Statistical approaches to explore clinical heterogeneity in psychosis

Islam, Atiqul

CHAPTER 1

General Introduction

1. Introduction

1.1. Psychosis (schizophrenia spectrum disorder)

Schizophrenia and related psychotic disorders are now referred to as Schizophrenia Spectrum Disorders (DSM-5) or –by some- as Psychosis spectrum syndrome (van Os, 2016). Schizophrenia and related psychotic disorders consist of multiple symptom dimensions (van Os et al., 2010). A particular group of these non-affective psychotic disorders is a complex, multidimensional chronic brain disorder with a lifetime prevalence of nearly 1.5% (Perala et al., 2007; van Os et al., 2010).

The early onset of the disease, along with its chronic course, makes schizophrenia a debilitating disorder for many patients and their relatives (Mueser and Jeste, 2008). The other psychotic disorders, *i.e.* schizoaffective disorder, schizophreniform disorder, psychotic disorder not otherwise specified and brief psychotic disorder, cause similar symptoms of psychosis, albeit with a shorter duration of illness. Causes of schizophrenia are still to be elucidated, although there are some well-known risk factors. For example, individuals with a first-degree relative with schizophrenia have a higher risk of developing the disorder, about tenfold compared to general population (Gejman et al., 2010). There is evidence that the level of familial clustering of psychotic disorder is higher when people are living in urban environment or belong to a minority group (van Os et al., 2010). It is now recognized that high heritability (80%) estimates from classical twin and adoption studies is not only due to genetic influence, but also underlying intra-familial environmental effects that are moderated by gene–environment interactions (Gejman et al., 2010; van Os et al., 2010; van Os and Kapur, 2009).

Psychotic disorder is characterized by an array of heterogeneous symptoms including delusions, hallucinations, disorganized speech or behavior, and impaired cognitive ability occurring for a significant period of time during at least one month period and associated with continuous problems over at least a six-month period (Perala et al., 2007; van Os and Kapur, 2009). These multiple dimensions have also been mentioned in DSM-5 (American Psychiatric Association, 2013). The delusions (i.e. mostly false beliefs rooted in the mind based on incorrect inference) and the hallucinations (i.e. sensory-driven incidents that involve hearing or seeing something that is not reality based) are often considered the cardinal features of the illness of psychosis, partly because they are easy to identify and greatly affect functioning and society.

The signs and symptoms of schizophrenia may vary dramatically from person to person, both in pattern and severity. Recently, there has been a debate on the nature of the negative symptoms. Negative symptoms comprise in dysfunction of communication, affect and emotion, socialization, capacity of pleasure and motivation (Stahl and Buckley, 2007). According to authors such as Liemburg et al. (2013), one can discriminate two dimensions of negative symptoms; i.e. 1) social amotivation (social, emotional withdrawal and reflects diminished interest in or affective commitment to the social environment) and 2) expressive deficit (blunted affect, poverty of speech and motor retardation, and reflects diminished expressive responsiveness in verbal and non-verbal communication). These authors have clearly demonstrated the clinical validity of such dimensions of negative symptoms in first-episode patients with a psychotic disorder in cross-sectional studies (Liemburg et al., 2013). However, whether this two-dimension approach also holds for longitudinal studies needs to be demonstrated.

Cognitive impairment is another dimension of symptoms. There has been a resurgence of interest in the cognitive alterations of schizophrenia. It is often stated that patients with a diagnosis of schizophrenia have a broad-based cognitive impairment of, on average, about 1 SD below the norm across a range of cognitive abilities (attention, speed of processing, working and long-term memory, executive function, and social cognition) (Kahn RS, 2013; van Os et al., 2010; van Os and Kapur, 2009; Fioravanti et al., 2005). Furthermore, the uniformity of the cognitive impairments has been questioned. Quee et al. (2014) found both severe impairment and normal functioning in non-affected siblings and a mixed profile group in between these two. Interestingly, these cognitive profiles correlated well with their affected family-member, suggesting that cognition is diverse and heterogeneous in people with psychotic disorders (Quee et al., 2014).

In line with the more integral approach on symptom dimensions, physical complaints have increasingly been considered as the "somatic dimension" of schizophrenia spectrum disorders. Comorbidity with somatic disorders has now been recognized as an important factor, leading to a 15-20 years shorter life expectancy for patients with schizophrenia. These comorbid health-conditions may contribute up to 60 percent of the three times excess of premature mortality in schizophrenia (De Hert et al., 2011; Parks et al., 2006; Vreeland, 2007). Indeed, patients with schizophrenia have up to 54 percent metabolic syndrome (Bruins et al., 2016) and a 2-3 fold higher risk of diabetes mellitus (Bushe and Holt, 2004; van Winkel et al., 2006) and cardiovascular diseases (Bresee et al., 2010; De Hert et al., 2009; Hennekens et al., 2005).

Thus, psychiatric symptoms (e.g. on positive and negative symptoms) amended with cognitive impairments, and somatic comorbidity are usually heterogeneous and differ in origin, structure and clinical expression. Although this notion has been widely known for a long time (Markova and Berrios, 1995), these differences are often overlooked both clinically and statistically in research. Ignoring differences in structure between symptoms has also naturally yielded biased homogeneous structure of symptoms. With the growing awareness of the heterogeneity of psychotic disorders, there is also a growing need in classical and model-based statistical clustering approaches to clarify the underlying structures.

The main aim of the thesis is to explore the heterogeneity in cognitive functioning and clinical symptoms in schizophrenia patients and their unaffected siblings using cross-sectional and longitudinal data. This aim is achieved by applying statistical methods, such as classical clustering, linear mixed effects and group-based trajectory modeling techniques.

1.2. Statistical Analysis for Heterogeneity

1.2.1. Heterogeneity

To illustrate heterogeneity, consider for example clinical trials where some patients do not response to the treatment and others do respond well under the same treatment plan. Therefore, modest clinical effects can sometimes be misleading because they may be composed of a mixture of significant benefits for some, no benefits for many and harm for a few. The same would hold true for schizophrenia patients, where individual differences would mask general patterns.

In general, each subject would have its own profile. The average profile of all subjects would give the idea of having just one population profile. The average profile may provide valuable

information, assuming that the population under study is to large extent homogenous. For longitudinal studies, Verbeke and Lesare (1996) demonstrated that this homogeneous population is then described by a single mean trajectory and variance-covariance matrix. However, this assumption is highly unrealistic when subgroups of populations exist. In psychiatry, different disease symptoms or diagnostic groups could be classified by different mean profiles. Ignoring the heterogeneity can produce biased estimates of the association parameters and their corresponding variance terms (Verbeke and Lesaffre, 1996). To break down the "seemingly homogeneous population" into more meaningful subgroups with similar profiles or patterns, one may need to quantify the heterogeneity.

One way to dissect heterogeneity of a group of patients with seemingly the same diagnosis is to use the positive and/or negative symptoms or cognitive impairment scores and apply clustering techniques to form homogeneous symptom subtypes (Dawes et al., 2011; Jablensky, 2006; Joyce and Roiser, 2007). Cluster analysis is used to classify objects into groups (or clusters/classes/components) such that objects within a group are more similar than between groups. Forming clusters depend on study design and analytical tools. In cross-sectional studies, where the subjects are independent in measuring the outcome, clustering techniques, like hierarchical clustering and K-means clustering, can address the heterogeneity and form homogeneous subtypes. In longitudinal studies psychiatric symptoms (on positive and negative symptoms or cognitive impairment) of patients (or siblings) may be heterogeneous over time. They are likely to originate from latent distinct trajectory patterns. Thus it is important to capture the individual trajectories and to understand what affects them. Mixed effects models are the leading statistical techniques to describe individual trajectories. Moreover, these potentially existing heterogeneity in disease course patterns, warrant a method which identifies the presence of unobserved heterogeneity and to form homogeneous subgroups of patients. This leads to finite mixture modeling (Schlattmann, 2009) which is designed to identify clusters of individuals following similar pattern of progression of outcomes over time (Jones and Nagin, 2007). In other words, the primary goal is to identify groups of patients that have similar patterns of symptoms over time. Some methods for dealing with heterogeneity for cross-sectional and longitudinal data are briefly explained below.

1.2.2. Classical clustering

Hierarchical clustering is a classical clustering technique, which is often used in psychiatry. It produces a nested (i.e. hierarchical) sequence of clusters that can be presented in a dendrogram (Figure 1). Agglomerative and divisive hierarchical clustering are ways to form nested clusters. Agglomerative clustering is a bottom-up method. It starts with partitioning all sample objects to individual or separate clusters and then successively merging the closest pair of clusters using some kind of similarity criterion. It ends when all objects are in one cluster (Figure 1). The advantage is that it can produce an informative ordering of the objects and produce small clusters, which may be helpful for discovery. Divisive clustering is the opposite of agglomerative method; it starts with all objects in one cluster and then iteratively split clusters which are most dissimilar until all objects end up in their own cluster (top-down approach) (Figure 1). The disadvantage of hierarchical clustering is that if an object becomes a member of any cluster, it will neither be removed from that cluster nor

be mixed up with objects of any other clusters. It may cause incorrect groups at an early stage (Fernández and Gómez, 2008).

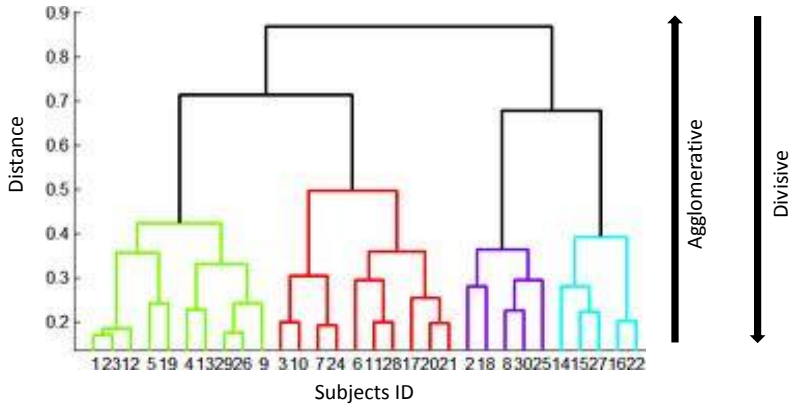

Figure 1: Dendrogram displaying number of clusters

In order to construct the clusters, there are different ways of measuring the distance between groups of objects such as single-linkage, complete-linkage, average-linkage and Ward's minimum variance method. The first three define the distance between clusters as the minimum, maximum and average distance between any two objects of the clusters. Ward's minimum variance method, on the other hand, finds the pair of clusters that leads to minimum increase in total within-cluster variance when merging two clusters as the distance measure (Ward, 1963).

Another popular set of clustering techniques are partitioning methods that attempts to divide the whole set of objects into a pre-defined number of clusters (Henry et al., 2005; Gordon, 1999; Borgen and Barnett, 1987; Hartigan and Wong, 1979). The most popular one is K-means clustering. It aims to minimize the sum of the squared distances between the objects and their cluster centers, by iteratively reallocating objects to the clusters until full convergence. To illustrate the procedure, let us assume that there are *k* clusters and choose *k* centroids in the data space of the objects (either randomly or by using hierarchical clustering). Then, assign each object to the closest centroid using a predetermined criterion (e.g. minimize the sum of the squared distances). When all objects are assigned to the *k* centroids, calculate the mean or median of the variables under study for the formed *k* groups. Then again assign each object to the new centroids and repeat the process until the groups do not change anymore.

Practically, clustering techniques provide little information about the cluster structure in the data. There is no unified approach on what essentially constitutes a cluster and no conclusive answer for choosing the number of clusters (Fernández and Gómez, 2008; Milligan and Cooper, 1985). Furthermore, another major difficulty is to estimate the threshold of the objective function (i.e. the within-cluster sum of squares) and the number of clusters when there is no information other than the observed values is available (Hartigan, 1975). This is true both for hierarchical and K-means clustering techniques. Since there is no *a priori* information on natural groupings or subtypes of patients or siblings on the basis of symptoms, I propose using hierarchical clustering to find the most appropriate number of clusters. The result of hierarchical clustering is a good option to obtain a priori

information and then use this solution as input for K-means clustering to finally form the subtypes of patients or siblings. However, one of the main problems in hierarchical clustering is to determine the true number of clusters since it produces an informative ordering of the objects and produce series of small clusters (Fernández and Gómez, 2008; Milligan and Cooper, 1985).

Over the past decades, several procedures/indices have been proposed for determining the number of clusters and testing the null hypothesis that there is no cluster structure in the dataset at all (Milligan and Cooper, 1985). The majority of existing indices does not test formally a null hypothesis but rather estimate a summary statistics that points towards an optimal number of clusters. These summary statistics are typically functions of the within clusters and the between clusters sum of squares. In addition, these indices are evaluated either locally or globally to determine the number of clusters with respect to the clustering algorithms. Global methods utilize entire dataset and maximize it as a function of the number of clusters. Most of the global methods are undefined for one cluster and hence there is no indication whether the data should be clustered at all. Local methods use individual pairs of clusters and test whether they would be merged or not (Tibshirani et al., 2001; Gordon, 1999). Several criteria or stopping rules for the indices have been provided in literature, but extensive simulation studies on realistic data were not conducted. Therefore, I will discuss these shortcoming and potential solutions in this dissertation.

1.2.3. Linear mixed models

Repeated measures on subjects are very common in health, social, behavioral and biological sciences. The major challenge in analyzing repeated measures data is the fact that the measurements on a subject are correlated. This correlation must be taken into account during the statistical analysis to obtain valid inference i.e. estimates of effect sizes for association between parameters and outcomes of interest. The correlation can often be captured by introducing random effects in the classical statistical analyses e.g. linear and logistic regression. These statistical models combine the components of fixed effects, random effects (e.g. random-intercept and random-slope), and repeated measurements in a single unified approach. These models are called linear mixed models (LMM) for continuous longitudinal outcomes (Laird and Ware, 1982; Verbeke and Molenberghs, 2000), and generalized linear mixed models (GLMM) for other type of outcomes (Breslow and Clayton, 1993; Molenberghs and Verbeke, 2006).

The mathematical equations for LMM are explained briefly here for longitudinal data and subject-specific time profiles. Let Y_{it} be the response measure (e.g. score of psychotic experiences) for subject i (=1, 2, …, *N*) measured at time T_{it}, t = 1, 2, …, n_i. The linear mixed model is simply defined in the matrix form as

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i b_i + \varepsilon_i$$

Where, $b_i \sim N(\mathbf{0}, D)$, $\varepsilon_i \sim N(\mathbf{0}, \Sigma_i)$, $\boldsymbol{Y}_i$ is the n_i-dimensional response vector for subject i, $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are the $(n_i \times p)$ and $(n_i \times q)$ design matrices of known covariates, $\boldsymbol{\beta}$ is the *p*-dimensional vector of population-average regression coefficients (fixed effects), b_i is the *q*-dimensional vector of random effects for subject i, ε_i is a n_i-dimensional vector of measurement error components. It is assumed that b_i and ε_i are independent. Conditional on the random effects b_i, the distribution of $\boldsymbol{Y}_i$ is given by $\boldsymbol{Y}_i | b_i \sim N(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i b_i, \Sigma_i)$. The inference is based on maximizing the likelihood function of the

marginal response $\boldsymbol{Y}_i$. More specifically, a subject-specific time trajectory of polynomial form can be defined as

$$Y_{it} = \beta_0 + \sum_{k=1}^{p}\sum_{r=0}^{q}(\beta_{rk}X_{ik} + b_{ri}) * T_{it}^{r} + \varepsilon_{it}$$

where, β_0 is the overall intercept, β_{rk} are the r^{th} (r = 0, 1, 2, …, q) polynomial form of time and k^{th} (r = 1, 2, …, p) fixed effects parameters, b_{0i} is the random-intercept, b_{ri} is the random-term for time order T_{it}^{r} of subject i, and ε_{it} is error disturbance term. The model parameters and variance components are estimated by either Maximum Likelihood (ML) or Restricted Maximum Likelihood (REML) estimation procedure (Verbeke and Molenberghs, 2000). Note that each subject has its own time profile in this model.

1.2.4. Group-based trajectory modeling

In literature, there are many approaches which are applied in many different situations to quantify homogeneous groups with their own shapes and patterns for longitudinal trajectories. To cluster subjects based on continuous longitudinal data, Verbeke and Lesaffre (1996) assumed a normal mixture in the distribution of random effects and applied their method to clustering of growth curves (Verbeke and Lesaffre, 1996). De la Cruz-Mesía et al (2008) proposed a mixture of nonlinear mixed models for describing nonlinear relationships across time and perform clustering of subjects on outcome (De la Cruz-Mesía et al., 2008). Other longitudinal approaches used to understand and group differences in developmental trajectories over time include latent class analysis which was primarily used for categorical or binary data (Vermunt and Magidson, 2003) or a semi-parametric mixture model that was appropriate for data with skewed distributions (Jones et al., 2001). The growth mixture modelling is one of the approaches when outcome variables are approximately normally distributed.

Group-based trajectory modeling (GBTM) is a semi-parametric statistical method for analyzing developmental trajectories, which means describing the evolution of an outcome over time (Nagin, 1999). GBTM is sometimes called latent class growth modeling (Andruff et al., 2009). GBTM is an application of finite mixture modeling and is designed to identify clusters of individuals following similar patterns of change of some behavioral, biological, physical outcome (e.g. cognition, negative symptom) over time (Jones and Nagin, 2007; Nagin, 2014). Traditional growth curve modeling techniques assume that subjects come from a single population and estimate a single trajectory that averages the individual trajectories of all subjects in a given sample. This average trajectory comprises the averaged intercept and slope for the entire sample. This approach captures individual differences by estimating random coefficients that represents the variability in the intercept and slope. But GBTM fixes the slope and intercept for the subgroup of individuals having a similar trajectory, given that individual differences are captured by the multiple trajectories included in the model (Andruff et al., 2009). GBTM is also a flexible statistical tool for identifying and summarizing the homogenous group of individuals and observing their development patterns over time in the form of both graphical and tabular way, which makes it easier to understand. Another important feature is that the GBTM also takes into account drop-out of participants over time (Haviland et al., 2011).

Mathematically, let Y_{it} be the longitudinal sequence of measurements (e.g. cognition or subdomains of negative symptoms) on individual *i* (= 1, 2, …, *N*) over time *t* (= 1, 2, …, *T*). The GBTM assumes that the population is composed of a mixture of *g* (=1, 2, …, *G*) underlying trajectory groups with marginal density $f(y_i) = \sum_g \pi_g p^g(y_i)$, where $p^g(y_i)$ is the density function of Y_i given its membership in group *g* and π_g is the probability of belonging to group *g*. The basic GBTM assumes that the random variables, Y_{it} are independent given the condition on membership in group *g*, therefore $p^g(y_i) = \prod_{t=1}^{T} p^{gt}(y_{it})$. The group membership probabilities, π_g are estimated by a multinomial *logit* function as $\pi_g = \exp(\theta_g)/\sum_{g=1}^{G} \exp(\theta_g)$, where θ_1 is standardized to zero so that estimation of each probability of π_g stays between 0 and 1.

In this thesis, I will apply the censored normal model (CNORM). The CNORM model is useful for modeling the conditional distribution of psychometric scale data, given group membership (Jones et al., 2001; Nagin and Tremblay, 2001). A normal distribution allowing for censoring is used because the data tend to cluster at the minimum (Min) and at the maximum (Max) of the scale. In our case, $p^{gt}(y_{it})$ is assumed to follow the censored normal distribution to accommodate the possibility of clustering at the value minimum and maximum. The likelihood of observing the data trajectory for individual *i*, given he/she belongs to group *g*, is given by

$$p^g(y_i) = \prod_{y_{it}=Min} \Phi\big((Min - \mu_{it}^g)/\sigma\big) \prod_{Min<y_{it}<Max} \frac{1}{\sigma}\varphi\big((y_{it} - \mu_{it}^g)/\sigma\big) \prod_{y_{it}=Max} \Big(1 - \Phi\big((Min - \mu_{it}^g)/\sigma\big)\Big)$$

Where, $\mu_{it}^g = \beta_0^g + \beta_1^g time_{it} + \beta_2^g time_{it}^2 + \beta_3^g time_{it}^3$ be the mean group time profile for the symptom/cognitive measurement in group *g* (Jones et al., 2001). Likewise growth curve modeling, a polynomial relationship is used to model the link between period and cognition/symptoms of individuals. The model assumes up to third-order polynomial relationship between μ_{it}^g and period (e.g. follow-up time) (Jones and Nagin, 2007).

1.3. Statistical Analysis for Associations

1.3.1. Modeling association

Many studies have shown a general pattern that patients are being more affected than normal controls with respect to reporting problems on functioning and symptoms on a population level. In majority of these studies non-affected siblings display values somewhere between these two groups (Quee et al., 2014; Krabbendam et al., 2005). To be able to compare groups of subject that are independent, Pearson chi-square (for categorical outcome) or ANOVA (for continuous outcome) is applicable. However, when subjects belong to the same family (as in patient-sibling studies), the group comparisons may be done using linear or generalized linear mixed effects models taking into account the familial correlation. This may lead to confirmation of familial liability, meaning that patients having more diseases than their unaffected siblings and healthy controls. One may also use the intra-cluster correlation coefficient (ICC) to calculate the measure of the relatedness/correlations (e.g. familial liability) of the subjects *within* the family.

1.3.2. Mixture distribution modeling

As mentioned earlier, siblings are at risk to develop psychotic experiences. A limited number of them make the transition to a clinical-defined psychosis over time. Psychotic experiences are also heterogeneous over time and this heterogeneity can possibly be explained by different factors. To learn more about the development of psychotic experiences over time, one can study factors that are known to predict psychosis. In this case the outcome is an increase or decrease of the number of psychotic experiences, measured on a numerical scale. This outcome is most likely not normally distributed, since a large number of siblings may have no psychotic experiences at all while a small number of siblings do actually report psychotic experiences. In other words, the distribution of the outcome is at least highly skewed to right. However, given the heterogeneous nature of the reported psychotic experiences, the distribution is more likely to be a bi-modal distribution (Figure 2). Such a distribution can be described by a mixture distribution of known parametric (like the normal) distributions.

For a unimodal distribution, the classical approach, for example LMM or GLMM, can be used to estimate the effect of the predictors on the outcome. This would still be the case even when the distribution of outcome is highly skewed. However, when there are concerns with a lot of zeros in the observed continuous outcome, the distribution of the outcome would display at least a bi-modal distribution. Classical statistical approaches may not lead to the correct estimates of the effects of the predictors. An alternative approach, which can deal with the excessive number of zeros, is then preferred to obtain unbiased estimates. The group of subjects with no psychotic experiences can be scored *zero*, while the group with experiences is labeled *one.* These two groups can be described with a Bernoulli distribution (scoring zero or one). Then the next step is to consider those subjects, who report experiences (i.e. nonzero), and describe their continuous outcome with a skewed distribution, like the lognormal distribution. The final step is then to combine these two models and to estimate the parameters for the probability of having a nonzero value (e.g. using generalized linear mixed effects models) together with the parameters for the skewed distribution (e.g. using a random effects lognormal model). The random effects in the binary and continuous parts are needed to address the repeated measurements over time (Tooze et al., 2002; Olsen and Schafer, 2001; Smith et al., 2015). This type of modeling is referred to as random effects mixture modeling and it enables to draw conclusions on the factors that may associate the binary part and/or the continuous part.

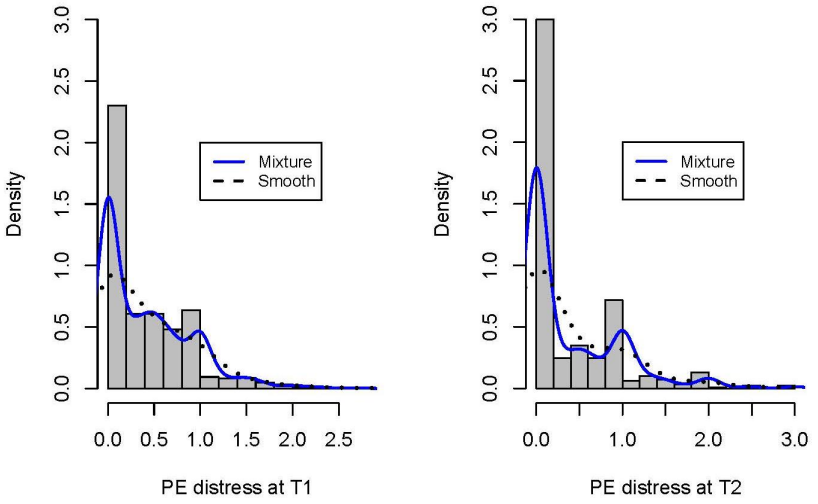

Figure 2: Empirical density, mixture density and smooth density of PE distress at time T1 and T2.

1.3.3. Dealing with missingness

Missing data are inevitable in any study, but in particular when data are collected repeatedly over time. The collection of a complete dataset on subjects is almost impossible. Depending on the nature of the study, missingness will appear in various forms. For instance, in a cross-sectional survey, missing data is usually of the form of item non-response. In this case a subject is not able or does not want to respond to a particular question or measurement. In longitudinal studies, attrition is the most common missing data problem. Here subjects drop out of the study prematurely before its termination and do not return. The pattern of attrition is an example of missing data for which the incompleteness can be ordered monotonically. Attrition, also referred to as dropout, may not be the only form of missing data in a longitudinal setting. For instance, a subject can miss several observation periods but eventually returns to the study. The latter type of missingness is often referred to as intermittent missingness.

To handle missing data, the mechanisms that lead to these missing values are of main importance. For instance, what drives the missingness patterns? To be more specific, is there a relationship between the missing data and the underlying values in the dataset. Three types of missing-data mechanisms have been defined in the literature. A process is said to be missing completely at random (MCAR) if the missingness is independent of both the unobserved and observed data. This implies that the probability of missing data on response is unrelated to the value of response itself or to the values of any other variables in the dataset. If the missingness depends on the observed data but it is independent of the unobserved values then the missing mechanism is said to be missing at random (MAR). MAR implies that the probability of having missing data is unrelated to the values that were missing, conditional on the observed values. If the process is neither MCAR nor MAR it is missing not at random (MNAR). The process then depends on the unobserved measurements. The assumption of MNAR does imply that the probability of a measurement being missing depends on the unobserved data (Rubin, 1976; Verbeke and Molenberghs, 2000; Little and Rubin, 2002; Allison, 2002).

Various simple methods, such as complete case analysis (CC) and last observation carried forward (LOCF), for handling missing data are available. Generally, simple methods such as complete case analysis (CC) work under the assumption that the missing mechanism is MCAR and in some special cases of MAR. A complete case analysis includes all subjects that would have all data recorded (relevant to the analysis). However, the method suffers from severe drawbacks: loss of valuable information, biased estimates, and inefficient estimates. The method of last observation carried forward (LOCF) replaces every missing value by the last observed value from the same subject. This method can be used both for monotone and non-monotone missing data but it is typically used in situations where incompleteness is due to attrition. Like other single imputation methods it overestimates the precision by treating imputed values and observed values on equal footing (Molenberghs and Verbeke, 2006; Beunckens et al., 2005). More importantly though even under the very strong assumption of MCAR, LOCF can be biased.

Multiple Imputation (MI) is currently one the most popular methods to deal with missingness under the MAR assumption. The idea of multiple imputation procedure is to replace each missing observation in the dataset with M plausible values, creating a set of M fully complete data sets. To analyze the data, each imputed data set is analyzed separately using conventional analysis method and programs. The results are then pooled in such a manner that the uncertainty in the imputed values averages out and disappears (Verbeke and Molenberghs, 2000; Little and Rubin, 2002; van Buuren, 2007). Maximum likelihood (ML) method sometimes provides valid inferences under MAR assumption. The approach gives appropriate estimates when the missingness occurs only in the (repeated) outcomes. If risk factors or predictors are missing ML may lead to biased estimates. Pattern mixture modeling, on the other hand, deals with missingness under the MNAR assumption (Verbeke and Molenberghs, 2000). It studies the statistical model conditional on the missing data indicators. ML, MI or Bayesian MI may provide biased estimates under MNAR (Schafer and Graham, 2002).

1.4. Database and software

To determine the aims of the thesis, data were obtained from the Genetic Risk and Outcome of Psychosis (GROUP) project, a longitudinal multicenter cohort study in the Netherlands and Belgium from April 2004 t/m in December 2013. The GROUP project provides a rich cohort data set on patients with schizophrenia, their unaffected siblings, and healthy controls at baseline, three and six years of follow-up. Patients were identified from a representative set of clinicians by screening their caseload and evaluating the patients to the inclusion criteria. Subsequently, a group of patients presenting consecutively at these services either as out-patients or in-patients were recruited for the study. In order to test hypotheses about the aetiology of non-affective psychosis, a cohort of family members e.g. siblings and parents with resilience for psychosis was being included. Controls were selected through a system of random mailings to addresses in the catchment areas of the cases. GROUP study examines vulnerability factors and protective factors for developing a psychotic disorder and the course thereof (Korver et al., 2012).

Statistical software such as Statistical Analysis System (SAS) version 9.4 and RStudio version 0.97.551/0.99.902 (R version 3.0.1/3.1.1) were used to perform all analyses throughout the thesis.

1.5. Outline of the thesis

The general aim of the thesis is to explore the heterogeneity in cognitive functioning and clinical symptoms in schizophrenia patients and their unaffected siblings using cross-sectional and longitudinal data. To this end, this thesis decomposes mainly two parts of statistical modeling. Part A contains statistical analysis for heterogeneity, here I will apply and evaluate classical clustering technique, linear and generalized linear mixed effects modeling, and group-based trajectory modeling. Part B includes statistical modeling for associations, here I will apply classical ordinal logistic regression, Cox-regression, generalized linear mixed models and mixtures of generalized linear mixed effects modeling.

From clinical perspectives, the current thesis describes a number of studies focusing on cognition, clinical symptoms, functional outcomes and disease outcomes in patients with psychosis, their unaffected siblings and healthy controls. Cognitive and symptoms heterogeneity have been characterized in this thesis.

In **chapter 2**, I investigate fourteen cluster indices to identify the correct number of subtypes for cross-sectional data. I use simulations that were based on a real case study with eight cognitive measures. I compared the indices on their performances for hierarchical clustering of subjects, while the simulations generated mixture distributions of multivariate normal distributions. I will show how well indices predict the simulated pre-defined number of clusters and also determine whether they can be used to decide if there would exist multiple subtypes at all.

In **chapter 3**, I apply the GBTM to identify homogeneous cognitive trajectories of patients with schizophrenia and their unaffected siblings over time. Distinct trajectories of composite cognitive functioning over time are identified and they distinguish different trajectories between patients and siblings, respectively. After finding these meaningful homogeneous trajectories of patients and siblings, I examine whether patients' profiles predict cognitive profiles of their unaffected sibling.

Chapter 4 also describes the application of GBTM to determine homogenous groups of patients based on symptoms (social amotivation and expressive deficits of negative symptom sub-domains) over time. The aim is to determine if these homogeneous groups contribute to the understanding of subdomains of negative symptoms and investigate if their specific trajectories impact functioning and quality of life. Additionally, the application of multiple imputation techniques are used to deal with missing data on outcomes and other covariates. Next, I move to the association of heterogeneous *outcomes* and candidate risk factors.

In **chapter 5**, the predictive value of neurocognitive and social cognitive measures on the course and impact of psychotic experiences in siblings of people with psychotic disorders is investigated using mixture distribution model. The methodology of mixture of generalized linear mixed effects modeling is also explained in this chapter. The application of MI techniques is also used here.

In **chapter 6**, I describe the heterogeneity, regarding somatic diseases and complaints among patients with psychotic disorders, their unaffected siblings and healthy population. I examine the effects of gender, age and familial liability on the prevalence of multimorbidity.

In **chapter 7**, I investigate the factors that contribute to DUP in a large sample that represents the treated prevalence of non-affective psychotic disorders. DUP is categorized into meaningful ordinal groups and the ordinal logistic regression is applied to identify important factors. Other statistical approaches are also used to confirm factors associated with DUP and discuss as well.

Chapter 8 finally synthesizes the main findings and discusses these from multiple perspectives: from methodological consideration and clinical implication. I conclude with some future perspective and suggestions for further research.

References

Allison, P.D. (2002). *Missing Data*. Thousand Oaks, CA: Sage Publications.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders (5th ed.)*. Washington, DC: American Psychiatric Publishing.

Andruff, H., Carraro, N., Thompson, A., Gaudreau, P., Louvet, B. (2009). Latent Class Growth Modelling: A Tutorial. *Tutorials in Quantitative Methods for Psychology* 5:11-24.

Beunckens, C., Molenberghs, G., Kenward, M.G. (2005). Direct likelihood analysis versus simple forms of imputation for missing data in randomized clinical trials. *Clinical trials (London, England)* 2:379-386.

Borgen, F.H., Barnett, D.C. (1987). Applying cluster-analysis in counseling psychology research. *Journal of counseling psychology* 34:456-468.

Bresee, L.C., Majumdar, S.R., Patten, S.B., Johnson, J.A. (2010). Prevalence of cardiovascular risk factors and disease in people with schizophrenia: a population-based study. *Schizophrenia research* 117:75-82.

Breslow, N.E., Clayton, D.G. (1993). Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association* 88:9-25.

Bruins, J., Pijnenborg, M.G., Bartels-Velthuis, A.A., Visser, E., van den Heuvel, E.R., Bruggeman, R., Jorg, F. (2016). Cannabis use in people with severe mental illness: The association with physical and mental health--a cohort study. A Pharmacotherapy Monitoring and Outcome Survey study. *Journal of psychopharmacology (Oxford, England)* 30:354-362.

Bushe, C., Holt, R. (2004). Prevalence of diabetes and impaired glucose tolerance in patients with schizophrenia. *The British journal of psychiatry.Supplement* 47:S67-71.

Dawes, S.E., Jeste, D.V., Palmer, B.W. (2011). Cognitive profiles in persons with chronic schizophrenia. *Journal of clinical and experimental neuropsychology* 33:929-936.

De Hert, M., Correll, C.U., Bobes, J., Cetkovich-Bakmas, M., Cohen, D., Asai, I., Detraux, J., Gautam, S., Moller, H.J., Ndetei, D.M., Newcomer, J.W., Uwakwe, R., Leucht, S. (2011). Physical illness in patients with severe mental disorders. I. Prevalence, impact of medications and disparities in health care. *World psychiatry : official journal of the World Psychiatric Association (WPA)* 10:52-77.

De Hert, M., Dekker, J.M., Wood, D., Kahl, K.G., Holt, R.I., Moller, H.J. (2009). Cardiovascular disease and diabetes in people with severe mental illness position statement from the European Psychiatric Association (EPA), supported by the European Association for the Study of Diabetes (EASD) and the European Society of Cardiology (ESC). *European psychiatry : the journal of the Association of European Psychiatrists* 24:412-424.

De la Cruz-Mesía, R., Quintana, F.A., Marshall, G. (2008). Model-based clustering for longitudinal data. *Computational Statistics & Data Analysis* 52:1441-1457.

Fernández, A., Gómez, S. (2008). Solving Non-uniqueness in Agglomerative Hierarchical Clustering Using Multidendrograms. *Journal of Classification* 25:43-65.

Fioravanti, M., Carlone, O., Vitale, B., Cinti, M.E., Clare, L. (2005). A meta-analysis of cognitive deficits in adults with a diagnosis of schizophrenia. *Neuropsychology review* 15:73-95.

Gejman, P.V., Sanders, A.R., Duan, J. (2010). The Role of Genetics in the Etiology of Schizophrenia. *The Psychiatric clinics of North America* 33:35-66.

Gordon, A.D. (1999). *Classification*. London, UK: Chapman & Hall/CRC Monographs on Statistics & Applied Probability.

Hartigan, J.A. (1975). *Clustering Algorithms. A Wiley Publication in Applied Statistics, John Wiley & Sons, New York.*

Hartigan, J.A., Wong, M.A. (1979). A K-Means Clustering Algorithm. *Applied Statistics* 28:100-108.

Haviland, A.M., Jones, B.L., Nagin, D.S. (2011). Group-based Trajectory Modeling Extended to Account for Nonrandom Participant Attrition. *Sociological Methods & Research* 40:367-390.

Hennekens, C.H., Hennekens, A.R., Hollar, D., Casey, D.E. (2005). Schizophrenia and increased risks of cardiovascular disease. *American Heart Journal* 150:1115-1121.

Henry, D.B., Tolan, P.H., Gorman-Smith, D. (2005). Cluster analysis in family psychology research. *Journal of family psychology : JFP : journal of the Division of Family Psychology of the American Psychological Association (Division 43)* 19:121-132.

Jablensky, A. (2006). Subtyping schizophrenia: implications for genetic research. *Molecular psychiatry* 11:815-836.

Jones, B.L., Nagin, D.S., Roeder, K. (2001). A SAS Procedure Based on Mixture Models for Estimating Developmental Trajectories. *Sociological Methods & Research* 29:374-393.

Jones, B.L., Nagin, D.S. (2007). Advances in Group-Based Trajectory Modeling and an SAS Procedure for Estimating Them. *Sociological Methods & Research* 35:542-571.

Joyce, E.M., Roiser, J.P. (2007). Cognitive heterogeneity in schizophrenia. *Current opinion in psychiatry* 20:268-272.

Kahn RS, K.R. (2013). Schizophrenia is a cognitive illness: Time for a change in focus. *JAMA Psychiatry* 70:1107-1112.

Korver, N., Quee, P.J., Boos, H.B.M., Simons, C.J.P., de Haan, L., GPOUPinvestigators, Investigators, G. (2012). Genetic Risk and Outcome of Psychosis (GROUP), a multi-site longitudinal cohort study focused on gene-environment interaction: objectives, sample characteristics, recruitment and assessment methods. *International Journal of Methods in Psychiatric Research* 21:205-221.

Krabbendam, L., Myin-Germeys, I., Hanssen, M., van Os, J. (2005). Familial covariation of the subclinical psychosis phenotype and verbal fluency in the general population. *Schizophrenia research* 74:37-41.

Laird, N.M., Ware, J.H. (1982). Random-effects models for longitudinal data. *Biometrics* 38:963-974.

Lieburg, E., Castelein, S., Stewart, R., van, d.G., Aleman, A., Knegtering, H. (2013). Two subdomains of negative symptoms in psychotic disorders: established and confirmed in two large cohorts. *Journal of psychiatric research* 47:718-25.

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

Markova, I.S., Berrios, G.E. (1995). Mental symptoms: are they similar phenomena? The problem of symptom heterogeneity. *Psychopathology* 28:147-157.

Milligan, G.W., Cooper, M.C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika* 50:159-179.

Molenberghs, G., Verbeke, G. (2006). *Models for Discrete Longitudinal Data*. New York: Springer-Verlag.

Mueser, K. T. and Jeste, D. V. (2008). Lavretsky, H: History of Schizophrenia as a Psychiatric Disorder
. In *Clinical Handbook of Schizophrenia.* 3-12. New York: Guilford Press.

Nagin, D.S., Tremblay, R.E. (2001). Parental and early childhood predictors of persistent physical aggression in boys from kindergarten to high school. *Archives of General Psychiatry* 58:389-394.

Nagin, D.S. (1999). Analyzing Developmental Trajectories: A Semiparametric, Group-Based Approach. *Psychological Methods* 4:139-157.

Nagin, D.S. (2014). Group-based trajectory modeling: an overview. *Annals of Nutrition & Metabolism* 65:205-210.

Olsen, M.K., Schafer, J.L. (2001). A Two-Part Random-Effects Model for Semicontinuous Longitudinal Data. *Journal of the American Statistical Association* 96:730-745.

Parks, J., Svendsen, D., Singer, P., and Foti, M. (eds) (2006). *Morbidity and mortality in people with serious mental illness*. Alexandria: Alexandria: National Association of State Mental Health Program Directors (NASMHPD) Medical Directors Council.

Perala, J., Suvisaari, J., Saarni, S.I., Kuoppasalmi, K., Isometsa, E., Pirkola, S., Partonen, T., Tuulio-Henriksson, A., Hintikka, J., Kieseppa, T., Harkanen, T., Koskinen, S., Lonnqvist, J. (2007). Lifetime prevalence of psychotic and bipolar I disorders in a general population. *Archives of General Psychiatry* 64:19-28.

Quee, P.J., Alizadeh, B.Z., Aleman, A., van den Heuvel, E.R., GROUP Investigators. (2014). Cognitive subtypes in non-affected siblings of schizophrenia patients: characteristics and profile congruency with affected family members. *Psychological medicine* 44:395-405.

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

Schafer, J.L., Graham, J.W. (2002). Missing data: our view of the state of the art. *Psychological methods* 7:147-177.

Schlattmann, P. (2009). *Medical Applications of Finite Mixture models*. Berlin, Germany: Springer.

Smith, V.A., Neelon, B., Preisser, J.S., Maciejewski, M.L. (2015). A marginalized two-part model for longitudinal semicontinuous data. *Statistical methods in medical research*1-24.

Stahl, S.M., Buckley, P.F. (2007). Negative symptoms of schizophrenia: a problem that will not go away. *Acta Psychiatrica Scandinavica* 115:4-11.

Tibshirani, R., Walther, G., Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society.Series B, Statistical methodology* 63:411-423.

Tooze, J.A., Grunwald, G.K., Jones, R.H. (2002). Analysis of repeated measures data with clumping at zero. *Statistical methods in medical research* 11:341-355.

van Buuren, S. (2007). Multiple Imputation of Discrete and Continuous Data by Fully Conditional Specification. *Statistical methods in medical research* 16:219-242.

van Os, J., Kapur, S. (2009). Schizophrenia. *Lancet (London, England)* 374:635-645.

van Os, J., Kenis, G., Rutten, B.P. (2010). The environment and schizophrenia. *Nature* 468:203-212.

van Os, J. (2016). "Schizophrenia" does not exist. *BMJ* 352:.

van Winkel, R., De Hert, M., Van Eyck, D., Hanssens, L., Wampers, M., Scheen, A., Peuskens, J. (2006). Screening for diabetes and other metabolic abnormalities in patients with schizophrenia and schizoaffective disorder: evaluation of incidence and screening methods. *The Journal of clinical psychiatry* 67:1493-1500.

Verbeke, G., Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. New York: Springer series in statistics.

Verbeke, G., Lesaffre, E. (1996). A Linear Mixed-Effects Model With Heterogeneity in the Random-Effects Population. *Journal of the American Statistical Association* 91:217-221.

Vermunt, J.K., Magidson, J. (2003). Latent class models for classification. *Computational Statistics & Data Analysis* 41:531-537.

Vreeland, B. (2007). Treatment decisions in major mental illness: weighing the outcomes. *The Journal of clinical psychiatry* 68 Suppl 12:5-11.

Ward, J.H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association* 58:236.

Part A: Statistical Analysis for Heterogeneity