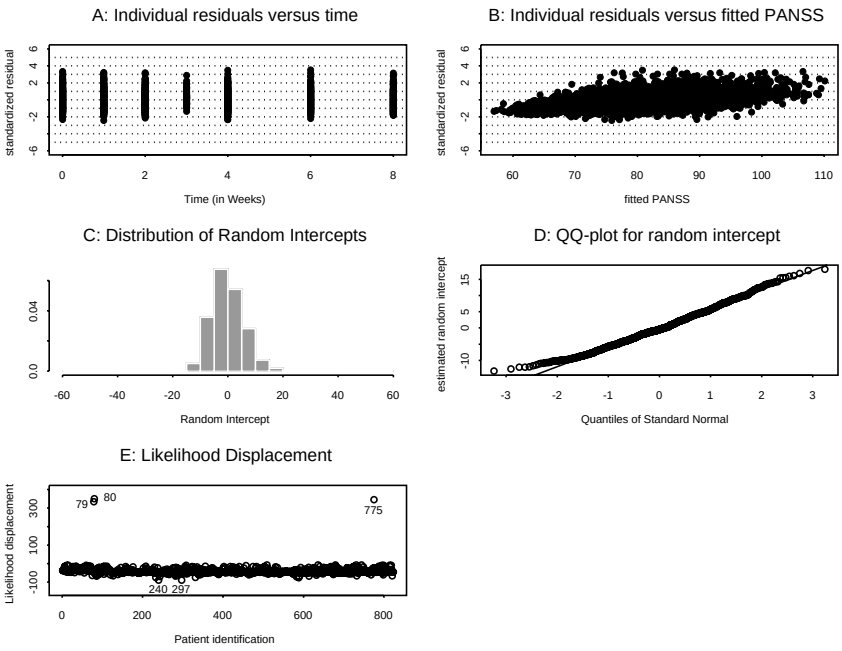


Figure 6.4: *Schizophrenia PANSS Data. Diagnostic plots for model 2.*

From (6.2) we can derive, the test-retest reliability for observations at time point $s$ and time point $t$ and lag time $u_{st}$ between them:

$$R(s,t) = \frac{\boldsymbol{z}_s D \boldsymbol{z}_t' + \tau^2 \exp(\frac{-u_{st}^2}{\rho^2})}{\sqrt{\boldsymbol{z}_s D \boldsymbol{z}_s' + \tau^2 + \sigma^2}\sqrt{\boldsymbol{z}_t D \boldsymbol{z}_t' + \tau^2 + \sigma^2}}. \tag{6.6}$$

Here, $\boldsymbol{z}_s$ is the design row in $Z$ corresponding to time $s$. Formula (6.6) can be used to calculate the different reliabilities for any specific time point and for any given time lag. Due to the questionable fit, that will not be presented here. Instead we will investigate a simpler model, by omitting the serial component.

## Model 4

Only the random intercept and the random slope are retained in (6.5). The estimated covariance parameters for the PANSS data are $\widehat{d}_{11} = 315.21$, $\widehat{d}_{12} = -8.01$, $\widehat{d}_{22} = 7.07$, $\widehat{\sigma}^2 = 79.63$. Subsequently, the reliability of measurement observed on time $s$ and

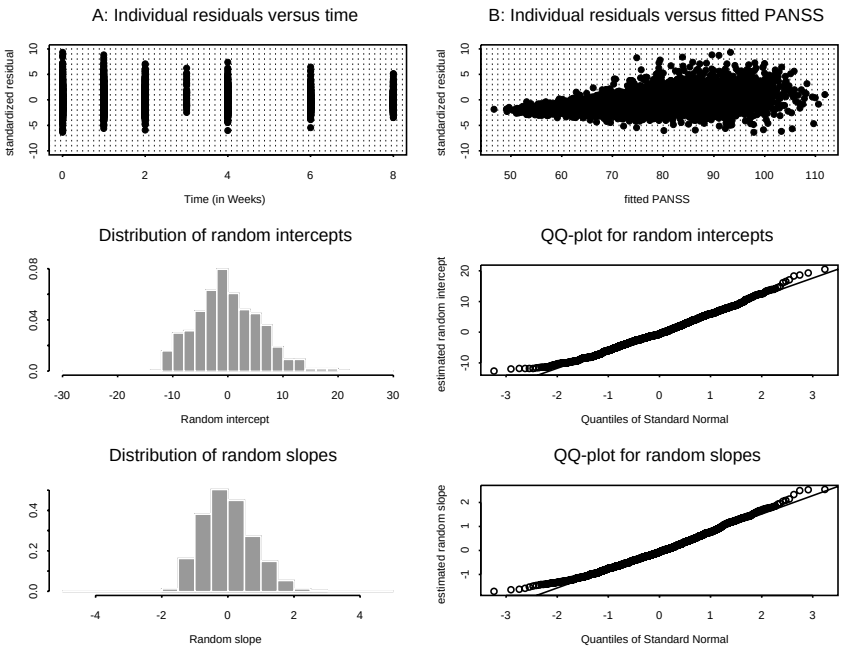


Figure 6.5: *Schizophrenia PANSS Data. Diagnostic plots for model 3.*

time $t$:

$$R(s,t) = \frac{\boldsymbol{z}_s D \boldsymbol{z}_t'}{\sqrt{\boldsymbol{z}_s D \boldsymbol{z}_s' + \sigma^2}\sqrt{\boldsymbol{z}_t D \boldsymbol{z}_t' + \sigma^2}}. \tag{6.7}$$

Table 6.1 displays the reliability coefficients estimated from model 4; only the upper diagonal is shown for this symmetric *test-retest reliability matrix*. Again we can observe that reliability is decreasing with increasing lag time. Another result that occurs is a slight increase in the reliability measure as time goes by, but for a fixed time lag.

Fig. 6.6 investigates the model diagnostics for this model and hints that the model fit has improved versus model 3. There are three influential observations for the variance components (240, 297 and 331) and three influential observations for the estimation of fixed effects (81, 86 and 88). After removing these influential measurements, the covariance parameters were estimated as $\widehat{d}_{11} = 310.22$, $\widehat{d}_{12} = -6.51$, $\widehat{d}_{22} = 6.49$, $\widehat{\sigma}^2 = 74.65$. The effect on estimation of the reliability coefficients is minimal. The largest difference is 0.03; e.g., test-retest reliability of observations on week 0 and

Table 6.1: *Schizophrenia PANSS Data. Estimated test-retest reliabilities using model 4.*

| | **Time point** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Time point** | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 0 | 0.80 | 0.79 | 0.76 | 0.72 | 0.68 | 0.62 | 0.57 | 0.52 | 0.47 |
| 1 | . | 0.79 | 0.79 | 0.76 | 0.73 | 0.69 | 0.65 | 0.61 | 0.57 |
| 2 | . | . | 0.80 | 0.79 | 0.78 | 0.75 | 0.72 | 0.69 | 0.66 |
| 3 | . | . | . | 0.81 | 0.81 | 0.80 | 0.78 | 0.75 | 0.73 |
| 4 | . | . | . | . | 0.82 | 0.82 | 0.82 | 0.80 | 0.79 |
| 5 | . | . | . | . | . | 0.84 | 0.84 | 0.84 | 0.83 |
| 6 | . | . | . | . | . | . | 0.86 | 0.86 | 0.86 |
| 7 | . | . | . | . | . | . | . | 0.87 | 0.88 |
| 8 | . | . | . | . | . | . | . | . | 0.89 |

week 8 increases from 0.47 to 0.5 after removal of the six influential observations.

### Summary of Various Models

Table 6.2 summarizes the parameter estimates and the log likelihood of the different models described above. Model 3 is the model with the largest likelihood, and would be the one of preference if one would rely purely on likelihood ratio testing. However, the diagnostic plots indicate that model 4 fits the data better, which is in line with the variogram where the random effect rather than the serial correlation dominates the within-subject correlation.

Note that our research is ancillary to the assessment of treatment effect. Indeed, by first considering an appropriate mean structure, one can concentrate on the variability structure, thus enabling the use of clinical trial data to study reliability.

## 6.3 Concluding Remarks

A body of research exists on reliability, especially in psychology and educational sciences. In the past decades the topic is also entered the field of health sciences and especially the psychiatric health sciences because of the inherent subjectivity

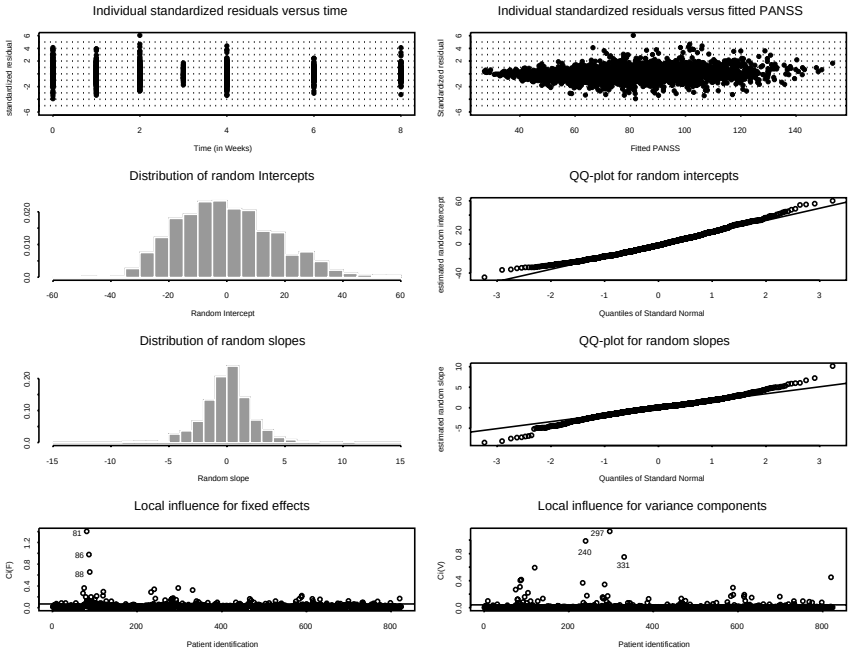


Figure 6.6: *Schizophrenia PANSS Data. Diagnostic plots for model 4.*

of the measures employed in this field. Test-retest reliability as one of the classical approaches typically deals with the problem of time: how to disentangle the measurement error from real fluctuations in what you are measuring ?

Wiley and Wiley (1970) were among the first authors to deal with this problem by assuming a linear relationship between two adjacent measurements. In this way also reliability will have different values at both moments of measurement. Tisak and Tisak (1986) also stressed the fact that reliability is not a fixed property of an instrument but changes with time. They proposed a method to calculate a time function of reliability. Dunn (1989) describes a method that uses components of variance in the calculation of reliability. He further extends this method to a mixed model to deal with rater effects by taking the rater into the model as a random effect. The mixed model methodology indeed allows a study of variance components and fixed effects simultaneously. The variance-covariance structure is typically decomposed further into three components: (1) measurement error (process with memory 0), (2) serial

Table 6.2: *Schizophrenia PANSS Data. Estimated variance component for models 1–4.*

| | | Estimates for various models | | | |
|---|---|---|---|---|---|
| Component | Par. | 1 | 2 | 3 | 4 |
| Var. rand. int. | $d_{11}$ | 311.00 | 103.21 | 47.24 | 315.21 |
| Cov. (rand. int., rand. slope) | $d_{12}$ | | | 13.65 | -8.01 |
| Var. rand. slope | $d_{22}$ | | | -0.10 | 7.07 |
| Serial process variance | $\tau^2$ | | 274.97 | 247.39 | |
| Serial process corr. par. | $\rho$ | | 6.38 | 5.82 | |
| Measurement error var. | $\sigma^2$ | 125.14 | 65.21 | 63.96 | 79.63 |
| $-2$ log likelihood | | 33870.7 | 33232.4 | 33192.2 | 33331.4 |

correlation (process with finite memory), and (3) random effects (accommodating hierarchies, infinite memory process). Such hierarchies arise due to repeated measurements over time. Other hierarchies could be accommodated as well. Indeed, even in our current work, hierarchy arises due to the fact that data come from five trials. A proper account of this calls for the incorporation of (meta-analytic and other) hierarchies into our modeling strategies. Some work exists to this effect and is known as generalizibility theory (Cronbach, 1963). The combination of this work with ours is the subject of next chapter.

While, for this reliability study, we are primarily interested in the variance components, mixed-model methodology provides an interesting opportunity to model the fixed effects as well. We do not have to make the unrealistic assumption that there is no change in a patients situation over time or with treatment. Instead, such changes can be incorporated into the model.

When using repeated measurements a third source of variation can be taken into account when calculating reliability, the so-called serial correlation. In this work, a method has been proposed that allows for serial correlation in the calculation of test-retest reliability, as well as random effects and measurement error.

The method was applied to the PANSS, a psychiatric rating scale for schizophrenia. Several models were applied: model 1 resulted in an overall test-retest reliability coefficient, averaging reliability across the 8 weeks, models 2-4 allowed us to study

the test-retest reliability as a function of time. We observed a gradual decrease of reliability with increasing time lag between measurements. As mentioned earlier, there are different possible scenarios to explain such effects, such as memory effect of the raters or other covariates that are not taken into account in the model. For the PANSS scale we obtained reliability estimates from almost 0.90 to 0.50. Up to a time interval of 5 weeks, the reliability does not go below 0.60, which is considerable. Another result that occurs quite consistently is a slight increase in the reliability measure as time goes by, but for a fixed time lag. The reason for this is most likely a learning effect in the raters. In a different setting, one might also encounter learning effects in the study subjects. Of course, other perhaps complementary explanations cannot be excluded.

The present method stresses once again the fact that reliability should not be perceived as a fixed quantity, but changes with circumstances and populations. Other covariates can be incorporated into the model to study their effect on error variance and on reliability. Modeling other sources of variation, like for example country or rater, is therefore an interesting topic. In psychometric theory, this is referred to as *generalizability theory* (Cronbach, 1963) as introduced in Section 4.2. In Chapter 7 we will explore generalizability using the same data set as in this Chapter.

A further important advantage of the present method is that it becomes possible to estimate trial-specific or population-specific reliability in clinical studies. This is especially true because, even in studies designed to assess reliability, it is difficult to exclude fluctuations in the true scores and furthermore these studies are often conducted with different populations and in different circumstances. Finally, when measurement sequences on a subset of respondents are incomplete, these data can still be used for analysis, unlike in the classical approaches. In our case, we have focused on population-level reliability. Should we calculate them trial-specific via model 1, we would obtain the values 0.72, 0.69, 0.72, 0.71, and 0.59.

While it seems variability in reliability over time could be ascribed to variability in study duration, we are protected against such spurious effects by the use of a likelihood framework, where shorter and longer sequences contribute to estimates at any time point, as in the missing data literature (Little and Rubin, 1987). Further, the strength of our methodology is that a proper time variable can be included into the mean and variance model, allowing us to combine studies of variable length.

Of course, some of the fluctuation observed in reliability estimates may be due purely to random noise, due to limited sample sizes. A clear perspective on this

can be obtained by calculating interval estimates, which can also be used to assess appropriate sample sizes.

When clinical trials are designed, only validated scales should be used. Therefore, validation should always happen before clinical trials are started. This should not prevent the statistician however from studying how well the scale actually performed during the trial: was the test-retest reliability indeed as predicted? Most often, the only focus is to estimating treatment effect, taking into account the observed variance, without investigating the latter. This is also a missed opportunity to increase knowledge about the scale: often the number of subjects and observations in studies designed to validate scales are rather low while the information coming from clinical trials can be very rich.

In this Chapter we focused on interval scaled Gaussian distributed data. In Chapter 8 we will extend methodology to any type of data.

# Chapter 7

# Generalizibility Estimation in Case of Interval Scaled Biomedical Data

Let us now apply the concepts of generalizability methodology as introduced in Section 4.2 on the pooled data described in Section 2.1. To demonstrate the concept, we will investigate impact of "country" on measurement error and reliability. First we will repeat the overall reliability analysis for the PANSS, ignoring country effects, using the framework derived in Section 6.1 and Equation (6.2). Subsequently we extract country effects by including country as a fixed effect in the model. Next we will investigate impact on reliability by country by applying the same model on each country separately, and study impact of a single country on overall reliability by omitting the country from the data. Finally we will assess overall impact of country via generalizability theory. The results of this chapter are published in the paper of Vangeneugden *et al* (2005).

## 7.1 Overall Reliability of PANSS Scale

We first applied a linear mixed model with a random intercept to analyze the total PANSS. Specializing notation in formula (3.5) to our particular setting:

$$Y_{PDT} = \mu + \mu_D + \mu_T + \mu_{DT} + b_P + \varepsilon_{PDTC}, \tag{7.1}$$

where $\mu_D$, $\mu_T$ and $\mu_{DT}$ denote the fixed effects for day, treatment and their interaction respectively, and $b_P$ denotes the random patient effect. The fitted variance components are $\widehat{d} = 311.00$ for the random intercept, and $\widehat{\sigma}^2 = 125.14$ for measurement error. Residuals show that the model fits the data reasonably well as shown in Figure 6.1. The ICC of reliability can by derived via equation (6.2):

$$R = \text{Corr}(Y_{PDT}, Y_{PD'T} \mid T, D, D') = \frac{d}{d + \sigma^2}. \tag{7.2}$$

For data containing two measurements per subject, this value equals the test-retest reliability of the instrument. For any series of repeated measurements, this value gives a global measure of the correlation between the measurements within subjects. For the PANSS data this global reliability measure yields a value of $\widehat{R} = 0.713$ (SE 0.012). The standard error is calculated using the delta method. If we first apply Fisher's variance-stabilizing transformation on $R$, $Z = 0.5[ln(1 + R) - ln(1 - R)]$ we obtain the following 95% CI $[0.688; 0.736]$.

## 7.2 Overall Reliability After Extracting Country Effects

Similar to the treatment and time effects as well as their interaction, the country effect can be extracted and a *corrected* overall reliability over time can be calculated. This can be done by including a fixed effect for country, and all second and third order interactions with country in (7.1). Then the fitted variance components are: $\widehat{d}^2 = 290.07$ and $\widehat{\sigma}^2 = 122.99$. The overall reliability can then be calculated from (7.2): $\widehat{R} = 0.702$. After applying Fisher's variance-stabilizing transformation on $R$, we obtain the following 95% CI $[0.676; 0.727]$.

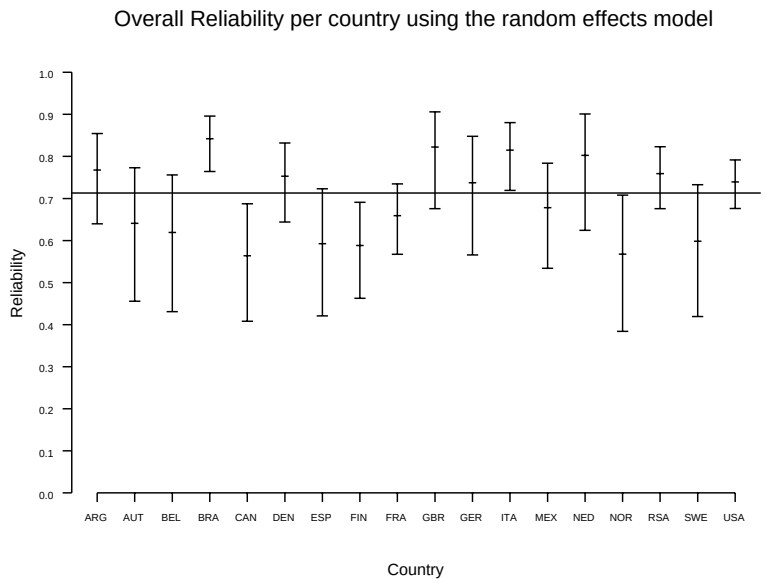


Figure 7.1: *Schizophrenia PANSS Data. Graphical representation of reliability of the total PANSS per country.*

## 7.3 Overall Reliability by Country

To study the impact of country, we apply the same model (7.1) to every subgroup containing the data for a specific country only. Furthermore, the weighted average of the reliability over all countries was calculated using the inverse of the variance of the reliability estimate as weight. Table 7.1 summarizes results and Figure 7.1 provides a graphical summary. The horizontal line represents the overall reliability ($R = 0.713$), the vertical lines represent 95% CI of the country-specific reliability (after Fisher's transformation).

From Table 7.1 and Figure 7.1 we observe that reliability is not very different between the countries: the highest reliability is observed in Brazil ($R = 0.842$) and the lowest in Canada ($R = 0.564$). Another way of looking at what these reliability coefficients represent is to look at the subject-specific residuals; the larger the measurement error, the lower the reliability. Figure 7.2 displays the residual profiles for full model (7.1), for the patients from Brazil and Canada separately. Residual

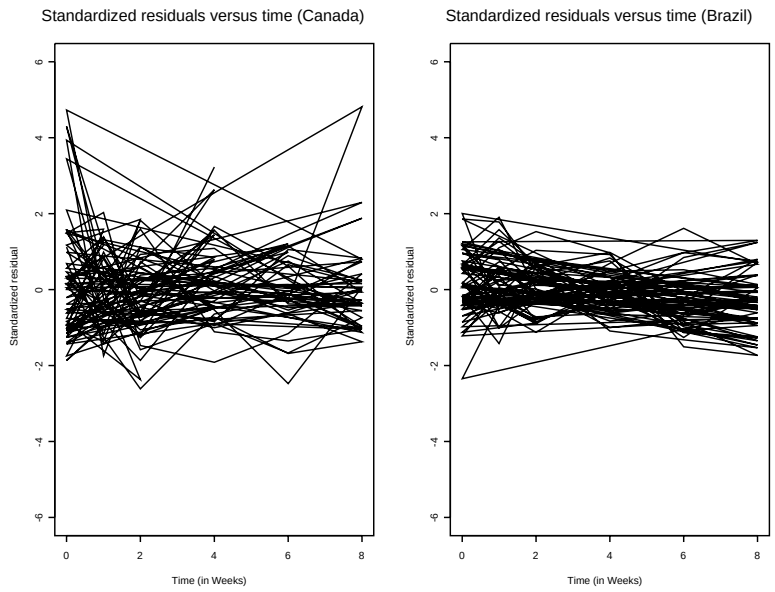


Figure 7.2: *Schizophrenia PANSS Data. Residuals profiles for Canada and Brazil.*

profiles from Brazil are more concentrated around 0 than the profiles from Canada. Additionally, we calculated the weighted average reliability across countries, where we used the inverse variance of the country specific reliability coefficient as weights (using the delta method). The result was an overall reliability of 0.735, with 95% CI $[0.635, 0.834]$ which is slightly higher than the overall reliability calculated in previous sections. If we apply Fisher's transformation first, the weighted average resulted in $R = 0.702$ with 95% CI $[0.577, 0.794]$.

## 7.4 Impact on Overall Reliability by Leaving Out a Country

Similar to what is often done in the calculation of Cronbach's alpha coefficient (which is also a reliability coefficient) to study internal consistency of a rating scale (Cronbach, 1951), we can study impact on overall reliability by leaving out the data of a specific country, per country. If the overall reliability increases, this would indicate

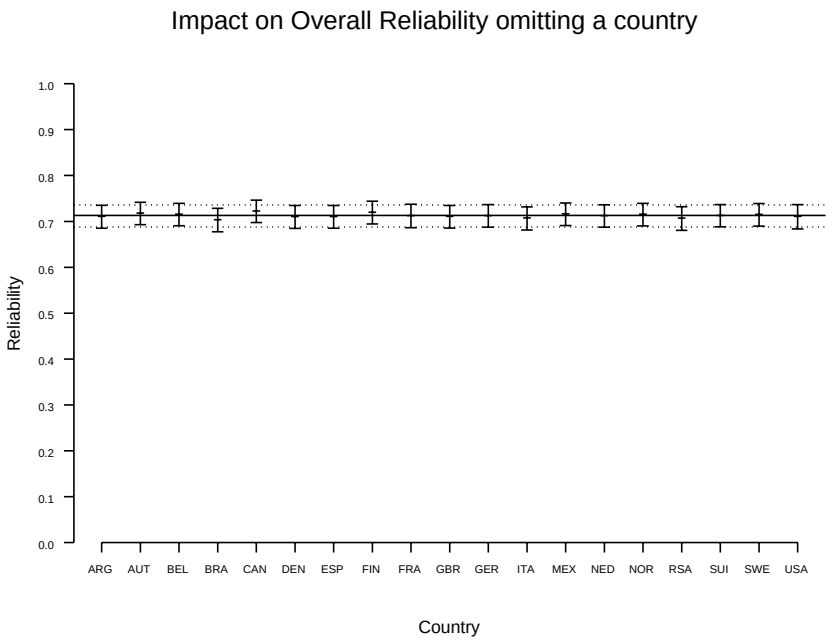


Figure 7.3: *Schizophrenia PANSS Data. Graphical depiction of overall reliability of the total PANSS omitting a specific country.*

poor reliability in the specific country. Table 7.1 and Figure 7.3 summarize the results. The horizontal line represents the overall reliability ($R = 0.713$) and the dashed lines represent upper and lower 95% CI of the overall reliability. We conclude that the impact of country on the overall reliability coefficient is ignorable.

## 7.5 Estimating Impact of Country: Generalizability Theory

Subgroup analysis by country as shown in the previous two sections can be elucidatory. Now, we want to quantify their effect on measurement error and calculate a generalizability coefficient, generalizing results across countries. If we use the data from the clinical trials as surrogate for a G-study, we could model time as either a fixed effect, and/or a random effect, or include serial correlation for time into the

model. We will start with the simpler model that considers time in days ($D$), treatment ($T$) and their interaction ($DT$) as fixed effects, country ($C$) and patient ($P$) as random effect, where patient is nested in country ($P(C)$):

$$Y_{PDTC} = \mu + \mu_D + \mu_T + \mu_{DT} + b_{P(C)} + b_C + \varepsilon_{PDTC}. \tag{7.3}$$

From this model we can calculate the overall test-retest reliability coefficient similar to (4.9):

$$\begin{aligned} R = \mathrm{Corr}(Y_{PDTC}, Y_{PD'TC} \mid T, D, D') &= \frac{\sigma_P^2 + \sigma_C^2}{\sigma_P^2 + \sigma_C^2 + \sigma_E^2} \\ &= \frac{293.8 + 16.0}{293.8 + 16.0 + 125.1} \\ &= 0.712 \end{aligned} \tag{7.4}$$

This test-retest reliability coefficient for any given country and time point follows directly from analyzing the clinical trial, similar as generalizability coefficients that are computed after design and analysis of a G-study. In the same spirit of D-studies, we can also generalize across countries: Although patients are nested within country in the trial setting, we assume (as a mind experiment) that patients can switch from one country to another:

$$\begin{aligned} R &= \mathrm{Corr}(Y_{PDTC}, Y_{PD'TC'} \mid C, C', T, D, D') \\ &= \frac{\sigma_P^2}{\sigma_P^2 + \sigma_C^2 + \sigma_E^2} = \frac{293.8}{293.8 + 16.0 + 125.1} = 0.676. \end{aligned} \tag{7.5}$$

Thus, generalizing across time points and countries, or taking account of impact of variance of country reduces the overall test-retest reliability from 0.713 to 0.676 for any given treatment. In this example, the price for setting up an international trial instead of a single country is rather small. The methodology can be easily extended to more complex situations including, for example, serial correlation or random time effects but also additional variables (e.g. sex).

If we allow for serial correlation, then we need to add $\omega_{PD}$ in (7.3), where $\omega_{PD}$ represent the serial effect for patient $P$ on day $D$. To investigate the presence of serial correlation, the variogram was created (not shown). The variogram is essentially flat. This implies that the largest component of variability is attributable to a random intercept, i.e., the within-unit correlation comes from a subject-specific intercept rather than from a serial correlation. However, there is a hint that a perhaps small

serial component may be present; we opt for a Gaussian serial process. Then $\Sigma_i$, the covariance matrix grouping the measurement error and serial components in (3.5), is defined by the matrix with elements:

$$\begin{aligned} \Sigma_{DD} &= \sigma_{DD} = \tau^2 + \sigma_E^2, \\ \Sigma_{DD'} &= \sigma_{DD'} = \tau^2 \exp(-u_{DD'}^2/\rho^2), \qquad D \neq D', \end{aligned}$$

where $\sigma_E^2$ denotes the measurement error variance and the exponential factor is the serial variance component with $u_{DD'}$ the time lag between measurements on days $D$ and $D'$. The estimated covariance parameters of this model, applied to the PANSS data, are $\widehat{\sigma_P^2} = 85.6$, $\widehat{\sigma_C^2} = 16.1$, $\widehat{\tau^2} = 277.2$, $\widehat{\rho} = 6.4$, and $\widehat{\sigma_E^2} = 65.3$. When we calculate the overall test-retest reliability coefficient, only generalizing across time points, similarly as we did in (7.4), we have:

$$\text{Corr}(Y_{PDTC}, Y_{PD'TC} \mid T, D, D') = \frac{\sigma_P^2 + \sigma_C^2 + \tau^2 \exp\left(\frac{-u_{DD'}^2}{\rho^2}\right)}{\sigma_P^2 + \sigma_C^2 + \tau^2 + \sigma_E^2}. \tag{7.6}$$

In analogy with (7.5), if we now generalize across time points and countries, we obtain:

$$\text{Corr}(Y_{PDTC}, Y_{PD'TC'} \mid C, C', T, D, D') = \frac{\sigma_P^2 + \tau^2 \exp\left(\frac{-u_{DD'}^2}{\rho^2}\right)}{\sigma_P^2 + \sigma_C^2 + \tau^2 + \sigma_E^2}, \tag{7.7}$$

which is a function of the time lag $u_{DD'}$ between two observations measured at time points $D$ and $D'$. For instance, the test-retest reliability (7.6) for two measurements observed with a time lag of two weeks is 0.796 while the generalizability coefficient (7.7) amounts to 0.759. For a time lag of 6 weeks we respectively have 0.491 and 0.454.

## 7.6 Concluding Remarks

The innovation this Chapter promotes is the introduction and implementation of valuable ideas from psychometrics into the area of clinical trials. The tool we use is the general linear mixed-effects model, including serial correlation. To our knowledge, this model has not been used in the psychometric literature to establish reliability or generalizability. Also, there seem to be no references discussing the calculation of reliability-generalizability coefficient based upon data from clinical trials.

Table 7.1: *S*chizophrenia PANSS Data. Reliability of the total PANSS per country and impact of country on overall reliability - Summary table.

| | | By country | | | | Omitting a country | |
|---|---|---|---|---|---|---|---|
| Country | # pat. | $d^2$ | $\sigma^2$ | Reliability | 95% CI | Reliability | 95% CI |
| ARG | 31 | 162.0 | 49.0 | 0.768 | [0.640 ; 0.854] | 0.711 | [0.685 ; 0.735] |
| AUT | 29 | 476.9 | 267.0 | 0.641 | [0.456 ; 0.773] | 0.718 | [0.693 ; 0.742] |
| BEL | 26 | 229.0 | 140.7 | 0.619 | [0.431 ; 0.756] | 0.716 | [0.690 ; 0.739] |
| BRA | 44 | 383.9 | 72.0 | 0.842 | [0.764 ; 0.896] | 0.704 | [0.677 ; 0.728] |
| CAN | 44 | 264.8 | 205.0 | 0.564 | [0.408 ; 0.688] | 0.723 | [0.698 ; 0.746] |
| DEN | 47 | 301.7 | 99.0 | 0.753 | [0.644 ; 0.832] | 0.711 | [0.685 ; 0.735] |
| ESP | 32 | 143.4 | 98.6 | 0.593 | [0.421 ; 0.723] | 0.711 | [0.685 ; 0.735] |
| FIN | 71 | 147.5 | 103.2 | 0.588 | [0.463 ; 0.691] | 0.720 | [0.694 ; 0.744] |
| FRA | 92 | 354.4 | 183.3 | 0.659 | [0.567 ; 0.735] | 0.713 | [0.686 ; 0.737] |
| GBR | 21 | 263.0 | 56.9 | 0.822 | [0.676 ; 0.906] | 0.711 | [0.685 ; 0.735] |
| GER | 25 | 249.9 | 88.9 | 0.738 | [0.566 ; 0.848] | 0.713 | [0.687 ; 0.737] |
| ITA | 39 | 356.9 | 81.0 | 0.815 | [0.719 ; 0.881] | 0.707 | [0.681 ; 0.732] |
| MEX | 36 | 376.2 | 178.6 | 0.678 | [0.534 ; 0.784] | 0.716 | [0.691 ; 0.740] |
| NED | 17 | 276.3 | 68.1 | 0.802 | [0.624 ; 0.901] | 0.713 | [0.687 ; 0.736] |
| NOR | 37 | 146.4 | 111.5 | 0.568 | [0.384 ; 0.708] | 0.716 | [0.690 ; 0.739] |
| RSA | 79 | 338.8 | 107.5 | 0.759 | [0.676 ; 0.823] | 0.707 | [0.681 ; 0.732] |
| SWE | 30 | 202.6 | 135.9 | 0.598 | [0.419 ; 0.733] | 0.715 | [0.690 ; 0.739] |
| USA | 122 | 323.6 | 114.1 | 0.739 | [0.676 ; 0.792] | 0.711 | [0.683 ; 0.736] |

| Analysis | Reliability/generalizability coefficient |
|---|---|
| uncontrolled for country | 0.713 [0.688; 0.736] |
| Adjusted for country as fixed effect | 0.702 [0.676; 0.727] |
| Weighted average across countries | 0.735 [0.635; 0.834] |
| Weighted average across countries (Fisher's) | 0.702 [0.577; 0.794] |
| Conditioning on country as random effect | 0.712 [0.686; 0.737] |
| Generalizing for country and test-retest | 0.676 [0.639; 0.709] |

The methodology we propose unifies two strands of validation technology: (1) reliability in the classical psychometric sense, including test-retest reliability and inter-rater agreement, and (2) generalizability theory used, for example, to assess the effect of country. All of this is done by embedding the classical linear models, used in reliability and generalizability theories, within a linear mixed-effects model framework. It is further shown that commonly used reliability and generalizability measures can be derived as (conditional) correlation coefficients. Apart from unification and resting on simple principles, our approach has several additional advantages: (1) the often strong and unrealistic assumption of steady state behavior, needed for classical reliability assessment, is not needed, thus allowing to use clinical trial data that have not been gathered explicitly with validation assessment in mind; (2) data with more than two measurements per patient can usefully be used irrespective of the particular measure under consideration; (3) data sequences do not need to be of equal length which is important, for example, when data are incomplete; (4) where applicable, serial correlation can be incorporated within the framework.

As stated in the encyclopedia of biostatistics of Armitage and Colton (1998), assessing observer reliability and agreement is essential for interpretation of clinical observations both in research and in medical practice. In general, improvement of observer reliability or agreement of clinical observations may have a lot of impact on the quality of health care. Tracing the sources and types of disagreements is the beginning of wisdom. Generalizability studies, which aim to determine the origin of the variation and their relative contribution to measurement errors, are most valuable in this respect. These studies can measure, among other things, the contribution of intra-observer and inter-observer variation to the total of measurement errors. This work has focused on the identification and on the measurement of factors that have an impact on measurements in a clinical trial. Subsequently, knowledge about the origin of the errors can help to improve the quality of the measurements. In the example above, a relatively high generalizability coefficient as determined by (7.5) suggested that country does not have a significant impact on the test-retest reliability and on measurement error. Thus, there is no need to reduce the influence of country on the PANSS for future trials. However, when we investigated the impact of baseline PANSS negative subtotal on measurement error, conclusions were different. We divided the baseline PANSS Negative subtotal in different categories. Subsequently, we derived the variance components and calculate the generalizability coefficient for baseline PANSS Negative subtotal similarly to the way it was done for country in (7.5). In this

analysis, the generalizability coefficient reduced to 0.52. This indicates that baseline PANSS Negative subtotal reduces the test-retest reliability. Subsequent analyses show that reliability is lowest in patients with a high baseline PANSS Negative subtotal. This means that patients with a higher deficit in social functions such as poverty of speech, apathy and emotional withdrawal are more difficult to rate, resulting in higher measurement error and lower test-retest reliability. In this case, the strategy could result in either additional training to rate patients with a high baseline negative subtotal, or in using a different scale in this type of patients.

Our approach is based on assessing generalizability after correcting for fixed effects, the most prominent one being treatment effect. When the treatments present in a trial are considered representative for a wider class of treatments, exactly as country or investigator do, treatment could be handled similarly.

Shavelson, Webb, and Rowley (1989) and Dunn (1989) have noticed that GT is not widely applied in psychological research and they both make a plea for a more extensive use. Shavelson surmises that the mathematical-technical development might be the main reason. Another reason might be that the cost to set-up G-studies for scale developers can be large. In this Chapter we have shown that the richness of clinical trial data can be used for this purpose.

Current Chapter focused on interval scaled measurements. In principle, the methodology can be extended to categorical data. In that case, the linear mixed model could be replaced by a generalized linear mixed model, allowing for instance for non-Gaussian data. This will be worked out in Chapter 8 and 9 respectively for reliability and generalizability.

# Chapter 8

# Reliability Estimation in Case of Binary Biomedical Data

In applied sciences, one is often confronted with the collection of hierarchical data or repeated measures, in particular longitudinal or clustered data. Methods for continuous such data are centered around the well-developed linear mixed effects model (LMM, Verbeke and Molenberghs, 2000); the same is true for software implementation. Drawing from the normal distribution, the LMM allows one to obtain marginal characteristics, such as marginal means, marginal covariate effects, and marginal correlation coefficients, in a very straightforward way. This is because the natural parameters in an LMM have a hierarchical and a marginal interpretation at the same time. Hence, deriving the intraclass correlation (ICC) from a random-intercept LMM is particularly straightforward as shown in (6.2) and coincides with the correlation from a compound-symmetric structure, the latter being the marginalization of the former. This makes the LMM is a flexible tool to study psychometric reliability based on longitudinal data, as in Chapter 6 and in Vangeneugden *et al.* (2004). Reliability reflects the amount of error inherent in any measurement and hence, in a general sense, how replication of the administration would give a different result (Streiner and Norman, 1995).

While also non-Gaussian outcomes are prominent, model formulation is less straightforward. One distinguishes between marginal and random-effects model families with now no easy relationship between both. An example of the marginal family is general-

ized estimating equations (GEE, Liang and Zeger, 1986), whereas the generalized linear mixed model (GLMM, Breslow and Clayton, 1993) is a well-known random-effects model. Whereas GEE is convenient and frequently used, it models the marginal regression function, treating the second and higher-order moments as nuisance. When the correlation is of primary scientific interest, e.g., when determining the ICC or studying reliability, a non-likelihood method like GEE has clear limitations. The GLMM has a full likelihood basis, but fails to produce the marginal correlations in an easy fashion, owing to a non-linear link function, as well as the mean-variance link (Molenberghs and Verbeke, 2005, Chapter 16). Due to the flexibility of the GLMM, it is a viable modeling candidate, even when the marginal correlation is of interest. We will show that the derivation of such correlations is generally feasible and derive the intra-class correlation coefficient of *reliability*. Note that, in classical terms, reliability is defined as the variance attributed to the difference among subjects divided by the total variance (Shrout and Fleiss, 1979) and therefore takes the form of the intra-class correlation coefficient as described in (4.3).

In this Chapter we will show how correlations can be derived by means of a GLMM, with particular attention to the reliability functions, operationalized by means of the ICC. It will be clear in what follows that, in the non-Gaussian case, reliability will no longer be constant, excepting special cases.

The work in this Chapter is described in Vangeneugden *et al.* (2008b), and partly in Molenberghs, Vangeneugden, and Laenen (2008).

## 8.1 Reliability Estimation in the General Linear Mixed Model Framework

In the GLMM setting introduced in Section 3.3, we can write the general model as follows: $\boldsymbol{Y}_i = \boldsymbol{\mu}_i + \boldsymbol{\varepsilon}_i$, where $\boldsymbol{\mu}_i$, the conditional mean, given the random effects, can be written as $\boldsymbol{\mu}_i = \boldsymbol{\mu}_i(\boldsymbol{\eta}_i) = h(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i)$, $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are known design matrices, $\boldsymbol{\beta}$ are fixed-effect parameters, $\boldsymbol{b}_i$ are random effects, and $h$ is a known link function. Finally, $\boldsymbol{\varepsilon}_i$ is the residual error component. We will now derive a general formula for the variance-covariance matrix of $\boldsymbol{Y}_i$ without any restriction on the distribution of the outcome variable nor on the complexity of the model, e.g., allowing for serial correlation or not. This maximizes the similarity with the case of continuous, normally distributed outcomes. However, a key distinction is that in the linear case there is

no mean-variance link, whereas here the residual variance will follow from the mean. The variance covariance matrix can be derived as follows:

$$\boldsymbol{V}_i = \text{Var}(\boldsymbol{Y}_i) = \text{Var}(\boldsymbol{\mu}_i + \boldsymbol{\varepsilon}_i) = \text{Var}(\boldsymbol{\mu}_i) + \text{Var}(\boldsymbol{\varepsilon}_i) + 2\text{Cov}(\boldsymbol{\mu}_i, \boldsymbol{\varepsilon}_i). \tag{8.1}$$

It is easy to show that $\text{Cov}(\boldsymbol{\mu}_i, \boldsymbol{\varepsilon}_i) = \text{Cov}[\text{E}(\boldsymbol{\mu}_i|\boldsymbol{b}_i), \text{E}(\boldsymbol{\varepsilon}_i|\boldsymbol{b}_i)] + \text{E}[\text{Cov}(\boldsymbol{\mu}_i, \boldsymbol{\varepsilon}_i|\boldsymbol{b}_i)] = 0$ since the first term is 0 and the second term equals $\text{E}\{\text{E}[\boldsymbol{\mu}_i - \text{E}(\boldsymbol{\mu}_i)](\boldsymbol{\varepsilon}_i)|\boldsymbol{b_i}\} = 0$ as $\boldsymbol{\mu}_i$ is a constant when conditioning on $\boldsymbol{b}_i$. For the first term in (8.1) we have, using a first-order Taylor series expansion around $\boldsymbol{b}_i = \boldsymbol{0}$:

$$\begin{aligned}
\text{Var}[\boldsymbol{\mu}_i] &= \text{Var}[\boldsymbol{\mu}_i(\boldsymbol{\eta_i})] = \text{Var}[\boldsymbol{\mu}_i(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i)] \\
&\cong \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{b}_i}\bigg|_{\boldsymbol{b}_i=0}\right) \text{Var}(\boldsymbol{b}_i) \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{b}_i}\bigg|_{\boldsymbol{b}_i=0}\right)' \\
&\cong \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\eta_i}}\frac{\partial \boldsymbol{\eta_i}}{\partial \boldsymbol{b}_i}\bigg|_{\boldsymbol{b}_i=0}\right) \boldsymbol{D} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\eta_i}}\frac{\partial \boldsymbol{\eta_i}}{\partial \boldsymbol{b}_i}\bigg|_{\boldsymbol{b}_i=0}\right)' \\
&\cong \boldsymbol{\Delta}_i\boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i'\boldsymbol{\Delta}_i',
\end{aligned} \tag{8.2}$$

where $\boldsymbol{\Delta}_i = \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\eta_i}}\Big|_{\boldsymbol{b}_i=0}$. For the second term in (8.1), we have:

$$\text{Var}(\boldsymbol{\varepsilon}_i) = \text{Var}[E(\boldsymbol{\varepsilon}_i|\boldsymbol{b}_i)] + E[\text{Var}(\boldsymbol{\varepsilon}_i|\boldsymbol{b}_i)] = E[\text{Var}(\boldsymbol{\varepsilon}_i|\boldsymbol{b}_i)] = \boldsymbol{\Phi}^{\frac{1}{2}}\boldsymbol{\Sigma}_i\boldsymbol{\Phi}^{\frac{1}{2}}, \tag{8.3}$$

where $\boldsymbol{\Phi}$ is a diagonal matrix with the overdispersion parameters along the diagonal. In case there are no overdispersion parameters, $\boldsymbol{\Phi}$ is set equal to the identity matrix. We can expand the variance function $\boldsymbol{\Sigma}_i$ so that

$$\text{Var}(\boldsymbol{\varepsilon}_i) = \boldsymbol{\Phi}^{\frac{1}{2}}\boldsymbol{A}_i^{\frac{1}{2}}\boldsymbol{R}_i\boldsymbol{A}_i^{\frac{1}{2}}\boldsymbol{\Phi}^{\frac{1}{2}}, \tag{8.4}$$

where $\boldsymbol{R}_i$ is the correlation matrix and $\boldsymbol{A}_i$ is a diagonal matrix containing the variances following from the generalized linear model specification of $\boldsymbol{Y}_{ij}$ given the random effects $\boldsymbol{b}_i = \boldsymbol{0}$, i.e., with diagonal elements $v(\mu_{ij}|\boldsymbol{b}_i = \boldsymbol{0})$. Using (8.2) and (8.4), we have the following expression for the variance-covariance matrix (8.1):

$$\boldsymbol{V}_i \cong \boldsymbol{\Delta}_i\boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i'\boldsymbol{\Delta}_i' + \boldsymbol{\Phi}^{\frac{1}{2}}\boldsymbol{A}_i^{\frac{1}{2}}\boldsymbol{R}_i\boldsymbol{A}_i^{\frac{1}{2}}\boldsymbol{\Phi}^{\frac{1}{2}}. \tag{8.5}$$

If the canonical link is used, we have $\boldsymbol{A}_i = \boldsymbol{\Delta}_i$ and (8.5) can be written as: $\boldsymbol{V}_i \cong \boldsymbol{\Delta}_i\boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i'\boldsymbol{\Delta}_i' + \boldsymbol{\Phi}^{\frac{1}{2}}\boldsymbol{\Delta}_i^{\frac{1}{2}}\boldsymbol{R}_i\boldsymbol{\Delta}_i^{\frac{1}{2}}\boldsymbol{\Phi}^{\frac{1}{2}}$. If in addition, conditional independence (no serial

correlation) is assumed, then (8.5) simplifies to: $\boldsymbol{V}_i \cong \boldsymbol{\Delta}_i \boldsymbol{Z}_i \boldsymbol{D} \boldsymbol{Z}_i' \boldsymbol{\Delta}_i' + \boldsymbol{\Phi}^{\frac{1}{2}} \boldsymbol{\Delta}_i \boldsymbol{\Phi}^{\frac{1}{2}}$. Further, if we reduce the random-effects part to a random-intercept model, i.e., $\boldsymbol{Z}_i = \boldsymbol{1}$ and $\boldsymbol{D} = d$, and (8.5) then reduces to $\boldsymbol{V}_i \cong \boldsymbol{\Delta}_i (d\boldsymbol{J}) \boldsymbol{\Delta}_i' + \boldsymbol{\Phi}^{\frac{1}{2}} \boldsymbol{\Delta}_i \boldsymbol{\Phi}^{\frac{1}{2}}$. Note that, if we have a normal distribution with the canonical identity link, $\boldsymbol{\Delta}_i$ reduces to the identity matrix $\boldsymbol{I}$ and $\boldsymbol{\Phi} = \sigma^2 \boldsymbol{I}$, in which case it follows that $\boldsymbol{V}_i$ reduces to $d\boldsymbol{J} + \sigma^2 \boldsymbol{I}$, with $\boldsymbol{J}$ a square $n_i$ dimensional matrix of ones, which is consistent with (6.3). Moreover, when we have a normal distribution with a general random-effects structure but without serial correlation, it is easy to show that $\boldsymbol{V}_i \cong \boldsymbol{Z}_i \boldsymbol{D} \boldsymbol{Z}_i' + \sigma^2 \boldsymbol{I}$ and that subsequently $\rho$ equals (6.2) when we leave out the serial correlation ($\tau$). This shows that (8.5) can be seen as a generalization of (6.2). While the above derivation is referred to as a first-order Taylor series expansion, the exact same expression follows if a second-order expansion is considered, owing to terms vanishing. Therefore, we are authorized to refer to it as a second-order Taylor series expansion, too. In the following section we will derive the marginal correlation in case of binary data when applying a random intercept model.

## 8.2 ICC for a Random-intercept Model for Binary Data

In this section, we will derive the formula for the ICC in case of a random intercept model for binomial data with a logit link and assuming no overdispersion. In this case, $\boldsymbol{V}_i$ reduces to $\boldsymbol{V}_i \cong \boldsymbol{\Delta}_i (d\boldsymbol{J}) \boldsymbol{\Delta}_i' + \boldsymbol{\Delta}_i = \boldsymbol{\Delta}_i (d\boldsymbol{J} + \boldsymbol{\Delta}_i^{-1}) \boldsymbol{\Delta}_i'$. Furthermore, $\boldsymbol{\Delta}_i$ is a diagonal matrix with $V_{ij}(0)$ as diagonal elements, where the variance function $V_{ij}(0) = \mu_{ij} \mid_{\boldsymbol{b}_i = 0} (1 - \mu_{ij} \mid_{\boldsymbol{b}_i = 0})$, and therefore $\boldsymbol{V}_i \cong \text{diag}(V_{ij}(0))[d\boldsymbol{J} + \text{diag}(V_{ij}(0))^{-1}]\text{diag}(V_{ij}(0))$. In other words, the variance-covariance matrix for subject $i$ is specified by the matrix with elements: $v_{ijj} = V_{ij}(0)[1 + V_{ij}(0)d]$, $v_{ijk} = dV_{ij}(0)V_{ik}(0), \quad (j \neq k)$. Based on these, we can determine a first-order approximation of the marginal correlation between time point $j$ and $k$, which is the intra class correlation coefficient of reliability:

$$\rho_{ijk} = \text{Corr}(Y_{ij}, Y_{ik}) = \frac{V_{ij}(0)V_{ik}(0)d}{\sqrt{\{V_{ij}(0)[1 + V_{ij}(0)d]\}\{V_{ik}(0)[1 + V_{ik}(0)d]\}}}. \tag{8.6}$$

This expression allows us to make a few simple but important observations. For any value of $V_{ij}(0)$ and $V_{ik}(0)$, $\rho_{ijk} = 0$ whenever $d = 0$, while $\rho_{ijk}$ tends to 1 when $d$ tends to $+\infty$. Even though this may seem obvious at first sight, especially because

it is similar to the behavior of the intraclass correlation in the classical linear model for continuous data, one must give proper reflection to the impact of the binary nature of our outcomes, since certain correlation coefficients in certain models are highly constrained. For example, the correlation coefficients in the Bahadur (1961) model, being of the Pearson type, are highly constrained (Aerts *et al.* 2002). These authors showed that in some realistic settings only a tiny interval around zero of allowable correlations remains. It is useful to realize such constraints already apply to the Pearson correlation in a simple two by two contingency table. A mild form of the Bahadur constraints survives in generalized estimating equations, especially those of the second order. The multivariate probit model (Molenberghs and Verbeke, 2005), on the other hand, is constrained only by the requirement that the tetrachoric correlations form a positive definite matrix. This advantage of the probit model is counterbalanced by its heavy computational burden. Also, the beta-binomial model (Molenberghs and Verbeke, 2005) allows for all non-negative correlations as well as moderate negative values (see also Molenberghs and Verbeke, 2005). The problem suffers from its inability to accommodate within-cluster covariates, such as time in longitudinal studies. Thus, the proposed modeling framework is at the same time flexible, relatively easy from a numerical point of view, and does not face the strong constraints like in, for example, the Bahadur (1961) model.

One might wonder why no negative correlations are allowed. Also this aspect is similar to the linear mixed model, where the random-intercepts model, when its full hierarchical interpretation is adopted, does not allow for negative correlations. Once attention is restricted to the marginal model, some negative correlation can occur as well. Indeed, the compound-symmetry model can produce negative correlations, as long as the overall correlation matrix, of the form $\sigma^2\boldsymbol{I}+d\boldsymbol{J}$, remains positive-definite. Note that, while this article focuses on the correlation coefficient, also in line with classical reliability approaches, other measures of association between the outcomes, such as the odds ratio model (Molenberghs and Verbeke, 2005) could be entertained. Arguably, this would require a fundamentally different approach, and is beyond the scope of this article.

## 8.3 Simulation Study

A reason for concern is the quality of approximation (8.6) since, unlike in the linear case, here a Taylor series expansion needs to be used. To provide a perspective on the

impact of this issue, we conducted a limited but insightful set of simulations. Precisely, we generated data from the Bahadur (1961) model, and then estimated the correlation coefficient using both generalized estimating equations (GEE, Liang and Zeger, 1986) and our proposed approach. While it ought to be noted that a correlation coefficient for non-continuous data is a model-dependent concept, the relative agreement between the coefficients resulting from the various models still sheds some light on the quality of the approximation.

The Bahadur model is defined in terms of the marginal probability $\pi_{ij} = E(Y_{ij}) = P(Y_{ij} = 1)$ and standardized deviations $\varepsilon_{ij} = (Y_{ij} - \pi_{ij})/\sqrt{\pi_{ij}(1-\pi_{ij})}$ and $e_{ij} = (y_{ij} - \pi_{ij})/\sqrt{\pi_{ij}(1-\pi_{ij})}$, where $y_{ij}$ is an actual value of the binary response variable $Y_{ij}$. Further, letting $\rho_{ij_1j_2} = E(\varepsilon_{ij_1}\varepsilon_{ij_2})$, $\rho_{ij_1j_2j_3} = E(\varepsilon_{ij_1}\varepsilon_{ij_2}\varepsilon_{ij_3})$, ..., $\rho_{i12\ldots n_i} = E(\varepsilon_{i1}\varepsilon_{i2}\ldots\varepsilon_{in_i})$, the general Bahadur model can be represented by the expression $f(\boldsymbol{y}_i) = f_1(\boldsymbol{y}_i)c(\boldsymbol{y}_i)$, where $f_1(\boldsymbol{y}_i) = \prod_{j=1}^{n_i} \pi_{ij}^{y_{ij}}(1-\pi_{ij})^{1-y_{ij}}$ and

$$c(\boldsymbol{y}_i) = 1 + \sum_{j_1<j_2} \rho_{ij_1j_2} e_{ij_1} e_{ij_2} + \sum_{j_1<j_2<j_3} \rho_{ij_1j_2j_3} e_{ij_1} e_{ij_2} e_{ij_3} + \ldots + \rho_{i12\ldots n_i} e_{i1} e_{i2} \ldots e_{in_i}.$$

For the purpose of our simulations, we will restrict this model to 2 and 3 measurements per subject, respectively. In the latter case, the three pairwise correlation will be set equal, while the third-order correlation will be set to zero. GEE can be viewed as a version of the Bahadur model where the higher-order correlations are left unspecified, and the pairwise correlation structure is considered a nuisance characteristic.

For the number of measurements equal to $n_i = n = 2$, the true correlations $\rho = 0.25$, 0.50, and 0.75 were considered, while for $n_i = n = 3$ we focused on $\rho = 0.20$, 0.40, and 0.60. For all six settings, 1000 datasets of size 200 patients were generated. For each such dataset, the pairwise correlation was estimated using both GEE and the proposed GLMM-based expression (8.6). Table 8.1 presents the results in terms of the simulation-averaged correlation together with its standard deviation. Not surprising, the agreement between GEE and the generating Bahadur model is excellent, since GEE can be viewed as a restricted-moment version of the Bahadur model. Importantly for our purposes, the behavior of the GLMM-based expression (8.6) is quite acceptable. While, as stated earlier, the correlation is model-dependent, it falls everywhere within the same range as the one of the generating model. Note that, for our approach when $n = 3$, we have three coefficients, one for each pair of measurements. It would, in principle, be possible to replace the three estimates with a

Table 8.1: *Results of the simulation study. n refers to the number of measurements per subject. 'True' is the correlation used in the generating Bahadur model. For both GEE and GLMM, the simulation-averaged correlation coefficients and their simulation standard deviations are reported.*

| | | GEE | GLMM | |
|---|---|---|---|---|
| $n = n_i$ | True $\rho$ | Est.(SD) | Coeff | Est.(SD) |
| 2 | 0.25 | 0.248 (0.07) | $\rho$ | 0.270 (0.08) |
| 2 | 0.50 | 0.499 (0.06) | $\rho$ | 0.554 (0.07) |
| 2 | 0.75 | 0.753 (0.05) | $\rho$ | 0.568 (0.30) |
| 3 | 0.20 | 0.199 (0.05) | $\rho_{12}$ | 0.225 (0.06) |
| | | | $\rho_{13}$ | 0.228 (0.06) |
| | | | $\rho_{23}$ | 0.237 (0.06) |
| 3 | 0.40 | 0.398 (0.05) | $\rho_{12}$ | 0.467 (0.06) |
| | | | $\rho_{13}$ | 0.474 (0.06) |
| | | | $\rho_{23}$ | 0.497 (0.06) |
| 3 | 0.60 | 0.598 (0.04) | $\rho_{12}$ | 0.666 (0.05) |
| | | | $\rho_{13}$ | 0.678 (0.05) |
| | | | $\rho_{23}$ | 0.723 (0.04) |

common one. Since this would come down in averaging the three correlations, it would further enhance stability. This is why we have chosen this somewhat more variable and therefore conservative presentation in terms of three separate coefficients.

Additionally, a simulation based on an actual GLMM was performed, using a simple random-effects model with $\boldsymbol{X}_i\boldsymbol{\beta} = \beta_0$. In this simulation, 10,000 datasets with 200 subjects were generated, each subject having 5 measurements as in the application of Section 8.4. Here, $\beta_0 = -1.61$ and the variance of the random intercept, $d = 6.57$, was taken as observed in the application. Also here, the pairwise correlation was estimated using both GEE as well as the proposed GLMM-based expression (8.6). For the GEE, the mean correlation and its standard deviation was observed to be 0.465 (s.d. 0.04) and for GLMM the results were very similar, leading to a mean correlation of 0.473 (s.d. 0.05). Note that the GLMM based correlation of the real

data was estimated to be 0.48.

Thus, we conclude that the correlation, based on GLMM, is a practically acceptable indication for association. In principle, it would be possible to further enhance performance using Monte-Carlo Markov Chain based methods, including the bootstrap. While such an approach would increase the computational burden somewhat, it certainly falls within the realm of practical feasibility.

## 8.4 Data Analysis

Let us now apply the concepts described above to the pooled schizophrenia data described in Section 2.1. We will calculate the ICC for response defined as obtaining either *very much improved* or *much improved* on the CGI of overall change versus baseline. The focus of this analysis is not to study treatment differences, but rather to investigate correlation between longitudinal binary data. To do so, we will calculate the ICC under different assumptions, with gradually increasing modeling complexity. For simplicity, we will focus on models with random intercepts and no serial correlation. Of course, as stated in Section 8.1, the extension to the more general case is straightforward but algebraically a bit more tedious.

### 8.4.1 Observed Response Rate and Correlation

Figure 8.1 displays the response rate of both treatment groups combined across time. We can see that the observed response rate increases over time from 0.15 at week 1 to 0.47 at week 8. Also note that only 490 from the 774 subjects who started treatment have an observed CGI score at week 8 due to attrition. Figure 8.2 illustrates the correlation of the different responses over time. Circles are drawn at different response level combinations (no, yes) in this matrix plot, exhibiting the correlation between the observed responses at different pairs of time points. The diameter of the circle is proportional to the number of subjects at each response combination, e.g., the large circle for response=No at Week 1 and response=No at Week 2 indicates that many subjects who did not respond at Week 1 also did not respond at Week 2. The larger the diameter of the circles are at the bisector line ($y = x$), the larger the correlation is between the same level of response at time point $j$ versus time point $j'$. From the first 2 rows of the plot, we can see that correlation is high if we compare Week 1 and 2, but that this correlation decreases slightly in time, when the lag time between

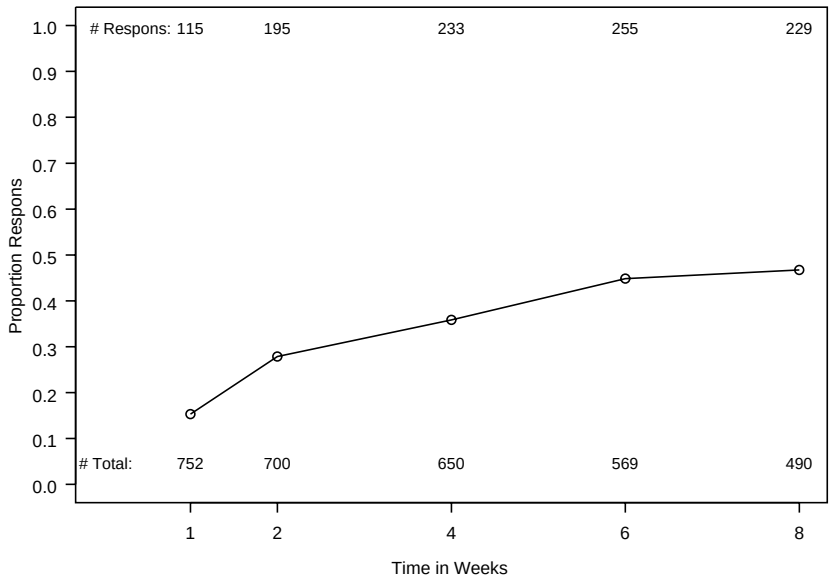


Figure 8.1: *Schizophrenia CGI Data. Graphical representation of observed response over time.*

observations is increased. On the other hand, the correlation between Week 6 and 8 is noticeably higher, as the diameter of the circles on the bisector line (same response at Week 6 and 8) are large and almost zero for circles not on the bisector line (different response on Week 6 and 8).

### 8.4.2 Initial Analysis

To exemplify computations, let us assume there are no covariates. Then, $\boldsymbol{X}_{is}\boldsymbol{\beta} = \beta$ is constant and (8.6) simplifies to: $V_{ij}(0) = V(0) = \exp(\beta)/(1+\exp(\beta))^2$ and $\rho_{ijk} = \rho = V(0)d/(1+V(0)d)$. When using this expression for a variety of subgroups and/or combination of times, a detailed picture can emerge but, as we will illustrate in what follows, it is possible and more elegant to incorporate the ICC into a fully specified model.

We can use the SAS procedure NLMIXED to fit this random-effects model, using adaptive Gaussian quadrature. Table 8.2 summarizes the results for a selection of subgroups. Before discussing these, let us note that subgroup analyses can rightfully be considered unsatisfactory by some. Therefore, we will revisit the concept of sub-

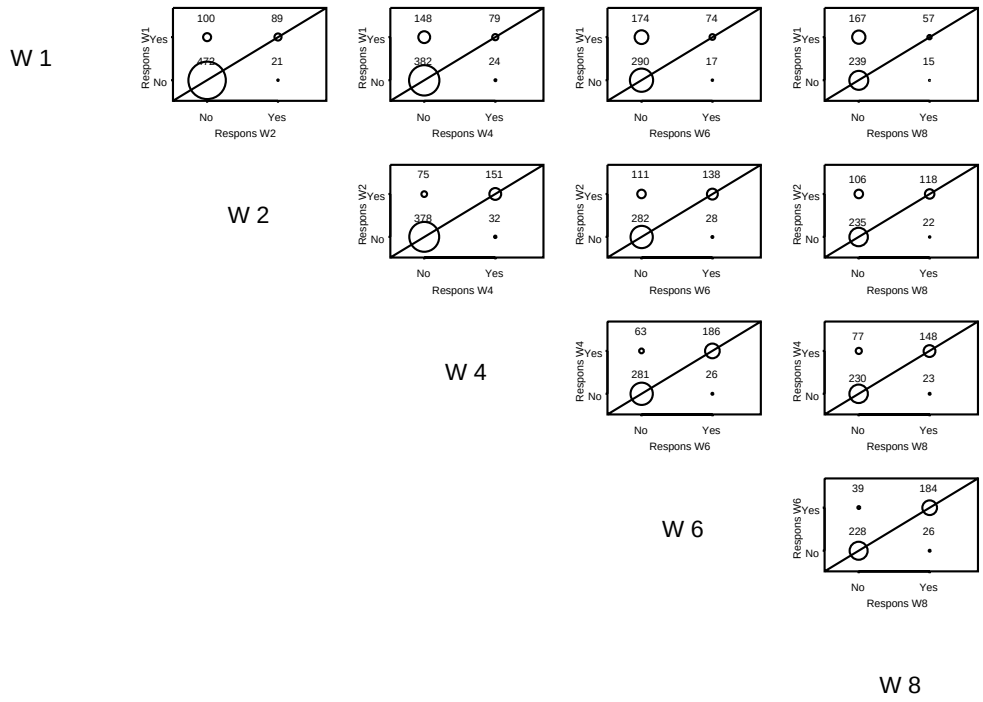


Figure 8.2: *Schizophrenia CGI Data. Graphical representation of the correlation of observed response over time.*

groups, but then in a more principled modeling approach, in Section 8.4.3. An added advantage of this approach is that the quality of the fit will be enhanced, owing to the high-quality approximation to the integration, required for likelihood evaluation. This is important, not only for the determination of the correlation coefficient, but also for other assessments, such as whether there is a significant treatment difference. Of course, one should be aware that reaching convergence with the NLMIXED procedure or related software for non-linear models is not straightforward. Tools exploiting linearity of the predictor are somewhat easier, but often based on poor approximations such as first-order PQL or MQL (Molenberghs and Verbeke, 2005). Such alternative procedures may be used, however, to obtain good starting values, upon which the use of the non-linear procedures becomes easier.

One observes that the ICC is somewhat larger in the risperidone treatment group. Additionally, we see that the ICC for observations measured at Week 1 and Week 8

Table 8.2: *Schizophrenia CGI Data. Summary of different subgroup analysis investigating impact time and treatment effect on reliability. Standard errors are calculated from the delta method.*

| | Intraclass correlation $\rho$ (SE) | | |
|---|---|---|---|
| time points included | combined treatments | risperidone | active control |
| all time points | 0.48 (0.026) | 0.55 (0.038) | 0.40 (0.035) |
| Week 1 and Week 8 | 0.11 (0.045) | 0.11 (0.066) | 0.10 (0.060) |
| Week 6 and Week 8 | 0.85 (0.026) | 0.87 (0.032) | 0.82 (0.043) |

Table 8.3: *Schizophrenia CGI Data. Overall ICC (SE) matrix, marginal over treatment. Standard errors are calculated from the delta method.*

| | **time** | | | |
|---|---|---|---|---|
| **Week** | 2 | 4 | 6 | 8 |
| 1 | 0.29 (0.029) | 0.33 (0.030) | 0.35 (0.029) | 0.35 (0.029) |
| 2 | 1 | 0.53 (0.032) | 0.57 (0.030) | 0.57 (0.029) |
| 4 | | 1 | 0.64 (0.027) | 0.65 (0.026) |
| 6 | | | 1 | 0.70 (0.024) |

is much smaller than the ICC measured from observations at Week 6 and Week 8. Here we should note that the ICC between Week 6 and 8 can truly be interpreted as an ICC of reliability in the psychometric sense. Indeed, the psychiatric condition of the patients was rather stable and did not change between Week 6 and 8: the mean total PANSS was 69.2 at Week 6 and 68.8 at Week 8. Also CGI response remained stable between Week 6 and 8: 412 (86%) out of the 477 subjects had the same response level as can be observed in Figure 8.2. It is in such stable conditions that test-retest reliability of scale is evaluated, and often with a two-week time interval (Streiner and Norman 1995, Chapter 8). The same is not true when comparing Week 1 (mean PANSS of 80.8) and Week 8; that is, the ICC between Week 1 and 8 cannot be interpreted as an ICC of reliability but merely a correlation between two time points. As discussed in Vangeneugden *et al.* (2004), appropriate models can

be used to model and extract time and treatment effects, which avoids the need to assume that there is no change in a patient's situation over time. Thus, by using an appropriate model with well chosen covariate effects, a trial population is, in a broad sense, standardized towards a general population. By correcting for covariates, it is assumed that the correlation structure of the residuals can be approximated by an exchangeable structure, captured via a random intercept. While this may be perceived as somewhat more subjective than when a dedicated reliability study is undertaken, the important advantage is that data already collected can be used, which may have important practical, economic, and even ethical advantages. It is important to note that, in case a random intercept is deemed insufficient to capture the correlation structure, more versatile random-effects structures can be used, whilst maintaining the idea behind the calculations for the marginal correlation coefficients. We will gradually take account of this, by first extracting time and then subsequently treatment effects. Of course, one ought not to forget that important but potentially complicated issues, such as dropout and non-compliance, may intervene. Since the method is likelihood-based, it is valid under the broad assumption of missingness at random, whereby missingness depends on observed outcomes and covariates but, given these, not further on unobserved outcomes. Likewise, when compliance issues intervene, it is important the covariates are chosen such that the causal interpretation of the resulting model be maintained. With good to perfect compliance, this is taken care of by virtue of randomization.

### 8.4.3 Accounting for Time and Treatment

If we adjust for time and ignore treatment, then $\rho$ can be derived via (8.6) and it is easy to show that $V_{ij}(0) = \exp(\beta_j)/(1+\exp(\beta_j))^2$, where $\boldsymbol{\beta}_j$ is the estimated coefficient of the indicator variable representing time $j$, when we use a model without an intercept in the fixed effects. The variation of the random effect was estimated to be $\widehat{d} = 10.04$ and this time we had $\widehat{\beta_{W1}} = -3.79$, $\widehat{\beta_{W2}} = -2.25$, $\widehat{\beta_{W4}} = -1.50$, $\widehat{\beta_{W6}} = -3.79$ and $\widehat{\beta_{W8}} = -0.41$. Table 8.3 provides the estimated intra-class correlation coefficient matrix. This is in line with the well-known relationship between marginal and random-effects regression parameters (Verbeke and Molenberghs, 2005), the correlations are determined by the random-intercept variance, together with the marginal probabilities factoring into the variance function: $\beta_j \cong \sqrt{1+0.346\ d}\cdot\text{logit}(p_j)$. Hence, these correlations are constant only in the simple case of a constant mean. Otherwise,

they are functions of the covariates. Note that, in case a random-intercepts model is deemed too simple, a more elaborate random-effects structure can be assumed, whilst maintaining the essence of the proposed calculations.

When exploring Table 8.3, correlations clearly vary considerably. This indicates that pairs of measurements early in the sequence are less reliable for one another than pairs later in the sequence. Indeed, one can realistically assume that measurements earlier in the sequence are more prone to variability than later on, when subjects are more adapted to the study protocol and/or learning effects have taken place. If we repeat this for each treatment group separately, we consistently have a higher correlation coefficient in the risperidone treated subjects. Note that the ICC between observations from Week 6 and Week 8 ($\rho = 0.70$) is lower as estimated in the previous section ($\rho = 0.85$). In the latter, however, only the subgroup of subjects with Week 6 and 8 was used, and if we apply the same model, accounting for time in this subgroup, then we have $\rho = 0.80$ instead of 0.70.

Jointly accounting for time and treatment produces a different ICC for each treatment group separately and also for each pair of time points. We allowed for interactions in the model. Table 8.4 and Figure 8.3 summarizes the results. Apart from the estimated ICC, also the empirical Pearson (product-moment) correlation coefficients are added. The agreement between both is reasonable, especially when it is taken into account that the ICC does, but the Pearson correlation does not take the effect of covariates into account. After adjusting for time and treatment, the ICC between observations at Week 1 and 8 increased from 0.11 (8.2) to 0.40 in the risperidone group.

## 8.5 Concluding Remarks

We proposed an approximation to calculate correlations from longitudinal data from generalized linear mixed models. Whilst for continuous, interval scaled data, derivation of correlations, such as the ICC of reliability is rather straightforward as derived in (6.2), it is more complex for other types of data. A general formula was derived using the GLMM. This formula could be used for interval, binary or other types of data, such as counts. For our case study, the reliability coefficient was derived for a binary response parameter, using a random-intercepts model. We observed that the correlation was higher between Week 6 and 8 as compared to Week 1 and Week 8. The slightly decreasing correlation, however, from Week 1 and Week 2 to Week 1 and

Table 8.4: *Schizophrenia CGI Data. The first entries represent the overall ICC of reliability (SE) matrix, accounting for treatment, time and their interaction. Standard errors are calculated from the delta method. The second entries are the ordinary Pearson correlation coefficients between the pairs of measurements.*

| **Week** | 2 | 4 | 6 | 8 |
|---|---|---|---|---|
| | | **risperidone** | | |
| 1 | 0.36 (.045) | 0.39 (.044) | 0.40 (.042) | 0.40 (.042) |
| | 0.52 | 0.41 | 0.33 | 0.27 |
| 2 | 1 | 0.62 (.036) | 0.64 (.033) | 0.64 (.032) |
| | | 0.65 | 0.52 | 0.53 |
| 4 | | 1 | 0.69 (.026) | 0.69 (.026) |
| | | | 0.70 | 0.61 |
| 6 | | | 1 | 0.71 (.023) |
| | | | | 0.75 |
| | | **active control** | | |
| 1 | 0.22 (.036) | 0.27 (.038) | 0.31 (.038) | 0.31 (.038) |
| | 0.52 | 0.34 | 0.33 | 0.27 |
| 2 | 1 | 0.42 (.046) | 0.48 (.043) | 0.49 (.041) |
| | | 0.59 | 0.49 | 0.43 |
| 4 | | 1 | 0.57 (.039) | 0.59 (.037) |
| | | | 0.66 | 0.57 |
| 6 | | | 1 | 0.67 (.029) |
| | | | | 0.70 |

Week 8 was not observed in the estimates. It should be noted that the random-effects model does also properly account for missing values due to attrition, provided the missing data are missing at random, which is not the case for the conventional *ad hoc* analyses. In contrast, classical methods such as the kappa statistics, can only include

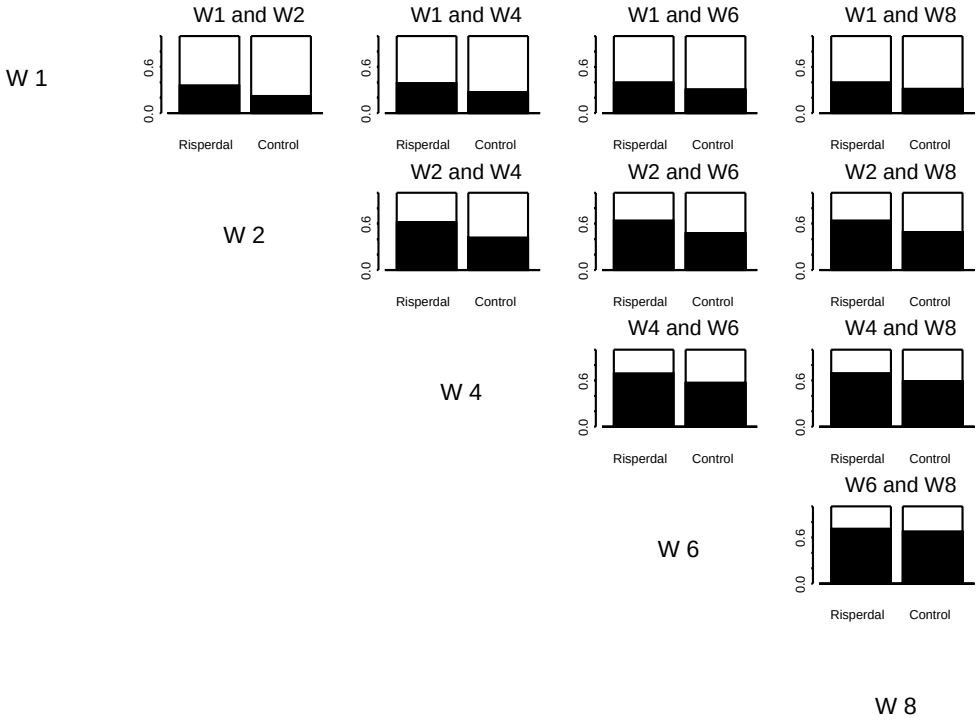


Figure 8.3: *Schizophrenia CGI Data. Estimated ICC using a random-intercept model including time, treatment and their interaction.*

paired observations. Another important advantage of the present method is that it becomes possible to estimate trial-specific or population-specific reliability. This is especially true because, even in studies designed to assess reliability, it is difficult to exclude fluctuations in the true scores and furthermore these studies are often conducted with different populations and in different circumstances. After extracting time and treatment effect and their interaction, clinical trial data can be used to make progress when studying test-retest reliability as a function of time. Indeed, reliability should not be perceived as a fixed quantity but changes with circumstances. Other covariates can be incorporated into the model to study their effect on error variance and on reliability. Modeling other sources of variation, like for example country or rater, is therefore an interesting topic for further research. In psychometric theory, this is referred to as generalizability theory and will be discussed in the next chapter.

Subgroup analyses using a simple model and more versatile models accounting for

time and treatment and their interaction suggested a higher ICC among subjects in the risperidone group than in subjects in the active control group, indicating that responses over time within the same subject were more consistent within the risperidone treatment group than in the active control group. The methodology can be used to derive population or trial-specific ICC of reliability in case of binary data. In particular, it extends the random intercepts model proposed in Chapter 6 and in Vangeneugden *et al.* (2004) to binary data.

The next step is to extend reliability testing to generalizability testing similar as was done in Chapter 7 versus Chapter 6 for interval scaled data, but this time using the GLMM framework.

# Chapter 9

# Generalizibility Estimation in Case of Binary Data

In this Chapter we extend generalizability to non-Gaussian outcomes; in particular, our focus will be on binary data. The results of this chapter are accepted for publication (Vangeneugden *et al.*, 2008a). Even in the univariate case, there are fundamental differences between Gaussian and non-Gaussian outcomes, since the latter usually require non-linear models, also exhibiting important differences in the relationship between mean and variance. Furthermore, repeated binary data are frequently encountered in clinical trials but pose challenges for model formulation. One distinguishes between marginal and random-effects model families and, unlike in the Gaussian situation, there is no easy relationship between both. An example of the marginal family is generalized estimating equations (GEE, Liang and Zeger, 1986), whereas the generalized linear mixed model (GLMM, Breslow and Clayton, 1993) introduced in Section 3.3 is likely the most prominent random-effects model (Molenberghs and Verbeke, 2005). Whereas GEE is convenient and frequently used, it models the marginal regression function, treating the second and higher-order moments as nuisance parameters, which limits its use when the correlation is of scientific interest, e.g., in view of the ICC. The GLMM, on the other hand, has a full likelihood basis, but fails to produce the marginal correlations in an easy fashion, owing to the presence of a non-linear link function, combined with a non-trivial mean-variance re-

lationship, forcing the variance to change with the mean and hence with the regressors (Molenberghs and Verbeke, 2005, Chapter 16). In spite of these considerations, we will show the GLMM provides a viable framework when correlations are of interest, with particular emphasis on the use of generalizability theory.

To fix ideas, let us give an example as to how the observed clinical trial data are typically decomposed, similarly as was done in Section 7.2:

$$Y_{PDT} = h(\mu + b_P + \mu_D + \mu_T + \mu_{DT}) + \varepsilon_{PDT}, \tag{9.1}$$

where $h(.)$ is a known link function. Further, $b_p$ denotes the random effect for patient $p = 1, \ldots, N$, $\mu_D$ the fixed time effect at day $d = 1, \ldots, n_p$, $\mu_T$ the fixed effect of treatment $t = 1, \ldots, T$, $\mu_{DT}$ their interaction. Finally, $\varepsilon_{PDT}$ refers to the residual error, the distribution of which is chosen in accordance with the outcome type. For example, when $Y_{PDT}$ is a binary indicator, it is customary to adopt for $h(\cdot)$ the antilogit function and for $\varepsilon_{PDT}$ the Bernoulli distribution with success probability $h(\mu + b_P + \mu_D + \mu_T + \mu_{DT})$. When other design levels are present, e.g., country or center, Model (9.1) can be extended in a straightforward fashion and various instances will be given in subsequent sections.

## 9.1 Correlation Between Two Observations Using the GLMM Framework

We will now derive a general formula for the correlation between two observations, within the GLMM framework. In the spirit of (9.1), and with notation consistent with Section 3.3, we can write the general model as:

$$Y_{PDT} = \mu_{PDT} + \varepsilon_{PDT}, \tag{9.2}$$

where

$$\mu_{PDT} = \mu_{PDT}(\eta_{PDT}) = h(\boldsymbol{x}'_{PDT}\boldsymbol{\beta} + \boldsymbol{z}'_{PDT}\boldsymbol{b}_{PDT}). \tag{9.3}$$

We group the errors $\varepsilon_{PDT}$ into a vector $\boldsymbol{\varepsilon}_P$, with variance-covariance matrix $\Sigma_P$. Further, consistent with earlier notation, $\boldsymbol{x}'_{PDT}$ and $\boldsymbol{z}'_{PDT}$ are vectors of fixed-effects and random-effects covariates, respectively; $\boldsymbol{\beta}$ is a vector of fixed-effects parameters and $\boldsymbol{b}_{PDT}$ is a vector of random effects, assumed to be zero-mean normally distributed with variance-covariance matrix $H$.

It is useful in what follows to decompose $\Sigma_P$ as:

$$\Sigma_P = \Phi^{\frac{1}{2}} A_P^{\frac{1}{2}} R_P A_P^{\frac{1}{2}} \Phi^{\frac{1}{2}},$$

where $\Phi$ is a diagonal matrix with the overdispersion parameters along the diagonal. In case there are no overdispersion parameters, $\Phi$ is set equal to the identity matrix. Further, $R_P$ is the correlation matrix, and $A_P$ is a diagonal matrix containing the variances following from the generalized linear model specification of $Y_{PDT}$ given the random effects $\boldsymbol{b}_{PDT} = \mathbf{0}$, i.e., with diagonal elements $v(\mu_{PDT}|\boldsymbol{b}_{PDT} = 0)$.

Model (9.3) allows for a variety of distributions for the outcome variable and a wide range of link functions, while the modeler has the freedom to include or leave out serial correlation. To calculate correlation $\text{Corr}(Y_{PDT}, Y_{P'D'T'})$, we repeat the general derivation of a general expression for the variance as in Section 8.1:

$$\begin{aligned} \text{Var}(Y_{PDT}) &= \text{Var}(\mu_{PDT} + \varepsilon_{PDT}) \\ &= \text{Var}(\mu_{PDT}) + \text{Var}(\varepsilon_{PDT}) + 2\text{Cov}(\mu_{PDT}, \varepsilon_{PDT}). \end{aligned} \tag{9.4}$$

It can be shown that (Molenberghs and Verbeke, 2005)

$$\begin{aligned} &\text{Cov}(\mu_{PDT}, \varepsilon_{PDT}) \\ &\quad = \text{Cov}[\text{E}(\mu_{PDT}|\boldsymbol{b}_{PDT}), \text{E}(\varepsilon_{PDT}|\boldsymbol{b}_{PDT})] + \text{E}[\text{Cov}(\mu_{PDT}, \varepsilon_{PDT}|\boldsymbol{b}_{PDT})] = 0, \end{aligned}$$

since the first term is zero and the second term equals

$$\text{E}\{\text{E}[\mu_{PDT} - \text{E}(\mu_{PDT})](\varepsilon_{PDT})|\boldsymbol{b}_{PDT}\} = 0$$

as $\mu_{PDT}$ is constant when conditioning on $\boldsymbol{b}_{PDT}$. For the first term in (9.4), we have:

$$\begin{aligned} \text{Var}(\mu_{PDT}) &= \text{Var}[\mu_{PDT}(\eta_{PDT})] = \text{Var}[\mu_{PDT}(\boldsymbol{x}'_{PDT}\boldsymbol{\beta} + \boldsymbol{z}'_{PDT}\boldsymbol{b}_{PDT})] \\ &\cong \left( \frac{\partial \mu_{PDT}}{\partial \boldsymbol{b}_{PDT}} \Big|_{\boldsymbol{b}_{PDT}=0} \right) \text{Var}(\boldsymbol{b}_{PDT}) \left( \frac{\partial \mu_{PDT}}{\partial \boldsymbol{b}_{PDT}} \Big|_{\boldsymbol{b}_{PDT}=0} \right)' \\ &\cong \left( \frac{\partial \mu_{PDT}}{\partial \eta_{PDT}} \frac{\partial \eta_{PDT}}{\partial \boldsymbol{b}_{PDT}} \Big|_{\boldsymbol{b}_{PDT}=0} \right) H \left( \frac{\partial \mu_{PDT}}{\partial \eta_{PDT}} \frac{\partial \eta_{PDT}}{\partial b_{PDT}} \Big|_{b_{PDT}=0} \right)' \\ &\cong \Delta_{PDT} \boldsymbol{z}'_{PDT} H \boldsymbol{z}_{PDT} \Delta'_{PDT}, \end{aligned}$$

where $\Delta_{PDT} = \frac{\partial \mu_{PDT}}{\partial \eta_{PDT}} \Big|_{b_{PDT}=0}$. Note that the above derivation is based on the delta method (Welsh 1996). For the second term in (9.4), we have:

$$\begin{aligned} \mathrm{Var}(\varepsilon_{PDT}) &= \mathrm{Var}[\mathrm{E}(\varepsilon_{PDT}|\boldsymbol{b}_{PDT})] + \mathrm{E}[\mathrm{Var}(\varepsilon_{PDT}|\boldsymbol{b}_{PDT})] \\ &= \mathrm{E}[\mathrm{Var}(\varepsilon_{PDT}|\boldsymbol{b}_{PDT})] = \left(\Phi^{\frac{1}{2}}\Sigma\Phi^{\frac{1}{2}}\right)_{PDT}, \end{aligned}$$

If the canonical link is used, we have $A_P = \Delta_P$ and then (9.4) becomes

$$\mathrm{Var}(\boldsymbol{Y}_P) \cong \Delta_P Z_P H Z_P' \Delta_P' + \Phi^{\frac{1}{2}} \Delta_P^{\frac{1}{2}} R_P \Delta_P^{\frac{1}{2}} \Phi^{\frac{1}{2}}. \tag{9.5}$$

To determine $\mathrm{Corr}(Y_{PDT}, Y_{P'D'T'})$, we still need to calculate $\mathrm{Cov}(Y_{PDT}, Y_{P'D'T'})$. Similar to the above, we have that $\mathrm{Cov}(\mu_{PDT}, \varepsilon_{P'D'T'}) = \mathrm{Cov}(\varepsilon_{PDT}, \mu_{P'D'T'}) = 0$. Therefore, we only need to derive $\mathrm{Cov}(\mu_{PDT}, \mu_{P'D'T'})$:

$$\begin{aligned} &\mathrm{Cov}(Y_{PDT}, Y_{P'D'T'}) \\ &= \mathrm{Cov}(\mu_{PDT}, \mu_{P'D'T'}) \\ &= \mathrm{Cov}[\mu_{PDT}(\boldsymbol{x}_{PDT}'\boldsymbol{\beta} + \boldsymbol{z}_{PDT}'\boldsymbol{b}_{PDT}), \mu_{P'D'T'}(\boldsymbol{x}_{P'D'T'}'\boldsymbol{\beta} + \boldsymbol{z}_{P'D'T'}'\boldsymbol{b}_{P'D'T'})] \\ &\cong \left(\frac{\partial \mu_{PDT}}{\partial \boldsymbol{b}_{PDT}} \Big|_{\boldsymbol{b}_{PDT}=0}\right) \mathrm{Cov}(\boldsymbol{b}_{PDT}, \boldsymbol{b}_{P'D'T'}) \left(\frac{\partial \mu_{P'D'T'}}{\partial \boldsymbol{b}_{P'D'T'}} \Big|_{\boldsymbol{b}_{P'D'T'}=0}\right)' \\ &\cong \left(\frac{\partial \mu_{PDT}}{\partial \eta_{PDT}} \frac{\partial \eta_{PDT}}{\partial \boldsymbol{b}_{PDT}} \Big|_{\boldsymbol{b}_{PDT}=0}\right) \mathrm{Cov}(\boldsymbol{b}_{PDT}, \boldsymbol{b}_{P'D'T'}) \\ &\quad \times \left(\frac{\partial \mu_{P'D'T'}}{\partial \eta_{P'D'T'}} \frac{\partial \eta_{P'D'T'}}{\partial \boldsymbol{b}_{P'D'T'}} \Big|_{\boldsymbol{b}_{P'D'T'}=0}\right)' \\ &\cong \Delta_{PDT} \boldsymbol{z}_{PDT}' \mathrm{Cov}(\boldsymbol{b}_{PDT}, \boldsymbol{b}_{P'D'T'}) \boldsymbol{z}_{P'D'T'} \Delta_{P'D'T'}'. \end{aligned} \tag{9.6}$$

The covariances $\mathrm{Cov}(b_{PDT}, b_{P'D'T'})$ depend on which of the random effects are common when correlating $Y_{PDT}$ and $Y_{P'D'T'}$. Using (9.5) and (9.6), we can calculate the correlation for any given situation, for any give GLMM. In the next section, we will derive the correlation for the case of binary data with random effects and without serial correlation. Note that, in the special case of Gaussian outcomes, (9.5) simply reduces to $\mathrm{Var}(\boldsymbol{Y}_P) = Z_P H Z_P' + R_P$.

The above calculations are general, in the sense that the variances in (9.5) and covariances in (9.6) allow for the flexible calculation of correlation coefficients, with (1) certain facets the same or different; (2) certain facets fixed (conditioned upon) or random; (3) correction for the presence of such fixed effects as treatment, time, country, baseline value, etc.; (4) for normally distributed outcomes based on linear mixed models or for binary data, count data, and other non-Gaussian data, using generalized linear mixed models. The price to pay is twofold. First, expressions (9.5) and (9.6) are approximate, except in the normal case and (2), related to the previous point, these expressions do not have the intuitive variance-component structure, or even 'averaging' structure, of classical reliability and generalizability coefficients. However, all classical expressions follow as special cases. In this sense, our framework allows for the calculation of conventional reliability and generalizability coefficients, their extensions to the non-normal case based on data from clinical trials or other data with measurements that are *a priori* not parallel, and even correlation coefficients that do not have a generalizability interpretation, but may be useful for other purposes.

## 9.2 Data Analysis

Let us now apply the concepts of reliability and generalizability to the pooled data described in Section 2.1. We will investigate the impact of 'country' on measurement error. Note that country can be seen as either a facet or an object of measurement. The generality of our approach allows for both views. Evidently, 'country' is of interest for this particular study, but the reader can easily substitute it with other variables, subject to his/her study of interest.

To illustrate the methods and underscore generality, we will consider country in five different roles, the analysis of which is all within reach by way of the modeling ideas developed in this manuscript. First, we will assess the overall reliability for a dichotomized version of CGI response, ignoring country effects. Second, country effects will be extracted by including country as a fixed effect into the model. Third, we will investigate the impact of country on reliability through application of the same model to each country separately. Fourth, we will also study the impact of a single country on overall reliability by leave-one-out ideas, i.e., by omitting one country at a time. Fifth, we will assess the overall impact of country via generalizability theory. Note that 'country' did not feature explicitly in previous Section. However,

the methodology is general and the facets generic. They can be replaced with those relevant in a particular case study.

### 9.2.1 Overall Reliability of CGI

First, we apply a simple random-intercept model, combined with fixed effects for treatment, time and their interaction. Hence, country does play a role in this analysis. With the logit link, (9.2) becomes:

$$Y_{PDT} = \frac{\exp(\mu + b_P + \mu_D + \mu_T + \mu_{DT})}{1 + \exp(\mu + b_P + \mu_D + \mu_T + \mu_{DT})} + \varepsilon_{PDT}, \tag{9.7}$$

where $\mu_D$, $\mu_T$, and $\mu_{DT}$ denote the fixed effects for day, treatment, and their interaction, respectively, and $b_P$ represents the random patient effect.

The overall correlation of observations within the same subject, on the same treatment, but on different time points, and conditioning on treatment and time points, can be expressed as $\mathrm{Corr}(Y_{PDT}, Y_{PD'T} \mid T, D, D')$. In this model, we have $Z = \mathbf{1}$ and $H = \sigma_P^2$, a scalar representing the variance of the random intercept, and since (9.7) does not include serial correlation we have that $R_P = I$. It is therefore possible to show that the variance covariance matrix (9.5) reduces to

$$\mathrm{Var}(Y_P) \cong \Delta_P(\sigma_P^2 J)\Delta_P' + \Phi\Delta_P = \Delta_P(\sigma_P^2 J + \Phi\Delta_P^{-1})\Delta_P',$$

where $\boldsymbol{J}$ is a rectangular matrix of ones. Furthermore, $\Delta_P$ is a diagonal matrix with $V_{PDT}(0)$ as diagonal elements, where the variance function $V_{PDT}(0) = \mu_{PDT}\,\big|_{b_{PDT}=0}$ $(1 - \mu_{PDT}\,\big|_{b_{PDT}=0})$, and therefore we have

$$\begin{aligned} &\mathrm{Var}(Y_{PDT}) \\ &\quad\cong\quad \mathrm{diag}(\mathrm{V_{PDT}}(0))[\sigma_{\mathrm{P}}^2\mathrm{J} + \Phi\mathrm{diag}(\mathrm{V_{PDT}}(0))^{-1}]\mathrm{diag}(\mathrm{V_{PDT}}(0)), \qquad (9.8) \\ &\mathrm{Cov}(Y_{PDT}, Y_{PD'T}) \\ &\quad\cong\quad \mathrm{diag}(\mathrm{V_{PDT}}(0))[\sigma_{\mathrm{P}}^2\mathrm{J}]\mathrm{diag}(\mathrm{V_{PD'T}}(0)). \qquad (9.9) \end{aligned}$$

Based on (9.8) and (9.9), we can determine a first-order approximation of the marginal correlation between time point $d$ and $d'$, which is the intraclass correlation coefficient of reliability:

$$\rho = \mathrm{Corr}(Y_{PDT}, Y_{PD'T}) = \frac{\sigma_1^2\sqrt{V_{PDT}(0)V_{PD'T}(0)}}{\sqrt{[\Phi_{PDT} + V_{PDT}(0)\sigma_2^2]\cdot[\Phi_{PD'T} + V_{PD'T}(0)\sigma_2^2]}}, \tag{9.10}$$

where $\sigma_1^2$ represents the covariance between the random effects and $\sigma_2^2$ is the variance resulting from the random effects. In this model, $\sigma_1^2 = \sigma_2^2 = \sigma_P^2$ since all other covariates are fixed effects. The delta method can be usefully applied to estimate the standard error:

$$\frac{\partial \rho}{\partial(\boldsymbol{\beta}, \boldsymbol{\lambda})} = \left(\frac{\partial(\boldsymbol{\eta}, \boldsymbol{\sigma}^2)}{\partial(\boldsymbol{\beta}, \boldsymbol{\lambda})}\right)\left(\frac{\partial(V_{PDT}(0), V_{PD'T}(0), \sigma_1^2, \sigma_2^2, \phi)}{\partial(\boldsymbol{\eta}, \boldsymbol{\sigma}^2)}\right) \times \left(\frac{\partial \rho}{\partial(V_{PDT}(0), V_{PD'T}(0), \sigma_1^2, \sigma_2^2, \phi)}\right). \tag{9.11}$$

Explicit expressions for the various components follow from straightforward linear algebra, as sketched in Appendix A. The SAS V9.1 procedure GLIMMIX was used to estimate $\Phi$, $\sigma_P^2$, and $V_{PDT}$. Details on the SAS implementations are provided in Appendix B. The reader interested in more ample details on the SAS implementations and output, can obtain such from the authors, upon simple request. Table 9.1(a) summarizes the results.

In case of continuous data, a single-measure overall intraclass correlation coefficient reliability would have been obtained (Chapter 7, Vangeneugden *et al.*, 2005). Here, for the binary data case, a separate intraclass coefficient of reliability is produced for each treatment group and each time point. From Table 9.1(a), we observe that the correlation is somewhat higher in the risperidone arm and that the correlation between week 1 and other time points is lower than the correlation between any two other time points that do not involve week 1. This non-constancy is, of course, not particular to this example but results from the non-Gaussian nature of the outcome.

### 9.2.2 Reliability of CGI Response Adjusting for Country

In Section 9.2.1, only treatment, time, and their interaction were included. Now, we will include countries as fixed effects, which will result in intraclass coefficients of reliability per treatment, time point, and country combination. Hence, country-specific analyses result. We will not present all coefficients but merely present the coefficients for one country, the U.S.A., in Table 9.1(b). Additionally, we list the ICC of reliability between weeks 6 and 8 in the risperidone group for all countries in Table 9.2. The results for the U.S.A. are consistent with the overall results, and when we investigate the correlation between weeks 6 and 8 in the risperidone group, we observe from column 3 in Table 9.2 that the ICC is rather stable across countries, the lowest correlation being for Austria (0.65, SE 0.09) and the highest for the U.S.A.,

Table 9.1: *Schizophrenia CGI Data. ICC matrices (SE), accounting for treatment, time and their interaction. Standard errors are calculated from the delta method. Five different situations are reported.*

| | **risperidone** | | | | **active control** | | | |
|---|---|---|---|---|---|---|---|---|
| **Week** | 2 | 4 | 6 | 8 | 2 | 4 | 6 | 8 |
| (a) Overall | | | | | | | | |
| 1 | .52(.04) | .55(.04) | .55(.04) | .55(.04) | .42(.04) | .47(.04) | .50(.04) | .50(.04) |
| 2 | 1 | .74(.02) | .74(.02) | .74(.02) | 1 | .61(.04) | .65(.03) | .66(.03) |
| 4 | | 1 | .78(.02) | .78(.02) | | 1 | .72(.03) | .73(.02) |
| 6 | | | 1 | .79(.01) | | | 1 | .78(.02) |
| (b) By country: U.S.A. | | | | | | | | |
| 1 | .52(.06) | .54(.06) | .54(.05) | .54(.05) | .38(.07) | .42(.07) | .46(.06) | .46(.06) |
| 2 | 1 | .73(.03) | .74(.03) | .74(.02) | 1 | .57(.06) | .62(.05) | .63(.05) |
| 4 | | 1 | .77(.02) | .77(.02) | | 1 | .69(.04) | .70(.04) |
| 6 | | | 1 | .78(.02) | | | 1 | .76(.02) |
| (c) Country as random effect: U.S.A. | | | | | | | | |
| 1 | .53(.05) | .55(.05) | .56(.05) | .56(.05) | .40(.06) | .45(.06) | .48(.05) | .48(.05) |
| 2 | 1 | .74(.03) | .75(.02) | .75(.02) | 1 | .59(.05) | .64(.04) | .65(.04) |
| 4 | | 1 | .78(.02) | .78(.02) | | 1 | .71(.03) | .72(.03) |
| 6 | | | 1 | .79(.02) | | | 1 | .77(.02) |
| (d) Generalized across countries: U.S.A. | | | | | | | | |
| 1 | .49(.05) | .51(.05) | .51(.05) | .51(.04) | .37(.05) | .41(.05) | .44(.05) | .45(.05) |
| 2 | 1 | .68(.03) | .69(.03) | .69(.03) | 1 | .55(.05) | .59(.04) | .60(.04) |
| 4 | | 1 | .72(.03) | .72(.03) | | 1 | .65(.04) | .66(.03) |
| 6 | | | 1 | .72(.03) | | | 1 | .71(.03) |
| (e) Generalized across baseline negative symptoms | | | | | | | | |
| 1 | .37(.13) | .38(.13) | .39(.13) | .39(.13) | .29(.10) | .32(.11) | .35(.12) | .35(.12) |
| 2 | 1 | .51(.18) | .52(.18) | .52(.18) | 1 | .43(.15) | .46(.16) | .46(.16) |
| 4 | | 1 | .54(.18) | .54(.18) | | 1 | .50(.17) | .51(.17) |
| 6 | | | 1 | .55(.19) | | | 1 | .54(.18) |

Sweden, and Spain (0.78, SE 0.02).

Table 9.2: *Schizophrenia CGI Data. Reliability by country and impact of country on overall reliability table. ICC $\rho$ (SE) between week 6 and 8 in risperidone, with (1) country as fixed effect, (2) country-specific analyzes, and (3) a given country omitted. (NA: not available by lack of data.)*

| Country | Number of patients | Country as fixed effect | By country | Omitting a given country |
|---|---|---|---|---|
| Argentina | 31 | 0.76 (0.04) | NA | 0.78 (0.02) |
| Austria | 29 | 0.65 (0.09) | 0.02 (0.04) | 0.78 (0.01) |
| Belgium | 26 | 0.76 (0.04) | NA | 0.78 (0.01) |
| Brazil | 44 | 0.73 (0.05) | 0.54 (0.14) | 0.79 (0.01) |
| Canada | 44 | 0.77 (0.02) | 0.76 (0.10) | 0.79 (0.01) |
| Denmark | 47 | 0.77 (0.02) | 0.65 (0.09) | 0.80 (0.01) |
| Spain | 32 | 0.78 (0.02) | 0.88 (0.07) | 0.79 (0.01) |
| Finland | 71 | 0.66 (0.07) | NA | 0.79 (0.01) |
| France | 92 | 0.77 (0.02) | 0.40 (0.11) | 0.81 (0.01) |
| Great Britain | 21 | 0.77 (0.03) | 0.91 (0.05) | 0.78 (0.01) |
| Germany | 25 | 0.73 (0.06) | NA | 0.78 (0.01) |
| Italy | 39 | 0.70 (0.07) | NA | 0.77 (0.02) |
| Mexico | 36 | 0.76 (0.03) | 0.92 (0.06) | 0.78 (0.02) |
| Netherlands | 17 | 0.74 (0.06) | 0.71 (0.37) | 0.78 (0.01) |
| Norway | 37 | 0.71 (0.06) | 0.91 (0.04) | 0.78 (0.01) |
| South Africa | 79 | 0.71 (0.05) | 0.80 (0.09) | 0.78 (0.02) |
| Sweden | 30 | 0.78 (0.02) | 0.94 (0.03) | 0.78 (0.01) |
| U.S.A. | 122 | 0.78 (0.02) | 0.75 (0.04) | 0.79 (0.02) |

### 9.2.3 Reliability of CGI by Country and Impact on Overall Reliability by Leaving Out a Country

When we apply the model to each country separately, we observe that the model did not always converge and estimates were less stable, especially and not surprisingly, in countries with few patients. Patients included in Finland had data up to week 6 only (Hoyberg *et al.*, 1993). The results are summarized in the third column of Table 9.2. A different way of investigating impact of country on reliability is by leaving out one country at a time. This is slightly less conventional from a classical generalizability

standpoint, but it is a useful analysis to assess how much a given country can weigh in on the analyses. If the overall reliability increases, this would provide evidence for a poor reliability in this specific country. The results are summarized in the fifth column of Table 9.2. Note that the impact was low for all countries, again suggesting that reliability is relatively consistent across countries.

### 9.2.4 Estimating Impact of Country via Generalizability Theory

Subgroup analysis by country as shown in the previous two sections can be enlightening. Now, we want to quantify their effect on measurement error and calculate a generalizability coefficient, thereby generalizing results across countries. We will add a random effect for country to the previous model, so that we have a model with time, treatment, and their interaction as fixed effects, and further country, indexed by $c$, and patient as random effects:

$$Y_{PDTC} = \frac{\exp(\mu + b_P + \mu_D + \mu_T + \mu_{DT} + b_c)}{1 + \exp(\mu + b_P + \mu_D + \mu_T + \mu_{DT} + b_c)} + \varepsilon_{PDTC}. \tag{9.12}$$

From (9.12) we can calculate the overall test-retest reliability coefficient as in Section 9.2.1, but this time accounting for country as a random effect instead of extracting it as a fixed effect. Then, $\sigma_1^2 = \sigma_2^2 = \sigma_P^2 + \sigma_C^2$ in (9.10). Table 9.1(c) shows that the results are consistent with the overall reliability coefficients.

This test-retest reliability coefficient for any given country and time point follows directly from analyzing the clinical trial, similar to generalizability coefficients that are computed after design and analysis of a G-study. In the spirit of D-studies, we can also generalize across countries. Indeed, although patients are nested within country in a clinical-trial setting, we assume, by way of a thought experiment, that patients could switch from one country to another, with the aim to evaluate the impact of country. We then have that $\sigma_1^2 = \sigma_P^2$ and $\sigma_2^2 = \sigma_P^2 + \sigma_C^2$, needed to calculate $\text{Corr}(Y_{PDTC}, Y_{PD'TC'})$ as in (9.10). Table 9.1(d) provides the ensuing ICC coefficients.

Thus, generalizing across time points and countries, or taking account of impact of variance of country, reduces the overall test-retest reliability approximately by 5%: for risperidone the decrease in reliability amounted to between 4–7% and for active control this was between 3–6%. In this situation, the price for setting up an international trial instead of a single country is rather small. This insight is relevant

and underscores the usefulness of the thought experiment. While, again, the 'country' aspect will be less relevant, or even irrelevant, to the reader's own study, our results indicate that it is possible to study the impact of generalizing over a given variable.

Evidently, the methodology can easily be extended to more complex situations including, for example, serial correlation or random time effects but also additional variables, such as, for example, age and sex of the patient.

### 9.2.5 Estimating Impact of Baseline PANSS Negative Subtotal on Reliability of CGI Response

In the computations reported above, a relatively high generalizability coefficient suggested that country does not have an important impact on the test-retest reliability and on measurement error. We now investigate the impact of baseline PANSS Negative subtotal on measurement error. We included a random intercept for baseline PANSS Negative subtotal instead of country in Model (9.12). Subsequently, we derived the variance components and calculated the generalizability coefficient for baseline PANSS Negative subtotal, similar to how it was done for country. In this analysis, the reduction in generalizability coefficient was more substantial: in the risperidone group between week 6 and 8, we have that the ICC reduces from 0.55 (SE 0.13) to 0.39 (SE 0.13) when generalizing across baseline negative subtotal. Full details are given in Table 9.1(e). This indicates that baseline PANSS Negative subtotal reduces the test-retest reliability. A clinical explanation for this phenomenon could be that patients with a higher deficit in negative symptoms at baseline, such as poverty of speech, apathy, or emotional withdrawal, are more difficult to evaluate, resulting in higher measurement error and lower test-retest reliability. A practical conclusion would be that additional training is needed for professionals having to rate patients with a high baseline negative subtotal or, even more invasive, in the recommendation to use a different scale in this type of patients. Such conclusions usefully illustrate how the methodology can be used, not only to assess the qualitative level of generalizability, but also how such results can impact the design of future studies.

## 9.3 Concluding Remarks

In this paper, we have extended classical reliability measures and associated estimation procedures in four important ways. First, fully longitudinal data can be used,

rather than paired measurements. Second, clinical trial data can be employed or, more generally data from other studies not expressly designed for the investigation of reliability, through adopting a modeling framework, obviating the need for parallel measurements. Third, the broad generalizability theory framework is invoked, encompassing the various classical reliability versions, such as inter-rater and test-retest reliability, and allowing for the study of such important factors' impact as day of measurement, rater, country, investigator, etc. Fourth, all calculations are conducted within the generalized linear mixed model paradigm, allowing one not only to accommodate all aforementioned aspects, but also to deal with Gaussian and non-Gaussian data alike. Specific emphasis was put on binary outcomes, but analogous computations for nominal, ordinal, or count data can be done as well. Unlike in the Gaussian case, the reliability and generalizability coefficients depend on the days, raters, countries, or whatever levels studied. This is due to the mean-variance link and the nonlinear nature of the model.

Of course, our calculations are based on a first-order approximation, the accuracy of which could be a cause of concern. Vangeneugden *et al.*, (2008b) have studied this issue and their results are surprisingly encouraging.

We would like to emphasize that we have focused on generalizability, with reliability as a special case. This implies that we have been less concerned with agreement. While the latter concept is also very important, it falls outside the scope of the current work.

This work was motivated by and applied to data from multi-country trial data collected in patients with chronic schizophrenia. Using the generalizability framework, we were able to establish that the reliability measures are rather stable across countries, and no single country has an undue effect on the overall reliability. Country-specific reliabilities varied in a usefully narrow range.

An important conclusion, never reached before, is that the price to pay for a multi-country study, rather than a single-country one, is a mere 5% in test-retest reliability. The ability to conduct multi-country studies is important in view of the availability of a larger pool of available patients, thereby reducing the length of the accrual period and/or increasing the sample size, and hence power.

# Chapter 10

# Marginal Correlation in Case of Count Data

In Chapter 8 we showed that the derivation of correlations based on GLMM is generally feasible if one is prepared to accept a Taylor-series based approximation. The results are general, not only across data types, such as continuous outcomes, binary or ordinal outcomes, and counts, but it applies also to multivariate repeated measures, where more than one sequence per subject is measured repeatedly, even with different data types for the various sequences.

One obvious though important case where the approximate results of Chapter 8 become exact (6.2) is for normally distributed outcomes, since then the GLMM reduces to the LMM. In this chapter, we will deal with another, somewhat less broadly studied special case: count data. To this effect, we will start from the modeling framework proposed by Molenberghs, Verbeke, and Demétrio (2008; henceforth abbreviated as MVD). These authors presented a model for longitudinal or otherwise hierarchical count data that simultaneously incorporates normal random effects in the linear predictor, as in any GLMM, as well as conventional overdispersion parameters. Overdispersion arises when the mean-variance relationship stemming from the posited generalized linear model (McCullagh and Nelder, 1989) is too restrictive. As such and especially with count data, it is a phenomenon that can occur even with univariate, cross-sectional data. One convenient way to incorporate overdispersion is through

gamma random effects, giving rise in the univariate case to the so-called negative binomial model (Breslow, 1984). Thus, whereas the normal random effects capture correlation between repeated measures and a portion of the overdispersion, the additional gamma random effects allow a more flexible incorporation of overdispersion, and hence potentially a better fitting model. Thus, the proposal by MVD, referred to as the *combined* model, generalizes at once the GLMM and negative-binomial models and therefore, *a fortiori*, the univariate Poisson model.

MVD also showed that, unlike the general GLMM case, the Poisson case allows for closed-form expressions for the mean vector, variance-covariance matrix, and even for the full joint probability vectors. This is true for the combined model and hence also for all of the aforementioned special cases, providing the opportunity to derive closed-form expressions for the within-unit correlation functions since it implies that there is no practical need for the Taylor-series based approximations mentioned earlier. Nevertheless, to provide additional insight, the closed-form expressions will be contrasted with their Taylor-series-based counterparts of various orders. The combined model and its submodels will be fitted to repeated epileptic-seizures data, which are known to exhibit considerable amounts of overdispersion, in addition to within-subject correlation.

## 10.1 Closed-form Derivation of the Correlation Function

As stated in the Section 3.4, MVD derived closed-form mean (3.20) and variance (3.21) for the combined model in the general, longitudinal context. These produce, as special cases, expressions for the negative binomial and the Poisson-normal models. Variance-covariance expression (3.21) renders straightforward the derivation of a closed-form correlation function expression.

In the general case of the combined model for longitudinal data with arbitrary fixed- and random-effects structures, the variance, deriving from (3.21) equals:

$$\begin{aligned} \mathrm{Var}(Y_{ij}) &= \phi_{ij} e^{\boldsymbol{x}'_{ij}\boldsymbol{\beta}+\frac{1}{2}\boldsymbol{z}_{ij}D\boldsymbol{z}'_{ij}} + \sigma_{i,jj} e^{2\boldsymbol{x}'_{ij}\boldsymbol{\beta}+2\boldsymbol{z}_{ij}D\boldsymbol{z}'_{ij}} \\ &+ \phi_{ij}^2 e^{2\boldsymbol{x}'_{ij}\boldsymbol{\beta}+\boldsymbol{z}_{ij}D\boldsymbol{z}'_{ij}} \left( e^{\boldsymbol{z}_{ij}D\boldsymbol{z}'_{ij}} - 1 \right). \end{aligned} \tag{10.1}$$

Likewise, the covariance can be written as:

$$\begin{aligned}\text{Cov}(Y_{ij},Y_{ik}) &= \phi_{ij}e^{\boldsymbol{x}'_{ij}\boldsymbol{\beta}+\frac{1}{2}\boldsymbol{z}_{ij}D\boldsymbol{z}'_{ij}}\left[\left(\frac{\sigma_{i,jk}}{\phi_{ij}\phi_{ik}}+1\right)e^{\frac{1}{2}(\boldsymbol{z}_{ij}D\boldsymbol{z}'_{ik}+\boldsymbol{z}_{ik}D\boldsymbol{z}'_{ij})}-1\right]\times \\ &\quad\times\phi_{ik}e^{\boldsymbol{x}'_{ik}\boldsymbol{\beta}+\frac{1}{2}\boldsymbol{z}_{ik}D\boldsymbol{z}'_{ik}}. \qquad (10.2)\end{aligned}$$

The correlation between two measurements $j$ and $k$ on the same experimental unit $i$ then is:

$$\text{Corr}(Y_{ij},Y_{ik})=\frac{\text{Cov}(Y_{ij},Y_{ik})}{\sqrt{\text{Var}(Y_{ij})\cdot\text{Var}(Y_{ik})}}. \qquad (10.3)$$

Because of its generality, it is hard to simplify (10.3), unless in specific cases. Of course, (10.1) and (10.2) simplify when we zoom in on the Poisson-normal case:

$$\text{Var}(Y_{ij}) = e^{\boldsymbol{x}'_{ij}\boldsymbol{\beta}+\frac{1}{2}\boldsymbol{z}_{ij}D\boldsymbol{z}'_{ij}}+e^{2\boldsymbol{x}'_{ij}\boldsymbol{\beta}+\boldsymbol{z}_{ij}D\boldsymbol{z}'_{ij}}\left(e^{\boldsymbol{z}_{ij}D\boldsymbol{z}'_{ij}}-1\right), \qquad (10.4)$$

$$\text{Cov}(Y_{ij},Y_{ik}) = e^{\boldsymbol{x}'_{ij}\boldsymbol{\beta}+\frac{1}{2}\boldsymbol{z}_{ij}D\boldsymbol{z}'_{ij}}\left(e^{\boldsymbol{z}_{ij}D\boldsymbol{z}'_{ik}}-1\right)e^{\boldsymbol{x}'_{ik}\boldsymbol{\beta}+\frac{1}{2}\boldsymbol{z}_{ik}D\boldsymbol{z}'_{ik}}. \qquad (10.5)$$

Likewise, they do for the negative-binomial case. There are two ways to approach this case. First, one can absorb the fixed effects into $\theta_{ij}$ and hence the conventional expression follows. However, let us opt for the second route, where (3.16) is maintained, only with $\boldsymbol{b_i}$ removed or, equivalently, $D=0$. Equivalently, we can start from (10.1) and (10.2), of course. At any rate, the variance and covariance can be written as:

$$\begin{aligned}\text{Var}(Y_{ij}) &= \phi_{ij}\mu_{ij}+\sigma_{i,jj}\mu_{ij}^2, \qquad (10.6)\\ \text{Cov}(Y_{ij},Y_{ik}) &= \mu_{ij}\mu_{ik}\sigma_{i,jk}. \qquad (10.7)\end{aligned}$$

Here, $\mu_{ij}=\exp(\boldsymbol{x}'_{ij}\boldsymbol{\beta})$. Evidently, (10.3) can be written in a convenient form as:

$$\text{Corr}(Y_{ij},Y_{ik})=\frac{\mu_{ij}\mu_{ik}\sigma_{i,jk}}{\sqrt{(\phi_{ij}\mu_{ij}+\mu_{ij}^2\sigma_{i,jj})\cdot(\phi_{ik}\mu_{ik}+\mu_{ik}^2\sigma_{i,kk})}}. \qquad (10.8)$$

Note that, when the fixed effects are subsumed into $\theta_{ij}$, the variance is written as $\mu_{ij}+\sigma_{i,jj}$ with covariance $\sigma_{i,jk}$, in which case the alternative, simple and conventional form for the correlation results:

$$\text{Corr}(Y_{ij},Y_{ik})=\frac{\sigma_{i,jk}}{\sqrt{(\mu_{ij}+\sigma_{i,jj})\cdot(\mu_{ik}+\sigma_{i,kk})}}. \qquad (10.9)$$

Additional insight can be obtained for special but important cases resulting from simplifying the mean and variance structures of the models. Let us show this for an

exchangeable structure, where $\boldsymbol{x}'_{ij}\boldsymbol{\beta} = \beta$, $\boldsymbol{z}_{ij} = 1$, and $D = d$. It is then sensible to also set $\phi_{ij} = \phi$, $\sigma_{i,jj} = \sigma^2$, and $\sigma_{i,jk} = \sigma^2\tau$, for $j \neq k$. For the combined model, we obtain the following simplifications of (10.1) and (10.2):

$$\begin{aligned}\text{Var}(Y_{ij}) &= \phi e^{\beta+\frac{1}{2}d} + \phi^2\left(e^{\beta+\frac{1}{2}d}\right)^2 \cdot \left[\left(\frac{\sigma^2}{\phi^2}+1\right)e^d - 1\right], \\ \text{Cov}(Y_{ij}, Y_{ik}) &= \phi^2\left(e^{\beta+\frac{1}{2}d}\right)^2 \cdot \left[\left(\frac{\tau\sigma^2}{\phi^2}+1\right)e^d - 1\right].\end{aligned}$$

As a result, the correlation becomes:

$$\text{Corr}(Y_{ij}, Y_{ik}) = \frac{\phi e^{\beta+\frac{1}{2}d}\left[\left(\frac{\tau\sigma^2}{\phi^2}+1\right)e^d - 1\right]}{1 + \phi e^{\beta+\frac{1}{2}d} \cdot \left[\left(\frac{\sigma^2}{\phi^2}+1\right)e^d - 1\right]}.$$

Considering the special case with only normally distributed random-effects, i.e., an exchangeable version of the conventional Poisson-normal model ($\phi_{ij} \equiv 1$ and $\sigma^2 \equiv 0$), simple algebra leads to:

$$\text{Corr}(Y_{ij}, Y_{ik}) = \frac{e^{\beta+\frac{1}{2}d}\left(e^d - 1\right)}{1 + e^{\beta+\frac{1}{2}d}\left(e^d - 1\right)}. \tag{10.10}$$

On the other hand, assuming that only the gamma-type random effects are present ($d = 0$), we derive:

$$\text{Corr}(Y_{ij}, Y_{ik}) = \frac{\mu^2\tau\sigma^2}{\phi\mu + \sigma^2\mu^2} = \frac{\tau\text{Var}(\lambda)}{E(\lambda) + \text{Var}(\lambda)},$$

where $\lambda = \phi\mu = \phi e^{\beta}$.

## 10.2 Taylor-series-based Derivation of the Correlation Function

Chapter 8 derived approximate expressions for the correlation function in the GLMM, including when multiple sequences on the same subject are observed. Given the absence, for the entirely general case, of closed-form expressions for the moments and hence, *a fortiori*, for the joint distribution, this is pragmatically a sensible way forward. In the Poisson case considered here, it is strictly speaking unnecessary to resort to such approximations. Nevertheless, we will derive the corresponding expressions for the Poisson-normal and the combined models. Here, not only a first-order but a general-order approximation will be derived. It is instructive to compare the various orders of approximations to the closed-form expressions, to study the quality of approximation.

### 10.2.1 General Derivation

We can usefully write the general model as $\boldsymbol{Y}_i = \boldsymbol{\mu}_i + \boldsymbol{\varepsilon}_i$, where $\boldsymbol{\mu}_i$, the conditional mean, given the random effects, can be written as $\boldsymbol{\mu}_i = h(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i)$, $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are known design matrices, $\boldsymbol{\beta}$ are fixed-effect parameters, $\boldsymbol{b}_i$ are random effects, and $h$ is a known link function. Finally, $\boldsymbol{\varepsilon}_i$ is the residual error component. As shown in Section 8.1 we have the following expression for the variance-covariance matrix:

$$\boldsymbol{V}_i \cong \boldsymbol{\Delta}_i\boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i'\boldsymbol{\Delta}_i' + \boldsymbol{\Phi}^{\frac{1}{2}}\boldsymbol{A}_i^{\frac{1}{2}}\boldsymbol{R}_i\boldsymbol{A}_i^{\frac{1}{2}}\boldsymbol{\Phi}^{\frac{1}{2}}. \tag{10.11}$$

When the canonical link is used, we have $\boldsymbol{A}_i = \boldsymbol{\Delta}_i$ and (10.11) can be written as: $\boldsymbol{V}_i \cong \boldsymbol{\Delta}_i\boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i'\boldsymbol{\Delta}_i' + \boldsymbol{\Phi}^{\frac{1}{2}}\boldsymbol{\Delta}_i^{\frac{1}{2}}\boldsymbol{R}_i\boldsymbol{\Delta}_i^{\frac{1}{2}}\boldsymbol{\Phi}^{\frac{1}{2}}$. If in addition, conditional independence (no serial correlation) is assumed, then (10.11) simplifies to: $\boldsymbol{V}_i \cong \boldsymbol{\Delta}_i\boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i'\boldsymbol{\Delta}_i' + \boldsymbol{\Phi}^{\frac{1}{2}}\boldsymbol{\Delta}_i\boldsymbol{\Phi}^{\frac{1}{2}}$. Further, if we reduce the random-effects part to a random-intercept model, i.e., $\boldsymbol{Z}_i = \boldsymbol{1}$ and $\boldsymbol{D} = d$, then (10.11) reduces to $\boldsymbol{V}_i \cong \boldsymbol{\Delta}_i(d\boldsymbol{J})\boldsymbol{\Delta}_i' + \boldsymbol{\Phi}^{\frac{1}{2}}\boldsymbol{\Delta}_i\boldsymbol{\Phi}^{\frac{1}{2}}$. The result is particularly simple to use in the case of normal outcomes of course. Chapter 8 applied the result to binary data and zoomed in on the random-intercept setting. This will be reviewed in Section 10.2.2, after which we turn to the count-data case in Section 10.2.3.

### 10.2.2 ICC for a Random-intercept Model for Binary Data

As a basis for comparison with the Poisson case to be dealt with next, let us first review the derivation in Chapter 8 of the ICC for a random-intercept model for binomial data with a logit link and assuming no overdispersion. In this case, $\boldsymbol{V}_i$ reduces to $\boldsymbol{V}_i \cong \boldsymbol{\Delta}_i(d\boldsymbol{J})\boldsymbol{\Delta}_i' + \boldsymbol{\Delta}_i = \boldsymbol{\Delta}_i(d\boldsymbol{J} + \boldsymbol{\Delta}_i^{-1})\boldsymbol{\Delta}_i'$. Furthermore, $\boldsymbol{\Delta}_i$ is a diagonal matrix with $V_{ij}(0)$ as diagonal elements, where the variance function $V_{ij}(0) = \mu_{ij}\,\big|_{\boldsymbol{b}_i=0}$ $(1 - \mu_{ij}\,\big|_{\boldsymbol{b}_i=0})$, and therefore $\boldsymbol{V}_i \cong \text{diag}(V_{ij}(0))[d\boldsymbol{J} + \text{diag}(V_{ij}(0))^{-1}]\text{diag}(V_{ij}(0))$. In other words, the variance-covariance matrix for subject $i$ is specified by the matrix with elements: $v_{ijj} = V_{ij}(0)[1 + V_{ij}(0)d]$, $v_{ijk} = dV_{ij}(0)V_{ik}(0)$, $(j \neq k)$. Based on these, we can determine a first-order approximation of the marginal correlation between time point $j$ and $k$, which is the intraclass correlation coefficient of reliability:

$$\rho_{ijk} = \text{Corr}(Y_{ij}, Y_{ik}) = \frac{V_{ij}(0)V_{ik}(0)d}{\sqrt{\{V_{ij}(0)[1 + V_{ij}(0)d]\}\{V_{ik}(0)[1 + V_{ik}(0)d]\}}}. \tag{10.12}$$

Note that, when $d = 0$, then $\rho_{ijk} = 0$, and when $d \to \infty$, then $\rho_{ijk} \to 1$. Thus, the full positive correlation range is attainable, quite unlike marginal models for correlated binary data, that experience restrictions on the correlation parameter space

to certain degrees. For a discussion, see Molenberghs and Verbeke (2005). No negative correlations can occur, which is entirely in line with the model's hierarchical nature, i.e., where $d$ is and remains interpretable as a variance. The related discussion for the case of linear mixed models can be consulted in Verbeke and Molenberghs (2000).

### 10.2.3 ICC for a Random-intercept Model for Count Data

Following similar logic as in the previous section, with a random intercept only, we can write for the count-data case:

$$\boldsymbol{Y}_i = \boldsymbol{\mu}_i + \boldsymbol{\varepsilon}_i = e^{X_i\boldsymbol{\beta}+\boldsymbol{J}_i b_i} + \boldsymbol{\varepsilon}_i,$$

where $\boldsymbol{J}_i$ is an $n_i$-dimensional vector of ones, and further assuming that there is no overdispersion, i.e., $\Phi = I$, it follows that

$$\begin{aligned} v_{ijj} &= e^{\boldsymbol{x}'_{ij}\boldsymbol{\beta}}\left(de^{\boldsymbol{x}'_{ij}\boldsymbol{\beta}}+1\right), \\ v_{ijk} &= de^{\boldsymbol{x}'_{ij}\boldsymbol{\beta}}e^{\boldsymbol{x}'_{ik}\boldsymbol{\beta}}, \qquad (j \neq k), \end{aligned}$$

producing the correlation-approximation:

$$\rho_{ijk} \simeq \frac{de^{\boldsymbol{x}'_{ij}\boldsymbol{\beta}}e^{\boldsymbol{x}'_{ik}\boldsymbol{\beta}}}{\sqrt{e^{\boldsymbol{x}'_{ij}\boldsymbol{\beta}}\left(de^{\boldsymbol{x}'_{ij}\boldsymbol{\beta}}+1\right)\cdot e^{\boldsymbol{x}'_{ik}\boldsymbol{\beta}}\left(de^{\boldsymbol{x}'_{ik}\boldsymbol{\beta}}+1\right)}}. \tag{10.13}$$

Note that, also here, (10.13) reduces to zero correlation for $d = 0$, and that $\rho_{ijk} \to 1$ when $d \to \infty$. Further assuming exchangeability, i.e., $\boldsymbol{x}'_{ij}\boldsymbol{\beta} = \beta$, (10.13) simplifies to

$$\rho \simeq \frac{de^{\beta}}{1+de^{\beta}}. \tag{10.14}$$

Evidently, we could also start from the explicit expression (10.10) and derive, by Taylor-series expansions for numerator and denominator separately:

$$\rho = \left[e^{\beta}\sum_{n=1}^{+\infty}\left(\frac{3^n-1}{n!\,2^n}\right)d^n\right] \Bigg/ \left[1+e^{\beta}\sum_{n=1}^{+\infty}\left(\frac{3^n-1}{n!\,2^n}\right)d^n\right] \tag{10.15}$$

Obviously, (10.14) immediately follows from (10.15) by restricting the Taylor series to the first order. To get a rough idea of the approximations' quality, assume $\beta = 0.6236$ and $d = 1.1792$, then $\rho = \rho_{[\infty]} = 0.8834$. The first order approximation is $\rho_{[1]} = 0.6875$, with subsequent values $\rho_{[2]} = 0.8274$ and $\rho_{[3]} = 0.8658$. The eighth

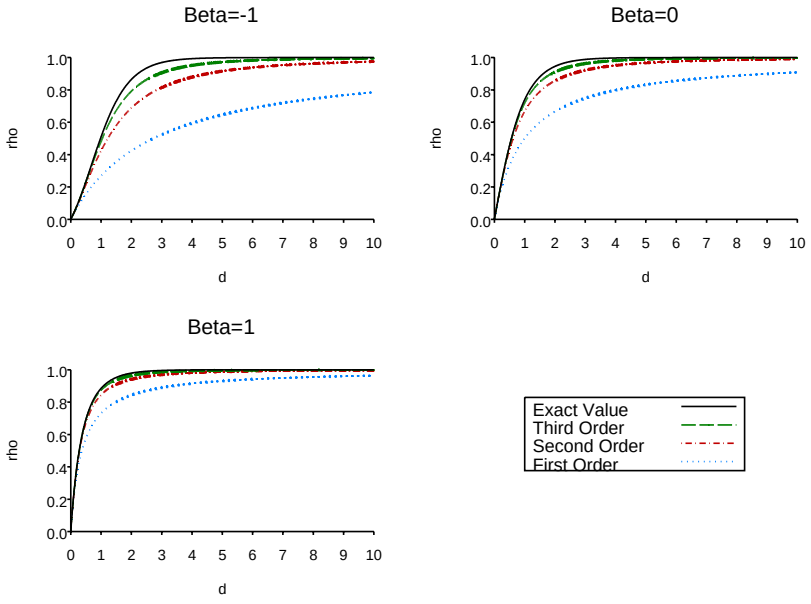


Figure 10.1: *Quality of Taylor-series approximations for the correlation function in the clustered exchangeable case.*

order is accurate to four decimal places while, if eight correct decimals are required, one has to go up to order 13. Figure 10.1 presents the quality of the first- to third-order approximations, for $\beta = -1$, 0, and 1, and for the range $[0, 10]$ for $d$. Apart from the obvious increase in quality with increasing order, it is also clear that all orders converge rather quickly to the correct value with increasing random-intercept variance. Given that the numerator and denominator of (10.15) differ only in the constant term, this is not surprising, since for increasing $d$ the leading terms, will rapidly dominate.

Switching to the combined model for the exchangeable case, the explicit form for the correlation becomes:

$$\rho = \frac{\phi\left(e^{\beta+\frac{1}{2}d}\right)\left[\left(\frac{\tau\sigma^2}{\phi^2}+1\right)e^d-1\right]}{1+\phi\left(e^{\beta+\frac{1}{2}d}\right)\left[\left(\frac{\sigma^2}{\phi^2}+1\right)e^d-1\right]}.$$

Similarly, a Taylor series expansion is:

$$\rho = \left[\phi e^{\beta} \sum_{n=1}^{+\infty} \left( \frac{3^n \tau \sigma^2 + (3^n - 1)\phi^2}{n!\, 2^n \phi^2} \right) d^n \right] \Bigg/ \left[ 1 + \phi e^{\beta} \sum_{n=1}^{+\infty} \left( \frac{3^n \sigma^2 + (3^n - 1)\phi^2}{n!\, 2^n \phi^2} \right) d^n \right].$$

## 10.3 Estimation

In the classical univariate overdispersion model, a common choice for the distribution of the parameter is (dropping the index $i$) $\lambda \sim \text{Gamma}(\alpha_1, \alpha_2)$, with density

$$f(\lambda) = \frac{1}{\alpha_2^{\alpha_1} \Gamma(\alpha_1)} \lambda^{\alpha_1 - 1} e^{-\lambda/\alpha_2},$$

where $\Gamma(\cdot)$ is the gamma function. Straightforward algebra produces

$$\text{P}(Y = y) = \left( \begin{array}{c} \alpha_1 + y - 1 \\ \alpha_1 - 1 \end{array} \right) \left( \frac{\alpha_1}{\alpha_2 + 1} \right)^y \left( \frac{1}{\alpha_2} \right)^{\alpha_1}.$$

The corresponding mean and variance are then given by $\alpha_1 \alpha_2$ and $\alpha_1 \alpha_2 (\alpha_2 + 1)$, respectively.

Turning to the models with normal random effects, they can be fitted by maximization of the marginal likelihood, obtained by integrating out the random effects from conditional densities of the form (3.9), in particular from their Poisson-normal form as specified by (3.10)–(3.12). The likelihood for $\boldsymbol{\beta}$, $D$, and $\phi$ takes the form

$$L(\boldsymbol{\beta}, D, \phi) = \prod_{i=1}^{N} \int \prod_{j=1}^{n_i} f_{ij}(y_{ij} | \boldsymbol{b_i}, \boldsymbol{\beta}, \phi)\ f(\boldsymbol{b_i} | D)\ d\boldsymbol{b_i}. \tag{10.16}$$

The key problem in maximizing (10.16) is the presence of $N$ integrals over the $q$-dimensional random effects $\boldsymbol{b_i}$. Generally, no closed-form solution exists, in which case one resorts to such methods as numerical integration or expansion techniques (Molenberghs and Verbeke, 2005).

In some special cases, these integrals can be resolved analytically. The best known example is the linear mixed effects model (Verbeke and Molenberghs, 2000). Fortunately, also the combined model, and hence the Poisson-normal as a special case,

lends itself to such analytic calculations. MVD derived the joint probability of $\boldsymbol{Y_i}$:

$$\begin{aligned}\mathrm{P}(\boldsymbol{Y_i} = \boldsymbol{y_i}) &= \sum_{\boldsymbol{t}} \left[ \prod_{j=1}^{n_i} \begin{pmatrix} y_{ij} + t_j \\ y_{ij} \end{pmatrix} \begin{pmatrix} \alpha_{1j} + y_{ij} + t_j - 1 \\ \alpha_{1j} - 1 \end{pmatrix} (-1)^{t_j} \alpha_{2j}^{y_{ij}+t_j} \right] \\ &\times \exp\left( \sum_{j=1}^{n_i} (y_{ij} + t_j) \boldsymbol{x}'_{ij} \boldsymbol{\beta} \right) \exp\left\{ \frac{1}{2} \left[ \sum_{j=1}^{n_i} (y_{ij} + t_j) \boldsymbol{z}'_{ij} \right] D \left[ \sum_{j=1}^{n_i} (y_{ij} + t_j) \boldsymbol{z}_{ij} \right] \right\},\end{aligned}$$

where the vector-valued index $\boldsymbol{t} = (t_1, \ldots, t_{n_i})$ ranges over all non-negative integer vectors and $\alpha_{j1}$ and $\alpha_{j2}$ are the gamma-distribution parameters for occasion $j$.

As one way forward for parameter estimation, MVD proceeded by what they termed partial marginalization, i.e., by integrating (3.15)–(3.19) over the gamma random effects only. The corresponding probability is:

$$\begin{aligned}&\mathrm{P}(Y_{ij} = y_{ij} | \boldsymbol{b_i}) \\ &= \begin{pmatrix} \alpha_{1j} + y_{ij} - 1 \\ \alpha_{1j} - 1 \end{pmatrix} \cdot \left( \frac{\alpha_{2j}}{1 + \kappa_{ij}\alpha_{2j}} \right)^{y_{ij}} \cdot \left( \frac{1}{1 + \kappa_{ij}\alpha_{2j}} \right)^{\alpha_{1j}} \kappa_{ij}^{y_{ij}}, \quad (10.17)\end{aligned}$$

where $\kappa_{ij} = \exp(\boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}_{ij}\boldsymbol{b_i})$. Note that, with this approach, we assume that the gamma random effects are independent within a subject. Hence, all correlation stems from the normal random effects. Recall that the general model does not force this restriction to hold. Now, it is easy to obtain the fully marginalized probability by numerically integrating the normal random effects out of (10.17) using tools, such as the SAS procedure NLMIXED, that allow for normal random effects in arbitrary, user-specified models.

It is important to realize that not all parameters may be simultaneously identifiable. For example, the gamma-distribution parameters $\alpha_{1j}$ and $\alpha_{2j}$ are not simultaneously identifiable when the linear-predictor part is also present, owing to aliasing with the intercept term. In the next section, when analyzing the epilepsy data, we will first assume that these parameters are independent of measurement occasion $j$ and further that $\alpha_1 \cdot \alpha_2 = 1$.

## 10.4 Analysis of the Epilepsy Data

MVD analyzed the data, introduced in Section 2.2. We will re-analyze the data but with different constraints on the gamma parameters, and in addition calculating

the within-subject correlations, using the results of Section 10.1. Let $Y_{ij}$ represent the number of epileptic seizures patient $i$ experiences during week $j$ of the follow-up period. Also, let $t_{ij}$ be the time point at which $Y_{ij}$ has been measured, $t_{ij} = 1, 2, \ldots$ until at most 27. Let us consider the combined model (3.15)–(3.19), with specific choices

$$\ln(\lambda_{ij}) = \begin{cases} (\beta_{00} + b_i) + \beta_{01} t_{ij} & \text{if placebo,} \\ (\beta_{10} + b_i) + \beta_{11} t_{ij} & \text{if treated,} \end{cases} \tag{10.18}$$

where the random intercept $b_i$ is assumed to be zero-mean normally distributed with variance $d$. We consider special cases (1) the ordinary Poisson model, (2) the negative-binomial model, (3) the Poisson-normal model, together with (4) the combined model. The SAS implementation is presented in MVD. Estimates (standard errors) are displayed in Table 10.1. The negative-binomial model and the Poisson-normal model both provide improved fits relative to the standard Poisson model, with the combined model providing a further, strong improvement. MVD showed that the effects of choosing the combined model was clearly seen in such key inferential parameters as the difference and the ratio of the slopes (not reproduced here). In particular, they established that, whereas the conventionally used and broadly implemented Poisson-normal model would suggest a significant effect of treatment, this is no longer true with the combined model.

Let us now turn to the correlation functions. Since the gamma random effects are assumed independent, we only need consider the Poisson-normal and combined cases. Obviously, since the fixed-effects structure is not constant but rather depends on time, we have to apply the general correlation function (10.3). In the Poisson-normal case, and for the placebo group, based on the parameter estimates in Table 10.1, we obtain:

$$\text{Corr}(Y(t), Y(s)) = \frac{35.58 \cdot 0.99^{t+s}}{\sqrt{(4.04 \cdot 0.99^t + 35.58 \cdot 0.97^t) \cdot (4.04 \cdot 0.99^s + 35.58 \cdot 0.97^s)}},$$

where $Y(t)$ represents the outcome for an arbitrary subject at time $t$. The smallest and largest values for the correlation functions, for both arms, and for both the Poisson-normal and combined models, are given in Table 10.2.

Within each model, there is relatively little difference between the placebo and treated groups, although the difference is a bit more pronounced in the combined model. Further, the correlation range within every group is relatively narrow. The most noteworthy feature, unquestionably, is the large discrepancy between both models. This is because the Poisson-normal model forces the correlation and overdisper-

Table 10.1: *Epilepsy Study. Parameter estimates and standard errors for the regression coefficients in (1) the Poisson model, (2) the negative-binomial model, (3) the Poisson-normal model, and (4) the combined model. Estimation was done by maximum likelihood using numerical integration over the normal random effect, if present.*

| | | Poisson | Negative-binomial |
|---|---|---|---|
| Effect | Parameter | Estimate (SE) | Estimate (SE) |
| Intercept placebo | $\beta_{00}$ | 1.2662 (0.0424) | 1.2594 (0.1119) |
| Slope placebo | $\beta_{01}$ | −0.0134 (0.0043) | −0.0126 (0.0111) |
| Intercept treatment | $\beta_{10}$ | 1.4531 (0.0383) | 1.4750 (0.1093) |
| Slope treatment | $\beta_{11}$ | −0.0328 (0.0038) | −0.0352 (0.0101) |
| Negative-binomial parameter | $\alpha_1$ | — | 0.5274 (0.0255) |
| Negative-binomial parameter | $\alpha_2 = 1/\alpha_1$ | — | 1.8961 (0.0918) |
| Variance of random intercepts | $d$ | — | — |
| −2 log-likelihood | | 11,590→ −1492 | −6755 |
| | | Poisson-normal | Combined |
| Effect | Parameter | Estimate (SE) | Estimate (SE) |
| Intercept placebo | $\beta_0$ | 0.8179 (0.1677) | 0.9112 (0.1755) |
| Slope placebo | $\beta_1$ | −0.0143 (0.0044) | −0.0248 (0.0077) |
| Intercept treatment | $\beta_0$ | 0.6475 (0.1701) | 0.6555 (0.1782) |
| Slope treatment | $\beta_2$ | −0.0120 (0.0043) | −0.0118 (0.0074) |
| Negative-binomial parameter | $\alpha_1$ | — | 2.4640 (0.2113) |
| Negative-binomial parameter | $\alpha_2 = 1/\alpha_1$ | — | 0.4059 (0.0348) |
| Variance of random intercepts | $d$ | 1.1568 (0.1844) | 1.1289 (0.1850) |
| −2 log-likelihood | | 6272= −6810(g.l.) | −7664 |

sion effects to stem from a single additional parameter, the random-intercept variance $d$. Thus, considerable overdispersion also forces the correlation to increase, arguably beyond what is consistent with the data. In the combined model, in contrast, there are *two* additional parameters, giving proper justice to both correlation and overdispersion effects. It was already clear from the above discussion and that in MVD that the combined model is an important improvement. This now clearly manifests itself in the correlation function, too.

Table 10.2: *Epilepsy Study. Observed smallest and largest values for the correlation function, for the Poisson-normal and combined models, and for both treatment arms. The time pair for which the values are observed is shown too.*

| | | Smallest value | | Largest value | |
|---|---|---|---|---|---|
| Model | Arm | $\rho$ | time pair | $\rho$ | time pair |
| Poisson-normal | placebo | 0.8577 | 26 & 27 | 0.8960 | 1 & 2 |
| Poisson-normal | treatment | 0.8438 | 26 & 27 | 0.8794 | 1 & 2 |
| Combined | placebo | 0.3041 | 26 & 27 | 0.3134 | 1 & 2 |
| Combined | treatment | 0.2234 | 1 & 2 | 0.3410 | 26 & 27 |

## 10.5 Concluding Remarks

Many inferential questions can be framed in terms of correlations between repeated or otherwise hierarchical measures taken on the same experimental unit. Such data can be conveniently modeled using random effects models, like linear and generalized linear mixed models. In contrast with the LMM case, the GLMM renders more difficult the derivation of such correlations. Chapter 8 derived approximate correlation expressions in a broad GLMM framework, where not only a single one but several repeated-measures sequences per unit may be recorded, perhaps of a different data type.

In this chapter, we considered the specific case of repeated Poisson data. Two aspects set our approach apart from the binary one. First, we considered an extended GLMM, where conventional normally distributed random effects in the linear predictor are supplemented with overdispersion effects, conveniently captured by, for example, gamma random effects. This model, owing to MVD, generalizes at once the Poisson-normal and negative-binomial models for count data. Second, as shown in MVD, the combined model, and hence its special cases, allows for explicit expressions, not only for the mean and variance functions, but also for the entire joint probability mass function. This is a major difference between the Poisson and other non-Gaussian cases, since it does allow for closed form derivations of the correlation function. As a consequence, the quality of the Taylor-series expansion based correlation functions, derived in Chapter 8, can be assessed by comparison with the closed-form expression.

We applied the methodology to data from a case study in patients with epileptic seizures. In synchrony with MVD, we have shown that the combined method improves the fit over its special cases. In addition, in this paper, we have shown that the correlations derived from the more conventional but also more restricted Poisson-normal model can be highly misleading. For this particular example, the Poisson-normal model suggests a high within-patient correlation among any two time points within any of the two treatment arms. However, the correlations stemming from the combined model are small to moderate.

One may wonder why a mixed-model approach is used when marginal correlation is of interest, since marginal models may produce correlation parameters more straightforwardly, perhaps even for the special case of Poisson data, in spite of the closed-form expressions derived here. However, marginal models come with their own issues. First, non-likelihood based methods such as GEE treat correlation as nuisance parameters only, whence they cannot be used for inferential purposes. Second, full likelihood methods may be highly prohibitive in terms of computational requirements. Third, the correlation ranges attainable in marginal models may be highly restricted, whereas we have seen here in a number of special but important cases, such as exchangeable clustered data, that the entire range of positive correlations can be reached. For a discussion of these and related matters, see Molenberghs and Verbeke (2005) and Molenberghs and Kenward (2007). Finally, fitting these models and hence deriving the correlations in practical terms is quite feasible, using what is termed by MVD partial marginalization. To this end, for example, the SAS procedure NLMIXED can be used.

# Chapter 11

# Estimating Criterion Validity Using the GLMM Framework and Concepts from Surrogate Marker Evaluation

In Chapters 6 to 10 we focused on correlation of repeated measures within a subject. The goal was to estimate reliability and also to study which factors influence reliability though generalizability theory. Often, one is confronted with the situation that multiple response variables are measured over time, sometimes referred to as a family of responses. These different response variables can but do not have to be of the same type. Sometimes, the goal is to estimate treatment effects in a multivariate way, i.e., jointly estimate treatment effects on the binary and the continuous responses. In that case, one not only needs to take account of the correlation within a subject for a specific single response, but also take account of the correlation between the different responses for a specific subject. One application in the psychometric literature is the situation where one wants to estimate the correlation of a certain response variable with a gold standard to establish *criterion validity*, as introduced in Section 4.3. In this chapter we want to provide tools to study criterion validity. In Section 11.1 we will extend the GLMM framework of Chapter 8 to derive correlation measures be-

tween 2 sequences of scales, and in Section 11.2 we will borrow techniques developed in the area of surrogate marker methodology to derive correlation measures.

## 11.1 Using the GLMM Framework to Study Criterion Validity

Suppose we want to study the correlation between a continuous interval scaled parameter $Y_{i1}$ and a binary response $Y_{i2}$, then a GLMM can be extended too, as described in Molenberghs and Verbeke (2005):

$$\begin{pmatrix} Y_{i1} \\ Y_{i2} \end{pmatrix} = \begin{pmatrix} \mu_1 + \lambda b_i + \alpha_1 X_i \\ \dfrac{\exp[\mu_2 + b_i + \alpha_2 X_i]}{1 + \exp[\mu_2 + b_i + \alpha_2 X_i]} \end{pmatrix} + \begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \end{pmatrix}.$$

Here, $\varepsilon_{i1}$ and $\varepsilon_{i2}$ are the error terms for the continuous and binary outcomes, respectively. Obviously, the first one will be normally distributed while the second one follows a Bernoulli distribution. We have included a scale parameter $\lambda$ in the continuous component of an otherwise random-intercept model, because the continuous and binary outcome are measured on a different scale. In this case, we have

$$\boldsymbol{Z}_i = \begin{pmatrix} \lambda \\ 1 \end{pmatrix}, \boldsymbol{\Delta}_i = \begin{pmatrix} 1 & 0 \\ 0 & v_{i2}(0) \end{pmatrix}, \boldsymbol{\phi} = \begin{pmatrix} \sigma^2 & 0 \\ 0 & 1 \end{pmatrix},$$

with $v_{i2}(0) = \mu_i \,|_{\boldsymbol{b}_i=0} \,(1-\mu_i \,|_{\boldsymbol{b}_i=0})$. Note that $\boldsymbol{Z}_i$ is not a design matrix in the strict sense, since it contains an unknown parameter. Nevertheless, it is useful to consider this decomposition, implying that (8.5) becomes

$$\boldsymbol{V}_i = \begin{pmatrix} \lambda^2 & v_{i2}(0)\lambda \\ v_{i2}(0)\lambda & v_{i2}(0)^2 \end{pmatrix} \tau^2 + \begin{pmatrix} \sigma^2 & 0 \\ 0 & v_{i2}(0) \end{pmatrix} = \begin{pmatrix} \lambda^2\tau^2 + \sigma^2 & v_{i2}(0)\lambda\tau^2 \\ v_{i2}(0)\lambda\tau^2 & v_{i2}(0)^2\tau^2 + v_{i2}(0) \end{pmatrix}.$$

Here, $\tau^2$ is the random-intercept variance. As a result, we have the following approximation for the marginal correlation: $\rho(\beta) = v_{i2}\lambda\tau^2/\sqrt{\lambda^2\tau^2+\sigma^2}\sqrt{v_{i2}^2\tau^2+v_{i2}}$, which we can now apply to the same data set described in Section 2.1 to estimate the correlation at Week 8 between the binary CGI response variable and the continuous response parameter defined as the total PANSS. Table 11.1 summarizes the results. We can conclude that there was a high correlation between the response variable defined by the CGI and the total PANSS indicating criterion validity of the derived CGI response and the total PANSS. This correlation was similar in both treatment

Table 11.1: *Schizophrenia Data. Parameter estimates (standard errors) for a bivariate joint GLMM analysis to estimate criterion validity between response and total PANSS at Week 8. The SAS procedure NLMIXED has been used. Standard errors are calculated using the delta method.*

| Endpoint | Effect | Parameter | Estimate | (SE) |
|---|---|---|---|---|
| Total PANSS | Intercept | $\mu_1$ | 68.98 | (1.59) |
| | Treatment | $\alpha_1$ | -0.41 | (2.06) |
| | Standard deviation | $\sigma_1$ | 13.83 | (0.43) |
| | Variation | $\sigma_1^2$ | 191.37 | (11.90) |
| | Inflation | $\lambda$ | -0.97 | (0.61) |
| Response (CGI) | Intercept | $\mu_2$ | -2.56 | (3.25) |
| | Treatment | $\alpha_2$ | 0.96 | (2.44) |
| Common parameters | R.I. st.dev. | $\tau$ | 16.84 | (10.73) |
| | R.I. var. | $\tau^2$ | 283.74 | (361.40) |
| | Corr. (control) | $\rho_{\text{cont}}$ | -0.74 | (0.026) |
| | Corr. (risperidone) | $\rho_{\text{ris}}$ | -0.75 | (0.022) |

groups. Note that the correlation ($-0.75$ in the risperidone group and $-0.74$ in the control group) is negative because higher PANSS values indicate a more psychotic condition and response was coded 1 if the CGI was equal to "very much improved" or "much improved". In the classical approach, often the Pearson or the Spearman correlation is calculated, including only data observed at Week 8 for both the binary response and the continuous PANSS score. Here, this resulted in $-0.59$ and $-0.61$, for Pearson's and Spearman's correlation, respectively.

While in this section we have considered two outcomes of a different type, hence restricting attention to a cross-sectional setting, it is perfectly possible to combine the longitudinal ideas of previous sections with the multivariate setting considered here, thus producing a flexible method that can handle multivariate longitudinal data. One can then distinguish between various types of correlations, e.g., within-sequence (referring to the reliability concept), between two different measurements taken at the same time (of relevance in marker evaluation), and even between different measurements at different times. Details on how such models can be built and fitted

are given in Molenberghs and Verbeke (2005, Chapter 24).

## 11.2 Criterion Validity and Surrogate Maker Methodology

In this section we will apply the methods of Sections (5.2) and (5.3) to the data described in Section 2.1. We will show how these methods can be used to investigate the criterion validity between the three scales of interest: PANSS, BPRS and CGI. We will successively consider the relationships between (i) PANSS and BPRS (Section 11.2.1), (ii) PANSS and CGI (Section 11.2.2) and (iii) BPRS and CGI (Section 11.2.3). The binary indicator for treatment ($Z_{ij}$) will be set to 0 for the conventional antipsychotic agents and to 1 for risperidone.

### 11.2.1 Relationship Between PANSS and BPRS

The relationship between PANSS and BPRS was studied first. Since the BPRS is derived from the PANSS by selecting some of its items, there is a natural link between these two scales. However, even though one could expect correlation due to one scale being a subset of the other, it remains to determine how large such association is. With our notation we assume PANSS plays the role of $S_1$ and BPRS plays the role of $S_2$. Figure 11.1 (a) shows a scatter plot of BPRS versus PANSS. Clearly, both scales are highly correlated. The Pearson's correlation coefficient equals $\rho = 0.96$. Throughout, the sample sizes of the units were used to weight the observations in the calculation of the $R^2$ values. Figure 11.1 (b) shows a plot of the treatment effects on the PANSS versus the treatment effects on the BPRS for the different units. These seem to be highly correlated.

Indeed, using the multi-trial method with country as unit of analysis we found highly conclusive values for the coefficients of determination at the trial *and* individual level. Since no clear "true endpoint" can be assigned, we calculated both $R^2_{b_i|a_i,m_{S_{2i}}} = 0.94$ (95% confidence interval: [0.82,0.97]) and $R^2_{a_i|b_i,m_{S_{1i}}} = 0.93$ (95% confidence interval: [0.83,0.97]). However, calculating the estimate (5.14) based on the reduced model we found $R^2_{b_i|a_i} = 0.93$ with 95% confidence interval $[0.86, 0.97]$, which is very close to the previous values but has the advantage of being symmetric in both scales. Its value indicates that not much would be gained in the precision of the prediction if instead of the full model the reduced model were used to predict the

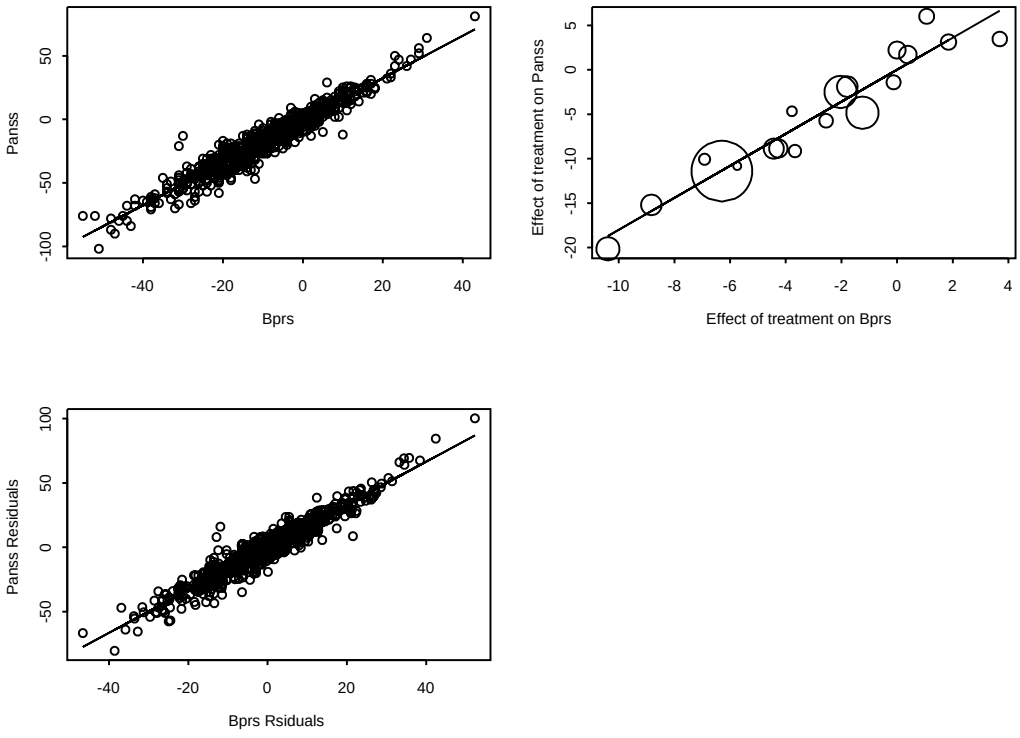


Figure 11.1: *Schizophrenia Data. (a) Scatter Plot of BPRS versus PANSS;(b) Treatment Effects on PANSS by Treatment Effects on BPRS. The size of each point is proportional to the number of patients examined by the corresponding investigator; (c) Plot of the residuals of BPRS versus PANSS.*

treatment effect. While in this case study, as well as in extensive simulations by Tibaldi *et al.* (2002), there is close agreement between full and reduced coefficients, it is very important not to take this for granted in different applications. When fitting the full model would turn out to be computationally intensive, then it is advisable to conduct an appropriate, perhaps even limited, simulation study. The individual coefficient of determination was calculated as $R^2_{\text{indiv}} = 0.92$ with 95% confidence interval $[0.90, 0.93]$. Note that this quantity is symmetric in both scales. Graphically this correlation is represented by the residual plot shown in Figure 11.1 (c).

### 11.2.2 Relationship Between PANSS and CGI

In this section we will study the symmetric relationship between PANSS ($S_1$) and CGI ($S_2$). The meta-analytic approach yielded $R^2_{b_i|m_{S_{2i}},a_i} = 0.81$ (95% confidence interval $[0.56, 0.93]$), $R^2_{a_i|m_{S_{1i}},b_i} = 0.83$ (95% confidence interval $[0.63, 0.96]$) at the trial level and $R^2_{\text{indiv}} = 0.78$ with 95% confidence interval $[0.68, 0.92]$ at the individual level. Clearly, these quantities indicate that the agreement between PANSS and CGI, is acceptably high. However it is important to notice that this value of $R^2_{\text{indiv}}$ is the squared correlation between the latent unobservable variable $\widetilde{S}_2$ and the observable scale $S_1$.

At the trial level, the situation is totally different. Marginally

$$\widetilde{S}_{2i}|Z_i \sim N\left(\mu_{\widetilde{S}_{2i}} + \alpha_i Z_i, \frac{1}{1-\rho^2}\right)$$

implying that $P\left(S_{2i} \leq r\right) = \Phi\left(\gamma^i_{0r} + \gamma^i_Z Z_i\right)$ where $\gamma^i_Z = -\sqrt{1-\rho^2}\alpha_i$. This formula shows the linear relationship between the treatment effect on the latent variable $\widetilde{S}_2$ and the treatment effect on the observable scale $S_2$ confirming the validity of our conclusions at the trial level.

Figure 11.2(a) shows boxplots for PANSS at the different categories of CGI. It is clear from the graph that it seems to be a positive correlation between both scales, the small decrease of the PANSS median at the end of the graph can be explained by the small sample size available at this point (only 6 patients). The Spearman correlation coefficient was 0.77 confirming our previous idea of a high positive correlation. Figure 11.2(b) plots the treatment effects on CGI versus the treatment effects on PANSS.

In addition, we calculated the $R^2$ measure at the trial level for the "reduced" model. This yielded $R^2_{b_i|a_i} = R^2_{a_i|b_i} = 0.81$ with 95% confidence interval $[0.56, 0.94]$ which coincides with the trial-level values obtained from the "full" model. Apart from the attractive feature that this quantity is symmetric in both scales, the result again indicates that not much would be gained in the precision of the treatment prediction if instead of the full model, the reduced model were used.

Based on the results of the above meta-analytic method, we are able to predict, for example, the treatment effect on the CGI response based on the observed treatment effect on PANSS (or vice versa). Each time the *ith* unit was removed from the validation process that was based only on the other 18. First, a linear model was fitted for PANSS at the *ith* unit and $\beta_i$ was estimated. Later, using formula (5.12)

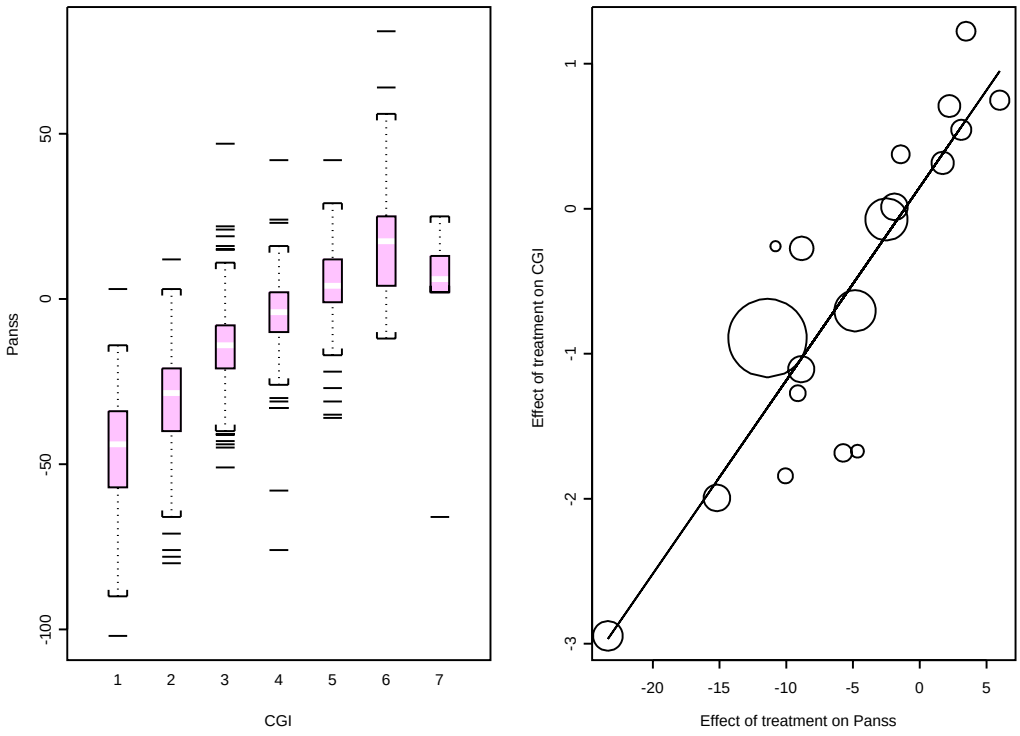


Figure 11.2: *Schizophrenia Data. a) Boxplots of PANSS by CGI; (b) Treatment Effects on PANSS by Treatment Effects on CGI. The size of each point is proportional to the number of patients.*

and the information available from the validation process $\alpha_i$, was estimated as well. Finally, treatment effect on CGI was predicted as $\widehat{\gamma}_Z^i = -\sqrt{1-\widehat{\rho}^2}\widehat{\alpha}_i$. Table 11.2 reports prediction intervals for the 19 units together with the number of patients per unit. In this table, $\hat{\beta}_0$ and $\widehat{\gamma_{Z0}^i}$ are values estimated from the data; $\widehat{\gamma_Z^i}$ is the predicted treatment effect on CGI, given its effect on PANSS. Clearly, in all cases, the predicted values for $\widehat{\gamma_{Z0}^i}$ agree reasonably well with the effects estimated from the data taking into account the standard deviation of this estimate.

Table 11.2: *Schizophrenia Data. Predictions for the treatment effects on CGI based on the observed treatment effects on PANSS. Estimates (standard errors) are shown. Here $\hat{\beta}_0$ is the treatment effects on PANSS estimated from the data, $\widehat{\gamma^i_{Z0}}$ is the treatment effects on CGI estimated from the data and $\widehat{\gamma^i_Z}$ is the prediction for the treatment effects on CGI.*

| Unit | # patients | $\hat{\beta}_0$ | $\widehat{\gamma^i_{Z0}}$ | $\widehat{\gamma^i_Z}$ |
|---|---|---|---|---|
| 1 | 31 | 3.47 ( 8.00) | -0.76 (0.40) | -0.24 (0.34) |
| 2 | 29 | -5.74 ( 8.27) | 0.76 (0.41) | 0.27 (0.78) |
| 3 | 26 | -9.14 ( 8.76) | 0.50 (0.41) | 0.50 (0.57) |
| 4 | 44 | -1.90 ( 6.72) | 0.06 (0.32) | 0.05 (0.25) |
| 5 | 44 | -15.19 ( 6.72) | 0.87 (0.33) | 0.88 (0.52) |
| 6 | 37 | 1.72 ( 7.82) | -0.18 (0.36) | -0.18 (0.37) |
| 7 | 32 | 6.00 ( 7.87) | -0.33 (0.38) | -0.46 (0.47) |
| 8 | 68 | -4.85 ( 5.92) | 0.44 (0.28) | 0.22 (0.28) |
| 9 | 49 | -23.37 ( 8.68) | 1.22 (0.47) | 1.40 (0.81) |
| 10 | 43 | -8.88 ( 6.79) | 0.49 (0.32) | 0.48 (0.46) |
| 11 | 21 | -4.66 ( 9.73) | 0.94 (0.49) | 0.19 (0.32) |
| 12 | 25 | -10.07 ( 9.12) | 1.00 (0.46) | 0.53 (0.38) |
| 13 | 39 | -8.84 ( 7.15) | 0.06 (0.33) | 0.50 (0.45) |
| 14 | 36 | 2.22 ( 7.42) | -0.28 (0.35) | -0.20 (0.54) |
| 15 | 17 | -10.81 (10.97) | 0.99 (0.58) | 0.66 (0.51) |
| 16 | 33 | 3.11 ( 9.48) | -0.10 (0.44) | -0.27 (0.32) |
| 17 | 69 | -2.50 ( 5.40) | 0.002 (0.25) | 0.09 (0.28) |
| 18 | 30 | -1.42 ( 8.20) | -0.24 (0.39) | 0.03 (0.49) |
| 19 | 128 | -11.41 ( 3.94) | 0.40 (0.18) | 0.65 (0.38) |

### 11.2.3 Relationship Between BPRS and CGI

When studying the relationship between CGI ($S_1$) and BPRS ($S_2$), we found similar results to the ones obtained in Section 11.2.2. This is not so surprising given the

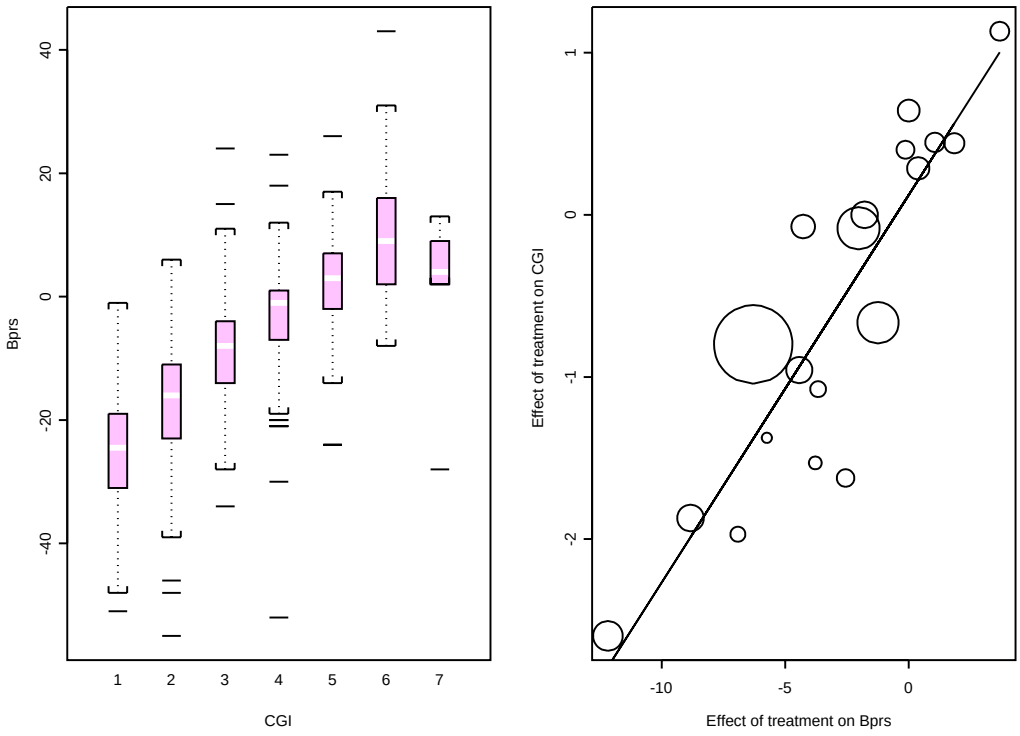


Figure 11.3: *Schizophrenia Data. a) Boxplots of BPRS by CGI; (b)Treatment Effects on CGI by Treatment Effects on BPRS. The size of each point is proportional to the number of patients examined by the corresponding investigator.*

strong relationship between BPRS and PANSS. Since results for the full and reduced models almost coincide, we only present the values for the reduced model here.

Using the meta-analytic approach we find a value of 0.78 for $R^2_{\text{trial}}$ with 95% confidence interval $[0.48, 0.93]$ and $R^2_{\text{indiv}} = 0.73$ with 95% confidence interval $[0.64, 0.84]$. Here, the same remarks of the previous section would be valid; hence it is possible that $R^2_{\text{indiv}}$ could be overestimating the real correlation between BPRS and CGI. Figure 11.3 (a)–(b), as before, show respectively the scatter plot of boxplots of BPRS for the different categories of CGI and the treatment effects on CGI by the treatment effects on BPRS, respectively.

# Chapter 12

# Case Study in Incomplete Data

This Chapter follows the development of Michiels *et al.* (2002). We will focus on two aspects of the modelling process: the exploratory phase (and likewise assessment of model fit), and the handling of incomplete sequences. We show how a suite of well-chosen plots can support the modelling task and how a simple sensitivity analysis can be conducted to assess the influence of dropout. While previous chapters focussed on reliability, generalizability and validity, here the focus is on estimation of treatment effects in presence of incomplete data.

## 12.1 Exploratory Analysis

Most books on longitudinal data discuss exploratory analysis. See, for example, Diggle, Liang, and Zeger (1994). However, most effort is spent to model building and formal aspects of inference. In this section, we present a selected set of plots to underpin the model building. We distinguish between two modes of display: (1) averaged over (sub)populations and (2) individual profiles. Both ways can be used to present three fundamental aspects of the longitudinal structure: (1) the average evolution; (2) the variance function, (3) the correlation structure. We will discuss some of the less frequently used displays in what follows.

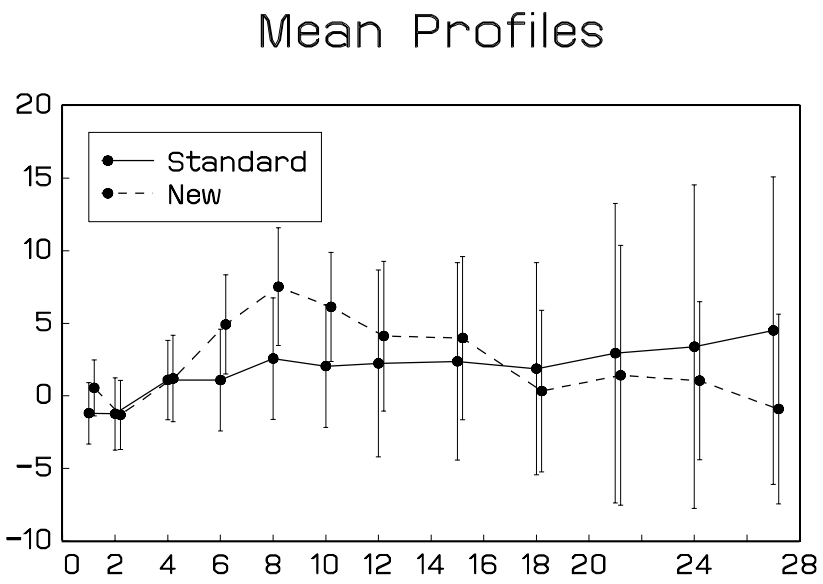


Figure 12.1: *The Vorozole Study. Mean profiles.*

The average evolution describes how the profile for a number of relevant subpopulations (or the population as a whole), evolves over time. The results of this exploration will be useful in order to choose a fixed-effects structure for the linear mixed model.

The mean profiles per treatment arm, as well as their 95% confidence intervals, are plotted in Figure 12.1. The average profiles indicate an increase over time which is slightly stronger for the vorozole group until month 14, and afterwards, the megestrol acetate group shows a slightly higher FLIC score. As can be seen from the confidence intervals, these differences are clearly not significant.

Owing to thinning of information with elapsing time, we decided to restrict attention to the first 2 years. This leads to a maximum of 13 observations per subject (month 1, 2, 4, 6, . . . , 24).

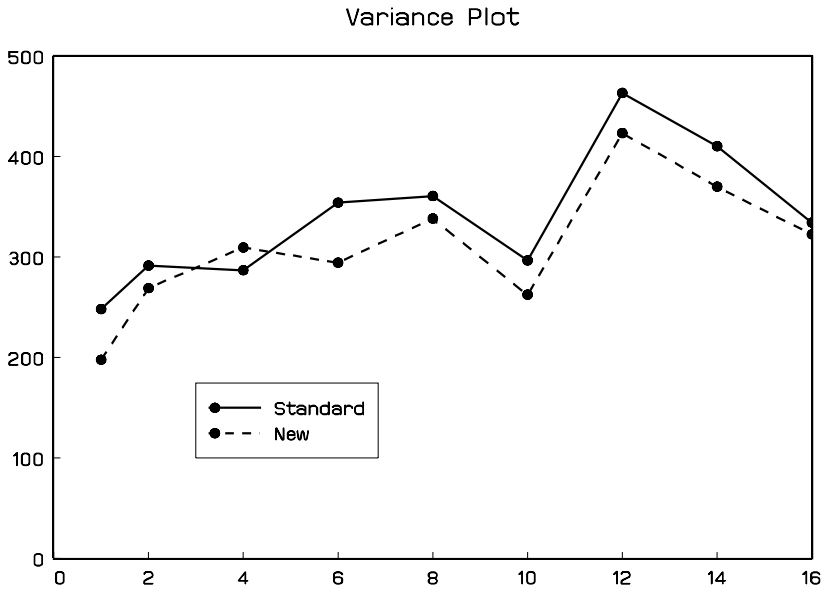


Figure 12.2: *The Vorozole Study. Variance function.*

In addition to the average evolution, the evolution of the variance is important to build an appropriate longitudinal model. Clearly, one has to correct the measurements for the fixed-effects structure and hence detrended values have to be used. These detrended values are merely the outcome values (change in FLIC-score), subtracted by the mean change, calculated at each time point separately. Again, two plots are of interest. The variance function is plotted in Figure 12.2. It seems to be relatively stable, except for a sharp decline near the end (at which point there are large dropout rates), and hence a constant variance model is a plausible starting point.

The correlation structure describes how measurements within a subject correlate. The correlation function depends on a pair of times and only under the assumption of stationarity does this pair of times simplify to the time lag only. This is important since many exploratory and modelling tools are based on this assumption.

A scatter plot matrix is given in Figure 12.3. The off-diagonal elements picture scatter plots of standardized residuals obtained from pairs of measurement occasions. The decay of correlation with time is studied by considering the evolution of the scatters with increasing distance to the main diagonal. Stationarity on the other hand implies that the scatter plots remain similar within diagonal bands *if measurement occasions are approximately equally spaced*. In addition to the scatter plots, we place

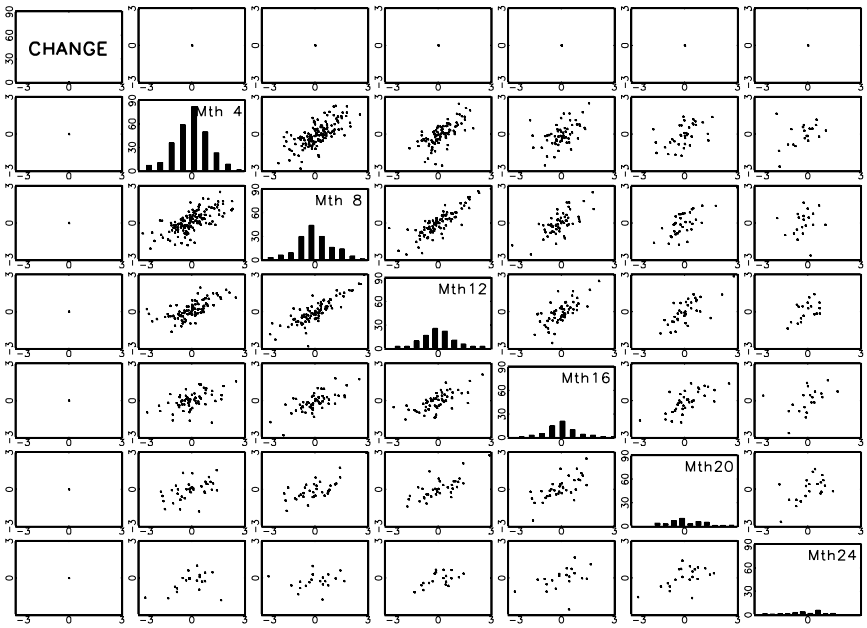


Figure 12.3: *The Vorozole Study. Scatter plot matrix for selected time points. The same vertical scale is used along the diagonal to display the attrition rate as well.*

histograms on the diagonal, capturing the variance structure including such features as skewness. If the axes are given the same scales, it is very easy to capture the attrition rate as well.

## 12.2 A Selection Model Formulation

First, a linear mixed model for the measurements of the form (3.5) is assumed. Secondly, we will model the dropout mechanism. We assume that incompleteness is due to dropout only, and that the first measurement $Y_{i1}$ is obtained for everyone. We refer to Section 3.5 for the general formulation and notation. For each subject $i$, denote $D_i$ to be the dropout indicator, one higher than the occasion of the last obtained measurement (3.23). The model for the dropout process is based on a logistic regression for the probability of dropout at occasion $j$, given the subject is still in the study. We denote this probability by $g(\boldsymbol{h}_{ij}, y_{ij})$ in which $\boldsymbol{h}_{ij}$ is a vector containing all responses observed up to but not including occasion $j$, as well as relevant covariates $w_{ik}$. We

then assume that $g(\boldsymbol{h}_{ij}, y_{ij})$ satisfies

$$\text{logit}[g(\boldsymbol{h}_{ij}, y_{ij})] = \text{logit}\left[\text{pr}(D_i = j | D_i \geq j, \boldsymbol{y}_i, W_i)\right] = \boldsymbol{h}_{ij}\boldsymbol{\psi}_0 + y_{ij}\psi_d \quad i = 1, \ldots, N, \tag{12.1}$$

where $\boldsymbol{\psi} = (\boldsymbol{\psi}_0', \psi_d)'$. When $\psi_d$ equals zero, the dropout model is random, and all parameters can be estimated using standard software since the measurement model for which we use a linear mixed model and the dropout model, assumed to follow a logistic regression, can then be fitted separately. If $\psi_d \neq 0$, the dropout process is assumed to be non-random.

Model (12.1) is now used to construct the dropout process:

$$f(d_i | \boldsymbol{y}_i, W_i, \boldsymbol{\psi}) = \begin{cases} \displaystyle\prod_{j=2}^{n_i} [1 - g(\boldsymbol{h}_{ij}, y_{ij})] & \text{for a completer } (d_i = n_i + 1), \\ \displaystyle\prod_{j=2}^{d-1} [1 - g(\boldsymbol{h}_{ij}, y_{ij})] g(\boldsymbol{h}_{id}, y_{id}) & \text{for a dropout } (d_i = d \leq n_i). \end{cases} \tag{12.2}$$

Several authors point to the sensitivity of this model to assumptions about the dropout process which are fundamentally not verifiable. See the discussion to Diggle and Kenward (1994) and Verbeke *et al.* (2001).

However, in the framework of a sensitivity analysis, it is useful to compare both MAR and MNAR versions of a selection model with pattern-mixture model counterparts. In the next section, we will indicate that, for the case of the Vorozole study, the substantive conclusions under both modelling frameworks are essentially the same.

## Application to the Vorozole Study

It is convenient to start under MAR, since then the measurement model and the dropout model can be fitted separately. Thereafter, the MNAR version will be considered.

Since we are modelling change versus baseline, all models are forced to pass through the origin. This is done by allowing the main covariate effects, but only through their interactions with time. The following covariates were considered for the measurement model: baseline value, treatment, dominant site, and time in months (up to a cubic time trend). Second order interactions were considered as well. Then, a backwards selection procedure was performed. For design reasons, treatment was kept

Table 12.1: *The Vorozole Study. Selection model parameter estimates (standard errors).*

| Effect | Parameter | Estimate (SE) |
|---|---|---|
| *Fixed-Effect Parameters:* | | |
| Time | $\beta_0$ | 7.78 (1.05) |
| Time∗baseline | $\beta_1$ | −0.065 (0.009) |
| Time∗treatment | $\beta_2$ | 0.086 (0.157) |
| Time$^2$ | $\beta_3$ | −0.30 (0.06) |
| Time$^2$∗baseline | $\beta_4$ | 0.0024 (0.0005) |
| *Variance Parameters:* | | |
| Random intercept | $d$ | 105.42 |
| Serial variance | $\tau^2$ | 77.96 |
| Serial association | $\lambda$ | 7.22 |
| Measurement error | $\sigma^2$ | 77.83 |

in the model in spite of its non-significance. An $F$ test for treatment effect produces a $p$ value of 0.5822. Apart from baseline, no other time-stationary covariates were kept. A quadratic time effect provided an adequate description of the time trend. Based on Figures 12.1, 12.2, 12.3 we have selected a covariance structure including random intercepts, a spatial Gaussian process and measurement error. The final model is presented in Table 12.1.

The total correlation between two measurements, one month apart, equals 0.696. The residual correlation, which remains after accounting for the random effects, is still equal to 0.491. The serial correlation, obtained by further ignoring the measurement error, equals $\rho = \exp(-1/7.22^2) = 0.981$. Fitted profiles are displayed in Figure 12.4.

For each treatment group, we obtain three sets of profiles. The fitted complete profile is the average curve that would be obtained, had all individuals been completely observed. If we use only those predicted values that correspond to occasions at which an observation was made, then the fitted incomplete profiles are obtained. The latter are somewhat above the former when the random effects are included, and somewhat

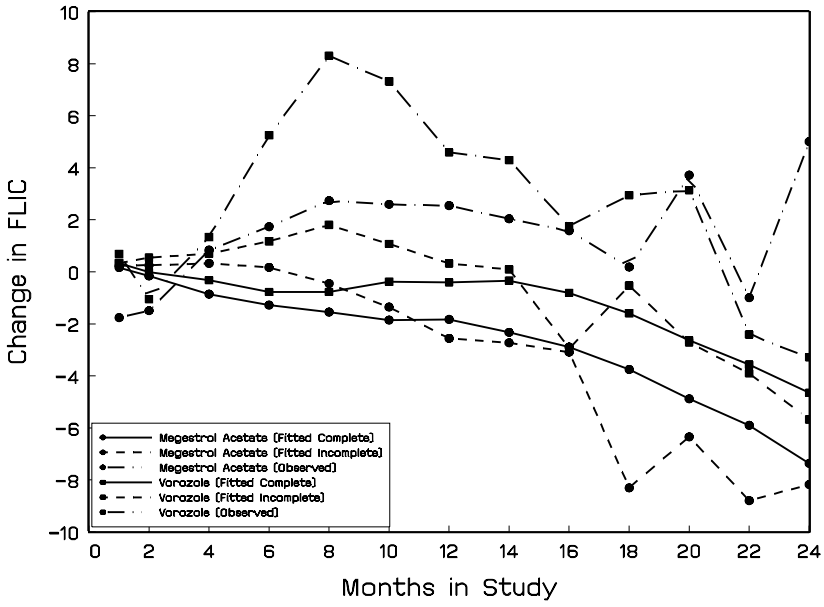


Figure 12.4: *The Vorozole Study. Fitted profiles (averaging the predicted means for the incomplete and complete measurement sequences, without the random effects).*

below when they are not, suggesting that individuals with lower measurements are more likely to disappear from the study. In addition, while the fitted complete curves are very close (the treatment effect was not significant), the fitted incomplete curves are not, suggesting that there is more dropout in the megestrol acetate arm than in the vorozole arm. This is in agreement with the dropout rate, displayed in Figure 2.4, and should not be seen as evidence of a bad fit. Finally, the observed curves, based on the measurements available at each time point, are displayed. These are higher than the fitted ones, but this should be viewed with the standard errors of the observed means in mind (see Figure 12.1).

Next, we will study factors which influence dropout. A logistic regression model, described by (12.1) and (12.2) is used. To start, we restrict attention to MAR processes, whence $\psi_d = 0$. The first model includes treatment, dominant site, baseline value, and the previous measurement but only the last two are significant, producing (estimate (SE)):

$$\text{logit}[g(\boldsymbol{h}_{ij})] = 0.080(0.341) - 0.014(0.003)\text{base}_i - 0.033(0.004)y_{i,j-1}. \qquad (12.3)$$

Diggle and Kenward (1994) and Molenberghs, Kenward, and, Lesaffre (1997) considered non-random versions of this model by including the current, possible unobserved measurement, such as in (12.1). This requires more elaborate fitting algorithms, since the missing data process is then non-ignorable, and hence (3.36) and (3.37) needs to be used. Diggle and Kenward (1994) used the simplex algorithm (Nelder and Mead, 1965), while Molenberghs, Kenward, and Lesaffre (1997) fitted their models with the EM algorithm (Dempster, Laird, and Rubin, 1977). The algorithm of Diggle and Kenward is implemented in Oswald (Smith, Robertson, and Diggle, 1996). With larger datasets such as this one, convergence can be painstakingly difficult and one has to worry about apparent convergence. Therefore, we first proceed in an alternative way. Both Diggle and Kenward (1994) and Molenberghs *et al.* (1994) observed that in informative models, dropout tends to depend on the increment, i.e., the difference between the current and previous measurements $y_{ij} - y_{i,j-1}$. Clearly, a very similar quantity is obtained as $y_{i,j-1} - y_{i,j-2}$, but a major advantage of such a model is that it fits within the MAR framework. In our case, we obtain (estimate (SE)):

$$\begin{aligned}
&\text{logit}[g(\boldsymbol{h}_{ij})] \\
&\quad = \quad 0.033(0.401) - 0.013(0.003)\text{base}_i + 0.012(0.006)y_{i,j-2} - 0.035(0.005)y_{i,j-1} \\
&\quad = \quad 0.033(0.401) - 0.013(0.003)\text{base}_i - 0.023(0.005)\frac{y_{i,j-2} + y_{i,j-1}}{2} \\
&\qquad\quad -0.047(0.010)\frac{y_{i,j-1} - y_{i,j-2}}{2}
\end{aligned} \tag{12.4}$$

indicating that both size and increment are significant predictors for dropout. We conclude that dropout increases with a decrease in baseline, in overall level of the outcome variable, as well as with a decreasing evolution in the outcome. In the next section, we will see that these conclusions can also be obtained from the pattern-mixture model building exercise. However, before we proceed, we will first consider the MNAR versions of the selection model.

Using Oswald, both dropout models (12.3) and (12.4) can be compared with their non-random counterparts, where $y_{ij}$ is added to the linear predictor. The first one becomes

$$\text{logit}[g(\boldsymbol{h}_{ij}, y_{ij})] = 0.53 - 0.015\text{base}_i - 0.076y_{i,j-1} + 0.057y_{ij} \tag{12.5}$$

while the second one becomes

$$\text{logit}[g(\boldsymbol{h}_{ij}, y_{ij})] = 1.38 - 0.021\text{base}_i - 0.0027y_{i,j-2} - 0.064y_{i,j-1} + 0.035y_{ij}. \tag{12.6}$$

Note that formal testing of dropout models (12.5) versus (12.3) and for (12.6) versus (12.4) should be approached with caution, see Molenberghs, Kenward, and Lesaffre (1997) and discussions to Diggle and Kenward (1994) by Rubin, by Little, and by Laird. Therefore, apart from the message that in this particular example, the MAR version of the selection model is sufficient to describe the data, it is a useful idea to compare the conclusions with the ones obtained from a pattern-mixture model, as was also suggested by Hogan and Laird (1997).

## 12.3 A Pattern-mixture Model Formulation

This family is based on the factorization defined in 3.26, where the conditional density of the measurements given the dropout pattern is combined with the marginal density describing the dropout mechanism. After initial mention of these models (Little and Rubin, 1987, Glynn, Laird, and Rubin, 1986) they are receiving more attention lately. It is generally believed that fitting pattern-mixture models is more honest in the sense that no *implicit* untestable assumptions are made and that they are computationally advantageous as well.

The dropout process (12.2) simplifies to $f(d_i|W_i, \boldsymbol{\psi})$ which is a, possibly covariate-corrected, model for the probability to belong to a particular pattern. Its components, $g(h_{ij})$, containing only covariates now, describe the dropout rate at each occasion.

The measurement model has to reflect dependence on dropout. In its most general form, this implies that (3.5) is replaced by

$$\begin{cases} \boldsymbol{Y_i} = X_i\boldsymbol{\beta}(d_i) + Z_i\boldsymbol{b_i} + \boldsymbol{\varepsilon_i} \\ \\ \boldsymbol{b_i} \sim N(\mathbf{0}, D(d_i)), \\ \\ \boldsymbol{\varepsilon_i} \sim N(\mathbf{0}, \Sigma_i(d_i)). \end{cases} \tag{12.7}$$

Thus, the fixed effects as well as the covariance parameters are allowed to change with dropout pattern and a priori no restrictions are placed on the structure of this change.

Model family (12.7) contains under identified members since it describes the full set of measurements in pattern $d_i$, even though there are not measurements after occasion $d_i - 1$. Little (1993, 1994) advocated the use of identifying restrictions which works well in relatively simple settings. Molenberghs *et al.* (1998) proposed a particular

set of restrictions for the monotone case which correspond to MAR. To avoid this problem, simplified (identified) models can be considered. The advantage is that the number of parameters decreases, which is generally an issue with pattern-mixture models. Hogan and Laird (1997) noted that in order to estimate the large number of parameters in general models, one has to make the awkward requirement that each dropout pattern is sufficiently "filled", in other words one has to require large numbers of dropouts. Note however that simplified models, qualified as "assumption rich" by Sheiner, Beale, and Dunne (1997), are also making untestable assumptions and therefore illustrate that even pattern-mixture models do not provide a free lunch. A main advantage however is that the need of assumptions and their implications are more obvious. For example, it is not possible to assume an unstructured time trend in incomplete patterns, except if one restricts attention to the time range from onset until dropout. In contrast, assuming a linear time trend allows estimation in all patterns containing at least two measurements.

In general, we distinguish between two types of simplifications to identify pattern-mixture models. First, trends can be restricted to functional forms supported by the information available within a pattern. The linear time trend discussed earlier is an example. Secondly, one can let the parameters vary across patterns in a parametric way. Thus, rather than estimating a separate time trend in each pattern, one could assume that the time evolution is unstructured in each pattern, but parallel across patterns. The available data can be used to assess whether such simplifications are supported within the time ranges for which there is information. Using the so-obtained profiles past the time of dropout still requires extrapolation.

## Application to Vorozole Study

In analogy with the exploration in the selection model context, it is natural to explore the data from a pattern-mixture point of view. To this end, plots per dropout pattern can be constructed. Figure 12.5 displays the averaged profiles per pattern.

Figure 12.5 clearly shows that pattern-specific profiles are of a quadratic nature with in most cases a sharp decline prior to dropout. Note that this is in line with the fitted dropout mechanism (12.4). Therefore, this feature needs to be reflected in the pattern-mixture model. In analogy with our selection model, the profiles are forced to pass through the origin. This is done by allowing only time main effect and interactions of other covariables with time in the model.

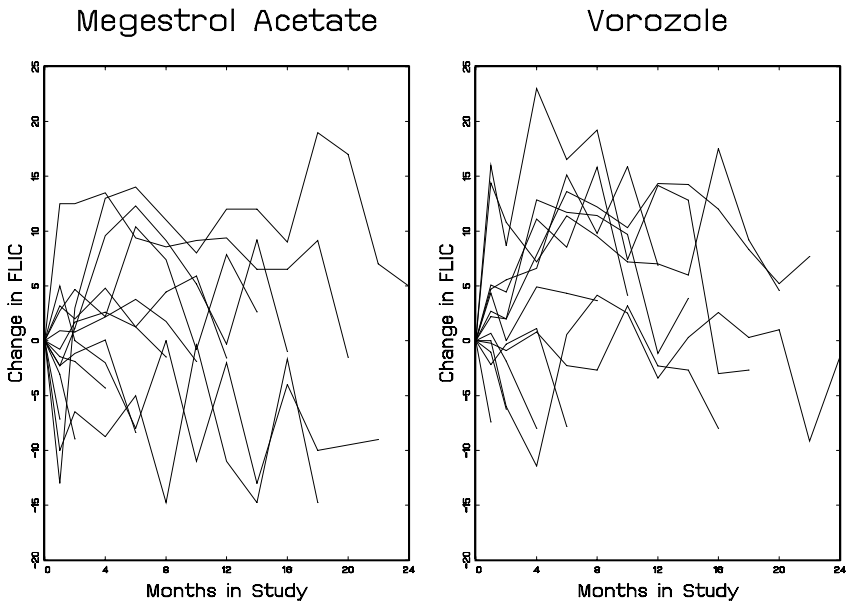


Figure 12.5: *The Vorozole Study. Mean profiles, per dropout pattern, grouped per treatment arm.*

The most complex pattern-mixture model we consider includes a different parameter vector for each of the observed patterns. This is done by including the interaction of all effects in the model with *pattern*, a factor variable calculated as 2+ the number of observations after baseline. We then proceed by backward selection in order to simplify the model. First, we found that the covariance structure is common to all patterns, encompassing random intercept, a serial exponential process, and measurement error.

For the fixed effects we proceeded as follows. A backward selection procedure, starting from a model that includes a main effect of time and time$^2$, as well as interactions of time with baseline value, treatment effect, dominant site and pattern, and the interaction of pattern with time$^2$. This procedure revealed main effects of time and time$^2$, as well as interactions of time with baseline value, treatment effect, and pattern, and the interaction of pattern with time$^2$. This reduced model can be found in Table 12.2. Note the sign difference of the time by treatment interaction between Tables 12.1 and 12.2. Of course, these models are drastically different and moreover the effects are not significant.

Table 12.2: *The Vorozole Study. Parameter estimates and standard errors for the first pattern-mixture model.*

| Effect | Estimate (SE) | Effect | Estimate (SE) |
|---|---|---|---|
| *Fixed-effect Parameters:* | | | |
| Time | 4.671 (0.844) | $\text{Time}^2$ | −0.034 (0.029) |
| Time∗Pattern 1 | −8.856 (2.739) | $\text{Time}^2$∗Pattern 1 | |
| Time∗Pattern 2 | −0.796 (2.958) | $\text{Time}^2$∗Pattern 2 | −1.918 (1.269) |
| Time∗Pattern 3 | −1.959 (1.794) | $\text{Time}^2$∗Pattern 3 | −0.145 (0.365) |
| Time∗Pattern 4 | 1.600 (1.441) | $\text{Time}^2$∗Pattern 4 | −0.541 (0.197) |
| Time∗Pattern 5 | 0.292 (1.295) | $\text{Time}^2$∗Pattern 5 | −0.107 (0.133) |
| Time∗Pattern 6 | 1.366 (1.035) | $\text{Time}^2$∗Pattern 6 | −0.181 (0.080) |
| Time∗Pattern 7 | 1.430 (1.045) | $\text{Time}^2$∗Pattern 7 | −0.132 (0.071) |
| Time∗Pattern 8 | 1.176 (1.025) | $\text{Time}^2$∗Pattern 8 | −0.118 (0.061) |
| Time∗Pattern 9 | 0.735 (0.934) | $\text{Time}^2$∗Pattern 9 | −0.083 (0.049) |
| Time∗Pattern 10 | 0.797 (1.078) | $\text{Time}^2$∗Pattern 10 | −0.078 (0.055) |
| Time∗Pattern 11 | 0.274 (0.989) | $\text{Time}^2$∗Pattern 11 | −0.023 (0.046) |
| Time∗Pattern 12 | 0.544 (1.087) | $\text{Time}^2$∗Pattern 12 | −0.026 (0.049) |
| Time∗Baseline | −0.031 (0.004) | Time∗Treatment | −0.067 (0.166) |
| *Variance Parameters:* | | | |
| Random intercept | 78.45 | | |
| Serial variance | 95.38 | | |
| Serial association | 8.85 | | |
| Measurement error | 73.77 | | |

As was the case with the selection model in Table 12.1, treatment effect is non-significant. Indeed, a single degree of freedom $F$ test yields a $p$ value of 0.69. Note

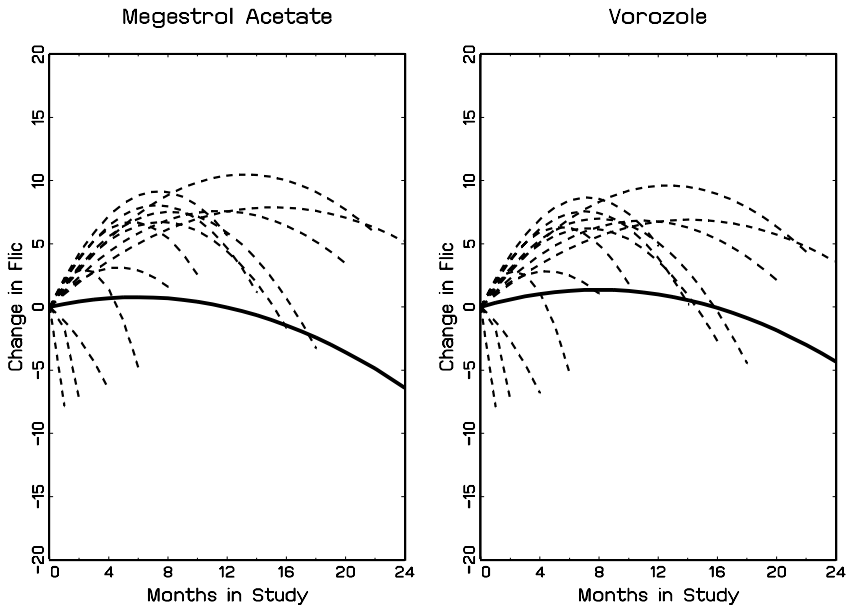


Figure 12.6: *The Vorozole Study. Fitted selection (solid line) and first pattern-mixture models (dashed lines).*

that such a test is possible since treatment effect does not interact with pattern, in contrast to the model which we will describe later. The fitted profiles are displayed in Figure 12.6. We observe that the profiles for both arms are very similar. This is due to the fact that treatment effect is not significant but perhaps also because we did not allow a more complex treatment effect. For example, we might consider an interaction of treatment with the square of time and, more importantly, an treatment effect which is pattern-specific. Some evidence for such an interaction is seen in Figure 12.5.

Our second, expanded model, allowed for up to cubic time effects, the interaction of time with dropout pattern, dominant site, baseline value and treatment, as well as their two- and three-way interactions. After a backward selection procedure, the effects included are time and time$^2$, the two-way interaction of time and dropout pattern, as well as three factor interactions of time and dropout pattern with (1) baseline, (2) group, and (3) dominant site. Finally, time$^2$ interacts with dropout pattern and with the interaction of baseline and dropout pattern. No cubic time effects were necessary, which is in agreement with the observed profiles in Figure 12.5. The parameter estimates of this model are displayed in Table 12.3. The model is graphically

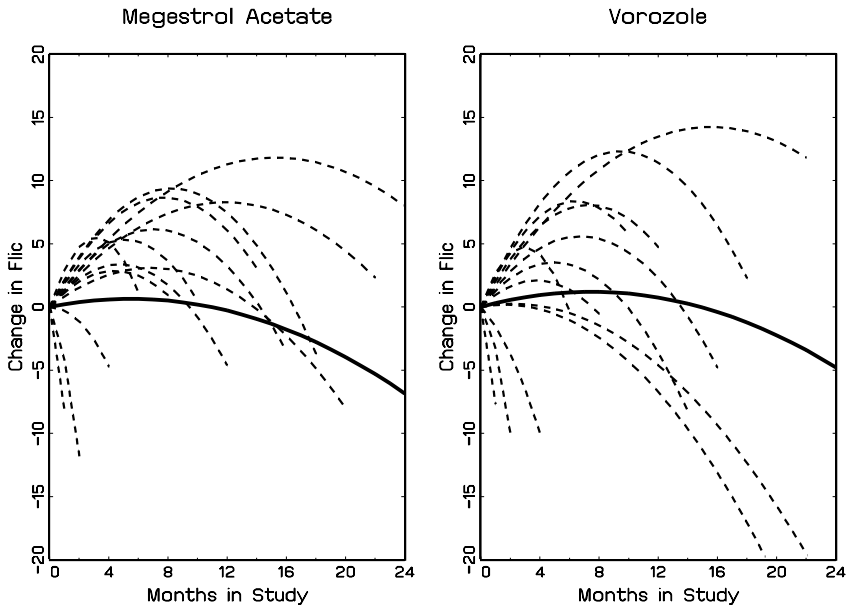


Figure 12.7: *The Vorozole Study. Fitted selection (solid line) and second pattern-mixture models (dashed lines).*

represented in Figure 12.7.

Because a pattern-specific parameter has been included, we have several options for the assessment of treatment. Since there are 13 patterns (remember we cut off the patterns at 2 years), one can test the global hypothesis, based on 13 degrees of freedom, of no treatment effect. We obtain $F = 1.25$, producing $p = 0.24$, indicating that there is no overall treatment effect. Each of the treatment effects separately is at a non-significant level. Alternatively, the *marginal* effect of treatment can be calculated, which is the weighted average of the pattern-specific treatment effects, with weights given by the probability of occurrence of the various patterns. Its standard error is calculated using a straightforward application of the delta method. This effect(SE) is equal to $-0.286(0.288)$ producing a $p$ value of 0.32, which is still non-significant.

While the 13 d.f. test is useful and appealing from a stratification point of view, it should be realized that one stratifies for a post-randomization variable and that therefore caution should be used. Such a test can play a role in a sensitivity assessment, but the marginal treatment effect, even under a PMM model, should not be omitted.

Table 12.3: *The Vorozole Study. Parameter estimates and standard errors for the second pattern-mixture model (part I). Each column represents an effect, for which a main effect is given, as well as interactions with the dropout patterns.*

| Fixed-effect parameters [estimate (SE)] | | | |
|---|---|---|---|
| Effect | Time | Time∗Baseline | $\text{Time}^2$ |
| Main | 5.468 (5.089) | −0.034 (0.040) | −0.271 (0.206) |
| Pattern 1 | 7.616 (21.908) | −0.119 (0.175) | |
| Pattern 2 | 44.097 (17.489) | −0.440 (0.148) | −18.632 (7.491) |
| Pattern 3 | 22.471 (10.907) | −0.218 (0.089) | −5.871 (2.143) |
| Pattern 4 | 10.578 (9.833) | −0.055 (0.079) | −1.429 (1.276) |
| Pattern 5 | 14.691 (8.424) | −0.123 (0.069) | −1.571 (0.814) |
| Pattern 6 | 7.527 (6.401) | −0.061 (0.052) | −0.827 (0.431) |
| Pattern 7 | −12.631 (7.367) | 0.086 (0.058) | 0.653 (0.454) |
| Pattern 8 | 14.827 (6.467) | −0.126 (0.053) | −0.697 (0.343) |
| Pattern 9 | 5.667 (6.050) | −0.049 (0.049) | −0.315 (0.288) |
| Pattern 10 | 12.418 (6.473) | −0.093 (0.051) | −0.273 (0.296) |
| Pattern 11 | 1.934 (6.551) | −0.022 (0.053) | −0.049 (0.289) |
| Pattern 12 | 6.303 (6.426) | −0.052 (0.050) | −0.182 (0.259) |
| Effect | $\text{Time}^2$∗Baseline | Time∗Treatment | |
| Main | 0.002 (0.002) | | |
| Pattern 1 | | 0.445 (5.095) | |
| Pattern 2 | 0.1458 (0.0644) | 0.867 (1.552) | |
| Pattern 3 | 0.0484 (0.0178) | −1.312 (0.808) | |
| Pattern 4 | 0.0080 (0.0107) | −0.249 (0.686) | |
| Pattern 5 | 0.0127 (0.0069) | −0.184 (0.678) | |
| Pattern 6 | 0.0058 (0.0036) | 0.527 (0.448) | |
| Pattern 7 | −0.0065 (0.0038) | 0.782 (0.502) | |
| Pattern 8 | 0.0052 (0.0029) | −0.809 (0.464) | |
| Pattern 9 | 0.0021 (0.0023) | −0.080 (0.443) | |
| Pattern 10 | 0.0016 (0.0024) | 0.331 (0.579) | |
| Pattern 11 | 0.0003 (00024) | −0.679 (0.492) | |
| Pattern 12 | 0.0015 (0.0021) | 0.433 (0.688) | |
| Pattern 13 | | −1.323 (0.706) | |

In summary, we obtain a non-significant treatment effect from all our different models, which gives more weight to this conclusion.

Given the fact that all patients are followed up until death or progression, dropout is an important concern. Since there is an intimate link between death/progression (seen as dropout) and quality of life (our response of interest), a careful assessment of the relation between both is important. Since quality of life is largely a consequence of the evolution of the disease, it is not unnatural to condition on an important aspect of it, i.e., death/progression which in our case is done by conditioning on the response pattern. This provides extra motivation to consider the PMM model in addition to the SEM model. Within such an homogenous group, the behavior is completely in line with intuition: profiles rise, reach a plateau, and then start decreasing, whereafter they drop out, as clearly seen in Figure 12.4. It can be seen as an advantage of PMM models that such representations, which are very insightful for the clinical researcher, are easily obtained. The SEM model on the other hand, concentrates more on the question which is of direct interest to the clinical trialist: the comparison of the treatment effect between the two groups, averaged over, or corrected for other covariates. In general, such an approach might help assess which risk factors are associated with dropout or dropout-pattern specific profiles. In this case, apart from treatment arm, mainly dominant tumor site is retained as a predictor variable.

## 12.4 Concluding Remarks

In this Chapter we have concentrated on total FLIC (i.e., change of the score versus baseline), a quality of life score measured in a multi-centric two arm study in postmenopausal women suffering from metastatic breast cancer. Since virtually all patients were followed up until disease progression or death, the amount of dropout is large. A very large group of patients drops out after just a couple of months.

While classically only selection models are fitted, pattern-mixture models can be seen as a viable alternative. We analyzed the data using both, leading to a sensitivity analysis. More confidence in the results can be gained if both models lead to similar conclusions.

The average profile in the selection model depends on the baseline value, as well as on time. The latter effect is mildly quadratic. There is no evidence for a treatment difference. However, it should be noted that the average profile found is the one that *would* have been observed, had no subjects dropped out, and under the additional as-

sumption that the MAR assumption is correct. Fitting non-random dropout models, in the sense of Diggle and Kenward (1994) is possible, but computationally difficult for a fairly large trial like this one. A separate study of the dropout mechanism revealed that dropout increases with three elements: (1) an unfavorable baseline score, (2) an unfavorable value at the previous month, as well as (3) an unfavorable change in value from the penultimate to the last obtained value.

A pattern-mixture model is fitted by allowing at first a completely separate parameter vector for each observed dropout pattern, which is then simplified by using standard model selection procedures, by considering whether effects are common to all patterns. A first pattern-mixture model features a common treatment effect, of which the assessment is then straightforward. A second model includes a separate treatment effect for each dropout pattern. This leads to two distinct test. The first one tests for equality of the whole treatment vector to be zero. The second one first calculates the marginal treatment effect from the vector of effects, by composing a weighted sum, where the weights are the multinomially estimated probabilities of the various patterns. In all cases, there is no treatment effect. However, a graphical display of the fitted profiles per pattern is enlightening, since it clearly confirms the trend detected in the selection models, that patients tend to drop out when their quality of life score is declining. Since this feature is usually coupled to an imminent progression or death, it should not come as a surprise. An important advantage of pattern-mixture models is that fitting them is more straightforward than non-random selection models. The additional calculations needed for the marginal treatment effect and its associated precision can be done straightforwardly using the delta method.

# Chapter 13

# Discussion and Further Research

When the biostatistician and the clinician are designing a new clinical study, they should have good information on the psychometric properties of the measurements that are planned to be done in clinical studies. Indeed, performing clinical studies is resource demanding and therefore it would irresponsible to built upon unreliable measurements. The strategy must be to use a scale, or measurements in general, which has been validated before and for which *reliability* (test-retest, inter-rater and internal consistency) and *validity* (content, construct and criterion) are established. The psychometric validation is usually done on a selected small sample from the population for which the scale is intended to be used. If the population of the trial is different, a new battery of reliability and validity testing might be warranted.

In this thesis, most of the focus was on quantifying reliability. Reliability reflects on the amount or measurement error which is inherent in any measurement, and therefore, reliability also reflects the extent to which a measurement instrument can differentiate among individuals. As differentiation between subjects randomized to different treatments is core business in clinical trials, it is obvious that reliability is of utmost importance! Generalizability Theory, as natural extension of reliability and Classical Theory can therefore be an extremely valuable tool to assess which factors influence reliability. As noted by Shalvelson, Webb, and Rowley (1989), "GT

is not widely applied in psychological research because of its formidable mathematical development". As noted by Dunn (1989), there is also a need for larger sample sizes, otherwise, many of the estimates of variance components will be practically worthless.

In this work, we proposed a framework to study *trial or population specific reliability* and *generalizability* based on longitudinal biomedical trial data. The goal is to use clinical trial data at hand and to evaluate psychometric properties of the measurement. The intention is certainly not to replace up-front validity and reliability testing but to stimulate post hoc evaluation on the performance of the scale or any other measurement of interest. The advantage is that clinical trialists can learn before embarking on new trials in a similar population whether they feel comfortable using the same scale again. These methods can also deliver a population-trial specific measure for reliability in case there is a need to confirm earlier reliability testing results; regulatory authorities might question reliability of the scale in the specific trial population. The measurements in clinical trials are often 'unstable', in a psychometric sense, due to present treatment and time effect. In contrast, in the classical theory setting, reliability testing is always done on patients in a steady state condition, resulting in 'parallel measurements' within the patients. Therefore, one of the biggest challenges was to find a way to extract these effects and to make the bridge to the classical reliability coefficient, a well known and established concept in psychometrics. This must not stop us however of trying to improve and learn for the future. We should indeed look beyond just merely the treatment difference and its *holly* p-value, but also look into variance and correlation. One big advantage that pharmaceutical companies have is that they often are sitting on a gold mine of clinical data as each regulatory submission requires at least 1000 patients. Here, the sample size is not an issue to study generalizability.

Investigators in the mental disorders traditionally have been more concerned with the psychometric properties of their measures than have their colleagues in other medical specialities. It is obvious however that also other types of measurements can benefit from studying its psychometric properties. For instance, are there subgroups of subjects for which the plasma RNA viral load for HIV is less reliable?

While the thesis focused on test-retest reliability, interrater reliability was not directly studied. This is natural as patients tend to be evaluated by only one investigator in the clinical trial setting. In GT however, we showed that impact of country can be evaluated using clinical trial data, similarly, also the impact of investigator could be assessed. Internal consistency was not addressed in this thesis. In my expe-

rience, internal consistency is assessed more regularly in clinical trials. This is done by the calculation of Cronbach's alpha coefficient using baseline measurements. It would be interesting however to study how this correlation between the items of a subscale evolves over time in future research.

The study of construct and content validity is difficult in the context of clinical trials. This requires a more fundamental evaluation to evaluate whether the scale measures what it purports to measure. The principles of criterion validity however can be applied to the clinical trial setting. Joint modelling in the GLMM framework and the techniques developed in surrogate maker validation offer more elaborate techniques and can give more insight then the simple evaluation of the Pearson correlation coefficient. This can be valuable to assess correlations between measurements and to evaluate whether changes in certain scales are correlated with more clinical tangible measurements. Again, as for reliability, this is not something we should only do to evaluate a new scale and its correlation with a more established golden standard. These techniques could also be used to investigate for instance the correlation between the emergence of mutations and plasma RNA viral load in clinical trials in HIV.

The common theme of this thesis was psychometric validation. The essence actually is the study of *correlation*, the correlation within a patient and between patients. It was clear that in case of Gaussian distributed data it was easy to obtain marginal correlation coefficients in a straightforward way via the LMM. This is related to the conjugacy between the normal distribution of random effects and the normal distribution of the measurements and the natural identity link. For count data, based on the work of Molenberghs, Verbeke, and Demétrio (2008), a closed formula was derived to calculate the correlation coefficient. Also here, there is conjugacy between Poisson distributed outcomes and gamma distributed random effects, even when normal random effects are additionally present; these, together with the natural link, the logarithmic one, produce the negative binomial model. In the binary case, however, in spite of conjugacy between binomial and beta distributions, leading to the well-known beta-binomial model, this property is destroyed when normal random effects are additionally present. This partly explains why there is no closed form for the variance-covariance functions, thus necessitating approximate correlation calculation. This provides motivation for further research for the binary case and to derive closed formulas for correlation based on the probit model, where there is conjugacy through the probit link. One more route for further research is to evaluate whether another approximation can provide an alternative to the Taylor-expansion employed in current

thesis.

The models used to estimate reliability, generalizability and validity are similar to the models used to estimate treatment effects. Fully longitudinal data were used instead of paired data to calculate the ICC. Since the methods are likelihood-based, they are valid under the broad assumption of missingness at random, whereby missingness depends on observed outcomes and covariates but, given these, not further on unobserved outcomes. As discussed in Chapter 12, different approaches and sensitivities should be done in case of incomplete data when longitudinal models are used to evaluate treatment effects. The same is true when we use these longitudinal models to estimate reliability, generalizability and validity. Further research is necessary on this topic.

# Appendix A

**Explicit Expressions for Components of (9.11)**

For notational simplicity, write $\pi_0 \equiv V_{pdt}(0)$ and $\pi_0' \equiv V_{pd't}(0)$. Further, note that $\boldsymbol{\eta} = (\eta_1 = \boldsymbol{x}_1'\boldsymbol{\beta}, \eta_2 = \boldsymbol{x}_2'\boldsymbol{\beta})'$, $\boldsymbol{\sigma}^2 = (\sigma_1^2, \sigma_2^2, \phi)'$, and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K, \phi)'$, where the $\lambda_k$ are the parameters figuring in the models for $\sigma_1^2$ and $\sigma_2^2$. We then obtain, for the first factor on the right hand side:

$$\frac{\partial(\boldsymbol{\eta}, \boldsymbol{\sigma}^2)}{\partial(\boldsymbol{\beta}, \boldsymbol{\lambda})} = \begin{pmatrix} \boldsymbol{x}_1' & 0 & 0 & 0 \\ 0 & \boldsymbol{x}_2' & 0 & 0 \\ 0 & \frac{\partial \sigma_1^2}{\partial \lambda_1} & \frac{\partial \sigma_2^2}{\partial \lambda_1} & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \frac{\partial \sigma_1^2}{\partial \lambda_K} & \frac{\partial \sigma_2^2}{\partial \lambda_K} & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

The second factor equals:

$$\frac{\partial(\pi_0, \pi_0', \sigma_1^2, \sigma_2^2, \phi)}{\partial(\eta_1, \eta_2, \sigma_1^2, \sigma_2^2, \phi)} = \left(\begin{array}{cc|c} \pi_0(1-\pi_0) & 0 & \mathbf{0}_{2,3} \\ 0 & \pi_0'(1-\pi_0') & \\ \hline \multicolumn{2}{c|}{\mathbf{0}_{3,2}} & I_{3,3} \end{array}\right).$$

To establish the elements of the third factor, first write $\rho = \sigma_1^2 \kappa_0^{1/2} \kappa_1^{-1/2} \kappa_2^{-1/2}$, with $\kappa_0 = \pi_0(1-\pi_0)\pi_0'(1-\pi_0')$, $\kappa_1 = \pi_0(1-\pi_0)\sigma_2^2 + \phi$, and $\kappa_2 = \pi_0'(1-\pi_0')\sigma_2^2 + \phi$. It then follows that

$$\begin{aligned}
\frac{\partial \rho}{\partial \pi_0} &= \frac{1}{2}\sigma_1^2 \kappa_0^{-1/2} \kappa_1^{-3/2} \kappa_2^{-1/2} (1-2\pi_0)\pi_0'(1-\pi_0')\phi, \\
\frac{\partial \rho}{\partial \pi_0'} &= \frac{1}{2}\sigma_1^2 \kappa_0^{-1/2} \kappa_1^{-1/2} \kappa_2^{-3/2} (1-2\pi_0')\pi_0(1-\pi_0)\phi, \\
\frac{\partial \rho}{\partial \sigma_1^2} &= \kappa_0^{1/2} \kappa_1^{-1/2} \kappa_2^{-1/2}, \\
\frac{\partial \rho}{\partial \sigma_2^2} &= -\frac{1}{2}\sigma_1^2 \kappa_0^{1/2} \kappa_1^{-3/2} \kappa_2^{-3/2} \left\{2\kappa_0 \sigma_2^2 + [\pi_0(1-\pi_0) + \pi_0'(1-\pi_0')]\,\phi\right\}, \\
\frac{\partial \rho}{\partial \phi} &= -\frac{1}{2}\sigma_1^2 \kappa_0^{1/2} \kappa_1^{-3/2} \kappa_2^{-3/2} \left\{\pi_0(1-\pi_0) + \pi_0'(1-\pi_0')]\sigma_2^2 + 2\phi\right\}.
\end{aligned}$$

# Appendix B

**SAS Implementation**

All data analyses have been conducted using the SAS procedure GLIMMIX to obtain parameter estimates and measures of precision. The correlation quantities derived have been obtained using user-defined code written in the SAS procedure IML. Here, we will provide some example code and a brief discussion. A full set of programs and output can be obtained from the authors' web pages.

First, a SAS program, using the procedure GLIMMIX, for the model of Section 9.2.1, with a random intercept and a scale parameter $\Phi$, taking into account treatment, time, as well as their interaction, is as follows:

```
ods output ParameterEstimates=datagen.model1estimates
           covparms=datagen.model1covparms
           covb=datagen.model1covb
           asycov=datagen.model1asycov;

proc glimmix data=cgi3 noclprint=26 asycov;
   class treat id  xtime_c;
   title 'Overall model with random patient to estimate
          overall reliability with scale parameter';
   model cgi_resp(event='1') = xtime_c|treat
         / covb dist=binary link=logit solution;
   random intercept / subject=id;
   random _residual_;
   run;
```

The coding is self-explanatory, in the sense that the fixed-effects structure involves time, treatment, and their interaction, and a random intercept is then added. The RANDOM `_residual_` statement ensures the scale, or overdispersion, parameter is included.

Second, to estimate the correlation and its associated standard error, SAS IML can be used. Let us exemplify this for the correlation between the first and second occasions, within the risperidone group.

```
proc iml;
   use datagen.model1estimates;
```

```
read all var {estimate} into beta;
use datagen.model1covparms;
read all var {estimate} into sigma;
use datagen.model1asycov;
read all var {CovP1,CovP2} into asycov;
close datagen.model1asycov;
use datagen.model1covb;
read all var {Col1,Col2,Col3,Col4,Col5,Col6,Col7,Col8,Col9,Col10,Col11,
              Col12,Col13,Col14,Col15,Col16,Col17,Col18} into covb;
close datagen.model1covb;
varint=sigma[1];
scale=sigma[2];
covarmatrix=block(covb,asycov);
zero=J(18,1,0);

*              Time     RX   Contr*Time Ris*time;
*         I  1 2 4 6 8  C R  1 2 4 6 8  1 2 4 6 8;
bw1ris=T({1  1 0 0 0 0  0 1  0 0 0 0 0  1 0 0 0 0});
bw2ris=T({1  0 1 0 0 0  0 1  0 0 0 0 0  0 1 0 0 0});
bw4ris=T({1  0 0 1 0 0  0 1  0 0 0 0 0  0 0 1 0 0});
bw6ris=T({1  0 0 0 1 0  0 1  0 0 0 0 0  0 0 0 1 0});
bw8ris=T({1  0 0 0 0 1  0 1  0 0 0 0 0  0 0 0 0 1});

bw1con=T({1  1 0 0 0 0  1 0  1 0 0 0 0  0 0 0 0 0});
bw2con=T({1  0 1 0 0 0  1 0  0 1 0 0 0  0 0 0 0 0});
bw4con=T({1  0 0 1 0 0  1 0  0 0 1 0 0  0 0 0 0 0});
bw6con=T({1  0 0 0 1 0  1 0  0 0 0 1 0  0 0 0 0 0});
bw8con=T({1  0 0 0 0 1  1 0  0 0 0 0 1  0 0 0 0 0});

pi1ris=exp(T(bw1ris)*beta)/(1+exp(T(bw1ris)*beta));
pi2ris=exp(T(bw2ris)*beta)/(1+exp(T(bw2ris)*beta));

k0_1_2_ris=(pi1ris*(1-pi1ris)*pi2ris*(1-pi2ris));
k1_1_2_ris=pi1ris*(1-pi1ris)*varint+scale;
k2_1_2_ris=pi2ris*(1-pi2ris)*varint+scale;

* correlation following equation (23)
r1_2_ris=(sqrt(k0_1_2_ris)*varint)/(sqrt(k1_1_2_ris)*sqrt(k2_1_2_ris));
```

```
D_rho_sigma_1_1_2_ris=sqrt(k0_1_2_ris)/(sqrt(k1_1_2_ris)*sqrt(k2_1_2_ris));
D_rho_sigma_2_1_2_ris=-0.5*varint*sqrt(k0_1_2_ris)*(k1_1_2_ris*k2_1_2_ris)**(-1.5)
                     *(2*k0_1_2_ris*varint
                     +scale*(pi1ris*(1-pi1ris)+pi2ris*(1-pi2ris)));
D_rho_po_1_2_ris=0.5*varint*(k0_1_2_ris*k2_1_2_ris)**(-0.5)
                *(k1_1_2_ris)**(-1.5)*(scale*pi2ris*(1-pi2ris)*(1-2*pi1ris));
D_rho_poa_1_2_ris=0.5*varint*(k0_1_2_ris*k1_1_2_ris)**(-0.5)
                 *(k2_1_2_ris)**(-1.5)*(scale*pi1ris*(1-pi1ris)*(1-2*pi2ris));
D_rho_phi_1_2_ris=-0.5*varint*sqrt(k0_1_2_ris)
                 *(k1_1_2_ris*k2_1_2_ris)**(-1.5)*(k1_1_2_ris+k2_1_2_ris);
F_ris_1_2=D_rho_po_1_2_ris*pi1ris*(1-pi1ris)*T(bw1ris)+D_rho_poa_1_2_ris*pi2ris
          *(1-pi2ris)*T(bw2ris) ||
D_rho_sigma_1_1_2_ris+D_rho_sigma_2_1_2_ris||D_rho_phi_1_2_ris ;

se_ris_1_2=sqrt(F_ris_1_2*covarmatrix*T(F_ris_1_2));

print "Estimated Correlation matix-Risperdal" cor_ris[format=8.2];
print "Standard error correlations-Risperdal" se_ris[format=8.2];
print "Estimated Correlation matix Control" cor_con[format=8.2];
print "Standard error correlations-Control" se_con[format=8.2];
quit;
```

The IML code is a little tedious, but otherwise reasonably straightforward. Different models require slightly modified coding of the GLIMMIX procedure, while the IML code needs adaptation as well.

The model of Section 9.2.2 requires replacement of two statement in the GLIMMIX code:

```
class treat id  xtime_c country;
model cgi_resp(event='1') = xtime_c|treat country
      / covb dist=bin link=logit solution;
```

The IML code is more extensive, since a specific contribution for each country is calculated.

The analyses of Section 9.2.3, by country on the one hand and using leave-on-country-out on the other hand, are done by applying macros:

```
%bycountry(inds=cgi3,land='ARG');
```

```
%countryout(land="ARG");
```

The program for Section 9.2.4, with country as random effect, is the same as the program for Section 9.2.1, i.e., the first one presented in this appendix, with simply the following statement added:

```
random intercept / subject=country;
```

in addition to the two RANDOM statements already present.

# References


Armitage, P. and Colton, T. (1998) *Encyclopedia of Biostatistics*. New York: Wiley.

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

Alonso, A., Geys, H., Molenberghs, G., and Vangeneugden, T. (2002) Investigating the criterion validity of psychiatric symptom scales using surrogate marker validation methodology. *Journal of Biopharmaceutical Statistics*, **12**, 161–179.

Bahadur, R.R. (1961) A representation of the joint distribution of responses to n dichotomous items. *Studies in item analysis and prediction,* H. Solomon (Ed.). Stanford mathematical studies in the social sciences VI. Stanford, CA: Stanford University Press.

Baker, S.G. and Kramer, B.S. (2003) A perfect correlate does not make a surrogate. *BioMed Central Medical Reseach Methodology*, **3**, 16.

Baker, S.G. and Laird, N.M. (1988) Regression analysis for categorical variables with outcome subject to non-ignorable non-response. *Journal of the American Statistical Association*, **83**, 62–69.

Baker, S.G., Rosenberger, W.F., and DerSimonian, R. (1992) Closed-form estimates for missing counts in two-way contingency tables. *Statistics in Medicine*, **11**, 643–657.

Bartholomew D.J. (1996) *Statistical Approach to Social Measurements*. London: Arnold.

Bartholomew D.J. and Knott M. (1999) *Latent Variable Models and Factor Analysis*. London: Arnold.

Bartko J.J. (1966) The intraclass correlation coefficient as a measure of reliability. *Psychological Reports*, **19**, 3–11.

Bell, M., Milstein, R., Beam-Goulet, J., Lysaker, P., and Cicchetti, D. (1992) The Positive and Negative Syndrome Scale and the Brief Psychiatric Rating Scale: Reliability, comparability, and predictive validity. *Journal of Nervous and Mental Disease* **180**, 723–728.

Blin, O., Azorin, J.M., and Bouhours, P. (1996) Anti psychotic and anxiolytic properties of risperidone, haloperidol and methotrimeprazine in schizophrenic patients. *Journal of Clinical Psychopharmacology*, **16**, 38–44.

Booth, J.G., Casella, G., Friedl, H., and Hobert, J.P. (2003) Negative binomial loglinear mixed models. *Statistical Modelling*, **3**, 179–181.

Brennan, R.L. (1992) *Elements of Generalizability Theory.* Iowa City, IA: ACT Publications.

Breslow, N. (1984) Extra-Poisson variation in log-linear models. *Applied Statistics*, **33**, 38–44.

Breslow, N.E. and Clayton, D.G. (1993) Approximate inference in generalized linear mixed models. *Journal of American Statistical Association*, **88**, 9–25.

Burzykowski, T., Molenberghs, G., and Buyse, M. (2005) *The Evaluation of Surrogate Endpoints.* New York: Springer.

Buyse, M. and Molenberghs, G. (1998) The validation of surrogate endpoints in randomized experiments. *Biometrics*, **54**, 1014–1029.

Buyse, M., Molenberghs, G., Burzykowski, T., Renard, D., and Geys, H. (2000) The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics*, **1**, 49–67.

Carmines, E.G. and Zeller, R.A. (1979) *Reliability and validity assessment.* Thousand Oaks, CA: Sage.

Conaway, M.R. (1992) The analysis of repeated categorical measurements subject to nonignorable nonresponse. *Journal of the American Statistical Association*, **87**, 817–824.

Conaway, M.R. (1993) Non-ignorable non-response models for time-ordered categorical variables. *Applied Statistics*, **42**, 105–115.

Cook, R.D. (1986) Assessment of local influence. *Journal of the Royal Statistical Society, Series B*, **48**, 133–169.

Cronbach, L.J. (1951) Coefficient alpha and the internal structure of tests. *Psychometrika*, **16**, 297–334.

Cronbach, L.J., Rajaratnam, N., and Gleser, G.C. (1963) Theory of Generalizability: a liberalization of reliability theory. *British Journal of Statistical Psychology*, **16**, 137–163.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, **39**, 1–38.

Diggle, P.J. (1989) Testing for random dropouts in repeated measurement data. *Biometrics*, **45**, 1255–1258.

Diggle, P.D. and Kenward, M.G. (1994) Informative dropout in longitudinal data analysis (with discussion). *Applied Statistics*, **43**, 49–93.

Diggle, P.J., Liang, K.Y., and Zeger, S.L. (1994) *Analysis of Longitudinal Data*. Clarendon Press: Oxford.

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

Dunn, G. (1989) *Design and Analysis of Reliability Studies: The statistical evaluation of measurement errors*. Oxford University Press: New York.

Dunn, G. (2000) *statistics in psychiatry*. Edward Arnold: London.

Faught, E., Wilder, B.J., Ramsay, R.E., Reife, R.A., Kramer, L.D., Pledger, G.W., and Karim, R.M. (1996) Topiramate placebo-controlled dose-ranging trial in refractory partial epilepsy using 200-, 400-, and 600-mg daily dosages. *Neurology*, **46**, 1684–1690.

Fleiss, J.L. (1986) *Design and Analysis of Clinical Experiments*. New York: John Wiley & Sons.

Freedman, L.S., Graubard, B.I., and Schatzkin, A. (1992) Statistical validation of intermediate endpoints for chronic diseases. *Statistics in Medicine*, **11**, 167–178.

Glynn, R.J., Laird, N.M., and Rubin, D.B. (1986) Selection modelling versus mixture modelling with nonignorable nonresponse, In *Drawing Inferences from Self-Selected Samples*, Ed. H. Wainer, pp. 115–142. New York: Springer Verlag.

Goss, P.E., Winer, E.P., Tannock, I.F., and Schwartz, L.H. (1999) Randomized phase III trial comparing the new potent and selective third-generation aromatase inhibitor vorozole with megestrol acetate in postmenopausal advanced breast cancer patients. *Journal of Clinical Oncology*, **17**, 52–63.

Gould, A.L. (1980) A new approach to the analysis of clinical drug trials with withdrawals. *B*iometrics, **36**, 721–727.

Greenlees, W.S., Reece, J.S., and Zieschang, K.D. (1982) Imputation of missing values when the probability of response depends on the variable being imputed. *Journal of the American Statistical Association*, **77**, 251–261.

Heckman, J.J. (1976) The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement*, **5**, 475–492.

Hinde, J. and Demétrio, C.G.B. (1998a) Overdispersion: Models and estimation. *Computational Statistics and Data Analysis*, **27**, 151–170.

Hinde, J. and Demétrio, C.G.B. (1998b) *Overdispersion: Models and Estimation.* São Paulo: XIII Sinape.

Hogan, J.W. and Laird, N.M. (1997) Mixture models for the joint distribution of repeated measures and event times. *Statistics in Medicine*, **16**, 239–258.

Hoyberg, O.J., Fensbo, C., Remvig, J., Lingjaerde, O.K., Slotei-Nielsen, M., and Salvesen, I. (1993) Risperidone versus perphenazine in the treatment of chronic schizophrenic patients with acute exacerbations. *Acta Psychiatrica Scandinavica*, **88**, 395–402.

Huttunen, M.O., Piepponen, T., Rantanen, H., Larmo, L., Nyholm, R., and Raitasuo, V. (1995) Risperidone versus zuclopenthixol in the treatment of acute schizophrenic episodes: a double-blind parallel-group trial. *Acta Psychiatrica Scandinavica*, **91**, 271–277.

Jansen, I., Beunckens, C., Molenberghs, G., Verbeke, G., and Mallinckrodt, C. (2006a) Analyzing incomplete discrete longitudinal clinical trial data. *Statistical Science*, **21**, 52–69.

Jansen, I., Hens, N., Molenberghs, G., Aerts, M., Verbeke, G., and Kenward, M.G. (2006b) The nature of sensitivity in missing not at random models. *Computational Statistics and Data Analysis*, **50**, 830–858.

Kay, S.R., Fiszbein, A., and Opler, L.A. (1987) The Positive and Negative Syndrome Scale (PANSS) for schizophrenia. *Schizophrenia Bulletin*, **13**, 261–276.

Kay, S.R., Opler, L.A., and Lindenmayer, J.P.(1988) Reliability and validity of the Positive and Negative Syndrome Scale for schizophrenics. *Journal of Psychiatric Research*, **23**, 99–110.

Kenward, M.G. and Molenberghs, G. (1999) Parametric models for incomplete continuous and categorical longitudinal studies data. *Statistical Methods in Medical Research*, **8**, 51–83.

Kenward, M.G., Molenberghs, G., and Thijs, H. (2003) Pattern-mixture models with proper time dependence. *Biometrika*, **90**, 53–71.

Laird, N.M. and Ware, J.H. (1982) Random effects models for longitudinal data. *Biometrics*, **38**, 963–974.

Laird, N.M., Lange, N., and Stram, D. (1987) Maximum likelihood computations with repeated measures: application of the EM algorithm. *Journal of the American Statistical Association*, **82**, 97–105.

Laird, N.M. (1994) Discussion to Diggle, P.J. and Kenward, M.G.: 'Informative dropout in longitudinal data analysis'. *Applied Statistics*, **43**, 84.

Lawless, J. (1987) Negative binomial and mixed Poisson regression. *The Canadian Journal of Statistics*, **15**, 209–225.

Lee, Y., Nelder, J.A., and Pawitan, Y. (2006) *Generalized Linear Models with Random Effects*. Boca Raton: Chapman & Hall/CRC.

Lesaffre, E and Verbeke, G. (1998) Local influence in linear mixed models. *Biometrics*, **54**, 570–582.

Liang, K.Y. and Zeger, S.L. (1986) Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika*, **73**, 13–22.

Little, R.J.A. and Rubin, D.B. (1987) *Statistical Analysis with Missing Data*. New York: Wiley.

Little, R.J.A. (1993) Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, **88**, 125–134.

Little, R.J.A. (1994) A class of pattern-mixture models for normal incomplete data. *Biometrika*, **81**, 471–483.

Little, R.J.A. (1994) Discussion to Diggle, P.J. and Kenward, M.G.: 'Informative dropout in longitudinal data analysis'. *Applied Statistics*, **43**, 78.

Little, R.J.A. (1995) Modeling the drop-out mechanism in repeated measures studies. *Journal of the American Statistical Association*, **90**, 1112–1121.

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

Littell, R.C., Milliken, G.A., Stroup, W.W., and Wolfinger, R.D. (1996) *SAS System for Mixed Models*, Cary, NC: SAS Institute Inc.

Marder, S.R. and Meibach, R.C. (1994) Risperidone in the treatment of schizophrenia . *American Journal of Psychiatry*, **151**, 825–835.

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

Molenberghs, G., Kenward, M.G., and Lesaffre, E. (1997) The analysis of longitudinal ordinal data with nonrandom dropout. *Biometrika*, **84**, 33–44.

Molenberghs, G., Michiels, B., Kenward, M.G., and Diggle, P.J. (1998) Missing data mechanisms and pattern-mixture models. *Statistica Neerlandica*, **52**, 153–161.

Molenberghs, G., Buyse, M., Geys, H., Renard, D., Burzykowski, T., and Alonso, A. (2002) Statistical challenges in the evaluation of surrogate endpoints in randomized trials. *Controlled clinical trials*, **23**, 607–625.

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

Molenberghs, G. and Kenward, M.G. (2007) *Missing data in Clinical Studies*. New-York: John Wiley.

Molenberghs, G., Verbeke, G., and Demétrio, C.G.B. (2007) An extended random-effects approach to modeling repeated, overdispersed count data, *Lifetime Data Analysis*, **13**(4), 513–531.

Molenberghs, G., Vangeneugden, T., and Laenen, A. (2007) Estimating reliability and generalizability from hierarchical biomedical data. *Journal of Biopharmaceutical Statistics*, **17**(4), 595–627.

Nelder, J.A. and Mead, R. (1965) A simplex method for function minimisation. *The Computer Journal*, **7**, 303–313.

Overall, J.E. and Gorham, D.R. (1962) The Brief Psychiatric Rating Scale. *Psychological Reports*, **10**, 799–812.

Park, T. and Brown, M.B. (1994) Models for categorical data with nonignorable nonresponse. *Journal of the American Statistical Association*, **89**, 44–52.

Peralta, V. and Cuesta, M.J. (1994) Psychometric properties of the Positive and Negative Syndrome Scale (PANSS) in schizophrenia. *Journal of Psychiatric Research*, **53**, 31–40.

Peuskens, J. and the Risperidone Study Group (1995) Risperidone in the treatment of chronic schizophrenic patients: a multi.tio.l, multicentre, double-blind, parallel-group study versus haloperidol. *British Journal of Psychiatry*, **166**, 712–726.

Prentice, R.L. (1989) Surrogate endpoints in clinical trials: definitions and operational criteria. *Statistics in Medicine*, **8**, 431–440.

Robins, J.M., Rotnitzky, A., and Scharfstein, D.O. (1998) Semiparametric regression for repeated outcomes with non-ignorable non-response. *Journal of the American Statistical Association*, **93**, 1321–1339.

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

Rotnitzky, A. and Robins, J.M. (1995) Semi-parametric estimation of models for means and covariances in the presence of missing data. *Scandinavian Journal of Statistics: Theory and Applications*, **22**, 323–334.

Rotnizky, A. and Robins, J.M. (1997) Analysis of semi-parametric regression models for repeated outcomes in the presence of missing data. *Statistics in Medicine*, **16**, 81–102.

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

Rubin, D.B. (1994) Discussion to Diggle, P.J. and Kenward, M.G.: 'Informative dropout in longitudinal data analysis'. *Applied Statistics*, **43**, 80–82.

Schipper, H., Clinch, J., McMurray, A., and Levitt, M. (1984) Measuring the quality of life of cancer patients: the Functional-Living Index-Cancer: development and validation. *Journal of Clinical Oncology*, **2**, 472–483.

Searle, S.R., Casella G., and McCulloch, C.E. (1992) *Variance Components*, New York: John Wiley & Sons.

Shavelson, R.J., Webb, N.M., and Rowley, G.L. (1989) Generalizability Theory. *American Psychologist*, **44**, 922-932.

Sheiner, L.B., Beal, S.L., and Dunne, A. (1997) Analysis of nonrandomly censored ordered categorical longitudinal data from analgesic trials. *Journal of the American Statistical Association*, **92**, 1235–1244.

Shoukri, M.M. (2004) *Measures of interobserver agreement.* London: Chapman & Hall.

Shrout, P.E. and Fleiss, J.L. (1979) Intraclass correlations: uses in assessing interater reliability. *Psychological Bulletin*, **86**, 420–428.

Smith, D.M., Robertson, B., and Diggle, P.J. (1996) *Object-oriented Software for the Analysis of Longitudinal Data in S.* Technical Report MA 96/192. Department of Mathematics and Statistics, University of Lancaster, LA1 4YF, United Kingdom.

Stasny, E.A. (1986) Estimating gross flows using panel data with nonresponse: an example from the Canadian Labour Force Survey. *Journal of the American Statistical Association*, **81**, 42–47.

Streiner, D.L. and Norman, G.R. (1995) *Health measurement scales.* Oxford: Oxford University Press.

Tibaldi, F., Cortinas Abrahantes, J., Molenberghs, G., Renard, D., Burzykowski, T., Buyse, M., Parmar, M., Stijnen, T., and Wolfinger, R. (2003) Simplified hierarchical linear models for the evaluation of surrogate endpoints. *Journal of Statistical Computation and Simulation*, **73**, 643–658.

Tisak, J. and Tisak, M.S. (1996) Longitudinal models of Reliability and Validity: A Latent Curve Approach, *Applied Psychological Measurement*, **20/3**, 275–288.

Vangeneugden, T., Laenen, A., Geys, H., Renard, D., and Molenberghs, G. (2004) Applying linear mixed models to estimate reliability in clinical trial data with repeated measurements. *Controlled Clinical Trials*, **25**, 13–30.

Vangeneugden, T., Laenen, A., Geys, H., Renard, D., and Molenberghs, G. (2005) Applying concepts of generalizability theory on clinical trial data to investigate sources of variation and their impact on reliability. *Biometrics*, **61**, 295–304.

Vangeneugden, T., Molenberghs, G., Laenen, A., Alonso, A., and Geys, H. (2008a) Generalizability in non-Gaussian longitudinal clinical trial data based on generalized linear mixed models. *Journal of Biopharmaceutical Statistics*, **00**, 000–000.

Vangeneugden, T., Molenberghs, G., Laenen, A., Geys, H., Beunckens, C., and Sotto, C. (2008b) Marginal correlation in longitudinal binary data based on generalized linear mixed models. Submitted for publication.

Vangeneugden, T., Molenberghs, G., Verbeke, G., and Demétrio, C.G.B. (2008c) Marginal correlation from an extended random-effects model for repeated and overdispersed counts. Submitted for publication.

Verbeke, G. and Molenberghs, G. (2000) *Linear Mixed Models for Longitudinal Data.* New York: Springer.

Verbeke, G., Lesaffre, E., Molenberghs, G., Thijs, H., and Kenward, M.G. (2001) Sensitivity analysis for non-random dropout: a local influence approach. *Biometrics* **57**(1), 7–14.

Welsh, A.H. (1996) *Aspects of Statistical Inference.* New York: Wiley.

Wiley, D.E. and Wiley, J.A. (1970) The estimation of measurement error in panel data. *American Sociological Review*, **35**, 112–117.

Wu, M.C. and Bailey, K.R. (1988) Analysing changes in the presence of informative right censoring caused by death and withdrawal. *Statistics in Medicine*, **7**, 337–346.

Wu, M.C. and Bailey, K.R. (1989) Estimation and comparison of changes in the presence of informative right censoring: conditional linear model. *Biometrics*, **45**, 939–955.

Wu, M.C. and Carroll, R.J. (1988) Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, **44**, 175–188.

# Samenvatting

In klinische studies wordt vaak gebruik gemaakt van schalen en vragenlijsten. Vooraleer schalen in gebruik worden genomen worden ze echter eerst gevalideerd in een test steekproef. De bedoeling van deze validatie is om de psychometrische eigenschappen van een schaal of meting te evalueren. Meer bepaald evalueert men de betrouwbaarheid ("*reliability*") van een schaal en de validiteit ("*validity*"). Meer specifiek, voor de betrouwbaarheid (Shrout en Fleiss, 1979) wordt nagegaan of de schaal reproduceerbare resultaten geeft als ze meerdere keren gemeten wordt door verschillende onderzoekers ("*interrater reliability*") of herhaaldelijk binnen éénzelfde persoon ("*test-retest reliability*"). Deze eigenschappen worden geëvalueerd via de zogenaamde "*Intraclass Correlation Coefficient*" (ICC, Bartko, 1966). Om de validiteit van een schaal te evalueren gaat men na of de schaal effectief meet wat de intentie is dat ze zou moeten meten, meer bepaald gaat het dan over inhoudelijke -, constructie- en criteriumvaliditeit (Carmines en Zeller, 1979). Al deze psychometrische evaluaties gebeuren op een aparte en vaak relatief kleine steekproef. Betrouwbaarheid en validiteit zijn echter geen vaste grootheden die gekoppeld zijn aan een meting of schaal, deze eigenschappen zijn afhankelijk van de populatie waarin ze gemeten wordt.

In deze thesis hebben we een kader ontwikkeld waarin we betrouwbaarheid en validiteit, meer specifiek criteriumvaliditeit, bijkomend evalueren op basis van longitudinale gegevens resulterende uit klinische studies. De bedoeling is te evalueren hoe betrouwbaar de schaal werkelijk was in de populatie, ingesloten in de klinische studie. Zijn er factoren die deze betrouwbaarheid verlagen? Dit laatste is het onderwerp van "*Generalizability Theory*" (GT, Cronbach, 1963). Met deze informatie zouden we bijvoorbeeld de schaal kunnen verbeteren of extra training geven daar waar de reproduceerbaarheid minder goed was. Via criteriumvaliditeit kunnen we nagaan hoe een schaal is gecorreleerd met andere metingen of schalen. Dit laatste kan helpen

om een beeld te vormen of bepaalde veranderingen in een schaal al dan niet klinisch relevant zijn door ze te correleren met klinische parameters.

Na de inleiding in Hoofdstuk 1 en de beschrijving van de gebruikte data doorheen de thesis in Hoofdstuk 2, hebben we eerst een overzicht gegeven van de klassieke modellen voor longitudinale gegevens (Hoofdstuk 3). Hierin introduceerden we onder meer het *Linear Mixed Model* (LMM, Verbeke en Molenberghs, 2000), het *General Linear Mixed Model* (GLMM, Molenberghs en Verbeke, 2000), enkele gecombineerde modellen voor het modelleren van telgegevens, en verder ook een introductie in de taxonomy en analyse van de data met ontbrekende gegevens (Molenberghs and Kenward, 2007). In Hoofdstuk 4 introduceerden we psychometrische concepten die doorheen de thesis geëvalueerd werden op longitudinale klinische gegevens, en tenslotte in Hoofdstuk 5 laten we bestaande concepten in de evaluatie van surrogaat parameters de revue passeren.

In Hoofdstuk 6 hebben we een algemene formule voorgesteld om de test-retest ICC af te leiden voor longitudinale normaal verdeelde gegevens resulterende uit klinische studies in schizofrene patiënten. Deze formule is vervolgens toegepast op de totale PANSS, een schaal die de beladenheid van psychotische symptomen meet. Hierbij werden een 4-tal specifieke modellen met variërende complexiteit aangepast. De toepassing toonde aan dat het mogelijk was om een studiepopulatie specifiek ICC te berekenen. Deze variëerde van 0.80 tot ongeveer 0.5 afhankelijk van de tijdsduur tussen 2 metingen en afhankelijk van het model. In Hoofdstuk 7 hebben dit uitgebreid tot GT voor continue normaal verdeelde gegevens. De bedoeling hier was om na te gaan welke factoren een invloed hadden op betrouwbaarheid en meetfout. Zo hebben we ondermeer aangetoond dat het feit dat studies in meerdere landen worden uitgevoerd, geen of nauwelijks invloed heeft op de betrouwbaarheid van de PANSS schaal. Andere factoren zoals de beladenheid van negatieve psychotische symptomen bij de start van de studie, hebben dan weer een negatieve invloed op de betrouwbaarheid.

In Hoofdstuk 8 hebben we een algemeen kader uitgewerkt waarin betrouwbaarheid kan afgeleid worden op basis van het GLMM. Dit laat toe om betrouwbaarheid benaderend te berekenen voor elk type data (continue, binair en telgegevens) en voor verschillende niveaus van complexiteit (bijvoorbeeld met en zonder seriële correlatie). Concreet hebben we dan via een Taylorreeksbenadering de betrouwbaarheid gemeten voor binaire respons parameter (klinische verbetering ja/neen op de CGI) vanuit dezelfde gegevens resulterende uit klinische studies in de schizofrenie. Analoog aan Hoofdstuk 7 hebben we betrouwbaarheid uitgebreid tot GT in Hoofdstuk 9, maar dit

maal gebruik makend van het bredere kader dat GLMM aanreikt. Opnieuw hebben we dit voor het specifieke geval van binaire respons parameters uitgewerkt op concrete schizofrenie gegevens. Dit leidde tot dezelfde conclusies, met name dat de factor land geen invloed, en de negatieve psychotische beladenheid wel een invloed heeft op de betrouwbaarheid van de binaire parameter klinische respons.

In een volgende stap hebben we de ICC afgeleid voor telgegevens. Dank zij het werk van Molenberghs, Verbeke, en Demetrio (2008) was het mogelijk om naast de benaderende formule op basis het GLMM model uit Hoofdstuk 7, een gesloten formule af te leiden van de test-retest ICC. Dit is vervolgens toegepast op klinische gegevens van een studie in epilepsie, meer bepaald op de parameter "aantal epilepsie aanvallen". Dit toonde aan dat de benadering vrij vlug convergeerde naar het resultaat van de gesloten formule. Een andere vastelling was dat het gecombineerde model een beter model fit gaf dan het Poisson-normale model. Hierdoor kunnen de schattingen voor de ICC resulterende uit dit laatste model misleidend zijn.

Vervolgens hebben we de aandacht gericht op criterium validiteit, met andere woorden op de correlaties tussen simultaan gemeten parameters. Eerst hebben we het GLMM kader (Molenberghs en Verbeke, 2005) gebruikt om de correlatie tussen binaire klinische respons parameter en de continue totale PANSS te schatten. En vervolgens hebben we technieken gebruikt uit de validatie van surrogaat parameters (Alonso et al.) om de correlatie te schatten via $R^2_{trial}$ en $R^2_{indiv}$ tussen de continue PANSS schaal en de BPRS schaal enerzijds en de correlatie tussen PANSS en CGI en BPRS en CGI respons anderzijds, dit maal als ordinale schaal. Uit al deze analyses bleek een sterke correlatie tussen alle drie de metingen, met bijvoorbeeld een correlatie van 0.75 tot 0.81, afhankelijk van de methode en de parameterisatie van de CGI (binair of ordinaal).

In voorgaande hoofdtukken hebben we gebruik gemaakt van klassieke longitudinale modellen. Deze modellen veronderstellen *Missing at Random* (MAR) waarbij onvolledigheid kan afhangen van geobserveerde respons waarden. In Hoofdstuk 12 hebben we een toepassing uitgewerkt van een analyse van een dataset met ontbrekende gegevens van de FLIC, een schaal die de levenskwaliteit in borstkanker meet. In deze studie werden patiënten gevolgd tot aan progressie (Michiels *et al.*, 2002). Het gevolg is dat er veel uitval was met als resultaat een groot aantal missende waarden. Hier hebben we zowel een SEM als een PMM uitgewerkt met als doel een sensitiviteitsanalyse aan te reiken die de robuustheid van de conclusies evalueert.